

# Learning Stabilizing Policies via an Unstable Subspace Representation

Leonardo F. Toso, Lintao Ye, and James Anderson

**Abstract**—We study the problem of learning to stabilize (LTS) a linear time-invariant (LTI) system. Learning a stabilizing control policy is a fundamental problem in its own right, as policy gradient (PG) methods for optimal control problems [1] assume access to an initial stabilizing controller—a condition that may be as difficult to satisfy in practice as learning an optimal controller itself. Existing work to the LTS problem often rely on the assumption that the drift matrix is diagonalizable or require a large amount of data to achieve stabilization. We propose a two-phase algorithm that first learns the *left unstable subspace* of the system and then applies a discount method on the learned subspace, targeting to stabilize only the system’s unstable modes. By focusing on the unstable dynamics of the system, the effective dimension of the control space is significantly reduced. We provide non-asymptotic guarantees for both phases, and show that this reduction in dimensionality leads to improved sample complexity. In fact, we demonstrate that when the number of unstable modes is much smaller than the state dimension, learning to stabilize via an unstable subspace representation can substantially accelerate the process of finding a stabilizing policy. Numerical experiments are provided to validate our theoretical findings and confirm the sample complexity reduction achieved by our approach.

## I. INTRODUCTION

In contrast to traditional control methods, model-free, policy gradient (PG) approaches offer two substantial advantages, i) they are straightforward to implement, and ii) they adapt easily to new tasks with little parameter tuning required. PG methods have been widely applied to solve reinforcement learning (RL) problems in unknown environments [2]. In several settings they have provably strong optimality guarantees: [3], [4]. As a result, there has been a lot of interest in applying these methods to control problems, including; linear quadratic regulators (LQR) in the offline setting [5]–[8], online setting [9], multi-task setting [10]–[12], and networked control setting [13]. An important milestone was achieved in [5], which showed that the PG methods for LQR exhibits a benign optimization landscape, enabling global convergence at a linear rate [14].

There is one major problem encountered when applying PG methods to control problems, namely, it is typically assumed that one has access to an initial stabilizing controller. For the most fundamental problem in control, that of finding a stabilizing controller for an unknown system, such an assumption precludes the use of PG methods. Motivated by this, several authors have studied the problem of learning to

stabilize (LTS) [15]–[20]. Two notable existing approaches that this work builds on are: discounted methods [16], [18], [20] and unstable subspace learning [19], [21], [22]. In the first approach, PG methods are used to solve a sequence of discounted LQR problems with a carefully chosen sequence of increasing discount factors. Since the policy gradients are estimated from data (system trajectories), this approach potentially suffers from a high sample complexity [20]. Leveraging the fact that a stabilizing controller only needs to stabilize the unstable modes of the system, the second approach identifies the unstable dynamics and constructs a stabilizing controller targeting to stabilize only the “small” and unstable part of the system. While focusing on the unstable dynamics may lead to sample complexity reduction, i.e., since the dimension of the control space is reduced, [19], [21] require the identification of the system’s unstable dynamics. In this work, we tackle the LTS problem using discount methods, with the goal of reducing sample complexity by operating on the *left unstable subspace* of the system.

The related work [22] applies policy optimization on the unstable subspace to learn a stabilizing controller; however, it does not provide finite-sample guarantees for either the unstable subspace representation learning or the resulting controller. In the context of representation learning, [23], [24] propose learning a low-rank representation of the system model to enable sample-efficient estimation and linear quadratic control. In contrast, this work focuses on a low-rank representation of the controller, namely, one that has a physical interpretation as it spans the unstable subspace. Moreover, our work addresses the following questions:

- To what extent can we guarantee the stability of a high-dimensional system by performing discounted 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 unstable subspace representation?

### A. Contributions

- **Sample complexity reduction:** By focusing on the unstable subspace, namely, the low-dimensional subspace associated with the  $\ell \in \mathbb{N}$  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 prove that such approach reduces the sample complexity of finding a stabilizing policy from  $\tilde{O}(d_X^2 d_U)$  [20] to  $\tilde{O}(\ell^2 d_U)$  (Theorem 4), with  $d_U$  being the number of inputs, which yields a significant sample efficiency reduction when  $\ell \ll d_X$ .
- **Learning the left unstable subspace:** We provide finite-sample guarantees for learning a representation of the left eigenspace associated with the system’s unstable modes by

LT and JA are with the Department of Electrical Engineering at Columbia University, New York, USA. Emails: {lt2879, james.anderson}@columbia.edu. LY is with the School of Artificial Intelligence and Automation at the Huazhong University of Science and Technology, Wuhan, China. Email: yelintao93@hust.edu.cn.

sampling trajectory data from an adjoint system (Theorem 1). We demonstrate that operating on the left unstable subspace enables us to control the error in the closed-loop spectral radius in terms of the error in estimated representation, which improves as more data is collected. This contrasts with prior work [19], [21], which focuses on recovering a basis for 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 on the right subspace, which inevitably incurs a source of bias that becomes significant for non-symmetric drift matrices.

• **Non-diagonalizable matrices:** Our results accommodate non-diagonalizable drift matrices. In contrast to [19], [21], which restricts analysis to diagonalizable systems, we leverage the Jordan normal form decomposition and establish that the left unstable subspace representation can be learned with a finite amount of samples (Lemma 4 and Theorem 1). This is in sharp contrast to [21], where the sample complexity scales inversely with the spectral gap between the unstable modes, this dependence becomes problematic when the system is non-diagonalizable, as the gap becomes zero and the amount of data grows prohibitively large.

#### B. Related Work

• **Learning to stabilize with system identification:** A natural idea to find a stabilizing controller for an unknown system is to first identify a model from data and then synthesize a controller. In [17], using this approach, the authors develop an algorithm with a sample complexity that scales with  $d_X$ . However, for unstable systems, such scaling is undesirable as the state-norm grows too quickly. To address this, work in [19] shows that a stabilizing controller can be designed by only identifying the unstable modes of the system, which leads to a sample complexity that scales linearly with respect to the number of unstable modes  $\ell \ll d_X$ .

• **Learning to stabilize with policy gradient:** An alternative approach based on reinforcement learning, is to learn a stabilizing controller *without* performing system identification. Recent work [16], [18] 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 algorithm for constructing a stabilizing controller. Initially, setting  $\gamma$  sufficiently small, PG methods are used to solve the damped LQR problem. Once a stabilizing controller is obtained,  $\gamma$  is incrementally increased, and the process is repeated as  $\gamma \rightarrow 1$ . Subsequent work [20] provides an explicit update rule of  $\gamma$  which allows for characterizing the sample complexity, which scales as  $\mathcal{O}(d_X^2 d_U)$ .

#### C. 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  $\ell_2$  norm,  $\|\cdot\|_{\psi_2}$  denotes the sub-Gaussian norm [25], 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.  $\kappa(A)$  denotes the condition number of the matrix with the eigenvectors of  $A$  as columns. We use the big-O notation  $\mathcal{O}(\cdot)$  to omit constants and  $\tilde{\mathcal{O}}(\cdot)$  to omit logarithmic factors in the argument.

## II. PROBLEM FORMULATION

Consider the discrete-time LTI system

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

where  $x_t \in \mathbb{R}^{d_X}$  is the state and  $u_t \in \mathbb{R}^{d_U}$  denotes the control input. We assume that the initial state  $x_0$  is drawn according to a zero mean and isotropic distribution, i.e.,  $\mathbb{E}[x_0] = 0$ ,  $\mathbb{E}[x_0 x_0^\top] = I$ . We further assume that for any drawn initial state,  $x_0^i$ , its  $\ell_2$  and sub-Gaussian norms are uniformly upper bounded, namely,  $\|x_0^i\| \leq \mu_0$  and  $\|x_0^i\|_{\psi_2} \leq \mu_\psi$ . Let  $\{\lambda_1, \lambda_2, \dots, \lambda_{d_X}\}$ , with  $|\lambda_1| \geq \dots \geq |\lambda_{d_X}|$ , denote the eigenvalues of  $A$ . We are interested in the setting where  $A$  is open-loop unstable, namely,  $\rho(A) \geq 1$ , with  $\ell < d_X$  unstable modes  $\{\lambda_1, \dots, \lambda_\ell\}$ . We make the assumption that (1) is controllable (and thus stabilizable). This is a standard assumption [18]–[20] and it guarantees that there exists a state feedback control gain  $K \in \mathbb{R}^{d_U \times d_X}$  (also referred to as policy) that stabilizes (1), i.e.,  $\rho(A + BK) < 1$ . We aim to design this controller, without knowing the system model  $(A, B)$  using a policy gradient method [5].

#### A. Discounted Linear Quadratic Regulator Problem

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

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

where  $\mathcal{K} := \{K \mid \rho(A + BK) < 1\}$  denotes the set of stabilizing controllers, and  $(Q, R)$  are positive definite matrices. We 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). It is well-known [16] that by rescaling the state  $x_t$  by  $\gamma^{t/2}$ , problem (2) is equivalent to

$$\min_{K \in \mathcal{K}^\gamma} \left\{ J^\gamma(K) := \mathbb{E} \left[ \sum_{t=0}^{\infty} \tilde{x}_t^\top (Q + K^\top R K) \tilde{x}_t \right] \right\}, \quad \text{subject to } \tilde{x}_{t+1} = (A^\gamma + B^\gamma K) \tilde{x}_t, \text{ with } \tilde{x}_t = \gamma^{t/2} x_t \quad (3)$$

where  $\mathcal{K}^\gamma := \{K \mid \rho(A^\gamma + B^\gamma K) < 1\}$ , with  $A^\gamma := \sqrt{\gamma}A$ ,  $B^\gamma := \sqrt{\gamma}B$ . Note that if  $\gamma$  is set sufficiently small, in particular,  $\gamma < 1/\rho^2(A)$ , then the trivial controller  $K \equiv 0$  is stabilizing for the underlying discounted LQR problem. However, such state feedback control gain is not stabilizing for the original system ( $\gamma = 1$ ). This fact, combined with an appropriate incremental update of  $\gamma$ , is what allows us to design a stabilizing controller by solving a sequence of discounted LQR problems. In particular, let  $\gamma_j$  denote the discount factor at iteration  $j \in \mathbb{N}$ , recent work [20] shows that by repeating the following iterates: While  $\gamma_{j+1} < 1$

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

This process guarantees that a stabilizing controller  $K \in \mathcal{K}$  is found within a finite number of iterations. Here,  $\xi \in (0, 1)$  is a decay factor,  $\bar{J}$  is a uniform bound of the discount LQR cost across iterations, and  $\alpha_j > 0$  is the discount factor update rate. We will elaborate on the role and selection of each of these quantities when we introduce our method for learning a stabilizing controller on the unstable subspace in Section IV. For now, it is important to highlight that the explicit discount method presented in [20] comes with a sample complexity that scales quadratically with the system's state dimension, i.e.,  $\tilde{\mathcal{O}}(d_X^2 d_U)$ , thus limiting its applicability for high-dimensional systems where data collection is difficult and thus scarce (e.g., robot manipulation [26]).

However, high-dimensional unstable systems often possess only a small number of unstable modes, i.e.,  $\ell \ll d_X$ . 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 provide a positive answer for this question in this paper. Next, we introduce a linear parameterization of  $K$  for the stabilizing the unstable modes of (1) independently from its stable dynamics.

#### B. 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  $\Phi$  as the “unstable subspace representation”. Hence,  $A$  can be decomposed as

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

and  $A_s = \Phi_\perp^\top A \Phi_\perp$ , where  $A_u$  represents the unstable dynamics of  $A$ , i.e., 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  $A$ , and  $\Delta$  represents the “coupling” of the stable and unstable dynamics due to the above decomposition. Note that  $\Delta \equiv 0$  when  $A$  is symmetric.

In addition, 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  $\Phi$ , namely,  $K = \theta \Phi^\top$ . Hence, the closed-loop system matrix  $A + BK$  can be written as

$$A + BK = \Omega \begin{bmatrix} A_u + B_u \theta & \\ \Delta + B_s \theta & A_s \end{bmatrix} \Omega^\top := \Omega \bar{A} \Omega^\top,$$

where  $B_u = \Phi^\top B$  and  $B_s = \Phi_\perp^\top B$ . From the above expression, we note that it suffices to stabilize the low-dimensional unstable dynamics described by  $(A_u, B_u)$  in order to guarantee that the stability of  $(A, B)$ . Hence, we reduce the problem of stabilizing  $(A, B)$  via  $K$ , to that of stabilizing  $(A_u, B_u)$  by finding a low-dimensional controller  $\theta$ , 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 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  $\Delta$  in the top-right block of the decomposition of  $A$ , as it appears

in [19], [21]. This disrupts the triangular structure of  $\bar{A}$  and thus  $\Delta$  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, when  $\|\Delta\|$  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 and it may even prevent us from stabilizing the system, as seen in [19], [21]. That is not the case when we use the left unstable subspace representation.

#### C. Low-Dimensional Discounted LQR Problem

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

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

where  $z_0$  is drawn from a zero mean and isotropic distribution since  $\Phi$  is orthonormal. Moreover, we write the discounted LQR problem on the unstable subspace as follows

$$\min_{\theta \in \Theta^\gamma} \left\{ J^\gamma(\theta, \Phi) := \mathbb{E} \left[ \sum_{t=0}^{\infty} z_t^\top (\Phi^\top Q \Phi + \theta^\top R \theta) z_t \right] \right\},$$

subject to  $z_{t+1} = (A_u^\gamma + B_u^\gamma \theta) z_t$ , (5)

where  $\Theta^\gamma := \{\theta \mid \sqrt{\gamma} \rho(A_u^\gamma + B_u^\gamma \theta) < 1\}$  denotes the set of stabilizing controllers for the damped low-dimensional dynamics, described by  $A_u^\gamma = \sqrt{\gamma} A_u$  and  $B_u^\gamma = \sqrt{\gamma} B_u$ . Let  $\nabla J^\gamma(\theta, \Phi)$  denote the gradient of  $J^\gamma(\theta, \Phi)$  with respect to  $\theta$ . Therefore, we have the following

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

with  $E_\theta = (R + B_u^{\gamma\top} P_\theta^\gamma B_u^\gamma) \theta + B_u^{\gamma\top} P_\theta^\gamma A_u^\gamma$ , and covariance of the low-dimensional state,  $\Sigma_\theta = \mathbb{E} \left[ \sum_{t=0}^{\infty} z_t z_t^\top \right]$ , where  $P_\theta^\gamma = \Phi^\top Q \Phi + \theta^\top R \theta + (A_u^\gamma + B_u^\gamma \theta)^\top P_\theta^\gamma (A_u^\gamma + B_u^\gamma \theta)$ . With a slight abuse of notation, we write  $J^\gamma(\theta) = J^\gamma(\theta, \Phi)$  and note that  $J^\gamma(\theta) = \text{Tr}(P_\theta^\gamma)$ . Below, we define the sublevel set of stabilizing controllers for the damped system  $(A_u^\gamma, B_u^\gamma)$ .

*Definition 1:* Given a discount factor  $\gamma \in (0, 1]$  and scalar  $\mu_s > 0$ . Let  $\mathcal{S}_\theta^\gamma$  denote a sublevel set of  $\Theta^\gamma$ , that is,  $\mathcal{S}_\theta^\gamma \subseteq \Theta^\gamma$ , with  $\mathcal{S}_\theta^\gamma := \{\theta \mid J^\gamma(\theta) - J^\gamma(\theta^*) \leq \mu_s (J^\gamma(\theta_0) - J^\gamma(\theta^*))\}$ .

Similarly, we define  $\mathcal{S}_K^\gamma$  as the sublevel set of  $\mathcal{K}^\gamma$ . We also use  $J_\star^\gamma$  to denote the optimal cost. Moreover, let  $\phi$ ,  $\nu_\theta$ ,  $L_\theta$ ,  $L_K$  and  $\mu_{\text{PL}}$  be positive constants. The following properties of the discounted LQR costs  $J^\gamma(\theta)$  and  $J^\gamma(K)$  hold in their stabilizing sublevel sets  $\mathcal{S}_\theta^\gamma$  and  $\mathcal{S}_K^\gamma$ , respectively.

*Lemma 1:* Given the stabilizing controllers  $K, K' \in \mathcal{S}_K^\gamma$  and  $\theta, \theta' \in \mathcal{S}_\theta^\gamma$ , it holds that  $\|\nabla J^\gamma(K)\| \leq \phi$ ,  $\|\theta\| \leq \nu_\theta$ , and

$$\begin{aligned} \|\nabla J^\gamma(\theta) - \nabla J^\gamma(\theta')\|_F &\leq L_\theta \|\theta - \theta'\|_F, \\ \|\nabla J^\gamma(K) - \nabla J^\gamma(K')\|_F &\leq L_K \|\theta - \theta'\|_F. \end{aligned}$$

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

*Remark 2:* We define  $\phi$ ,  $\nu_\theta$ ,  $L_\theta$ ,  $L_K$ , and  $\mu_{\text{PL}}$  as the uniform bound over the set of all stabilizing controllers, i.e., either  $\mathcal{S}_\theta^\gamma$  or  $\mathcal{S}_K^\gamma$ , for any discount factor  $\gamma \in (0, 1]$ .

We conclude this section by recalling that our setting is model-free and thus  $\Phi$  cannot be computed directly. Next, we demonstrate that we can recover an accurate estimation  $\hat{\Phi}$ , denoted by  $\hat{\Phi}$ , when a sufficient amount of trajectory data is collected. The quality of the estimate is quantified via the subspace distance between the subspaces spanned by the columns of  $\hat{\Phi}$  and  $\Phi$ , as defined in [27].

*Definition 2:* Let  $\hat{\Pi} = \hat{\Phi}\hat{\Phi}^\top$  and  $\Pi = \Phi\Phi^\top$  be orthogonal projectors onto the subspace spanned by columns of  $\hat{\Phi}$  and  $\Phi$ , respectively. The subspace distance between  $\Phi$  and  $\hat{\Phi}$  is  $d(\hat{\Phi}, \Phi) \triangleq \|\hat{\Phi}^\top \Phi_\perp\| = \|\Pi - \hat{\Pi}\| \in [0, 1]$ .

### III. LEARNING THE LEFT UNSTABLE REPRESENTATION

To learn an estimation of the left unstable subspace representation, we proceed by first collecting data from the adjoint system of (1), i.e., the system with system matrix  $A^\top$  [28]. For this purpose, we can simply perform element-wise computations with the adjoint operator while playing with (1) accordingly. In particular, note that for any real-valued matrix  $A \in \mathbb{R}^{d_x \times d_x}$  and vectors  $x \in \mathbb{R}^{d_x}$ ,  $y \in \mathbb{R}^{d_x}$ , we can write  $\langle Ax, y \rangle = \langle x, A^\top y \rangle$ . Therefore, by playing (1) with a zero input  $u_0 \equiv 0$  and appropriate initial condition,  $x_0 = e_i$ , where  $\{e_i\}_{i=1}^{d_x}$  is the  $i$ -th canonical basis vector of  $\mathbb{R}^{d_x}$ , we collect and store  $e_i^+ = Ae_i$ , to obtain

$$(A^\top x)_i := \langle e_i, A^\top x \rangle = \langle e_i^+, x \rangle,$$

for all  $i \in [d_x]$ , which implies  $x_{t+1} = [x_t^\top e_1^+ \dots x_t^\top e_{d_x}^+]^\top$ . Hence, the next adjoint state  $x_{t+1}$  can be derived from its previous  $x_t$  and vectors  $\{e_i^+\}_{i=1}^{d_x}$  which are data samples obtained by playing with the original system (1) accordingly.

Suppose that we collect  $T$  data samples from the adjoint system and stored them in  $D = [x_1, x_2, \dots, x_T] \in \mathbb{R}^{d_x \times T}$ . We proceed by computing the singular value decomposition of  $D$ , i.e.,  $D = U\Sigma V^\top$ . An estimation of the orthonormal basis of the right unstable subspace of  $A^\top$  (or left unstable subspace of  $A$ ) is obtained with the top  $\ell$  columns of  $U$ , namely,  $\hat{\Phi} = [u_1, \dots, u_\ell]$ . We now demonstrate that  $d(\hat{\Phi}, \Phi)$  is sufficiently small when  $T$  is sufficiently large. To establish this result, we leverage a similar approach as discussed in [21, Theorem 5.1], with two key distinctions: our setting accommodates non-diagonalizable drift matrices  $A$ , and our estimation targets the left unstable subspace representation.

Let  $\Psi \in \mathbb{R}^{d_x \times d_x - \ell}$  be an orthonormal basis for the left stable subspace of  $A$ . We also define  $\Xi = [\Phi \ \Psi]$ , which contains the left eigenvectors of the unstable and stable modes of  $A$  (potentially being *generalized* eigenvectors due to the defective nature of  $A$ ). Hence, there exists matrices  $\Lambda_u \in \mathbb{R}^{\ell \times \ell}$  and  $\Lambda_s \in \mathbb{R}^{d_x - \ell \times d_x - \ell}$  that has the same spectrum as the Jordan blocks of the unstable and stable modes of  $A$ , respectively. As a result, we can write

$$A^\top [\Phi \ \Psi] = [\Phi \ \Psi] \begin{bmatrix} \Lambda_u & \\ & \Lambda_s \end{bmatrix},$$

and define  $\Xi^{-1} := S = [S_1^\top \ S_2^\top]^\top$  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  $D_2 = S_2 D$ . Note that  $D$  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 consider  $D_u$  with 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  $\ell$  columns of  $U$ , whereas  $\Pi$  projects onto the subspace spanned by the columns of  $\Phi U_1$ . We use the following lemma to characterize the distance between these subspaces.

*Lemma 3:* Let  $\sigma_\ell$  be the  $\ell$ -th singular value of  $D_u$  and  $\hat{\sigma}_{\ell+1}$  the  $\ell + 1$ -th singular value of  $D$ . Then, it holds that

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

The proof follows directly from Davis-Kahan Theorem [29] with the the following upper bound on  $\|D_2\|$ .

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

where we refer the reader to the appendix of our technical report [30] for more details. The only piece left to characterize  $d(\hat{\Phi}, \Phi)$  is the scaling of  $\sigma_\ell$  with respect to  $T$ .

*Lemma 4:* 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

$$\sigma_\ell \geq \frac{C_\sigma |\lambda_\ell|^{\frac{T}{2}} |\lambda_1|}{l^2 \sqrt{|\lambda_1|^2 - 1}}, \text{ w.p. } 1 - 4\delta_\sigma, \text{ where } C_\sigma = \mathcal{O}(1).$$

A detailed proof of this is deferred to [30]. Note that when  $T$  grows, the subspace distance  $d(\hat{\Phi}, \Phi) = \frac{\mathcal{O}(\sqrt{T})}{\mathcal{O}(|\lambda_\ell|^{T/2}) - \mathcal{O}(\sqrt{T})}$  goes to zero, with high probability. Next, we formalize the non-asymptotic guarantees of learning  $\hat{\Phi}$ .

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

We now take a moment to analyze this result. First, observe that the required number of samples  $T$  depends only *logarithmically* on the problem dimensions  $\ell$  and  $d_x$ . The main bottleneck in learning the left unstable subspace arises when the least unstable mode is close to marginal stability, i.e., when  $|\lambda_\ell| \approx 1$ . Conversely, Theorem 1 implies that estimation becomes easier as the system becomes more explosive, in particular, when  $|\lambda_\ell| \gg 1$ .

In addition, while the constant  $C_\sigma$  does not scale with  $\ell$  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 matrix takes the form  $\Lambda_i = \text{diag}(\lambda_i, \dots, \lambda_i) + \bar{N}_i$ , for all  $i \in [n]$ , where  $\bar{N}_i$  is a nilpotent matrix with ones on the first superdiagonal, if the geometric multiplicity of  $\lambda_i$ , denoted by  $\text{gm}(\lambda_i)$ , is equal to one. We note that, as discussed in [31], the estimation of unstable dynamics when the geometric multiplicity of the unstable eigenvalues is greater than one may become inconsistent. In our setting,



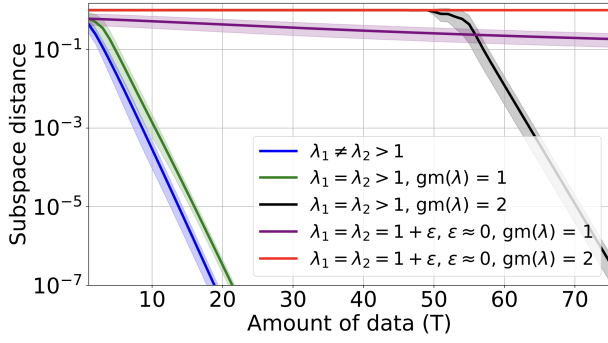


Fig. 1. Subspace distance  $d(\hat{\Phi}, \Phi)$  with respect to the amount of data ( $T$ ) for an illustrative example with  $d_X = 3$  states and  $\ell = 2$  unstable modes.

this effect deflates  $C_\sigma$  which in turn leads to an increase in the required number of samples  $T$  when  $A$  contains unstable modes with geometric multiplicity greater than one.

Figure 1, illustrates these trends predicted in our results 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 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 matrix with  $\text{gm}(\lambda) = 1$  (green curve). In contrast, when the least unstable mode is near marginal stability, successful subspace recovery becomes infeasible.

*Remark 3:* We emphasize that the guarantees for learning right unstable subspace of  $A$  presented in [21] 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. Moreover, this dependence on the spectral gap appears to be counterintuitive and does not align with the results illustrated in Figure 1.

#### IV. LTS ON THE UNSTABLE SUBSPACE

In this section, we introduce our approach for learning to stabilize by operating on the unstable subspace. Our method combines the unstable subspace representation learning, discussed previously, with the discounted method performed on that subspace. In particular, we aim to learn a low-dimensional controller  $\theta \in \mathcal{S}_\theta^1$  that stabilizes the unstable dynamics  $(A_u, B_u)$ . This is achieved by solving a sequence of discounted LQR problems via policy gradient, which requires the access to  $\nabla J^\gamma(\theta, \Phi)$ . However, since we operate in model-free, we cannot directly compute this gradient. Therefore, we use the zeroth-order gradient estimation [32], [33] to compute an estimation of the gradient, denoted by  $\hat{\nabla} J^\gamma(\theta, \hat{\Phi})$ . This estimation is performed by collecting trajectory data from the original system (1), which is “encoded” onto the estimated left unstable subspace via  $\hat{\Phi}$ , i.e.,  $z_t = \hat{\Phi}^\top x_t$ . Before presenting the zeroth-order gradient estimation method and its guarantees, we provide an upper bound on the error between  $\nabla J^\gamma(\theta, \Phi)$  and  $\nabla J^\gamma(\theta, \hat{\Phi})$  below.

*Lemma 5:* Suppose that  $\theta \in \mathcal{S}_\theta^\gamma$ . Then, it holds that

$$\|\nabla J^\gamma(\theta, \Phi) - \nabla J^\gamma(\theta, \hat{\Phi})\|_F \leq C_\Phi d(\hat{\Phi}, \Phi),$$

with  $C_\Phi = \sqrt{\ell} \left( (L_K \nu_\theta + \phi) \sqrt{2\ell} + \phi \right)$ .

Therefore, the error in the gradient incurred by the misspecified representation,  $\hat{\Phi}$ , can be made arbitrarily small, provided that  $d(\hat{\Phi}, \Phi)$  is sufficiently small. The proof of this lemma follows from Lemmas 1 and 2, combined with the upper bound  $\|\hat{\Phi} - \Phi\| \leq \sqrt{2\ell} d(\hat{\Phi}, \Phi)$  from [19, Corollary 5.3]. Additional details can be found in [30].

#### A. 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 [5], [6]. Next, we define the two-point zeroth-order estimation and its guarantees.

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

where  $U_i$  is randomly drawn from a uniform distribution on the sphere  $\sqrt{\ell} d_U \mathbb{S}^{\ell d_U - 1}$  and  $\theta_{1,i} = \theta + r U_i$ ,  $\theta_{2,i} = \theta - r U_i$ . Here,  $r > 0$  denotes the smoothing radius and  $n_s$  the number of rollouts. Moreover, let  $\tau > 0$  be the time horizon, the finite time horizon value function  $V^{\gamma, \tau}(\theta, z_0)$  is defined as follows

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

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 collected by playing (1) with  $u_t = \theta^\top \hat{\Phi}^\top x_t$ .

*Lemma 6:* Let  $\zeta$  and  $\varepsilon_\tau$  be positive scalars. Suppose that we have  $n_s = \mathcal{O}(\zeta^4 \mu_\psi^4 \log^6(\ell)) \ell$ ,  $\tau = \mathcal{O}(\log(1/\varepsilon_\tau))$  and  $r = \mathcal{O}(\sqrt{\varepsilon_\tau})$ . Then, it holds that

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

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

with probability  $1 - c_1(\ell^{-\zeta} + n_s^{-\zeta} - n_s e^{-\ell/8} - e^{-c_2 n_s})$ .  $c_1$  and  $c_2$  are positive constants, and  $C_{\text{est},1}$ ,  $C_{\text{est},3}$ , and  $C_{\text{est},4}$  scales as  $\mathcal{O}(d_U \ell \log^2(\ell))$ , and  $C_{\text{est},2} = \mathcal{O}(1)$ .

The proof for this lemma follows from [14, Section V] and Lemma 5. Lemma 6, tells us that if we have an accurate estimation of  $\Phi$ , and we set  $r$  sufficiently small, and  $\tau$  and  $n_s$  sufficiently large, then  $\|\hat{\nabla} J^\gamma(\theta, \hat{\Phi})\|_F^2 = \mathcal{O}(\|\nabla J^\gamma(\theta)\|_F^2)$  and  $\langle \nabla J^\gamma(\theta, \Phi), \hat{\nabla} J^\gamma(\theta, \hat{\Phi}) \rangle = \mathcal{O}(\|\nabla J^\gamma(\theta, \Phi)\|_F^2)$ , with high probability. This is crucial to establishing linear convergence of PG in each discounted LQR problem we solve.

Moreover, let  $\hat{J}^{\gamma, \tau}(\theta, \hat{\Phi}) = \frac{1}{n_c} \sum_{i=1}^{n_c} V^{\gamma, \tau}(\theta, z_0^i)$  be the estimated cost with  $n_c$  rollouts. We provide the following lemma to control the error between  $J^\gamma(\theta)$  and  $\hat{J}^{\gamma, \tau}(\theta, \hat{\Phi})$ .

*Lemma 7:* Given  $\theta \in \mathcal{S}_\theta^\gamma$  and  $\delta_\tau \in (0, 1)$ . Suppose that

$$\tau \geq \tau_0 := \frac{J^\gamma(\theta, \hat{\Phi})}{\sigma_{\min}(Q)} \log \left( \frac{8(J^\gamma(\theta, \hat{\Phi}))^2 \mu_0^2}{\sigma_{\min}(Q) J^\gamma(\theta)} \right),$$

$n_c \geq 8\mu_0^2 \log(2/\delta_\tau)$ , and  $d(\hat{\Phi}, \Phi) \leq J^\gamma(\theta)/(4\ell\sqrt{\ell}C_{\text{cost}})$ . Then,  $|\hat{J}^{\gamma, \tau}(\theta, \hat{\Phi}) - J^\gamma(\theta)| \leq \frac{1}{2} J^\gamma(\theta)$ , w.p.  $1 - \delta_\tau$ .

Here  $C_{\text{cost}}$  is polynomial in the problem parameters  $\|A\|$ ,  $\|B\|$ ,  $\|Q\|$ ,  $\|R\|$  and  $\nu_\theta$ . The proof is deferred to [30].

### B. Discounted LQR on the Unstable Subspace

With the results of the gradient and cost estimation in place, we are now equipped to present our discounted LQR method on the unstable subspace for learning to stabilize the system's unstable dynamics. First, we assume to have access to the upper bound  $|\lambda_1| \leq \bar{\lambda}_1$ . This is necessary to initialize the discount factor as  $\gamma_0 < 1/\bar{\lambda}_1^2$ , ensuring that  $\theta_0 \equiv 0$  stabilizes the corresponding damped system.

To ensure that the discount factor reaches one within a finite number of iterations, we adopt the explicit discount method proposed in [20]. In particular,  $\gamma_{j+1} = (1 + \xi\alpha_j)\gamma_j$ , where  $\xi \in (0, 1)$  is the decay factor and  $\alpha_j$  is given by

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

where this update rule follows from the Lyapunov stability analysis of the low-dimensional damped system, i.e., let  $V(z_t) = z_t^\top P_\theta^\gamma z_t$  be a quadratic Lyapunov function for  $z_{t+1} = \sqrt{\gamma_{j+1}}(A_u + B_u \theta)z_t$ , and  $\Delta V = V(z_{t+1}) - V(z_t)$ .

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

which implies that  $\frac{\gamma_{j+1}}{\gamma_j} (P_\theta^\gamma - \Phi^\top Q \Phi - \theta^\top R \theta) - P_\theta^\gamma \prec 0$  ensures  $\sqrt{\gamma_{j+1}}(A_u + B_u \theta) < 1$ . A sufficient condition is

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

where the last inequality follows from Bauer-Fike Theorem [34] and  $d(\hat{\Phi}, \Phi) \leq \frac{\sigma_{\min}(\hat{\Phi}^\top Q \hat{\Phi} + \theta^\top R \theta)}{4\|Q\|\sqrt{2\ell\kappa}(\hat{\Phi}^\top Q \hat{\Phi} + \theta^\top R \theta)}$ . Therefore, by using Lemma 7, we recover (6). Moreover, as discussed in [20], the decay factor  $\xi \in (0, 1)$  ensures that the updated controller  $\theta_{j+1}$  “strongly” stabilizes the damped system  $(A_u^{\gamma_{j+1}}, B_u^{\gamma_{j+1}})$ . In the next section, we show the role of  $\xi$  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 to learn a stabilizing controller for (1) by operating on the unstable subspace of the underlying system. 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 unstable subspace representation, which in turn defines the discounted LQR problem over the estimated 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  $\theta_j$  for the damped system with factor  $\gamma_j$ , we perform  $N$  policy gradient iterations with step size  $\eta$  using the estimated gradient  $\hat{\nabla} J^{\gamma_j}(\theta, \hat{\Phi})$  (lines 5 and 6). The number of iterations  $N$  is set to ensure that  $J^{\gamma_j}(\theta) \leq \bar{J}$ . We make the condition on  $N$  explicit in the next section. The discount factor is updated using (6) (line 8). Finally, Algorithm 1 returns  $K = \theta_{j+1} \hat{\Phi}^\top$ .

### 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 the estimated representation  $\hat{\Phi} = [u_1, \dots, u_\ell]$  be the top  $\ell$  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:      $\bar{\theta}_{n+1} = \bar{\theta}_n - \eta \hat{\nabla} J^{\gamma_j}(\bar{\theta}_n, \hat{\Phi})$ 
7:   Let  $\theta_{j+1} = \bar{\theta}_N$  and compute  $\alpha_j$  as in Eq. (6)
8:   Update  $\gamma_{j+1} = (1 + \xi\alpha_j)\gamma_j$  and  $j \leftarrow j + 1$ 
9: Output:  $K = \theta_{j+1} \hat{\Phi}^\top$ 

```

---

In the following section, we provide the required conditions on  $n_s, n_c, r, \tau, T, N$ , and  $\eta$  to guarantee that  $K \in \mathcal{S}_K^1$ , i.e.,  $K$  is a stabilizing controller for the original system (1).

### V. SAMPLE COMPLEXITY ANALYSIS

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

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

$$\bar{J} := \max\{2J_\star^1, J^{\gamma_0}(0)\}, \text{ and } \underline{\alpha} := \frac{3\sigma_{\min}(Q)}{2\bar{J} - 3\sigma_{\min}(Q)}.$$

*Theorem 2:* Given  $\delta_\tau \in (0, 1)$ ,  $\delta_\sigma \in (0, 1)$ , and  $\zeta > 0$ . Suppose that  $n_s = \mathcal{O}(\zeta^4 \mu_\psi^4 \log^6(\ell))$ ,  $n_c = \mathcal{O}(\log(1/\delta_\tau))$ ,  $\tau = \mathcal{O}(\log(1/\varepsilon_\tau) + \tau_0)$ ,  $r = \mathcal{O}(\sqrt{\varepsilon_\tau})$ , and the number of adjoint samples,  $T = \mathcal{O}\left(\log\left(\frac{\ell^7(d_X - \ell)\mu_0}{(1 - |\lambda_{\ell+1}|)\varepsilon_{\text{dist}}\delta_\sigma^3}\right)\right)$ , with

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

Then, with the number of PG iterations and step-size

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

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

In the above theorem,  $\bar{c}_1$  and  $\bar{c}_2$  are positive constants and  $C_{\text{dist},1}$  and  $C_{\text{dist},2}$  are polynomial in the problem parameters  $\|A\|, \|B\|, \|Q\|, \nu_\theta, L_\theta, L_K, \phi, \ell$  and  $d_U$ .

Theorem 2 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  $\tau$ , and number of iterations  $N$  are set large enough, with  $r$  and  $\eta$  sufficiently small, the algorithm produces a low-dimensional controller  $\theta_{j+1} \in \mathcal{S}_\theta^1$ . When lifted through  $\hat{\Phi}$ , this controller stabilizes (1). Additionally, our results highlight a key challenge: 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.,  $|\lambda_{\ell+1}| \approx 1$ . We emphasize that, in contrast to [22], our work is the first to provide the non-asymptotic guarantees for learning to stabilize LTI systems via their unstable subspace representation. The proof of Theorem 2 is presented in details in our technical report [30]. Below, 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 demonstrate this uniform bound, we use Lemmas 1, 2 and 6) to determine the number of PG iterations  $N$  we need to obtain  $J^{\gamma_j}(\theta) \leq \bar{J}$ . A preliminary condition on  $d(\hat{\Phi}, \Phi)$ , and on the estimation parameters comes from this step, where we make their corresponding error terms to scale as  $\mathcal{O}(\bar{J} - J_\star^\gamma)$ .

The uniform bound on the cost implies that  $\alpha_j$  is uniformly lower bounded as  $\alpha_j \geq \underline{\alpha} := \frac{3\sigma_{\min}(Q)}{2\bar{J} - 3\sigma_{\min}(Q)}$ . This implies that  $\gamma_j$  reaches one within  $\log(1/\gamma_0)/\log(1 + \xi\underline{\alpha})$  iterations. Hence, since  $\sqrt{(1 + \alpha_j)\gamma_j}\rho(A_u + B_u\theta_{j+1}) < 1$ , then  $\sqrt{(1 + \xi\alpha_j)\gamma_j}\rho(A_u + B_u\theta_{j+1}) < \bar{\lambda}_\theta$ . Note that  $\bar{\lambda}_\theta$  depends on the decay factor  $\xi$ , and we make it within  $(0, 1)$  to guarantee that the spectral radius of the closed-loop low-dimensional system is much smaller than one. Then,  $\rho(A_u + B_u\theta_{j+1}) < \bar{\lambda}_\theta$  follows from a simple induction step.

The last step of the proof is to show that  $K = \theta_{j+1}\hat{\Phi}^\top$  stabilizes  $(A, B)$ . 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 [35] and the generalized Bauer-Fike theorem [36]. In particular, the exponential dependence on  $\ell$  showing up in  $\varepsilon_{\text{dist}}$  follows from the generalized Bauer-Fike theorem, since  $A_u + B_u\theta_{j+1}$  and  $A_s$  are non-diagonalizable.

Next, we characterize the sample complexity of Algorithm 1. 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  $\mathcal{S}_c := j(n_c + n_s N)$ .

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

This result demonstrates the sample complexity reduction achieved by LTS on the unstable subspace. In contrast to [20], where it scales as  $\log(\rho(A))\tilde{\mathcal{O}}(d_\chi^2 d_U)$ , 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 note that in Algorithm 1, we also collect samples from the adjoint system, which scales with  $T$  and  $d_\chi$ , where the later is due to the element-wise computations with the adjoint operator. However, we emphasize that for a “regular” system with the least unstable and stable modes strictly away from marginal stability, and  $\text{gm}(\lambda) = 1$  for the unstable modes, then  $T$  should be negligible. As a result,  $\tilde{\mathcal{O}}(\ell^2 d_U) + \mathcal{O}(d_\chi)$  is smaller than  $\tilde{\mathcal{O}}(d_\chi^2 d_U)$ , when  $\ell \ll d_\chi$  (i.e., our regime of interest).

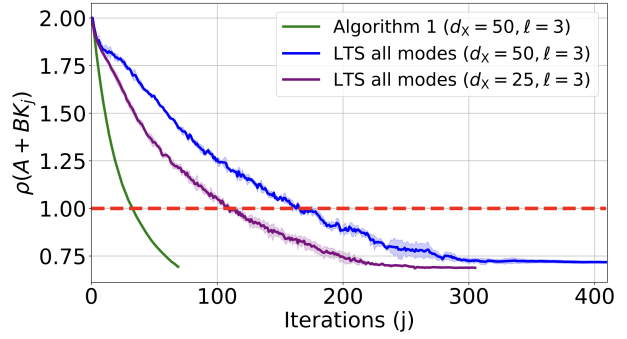


Fig. 2. Closed-loop spectral radius,  $\rho(A + BK)$ , with respect to iteration.

## VI. NUMERICAL VALIDATION

Numerical experiments are now presented to validate and illustrate our theoretical guarantees<sup>1</sup>. We demonstrate that when  $\ell \ll d_\chi$ , learning a stabilizing controller requires significantly less data when using Algorithm 1, compared to stabilizing all the modes in the  $d_\chi$ -dimensional space. To illustrate the impact of the ambient dimension  $d_\chi$ , we consider a nominal system  $(A_0, B_0)$  with four states, two inputs, and  $\ell = 3$  unstable modes, where  $|\lambda_1| = |\lambda_2| = 2$  and  $|\lambda_3| = 1.33$ , with  $\text{gm}(\lambda_1) = 1$ . The system is augmented by adding random stable modes, resulting in a higher-dimensional system with  $d_\chi$  states and  $d_U = 2$  inputs, while preserving the original three unstable modes of  $(A_0, B_0)$ . We use  $T = 40$  for learning the unstable representation. Figures 2 and 3 depict the mean and standard deviation over 10 independent runs. Additional details are provided in [30].

Figure 2 shows the spectral radius of the closed-loop system,  $\rho(A + BK_j)$ , with respect to the iteration count  $j$ , for three different cases: 1) (green curve) Algorithm 1, where we learn an unstable subspace representation and apply the discounted LQR method to stabilize only the unstable modes of a system with  $d_\chi = 50$  states; 2) (blue curve) applying the discounted LQR method to stabilize the full dynamics of a system with  $d_\chi = 50$  states; and 3) (purple curve) learning to stabilize all modes of a system with  $d_\chi = 25$  states.

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 consequently the number of samples required to learn a stabilizing controller. Remarkably, this reduction holds even when compared to learning to stabilize all modes of a system with half the dimensionality of the ambient system that Algorithm 1 is applied to. A similar trend is evident in Figure 3, which plots the discount factor as a function of iteration. These results validate our theoretical guarantees (Theorem 2 and Corollary 1), which predict the sample complexity reduction achieved by LTS directly on the unstable subspace.

## VII. CONCLUSIONS AND FUTURE WORK

We studied the problem of learning to stabilize an LTI system. To solve this problem, we first proposed a sample efficient method to learn the left unstable space of the

<sup>1</sup>Code for reproduction can be downloaded at <https://github.com/jd-anderson/LTS-unstable-representation>.



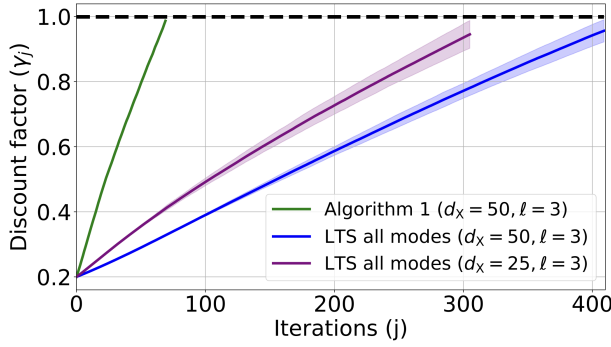


Fig. 3. Discount factor  $\gamma_j$  with respect to iteration ( $j$ ).

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 subspace distance between the learned and true representation is small enough, the discount method on the unstable subspace returns a stabilizing policy for the original system within a finite number of iterations. Compared to existing works, our approach works for 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 that share a common unstable subspace representation and designing online algorithms that continuously update the learned unstable subspace when new data samples of the system become available.

## REFERENCES

- [1] B. Hu, K. Zhang, N. Li, M. Mesbahi, M. Fazel, and T. Başar, “Toward a Theoretical Foundation of Policy Optimization for Learning Control Policies,” *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 6, pp. 123–158, 2023.
- [2] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, “Policy gradient methods for reinforcement learning with function approximation,” *Advances in neural information processing systems*, vol. 12, 1999.
- [3] A. Agarwal, S. M. Kakade, J. D. Lee, and G. Mahajan, “On the theory of policy gradient methods: Optimality, approximation, and distribution shift,” *Journal of Machine Learning Research*, vol. 22, no. 98, pp. 1–76, 2021.
- [4] J. Bhandari and D. Russo, “Global optimality guarantees for policy gradient methods,” *Operations Research*, vol. 72, no. 5, pp. 1906–1927, 2024.
- [5] M. Fazel, R. Ge, S. Kakade, and M. Mesbahi, “Global convergence of policy gradient methods for the linear quadratic regulator,” in *International conference on machine learning*. PMLR, 2018, pp. 1467–1476.
- [6] D. Malik, A. Pananjady, K. Bhatia, K. Khamaru, P. Bartlett, and M. Wainwright, “Derivative-free methods for policy optimization: Guarantees for linear quadratic systems,” in *The 22nd international conference on artificial intelligence and statistics*. PMLR, 2019, pp. 2916–2925.
- [7] B. Gravell, P. M. Esfahani, and T. Summers, “Learning optimal controllers for linear systems with multiplicative noise via policy gradient,” *IEEE Transactions on Automatic Control*, vol. 66, no. 11, pp. 5283–5298, 2020.
- [8] H. Mohammadi, A. Zare, M. Soltanolkotabi, and M. R. Jovanović, “Convergence and sample complexity of gradient methods for the model-free linear-quadratic regulator problem,” *IEEE Transactions on Automatic Control*, vol. 67, no. 5, pp. 2435–2450, 2021.
- [9] A. B. Cassel and T. Koren, “Online policy gradient for model free learning of linear quadratic regulators with  $\sqrt{T}$  regret,” in *International Conference on Machine Learning*. PMLR, 2021.
- [10] H. Wang, L. F. Toso, A. Mitra, and J. Anderson, “Model-free Learning with Heterogeneous Dynamical Systems: A Federated LQR Approach,” *arXiv preprint arXiv:2308.11743*, 2023.
- [11] L. F. Toso, D. Zhan, J. Anderson, and H. Wang, “Meta-learning linear quadratic regulators: a policy gradient maml approach for model-free lqr,” in *6th Annual Learning for Dynamics & Control Conference*. PMLR, 2024, pp. 902–915.
- [12] D. Zhan, L. F. Toso, and J. Anderson, “Coreset-based task selection for sample-efficient meta-reinforcement learning,” *arXiv preprint arXiv:2502.02332*, 2025.
- [13] A. Mitra, L. Ye, and V. Gupta, “Towards model-free lqr control over rate-limited channels,” *arXiv preprint arXiv:2401.01258*, 2024.
- [14] H. Mohammadi, M. Soltanolkotabi, and M. R. Jovanović, “On the linear convergence of random search for discrete-time LQR,” *IEEE Control Systems Letters*, vol. 5, no. 3, pp. 989–994, 2020.
- [15] S. Lale, K. Azizzadenesheli, B. Hassibi, and A. Anandkumar, “Explore more and improve regret in linear quadratic regulators,” *arXiv preprint arXiv:2007.12291*, vol. 31, p. 32, 2020.
- [16] A. Lamperski, “Computing stabilizing linear controllers via policy iteration,” in *2020 59th IEEE Conference on Decision and Control (CDC)*. IEEE, 2020, pp. 1902–1907.
- [17] X. Chen and E. Hazan, “Black-box control for linear dynamical systems,” in *Conference on Learning Theory*. PMLR, 2021, pp. 1114–1143.
- [18] J. Perdomo, J. Umenberger, and M. Simchowitz, “Stabilizing dynamical systems via policy gradient methods,” *Advances in neural information processing systems*, vol. 34, pp. 29 274–29 286, 2021.
- [19] Y. Hu, A. Wierman, and G. Qu, “On the sample complexity of stabilizing lti systems on a single trajectory,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 16 989–17 002, 2022.
- [20] F. Zhao, X. Fu, and K. You, “Convergence and sample complexity of policy gradient methods for stabilizing linear systems,” *IEEE Transactions on Automatic Control*, 2024.
- [21] Z. Zhang, Y. Nakahira, and G. Qu, “Learning to Stabilize Unknown LTI Systems on a Single Trajectory under Stochastic Noise,” *arXiv preprint arXiv:2406.00234*, 2024.
- [22] S. W. Werner and B. Peherstorfer, “System stabilization with policy optimization on unstable latent manifolds,” *Computer Methods in Applied Mechanics and Engineering*, vol. 433, p. 117483, 2025.
- [23] T. T. Zhang, L. F. Toso, J. Anderson, and N. Matni, “Sample-efficient linear representation learning from non-IID non-isotropic data,” in *The Twelfth International Conference on Learning Representations*, 2024.
- [24] B. D. Lee, L. F. Toso, T. T. Zhang, J. Anderson, and N. Matni, “Regret analysis of multi-task representation learning for linear-quadratic adaptive control,” *arXiv preprint arXiv:2407.05781*, 2024.
- [25] R. Vershynin, *High-dimensional probability: An introduction with applications in data science*. Cambridge university press, 2018, vol. 47.
- [26] A. Billard and D. Kragic, “Trends and challenges in robot manipulation,” *Science*, vol. 364, no. 6446, p. eaat8414, 2019.
- [27] G. W. Stewart and J.-g. Sun, “Matrix perturbation theory,” *Academic press*, 1990.
- [28] O. Kouba and D. S. Bernstein, “What is the adjoint of a linear system? [lecture notes],” *IEEE Control Systems Magazine*, vol. 40, no. 3, pp. 62–70, 2020.
- [29] C. Davis and W. M. Kahan, “The rotation of eigenvectors by a perturbation. III,” *SIAM Journal on Numerical Analysis*, vol. 7, no. 1, pp. 1–46, 1970.
- [30] L. F. Toso, L. Ye, and J. Anderson, “Learning Stabilizing Policies via an Unstable Subspace Representation,” 2025. [Online]. Available: <https://github.com/jd-anderson/LTS-unstable-representation/blob/main/LTS.pdf>
- [31] T. Sarkar and A. Rakhlin, “Near optimal finite time identification of arbitrary linear dynamical systems,” in *International Conference on Machine Learning*. PMLR, 2019, pp. 5610–5618.
- [32] A. D. Flaxman, A. T. Kalai, and H. B. McMahan, “Online convex optimization in the bandit setting: gradient descent without a gradient,” *arXiv preprint cs/0408007*, 2004.
- [33] J. C. Spall, *Introduction to stochastic search and optimization: estimation, simulation, and control*. John Wiley & Sons, 2005.
- [34] F. L. Bauer and C. T. Fike, “Norms and exclusion theorems,” *Numerische Mathematik*, vol. 2, no. 1, pp. 137–141, 1960.
- [35] R. Mathias, “Quadratic residual bounds for the hermitian eigenvalue problem,” *SIAM journal on matrix analysis and applications*, vol. 19, no. 2, pp. 541–550, 1998.
- [36] G. H. Golub and C. F. Van Loan, *Matrix computations*. JHU press, 2013.



## VIII. APPENDIX

### A. Auxiliary Results

*Lemma 8 (Young's inequality):* Given any two real-valued matrices  $A, B \in \mathbb{R}^{n \times m}$ . It holds that

$$\|A + B\|_2^2 \leq (1 + \beta)\|A\|_2^2 + \left(1 + \frac{1}{\beta}\right)\|B\|_2^2 \leq (1 + \beta)\|A\|_F^2 + \left(1 + \frac{1}{\beta}\right)\|B\|_F^2, \quad (7)$$

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

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

*Theorem 3 (Davis-Kahan [29]):* Let  $\Sigma$  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  $\ell$  eigenvectors of  $\Sigma$  and  $\Sigma + \Delta$ , respectively. Then, it holds that

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

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

*Lemma 9 (Generalized Bauer-Fike [36]):* Let  $Q^\top A Q = D + N$  be the Schur decomposition of  $A \in \mathbb{R}^{d \times d}$ ,  $D$  is diagonal and  $N$  upper triangular with zero diagonal. Then, it holds that

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

*Lemma 10 (Block perturbation bound):* For any 2-by-2 block matrices  $M$  and  $N$  in the form

$$M = \begin{bmatrix} M_1 & 0 \\ 0 & M_2 \end{bmatrix}, N = \begin{bmatrix} 0 & N_1 \\ N_2 & 0 \end{bmatrix},$$

it holds that  $|\rho(M + N) - \rho(M)| \leq C_{\text{gap}} \|N_1\| \|N_2\|$ , where  $C_{\text{gap}} = \frac{\kappa(M)\kappa(M+N)}{\min_i \{\text{gap}_i(M)\}}$ . Here  $\kappa(M)$  and  $\kappa(M + N)$  denote the condition number of  $M$  and  $M + N$ , respectively, and  $\text{gap}_i$  is the (bipartite) spectral gap around  $\lambda_i$  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 for non-symmetric matrices in [35, Theorem 5]. ■

### B. Discounted LQR problem

We define the discounted LQR problem as follows

$$\text{minimize}_{K \in \mathcal{K}} \left\{ J^\gamma(K) := \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t x_t^\top (Q + K^\top R K) x_t \right] \right\}, \text{ subject to (1) with } u_t = K x_t, \quad (9)$$

where the expectation is taken with respect to the initial condition. In addition,  $Q$  and  $R$  are positive definite matrices and  $\gamma \in (0, 1]$  denotes the discount factor. The above discounted LQR problem is equivalent to solve

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

where  $A^\gamma = \sqrt{\gamma}A$ ,  $B^\gamma = \sqrt{\gamma}B$ .

*Definition 3 (Set of stabilizing controllers):* Given a discount factor  $\gamma$ , the set of stabilizing controllers of the damped system  $(A^\gamma, B^\gamma)$  is  $\mathcal{K}^\gamma := \{K \mid \sqrt{\gamma}\rho(A + BK) < 1\}$ .

Given a discount factor  $\gamma \in (0, 1]$  and  $K \in \mathcal{K}^\gamma$  the discounted LQR cost and its gradient are defined as follows

$$J^\gamma(K) := \text{Tr}((Q + K^\top RK)\Sigma_K^\gamma), \quad \nabla J^\gamma(K) := 2E_K^\gamma \Sigma_K^\gamma, \quad \Sigma_K^\gamma := \sum_{t=0}^{\infty} \mathbb{E}[x_t x_t^\top], \quad (11)$$

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

### C. Linear Decomposition of the Control Policy

We consider the linear decomposition of the controller  $K = \theta \Phi^\top$ , with parameter vector  $\theta \in \mathbb{R}^{d_u \times \ell}$  and representation  $\Phi \in \mathbb{R}^{d_x \times \ell}$ , where  $\Phi$  has orthonormal columns. In particular, the columns of  $\Phi$  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  $\Phi$ , i.e.,  $x_t = \Phi z_t$ . Therefore, we can write

$$z_{t+1} = A_u z_t + B_u u_t, \quad \text{where } A_u = \Phi^\top A \Phi, \quad B_u = \Phi^\top B, \quad \text{and } u_t = \theta z_t, \quad (12)$$

and we write the discounted LQR problem on the low-dimensional dynamics (12) as follows

$$\text{minimize}_{\theta \in \Theta} \left\{ J^\gamma(\theta, \Phi) := \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t z_t^\top (\Phi^\top Q \Phi + \theta^\top R \theta) z_t \right] \right\}, \quad \text{subject to } z_{t+1} = (A_u + B_u \theta) z_t, \quad (13)$$

or equivalently

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

where  $A_u^\gamma = \Phi^\top A^\gamma \Phi$ ,  $B_u^\gamma = \Phi^\top B^\gamma$ ,  $\Theta := \{\theta \mid \rho(A_u + B_u \theta) < 1\}$ , 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  $\theta$ . Then, we have the following

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

where we can write

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

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

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

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

Let us now proceed to upper bound  $\|\nabla J^\gamma(\theta, \Phi) - \nabla J^\gamma(\theta, \hat{\Phi})\|$ , where  $\hat{\Phi}$  is an estimation of  $\Phi$ . To do so, we write

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

where (i) follows from Lemma 1 and the Definition 2. Moreover, (ii) also follows from Lemma 1. Therefore, by leveraging [19, Corollary 5.3] we have that  $\|\widehat{\Phi} - \Phi\| \leq \sqrt{2\ell}d(\widehat{\Phi}, \Phi)$ , which implies

$$\left\| \nabla J^\gamma(\theta, \Phi) - \nabla J^\gamma(\theta, \widehat{\Phi}) \right\| \leq \left( (L_K \nu_\theta + \phi) \sqrt{2\ell} + \phi \right) d(\widehat{\Phi}, \Phi), \quad (15)$$

or in the Frobenius norm

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

#### D. Gradient and Cost Estimation

We recall that we operate in model-free, i.e., we do not have access to system matrices  $(A, B)$ , thus we need to estimate the gradient  $\nabla J^\gamma(\theta, \widehat{\Phi})$ . We proceed by defining the two-point zeroth-order estimation and its guarantees.

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

where  $U_i$  is randomly drawn from a uniform distribution on the sphere  $\sqrt{\ell d_0} \mathbb{S}^{\ell d_0 - 1}$ , and  $\theta_{1,i} = \theta + rU_i$ ,  $\theta_{2,i} = \theta - rU_i$ . Note that the initial condition of the low-dimensional system  $z_0^i$  is also distributed according to a zero-mean isotropic distribution, since  $x_0^i$  is zero-mean and isotropic and  $\Phi$  is has orthonormal columns. Moreover, let  $\tau > 0$  be the time horizon, the finite time horizon value function  $V^{\gamma, \tau}(\theta, z_0)$  is defined as follows

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

where  $\{z_t\}_{t=0}^{\tau-1} = \{\widehat{\Phi}^\top x_t\}_{t=0}^{\tau-1}$  and  $\{x_t\}_{t=0}^{\tau-1}$  is the trajectory data of  $x_{t+1} = Ax_t + Bu_t$  when  $u_t = \theta \widehat{\Phi}^\top x_t$ . Moreover, let  $\widehat{\nabla} J^\gamma(\theta, \widehat{\Phi}) := \frac{1}{n_s} \sum_{i=1}^{n_s} \langle \nabla V^{\gamma, \tau}(\theta, z_0^i), U_i \rangle U_i$  be the unbiased estimate of  $\nabla J^\gamma(\theta, \widehat{\Phi})$  where the infinite horizon cost is  $V^\gamma(\theta, z_0) := \sum_{t=0}^{\infty} \gamma^t z_t^\top \left( \widehat{\Phi}^\top Q \widehat{\Phi} + \theta^\top R \theta \right) z_t$ . We note that  $\mathbb{E}[\widehat{\nabla} J^\gamma(\theta, \widehat{\Phi})] = \nabla J^\gamma(\theta, \widehat{\Phi})$  and refer the reader to [14, Section IV and V] for more details.

*Lemma 11 (Zeroth-order Estimation Bias [14]):* Suppose that the time horizon and smoothing radius satisfy  $\tau = \mathcal{O}(\log(1/\varepsilon))$  and  $r \leq \mathcal{O}(\sqrt{\varepsilon})$ , respectively. Then, it holds that  $\|\widehat{\nabla} J^\gamma(\theta, \widehat{\Phi}) - \widehat{\nabla} J^\gamma(\theta, \widehat{\Phi})\|_F \leq \varepsilon$ .

*Lemma 12 (Proposition 3 and 4 of [14]):* Let  $\mu_1$  and  $\mu_2$  be two positive scalars, and  $\mathcal{E}_1$  and  $\mathcal{E}_2$  be the following events

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

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

Here  $c_1$  and  $c_2$  are positive constants,  $\|z_0\|_{\psi_2} \leq \mu_\psi$  where  $\|\cdot\|_{\psi_2}$  denotes the sub-Gaussian norm. By using Lemmas 11 and 12 we can write

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

where (i) follows from Young's inequality (7) with  $\beta = 1$  and (ii) from (16). We can also use Young's inequality (7) to write  $\left\| \widehat{\nabla} J^\gamma(\theta, \widehat{\Phi}) \right\| \geq -\left\| \widehat{\nabla} J^\gamma(\theta, \widehat{\Phi}) - \widehat{\nabla} J^\gamma(\theta, \widehat{\Phi}) \right\|_F^2 + \frac{1}{2} \left\| \widehat{\nabla} J^\gamma(\theta, \widehat{\Phi}) \right\|_F^2$ , which implies that

$$\begin{aligned} \left\| \widehat{\nabla} J^\gamma(\theta, \widehat{\Phi}) \right\|_F^2 &\leq 4\mu_2 \left\| \nabla J^\gamma(\theta, \Phi) \right\|_F^2 + 2 \left\| \widehat{\nabla} J^\gamma(\theta, \widehat{\Phi}) - \widehat{\nabla} J^\gamma(\theta, \widehat{\Phi}) \right\|_F^2 + 4\mu_2 \ell \left( (L_K \nu_\theta + \phi) \sqrt{2\ell} + \phi \right)^2 d(\widehat{\Phi}, \Phi)^2 \\ &\stackrel{(i)}{\leq} 4\mu_2 \left\| \nabla J^\gamma(\theta, \Phi) \right\|_F^2 + 2\varepsilon^2 + 4\mu_2 \ell \left( (L_K \nu_\theta + \phi) \sqrt{2\ell} + \phi \right)^2 d(\widehat{\Phi}, \Phi)^2, \end{aligned} \quad (17)$$

where (i) is due to Lemma 11. Similarly, we can write

$$\left\langle \nabla J^\gamma(\theta, \widehat{\Phi}), \widehat{\nabla} J^\gamma(\theta, \widehat{\Phi}) \right\rangle \geq \mu_1 \left\| \nabla J^\gamma(\theta, \widehat{\Phi}) \right\|_F^2$$



$$\begin{aligned}
&\geq \frac{\mu_1}{2} \|\nabla J^\gamma(\theta, \Phi)\|_F^2 - \|\nabla J^\gamma(\theta, \hat{\Phi}) - \nabla J^\gamma(\theta, \Phi)\|_F^2 \\
&\geq \frac{\mu_1}{2} \|\nabla J^\gamma(\theta, \Phi)\|_F^2 - \ell \left( (L\nu_\theta + \phi)\sqrt{2\ell} + \phi \right)^2 d(\hat{\Phi}, \Phi)^2,
\end{aligned} \tag{18}$$

along with

$$\begin{aligned}
\langle \nabla J^\gamma(\theta, \hat{\Phi}), \tilde{\nabla} J^\gamma(\theta, \hat{\Phi}) \rangle &= \langle \nabla J^\gamma(\theta, \Phi), \hat{\nabla} J(\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} \|\hat{\nabla} J^\gamma(\theta, \hat{\Phi})\|_F^2 + \frac{1}{2\beta} \|\nabla J^\gamma(\theta, \hat{\Phi}) - \nabla J^\gamma(\theta, \Phi)\|_F^2 + \frac{\beta}{2} \|\nabla J^\gamma(\theta, \hat{\Phi})\|_F^2 \\
&\quad + \frac{1}{2\beta} \|\tilde{\nabla} J^\gamma(\theta, \hat{\Phi}) - \hat{\nabla} J^\gamma(\theta, \hat{\Phi})\|_F^2 \\
&\leq \langle \nabla J^\gamma(\theta, \Phi), \hat{\nabla} J^\gamma(\theta, \hat{\Phi}) \rangle + \frac{\beta}{2} \|\hat{\nabla} J^\gamma(\theta, \hat{\Phi})\|_F^2 + \frac{1}{2\beta} \|\nabla J^\gamma(\theta, \hat{\Phi}) - \nabla J^\gamma(\theta, \Phi)\|_F^2 \\
&\quad + \beta \|\nabla J^\gamma(\theta, \Phi)\|_F^2 + \beta \|\nabla J^\gamma(\theta, \Phi) - \nabla J^\gamma(\theta, \hat{\Phi})\|_F^2 + \frac{1}{2\beta} \|\tilde{\nabla} J^\gamma(\theta, \hat{\Phi}) - \hat{\nabla} J^\gamma(\theta, \hat{\Phi})\|_F^2 \\
&\leq \langle \nabla J^\gamma(\theta, \Phi), \hat{\nabla} J^\gamma(\theta, \hat{\Phi}) \rangle + \frac{\beta}{2} \|\hat{\nabla} J^\gamma(\theta, \hat{\Phi})\|_F^2 + \frac{\ell}{2\beta} \left( (L_K\nu_\theta + \phi)\sqrt{2\ell} + \phi \right)^2 d(\hat{\Phi}, \Phi)^2 \\
&\quad + \beta \|\nabla J^\gamma(\theta, \Phi)\|_F^2 + \beta \ell \left( (L_K\nu_\theta + \phi)\sqrt{2\ell} + \phi \right)^2 d(\hat{\Phi}, \Phi)^2 + \frac{\varepsilon^2}{2\beta} \\
&\stackrel{(i)}{\leq} \langle \nabla J^\gamma(\theta, \Phi), \hat{\nabla} J^\gamma(\theta, \hat{\Phi}) \rangle + \beta(2\mu_2 + 1) \|\nabla J^\gamma(\theta, \Phi)\|_F^2 + \varepsilon^2 \left( 1 + \frac{1}{2\beta} \right) \\
&\quad + \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(\hat{\Phi}, \Phi)^2,
\end{aligned} \tag{19}$$

where (i) follows from (17). Therefore, by setting  $\beta = \frac{\mu_1}{4(2\mu_2+1)}$  and applying (19) in (18), we have the following lemma.

**Lemma 13:** Given positive scalars  $\mu_1, \mu_2$  and  $\zeta$ . Suppose that  $n_s = \mathcal{O}(\zeta^4 \mu_\psi^4 \log^6(\ell)\ell)$ ,  $\tau = \mathcal{O}(\log(1/\varepsilon))$  and  $r = \mathcal{O}(\sqrt{\varepsilon})$ . Then, it holds that

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

with probability  $1 - c_2(\ell^{-\zeta} + n_s^{-\zeta} - n_s e^{-\ell/8} - e^{-c_3 n_s})$ , where  $c_4 = 1 + \frac{2(2\mu_2+1)}{\mu_1}$  and  $c_5 = \frac{2\mu_2\mu_1}{4(2\mu_2+1)} + \frac{2(2\mu_2+1)}{\mu_1} + \frac{\mu_1}{4(2\mu_2+1)}$ .

**1) Cost Estimation Error:** We conclude this subsection by revisiting Lemma 5 from [20] that controls the error in the cost estimation. Let  $\hat{J}^{\gamma, \tau}(\theta, \hat{\Phi}) = \frac{1}{n_c} \sum_{i=1}^{n_c} V^{\gamma, \tau}(\theta, z_0^i)$  be the estimated cost with  $n_c$  samples and  $z_0^i$  denoting a random draw of the low-dimensional initial state.

**Lemma 14:** Given  $\theta \in \mathcal{S}_\theta^\gamma$  and  $\delta_\tau \in (0, 1)$ . Suppose that the time horizon  $\tau$ , number of rollouts  $n_c$ , and subspace distance  $d(\hat{\Phi}, \Phi)$  satisfy

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

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

*Proof:* To prove this lemma we can first write

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

where we can use [20, Lemma 5] to control the first term, i.e., if  $\tau$  and  $n_c$  are set according to the condition of Lemma 14, 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  $|J^\gamma(\theta, \hat{\Phi}) - J^\gamma(\theta)| \leq \ell \|\hat{P}_\theta^\gamma - P_\theta^\gamma\|$ , where we can use the bound for the perturbation of the Lyapunov equation, presented in [11, Proof of Lemma 4] to obtain the following

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

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 in  $\|A\|, \|B\|, \|Q\|, \|R\|, \nu_\theta$ . Therefore, by using [19, Corollary 5.3], we can write  $|J^\gamma(\theta, \hat{\Phi}) - J^\gamma(\theta)| \leq C_{\text{cost}} \ell \sqrt{\ell} d(\hat{\Phi}, \Phi)$ , which completes the proof by setting  $d(\hat{\Phi}, \Phi) \leq \frac{J^\gamma(\theta)}{4\ell\sqrt{\ell}C_{\text{cost}}}$ .  $\blacksquare$

### E. Learning the Left Unstable Subspace Representation

1) *Estimation*: 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 the singular value decomposition of  $D$ , i.e.,  $D = U\Sigma V^\top$ , 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  $\Sigma = \text{diag}(\hat{\sigma}_1, \dots, \hat{\sigma}_{d_x}) \in \mathbb{R}^{d_x \times d_x}$ . We then construct the orthonormal basis for the left unstable subspace with the first  $\ell$  columns of  $U$ , i.e.,  $\hat{\Phi} = [u_1, \dots, u_\ell]$ . Next, let  $\hat{\Pi} = \hat{\Phi}\hat{\Phi}^\top$  and  $\Pi = \Phi\Phi^\top$  denote the projectors onto the subspace spanned by the columns of  $\hat{\Phi}$  and  $\Phi$ , respectively.

We proceed to demonstrate 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 [21, Theorem 5.1], where the main difference in our setting is that we consider a non-diagonalizable  $A$  as well as we construct the basis for the left unstable subspace of  $A$ . Since we assume  $A$  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 eigenspace of  $A^\top$ , respectively. Therefore, we can write

$$A^\top P = P \begin{bmatrix} \tilde{T}_u & 0 \\ 0 & \tilde{T}_s \end{bmatrix},$$

with  $\tilde{P} = [\tilde{\Phi} \ \tilde{\Psi}] \in \mathbb{R}^{d_x \times d_x}$ ,  $\tilde{T}_u \in \mathbb{R}^{\ell \times \ell}$ , and  $\tilde{T}_s \in \mathbb{R}^{d_x - \ell \times d_x - \ell}$ . Note that  $\tilde{T}_u$  has the same spectrum of the matrix composed of Jordan blocks corresponding to the unstable eigenvalues of  $A$  and  $\tilde{T}_s$  has the spectrum of its stable counterpart. By orthonormalizing the basis matrices  $\tilde{\Phi}$  and  $\tilde{\Psi}$  with a thin QR decomposition we obtain

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

with  $R_\Phi$  and  $R_\Psi$  being the upper triangular matrices for the QR decomposition of  $\tilde{\Phi}$  and  $\tilde{\Psi}$ , respectively. Their inverses exist due to the full column rankness of  $\tilde{\Phi}$  and  $\tilde{\Psi}$ . Moreover, we have that the spectra of  $\tilde{T}_u$  and  $T_u$  are the same as well as the spectra of  $\tilde{T}_s$  and  $T_s$ . Let  $\Xi = [\Phi \ \Psi]$  and  $S = [S_1^\top \ S_2^\top]^\top := \Xi^{-1}$ . Then, we can write the following

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

where  $D_1 = S_1 D$  and  $D_2 = S_2 D$ . We note that  $D$  is composed of  $D_u = \Phi D_1$  that comes from the left unstable subspace of  $A$  and  $D_s = \Psi D_2$  that depends on the left stable subspace of  $A$ . Firstly, let us consider  $D_u$  and write the SVD 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  $V_1 \in \mathbb{R}^{T \times d_x}$ .

We know that  $\hat{\Pi}$  is the projector onto the subspace spanned by the first  $\ell$  columns of  $U$ , while  $\Pi$  is the projector onto the columns of  $\Phi U_1$ . In order to use the Davis-Kahan theorem (i.e., Theorem 3) to control  $\|\hat{\Pi} - \Pi\|$ , we first write the following equivalent 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  $[v_i^\top, \pm u_i^\top]^\top$  for all  $i \in [d_x]$ . Moreover, we let  $\{\sigma_j\}_{j=1}^\ell$  denote the top  $\ell$  eigenvalues of  $\mathcal{D}_u$  which are the singular values of  $D_u$ . Therefore, we can use Theorem 3 to write

$$d(\hat{\Phi}, \Phi) = \|\hat{\Pi} - \Pi\| \leq \frac{\sqrt{2\ell} \|D_s\|}{\sigma_\ell - \hat{\sigma}_{\ell+1}} = \frac{\sqrt{2\ell} \|\Psi D_2\|}{\sigma_\ell - \hat{\sigma}_{\ell+1}} \leq \frac{\sqrt{2\ell} \|D_s\|}{\sigma_\ell - \hat{\sigma}_{\ell+1}} = \frac{\sqrt{2\ell} \|D_2\|}{\sigma_\ell - \hat{\sigma}_{\ell+1}}, \quad (20)$$

where we can upper bound  $\|D_2\|$  as follows

$$\|D_2\| \leq \sqrt{T} \|D_2\|_1 \leq \sqrt{T} \sum_{i=\ell+1}^{d_x} \sum_{t=1}^T |\lambda_i|^t \|x_0\| \leq \sqrt{T} \sum_{i=\ell+1}^{d_x} \sum_{t=1}^\infty |\lambda_i|^t \mu_0 \stackrel{(i)}{\leq} \frac{\sqrt{T}(d_x - \ell) \mu_0}{1 - |\lambda_{\ell+1}|}, \quad (21)$$

where (i) is due to the fact that  $\{\lambda_i\}_{i=\ell+1}^{d_X}$  are the stable eigenvalues of  $A$ . Then, by combining (20) and (21) we have

$$d(\hat{\Phi}, \Phi) \leq \frac{\sqrt{2\ell}\sqrt{T}(d_X - \ell)\mu_0}{(\sigma_\ell - \hat{\sigma}_{\ell+1})(1 - |\lambda_{\ell+1}|)}, \quad (22)$$

we now proceed to control  $\sigma_\ell$  and  $\hat{\sigma}_{\ell+1}$ . We first recall that  $\sigma_\ell$  is the  $\ell$ -th top singular value of  $D_1$ . The following lemma from [21] provides a high probability lower bound on such quantity.

*Lemma 15:* 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

$$\sigma_\ell \geq \frac{C_\sigma |\lambda_\ell|^{\frac{T}{2}} |\lambda_1|}{\ell^2 \sqrt{|\lambda_1|^2 - 1}}, \text{ with probability } 1 - 4\delta_\sigma, \text{ where } C_\sigma = \mathcal{O}(1).$$

*Proof:* The proof is similar to the one in [21, Lemma A.1]. The main difference is the way we control the lower bound of the following function

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

where  $S_\ell(1) = \{v \in \mathbb{R}^\ell \mid \min_{1 \leq i \leq \ell} |v_i| \geq 1\}$  is the outbox set as defined in [31], and  $\Lambda_u$  is the Jordan matrix composed with the Jordan blocks of the unstable modes of  $A$ , i.e., the spectrum of  $A_u$ . Then, we can write

$$\phi_{\min}(A_u, T) = \sqrt{\inf_{v \in S_\ell(1)} \sigma_{\min}(H(v)H(v)^\top)}, \text{ with } H(v) = [v \quad \Lambda_u^{-1}v \quad \Lambda_u^{-2}v, \dots, \Lambda_u^{-T+1}v] = \tilde{H}\tilde{V},$$

where  $\tilde{H} = [I \quad \Lambda_u^{-1} \quad \Lambda_u^{-2}, \dots, \Lambda_u^{-T+1}]$  and  $\tilde{V}$  being a  $\ell T \times T$  matrix with the  $v$  vectors placed accordingly. Hence,

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

where  $\|\tilde{H}^\dagger\| = 1/\sqrt{\sigma_{\min}(\tilde{H}\tilde{H}^\top)}$  and  $\sigma_{\min}(\tilde{H}\tilde{H}^\top) = \sigma_{\min}(\sum_{t=0}^{T-1} \Lambda_u^{-t} \Lambda_u^{-t, \top}) \geq \sum_{t=0}^{T-1} \lambda_{\min}(\Lambda_u^{-t} \Lambda_u^{-t, \top})$ . Therefore, we can write

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

and the rest of the proof follows from combining such result with [21, Lemma 16] by adapting to the setting without noise and with non-zero initial condition.  $\blacksquare$

On the other hand, we recall that  $\hat{\sigma}_{\ell+1}$  is the  $\ell+1$ -th singular value of  $D$  which is captured by its stable part  $D_s = \Psi D_2$ . This implies the following upper bound

$$\hat{\sigma}_{\ell+1} \leq \|D_2\| \leq \frac{\sqrt{T}(d_X - \ell)\mu_0}{1 - |\lambda_{\ell+1}|}, \quad (23)$$

where the second inequality follows from (21). Then, by combining Lemma 15 and (23) in (22) we have that

$$\begin{aligned} d(\hat{\Phi}, \Phi) &\leq \frac{\sqrt{2\ell}\sqrt{T}(d_X - \ell)\mu_0/(1 - |\lambda_{\ell+1}|)}{\frac{C_\sigma |\lambda_\ell|^{\frac{T}{2}} |\lambda_1|}{\ell^2 \sqrt{|\lambda_1|^2 - 1}} - \sqrt{T}(d_X - \ell)\mu_0/(1 - |\lambda_{\ell+1}|)} = \frac{\sqrt{2}\ell^{5/2}\sqrt{T}(d_X - \ell)\mu_0}{(1 - |\lambda_{\ell+1}|)C_\sigma |\lambda_\ell|^{T/2} - \sqrt{T}\ell^2(d_X - \ell)\mu_0} \\ &\stackrel{(i)}{\leq} \frac{2\ell^{5/2}\sqrt{T}(d_X - \ell)\mu_0}{(1 - |\lambda_{\ell+1}|)C_\sigma |\lambda_l|^{T/2}}, \end{aligned}$$

where (i) is due to the selection of  $T$  according to  $T \geq 2\log\left(\frac{2\ell^2(d_X - \ell)\mu_0}{(1 - |\lambda_{\ell+1}|)C_\sigma}\right)/\log|\lambda_l|$ . We now proceed to determine the condition of  $T$  to guarantee that  $d(\hat{\Phi}, \Phi) \leq \varepsilon$ , for some small  $\varepsilon$ .

$$\frac{2\ell^{5/2}\sqrt{T}(d_X - \ell)\mu_0}{(1 - |\lambda_{\ell+1}|)C_\sigma |\lambda_l|^{T/2}} \leq \varepsilon, \text{ which implies } T \geq \log\left(\frac{2\ell^{5/2}(d_X - \ell)\mu_0}{(1 - |\lambda_{\ell+1}|)C_\sigma \delta_\sigma \varepsilon}\right)/\log|\lambda_l|,$$

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



### F. Stabilizing Only the Unstable Modes

Given  $\hat{\Phi}$ , we now turn our attention to design a low-dimensional control gain  $\theta$  that stabilizes the low-dimensional system  $(A_u, B_u)$ , namely, the unstable dynamics of  $A$ . For this purpose, we leverage the explicit discount LQR method presented in [20], where the 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 discounted factor needs to guarantee  $\sqrt{\gamma_{j+1}}\rho(A_u + B_u\theta_{j+1}) < 1$  with  $\gamma_{j+1} > \gamma_j$ .

**Lemma 16:** Given a discount factor  $\gamma \in (0, 1]$ , a decay factor  $\xi \in (0, 1)$ , and a controller  $\theta$ , such that  $\sqrt{\gamma}\rho(A_u + B_u\theta) < 1$ . Suppose that  $\tau$  and  $n_c$  satisfy the conditions of Lemma 14, and suppose  $\gamma_+ = (1 + \xi\alpha)\gamma$ , with

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

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

*Proof:* We consider the quadratic Lyapunov function  $V(z_t) = z_t^\top P_\theta^\gamma z_t$  for the damped system  $z_{t+1} = \sqrt{\gamma_+}(A_u + B_u\theta)z_t$  and write the following

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

where (i) follows from the definition of  $P_\theta^\gamma$ . Hence, by guaranteeing that  $\frac{\gamma_+}{\gamma} (P_\theta^\gamma - \Phi^\top Q\Phi - \theta^\top R\theta) - P_\theta^\gamma \prec 0$  we ensure that  $\sqrt{\gamma_+}\rho(A_u + B_u\theta) < 1$ . By taking the trace from both sides of the former inequality, we have the sufficient condition

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

where the last inequality follows from Bauer-Fike Theorem [34] with  $d(\hat{\Phi}, \Phi) \leq \frac{\sigma_{\min}(\hat{\Phi}^\top Q\hat{\Phi} + \theta^\top R\theta)}{4\|Q\|\sqrt{2\ell\kappa}(\hat{\Phi}^\top Q\hat{\Phi} + \theta^\top R\theta)}$ . In particular, since we assume that the initial state is isotropic we have that  $J^\gamma(\theta, \Phi) = \text{Tr}(P_\theta^\gamma)$  which implies that

$$\begin{aligned} \gamma_+ &\leq \left( 1 + \frac{\frac{3}{2}\sigma_{\min}(\hat{\Phi}^\top Q\hat{\Phi} + \theta^\top R\theta)}{J^\gamma(\theta, \Phi) - \frac{3}{2}\sigma_{\min}(\hat{\Phi}^\top Q\hat{\Phi} + \theta^\top R\theta)} \right) \gamma \\ &\stackrel{(i)}{\leq} \left( 1 + \frac{3\sigma_{\min}(\hat{\Phi}^\top Q\hat{\Phi} + \theta^\top R\theta)}{\frac{4}{3}\hat{J}^{\gamma, \tau}(\theta, \hat{\Phi}) - 3\sigma_{\min}(\hat{\Phi}^\top Q\hat{\Phi} + \theta^\top R\theta)} \right) \gamma = (1 + \alpha)\gamma, \end{aligned} \quad (24)$$

where (i) is due to Lemma 14. As discussed in [20, 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.  $\blacksquare$

We now proceed to show that for a sufficiently large amount of PG iterations  $N$ , the LQR cost is uniformly bounded, i.e.,  $J^{\gamma_j}(\theta_{j+1}) \leq \bar{J}$ . For this purpose, given  $\theta_j = \bar{\theta}_0 \in \mathcal{S}_\theta^\gamma$ , we invoke Lemma 1 and 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  $\theta_\star^\gamma$  being the optimal controller of the corresponding discounted LQR problem with discount factor  $\gamma$ . In addition, (i) is due to Lemma 13 and (ii) follows from selecting the step-size according to  $\eta \leq$

$\min \left\{ \frac{\mu_1}{16\mu_2 L_\theta}, \frac{c_4}{2L_\theta \mu_2}, \frac{c_5}{L_\theta} \right\}$ . We use Lemma 2 in (iii). Therefore, we can add and subtract  $J^\gamma(\bar{\theta}^*)$  from both sides to 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

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

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

$$\epsilon \leq \sqrt{\frac{\mu_1(\bar{J} - J_\star^\gamma)}{48\mu_{\text{PL}}c_5}}, \quad d(\hat{\Phi}, \Phi) \leq \sqrt{\frac{\mu_1(\bar{J} - J_\star^\gamma)}{48\mu_{\text{PL}}c_4 \ell \left( (L_K \nu_\theta + \phi)\sqrt{2\ell} + \phi \right)^2}}, \quad \text{and } N \geq \frac{8\mu_{\text{PL}}}{\eta\mu_1} \log \left( \frac{3(J^\gamma(\bar{\theta}_0) - J_\star^\gamma)}{\bar{J} - J_\star^\gamma} \right), \quad (25)$$

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  $\tau$  and  $n_c$  according to Lemma 14 to guarantee that  $\left| \hat{J}^{\gamma_j, \tau}(\theta_{j+1}, \hat{\Phi}) - J^{\gamma_j}(\theta_{j+1}) \right| \leq \frac{1}{2} J^{\gamma_j}(\theta_{j+1})$ , we can write

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

where we can use such lower bound on  $\alpha_j$  to unroll the discount factor update over  $M$  iterations of Algorithm 1 and write

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

which implies that Algorithm 1 finds a stabilizing controller  $\theta_M \in \mathcal{S}_\theta^1$ , i.e.,  $\gamma_M = 1$ , within  $\frac{\log(1/\gamma_0)}{\log(1+\xi\underline{\alpha})}$  iterations. Moreover, from (24) we know that  $\sqrt{(1+\alpha_j)\gamma_j} \rho(A_u + B_u \theta_{j+1}) < 1$ , which can be used to write

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

which guarantees that after  $M$  iterations of the discounted LQR method, it returns a 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 if  $J^{\gamma_j}(\theta_{j+1}) \leq \bar{J}$  and  $\left| \hat{J}^{\gamma_j, \tau}(\theta_{j+1}, \hat{\Phi}) - J^{\gamma_j}(\theta_{j+1}) \right| \leq \frac{1}{2} J^{\gamma_j}(\theta_{j+1})$  hold for the  $j$ -th iteration of Algorithm 1, then it is also true for the subsequent iteration, with high probability. This result guarantees that the lower bound on  $\alpha_j$  is valid for every iteration, thus  $\rho(A_u + B_u \theta_{j+1}) < \bar{\lambda}_\theta$  after a sufficiently large amount of iterations.

*Lemma 17:* If  $J^{\gamma_j}(\theta_{j+1}) \leq \bar{J}$  and  $\left| \hat{J}^{\gamma_j, \tau}(\theta_{j+1}, \hat{\Phi}) - J^{\gamma_j}(\theta_{j+1}) \right| \leq \frac{1}{2} J^{\gamma_j}(\theta_{j+1})$ , 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 follows similarly from [20, Lemma 7] with our adapted definitions of  $\underline{\alpha}$ , and  $\bar{\lambda}_\theta$ .  $\blacksquare$

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 (25) we know that  $J^{\gamma_j}(\theta_{j+1}) \leq \bar{J}$  if

$$N \geq \frac{8\mu_{\text{PL}}}{\eta\mu_1} \log \left( \frac{2\bar{J}^2}{(1-\xi)\sigma_{\min}(Q)J_\star^1} \right), \quad \epsilon \leq \sqrt{\frac{\mu_1 J_\star^1}{48\mu_{\text{PL}}c_5}}, \quad \text{and } d(\hat{\Phi}, \Phi) \leq \sqrt{\frac{\mu_1 J_\star^1}{48\mu_{\text{PL}}c_4 \ell \left( (L\nu_\theta + \phi)\sqrt{2\ell} + \phi \right)^2}},$$

with probability  $1 - (\delta + c_1 N(\ell^{-\zeta} + n_s^{-\zeta} - n_s e^{-\ell/8} - e^{-c_2 n_s}))$ . We complete our analysis by union bounding over all iterations  $j$  of Algorithm 1.

### G. Lifting the Controller

Given  $\theta$  that stabilizes  $(A_u, B_u)$ , i.e.,  $\rho(A_u + B_u\theta) < 1$ , we proceed to prove that  $\rho(A + B\theta\hat{\Phi}^\top) < 1$ . For this purpose, we can write

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

then, we can use Lemma 10 to obtain

$$\begin{aligned} \rho(A + B\theta\hat{\Phi}) &\leq \max \left\{ \rho(A_u + B_u\theta\hat{\Phi}^\top \Phi), \rho(A_s + B_s\theta\hat{\Phi}^\top \Phi_\perp) \right\} + C_{\text{gap}} \|B_u\theta\hat{\Phi}^\top \Phi_\perp\| \|\Delta + B_s\theta\hat{\Phi}^\top \Phi\| \\ &\leq \max \left\{ \rho(A_u + B_u\theta\hat{\Phi}^\top \Phi), \rho(A_s + B_s\theta\hat{\Phi}^\top \Phi_\perp) \right\} + C_{\text{gap}} \|B\| \nu_\theta (\|A\| + \|B\| \nu_\theta) d(\hat{\Phi}, \Phi). \end{aligned} \quad (26)$$

We observe that the second term in the above expression is in the order of the subspace distance. Therefore, we can make it sufficiently small in order 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  $\Phi$  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 when  $A$  is almost symmetric (i.e., when  $A$  is easily decomposable into the stable and unstable modes). Let us now proceed to control  $\rho(A_u + B_u\theta\hat{\Phi}^\top \Phi)$  and  $\rho(A_s + B_s\theta\hat{\Phi}^\top \Phi_\perp)$ .

$$\begin{aligned} \rho(A_u + B_u\theta\hat{\Phi}^\top \Phi) &= \rho(A_u + B_u\theta + B_u\theta(\hat{\Phi}^\top \Phi - I)) \\ &\stackrel{(i)}{\leq} \rho(A_u + B_u\theta) + \max \left\{ \|B_u\theta(\hat{\Phi}^\top \Phi - I)\| C_{\text{bf},1}, \left( \|B_u\theta(\hat{\Phi}^\top \Phi - I)\| C_{\text{bf}} \right)^{1/\ell} \right\} \\ &\stackrel{(ii)}{\leq} \rho(A_u + B_u\theta) + (\|B\| \nu_\theta C_{\text{bf},1})^{1/\ell} d(\hat{\Phi}, \Phi)^{1/\ell} \leq \bar{\lambda}_\theta + (\|B\| \nu_\theta C_{\text{bf},1})^{1/\ell} d(\hat{\Phi}, \Phi)^{1/\ell}, \end{aligned} \quad (27)$$

where  $\bar{\lambda}_\theta := \rho(A_u + B_u\theta)$ , and (i) follows from Lemma 9 with  $C_{\text{bf},1}$  being a constant that depends on the Schur decomposition of  $A_u + B_u\theta$ . (ii) follows from Lemma 1 and  $d(\hat{\Phi}, \Phi) \leq \frac{1}{\|B\| \nu_\theta C_{\text{bf},1}}$ . Similarly, we have that

$$\rho(A_s + B_s\theta\hat{\Phi}^\top \Phi_\perp) \leq |\lambda_{\ell+1}| + (\|B\| \nu_\theta C_{\text{bf},2})^{1/\ell} d(\hat{\Phi}, \Phi)^{1/\ell}, \quad (28)$$

where  $C_{\text{bf},2}$  depends on the Schur decomposition of  $A_s$ . In addition, we require that  $d(\hat{\Phi}, \Phi) \leq \frac{1}{\|B\| \nu_\theta C_{\text{bf},2}}$ . Therefore, we combine (27) and (28) in (26) 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

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

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

**Theorem 4 (Main Result):** Given positive scalars  $\delta_\tau \in (0, 1)$ ,  $\delta_\sigma \in (0, 1)$  and  $\zeta > 0$ . Suppose that the method's parameters are selected as following.

- Gradient and cost estimation parameters:

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



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

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

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

- Time horizon:

$$T = \mathcal{O} \left( \log \left( \frac{\ell^{15/2}(d_{\text{X}} - \ell)\mu_0}{(1 - |\lambda_{\ell+1}|)\varepsilon\delta_\sigma^3} \right) / \log(|\lambda_\ell|) \right),$$

- Algorithm 1 parameters:

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

with  $\bar{J} := \max\{2J_\star^1, J^{\gamma_0}(0)\}$ . Then after  $M \geq \frac{\log(1/\gamma_0)}{\log(1+\xi\alpha)}$  iterations, with  $\underline{\alpha} := \frac{3\sigma_{\min}(Q)}{2\bar{J}-3\sigma_{\min}(Q)}$ , it holds that  $K = \theta_M \hat{\Phi}^\top \in \mathcal{K}$ , i.e.,  $\rho(A + B\theta_M \hat{\Phi}^\top) < 1$ , with probability  $1 - \bar{\delta}$  with  $\bar{\delta} := \delta_\sigma + M(\delta_\tau + \bar{c}_1 N(\ell^{-\zeta} + n_s^{-\zeta} - n_s e^{-\ell/8} - e^{-\bar{c}_2 n_s}))$ .

### H. Sample Complexity

We now proceed to characterize the sample complexity of Algorithm 1 as well as the benefit of learning to stabilize on the unstable subspace. We can quantify the sample complexity by the total number of data points  $x_t$  we query from the system (1) and its adjoint, that is,  $\mathcal{S}_c := \mathcal{S}_c^1 + \mathcal{S}_c^2$ , where  $\mathcal{S}_c^1 := M(n_c + n_s N)\tau$  includes the samples used in the discounted LQR for the low-dimensional system and  $\mathcal{S}_c^2 := T + d_{\text{X}}$  corresponds to the number of data points we need for estimating the left unstable subspace of  $A$ . We emphasize that the extra  $d_{\text{X}}$  term comes from sampling from the adjoint system through element-wise computations via the adjoint operator, as discussed in Section III.

*Corollary 2:* Let the arguments of Theorem 4 hold. Then, Algorithm 1 returns a stabilizing policy for system (1) with

$$\mathcal{S}_c = \underbrace{\log(\rho(A))\tilde{\mathcal{O}}(\ell^2 d_{\text{U}})C_{\text{sc},1}}_{\text{discounted LQR on the low-dimensional dynamics}} + \underbrace{\mathcal{O} \left( \log \left( \frac{\ell^{15/2}(d_{\text{X}} - \ell)C_{\text{sc},2}}{(1 - |\lambda_{\ell+1}|)(1 - \max\{\bar{\lambda}_\theta, |\lambda_{\ell+1}|\})^\ell} \right) \right)}_{\text{learning the left unstable subspace}} + \underbrace{\mathcal{O}(d_{\text{X}})}_{\text{sampling from the adjoint system}}$$

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

We note that the sample complexity is dominated by  $\tilde{\mathcal{O}}(\ell^2 d_{\text{U}}) + \mathcal{O}(d_{\text{X}})$  which scales much slower than  $\tilde{\mathcal{O}}(d_{\text{X}}^2 d_{\text{U}})$  when the number of unstable modes is much smaller than the number of states of the system, i.e.,  $\ell \ll d_{\text{X}}$ .

### I. Numerical Experiments - Supplementary Information

For the numerical experiments provided in Section VI, we consider the following nominal system

$$A_0 = \begin{bmatrix} 2 & 0.25 & 0 & 0 \\ 0 & 2 & -0.63 & 0 \\ 0 & 0 & 1 & 0.25 \\ 0 & 0 & 0.44 & 1 \end{bmatrix}, \quad B_0 = \begin{bmatrix} 0 & 1 \\ 0.13 & 0.5 \\ 0 & 0.75 \\ -0.063 & 1 \end{bmatrix}$$

which is augmented such that  $A = \text{blkdiag}(A_0, \frac{1}{2}(\tilde{A}/\|\tilde{A}\|))$  and  $B = [B_0^\top \quad \tilde{B}^\top]^\top$ , where  $\tilde{A} \in \mathbb{R}^{d_{\text{X}}-4}$  is a diagonal matrix with entries drawn from a standard normal distribution, and  $\tilde{B} \in \mathbb{R}^{(d_{\text{X}}-4) \times d_{\text{U}}}$  is a random matrix with i.i.d. Gaussian entries. Moreover, the horizon length for the left unstable subspace representation learning is set  $T = 40$ , which is enough to guarantee  $d(\hat{\Phi}, \Phi) = 1.23 \times 10^{-9}$ . The zeroth-order estimation and Algorithm 1 parameters are  $\gamma = 0.2$ ,  $n_s = 20$ ,  $n_c = 100$ ,  $\tau = 50$ ,  $r = 1 \times 10^{-5}$ ,  $\eta = 1 \times 10^{-4}$  and  $\xi = 0.9$ .