# Three Problems in the Control and Identification of Structured Linear Systems

by

Han Feng

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Industrial Engineering and Operations Research

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Associate Professor Javad Lavaei, Chair
Professor Venkat Anantharam
Associate Professor Anil Aswani
Assistant Professor Paul Grigas

Spring 2022

Three Problems in the Control and Identification of Structured Linear Systems

# Abstract

Three Problems in the Control and Identification of Structured Linear Systems

by

Han Feng

Doctor of Philosophy in Industrial Engineering and Operations Research

University of California, Berkeley

Associate Professor Javad Lavaei, Chair

We study three problems in the control and identification of structured linear systems. The structure is first manifested as sparsity pattern constraints on the system or control matrices, which complicate the feasible set of the optimal decentralized control problem. We find that the feasible set can be not only disconnected but also have a large number of connected components, which greatly limits the application of local search optimization algorithms. The issue of connectivity is addressed in the second problem, where we design homotopy paths that reduce the number of local minima of the optimal decentralized control problem. Finally, we study an identification scheme based on $l_1$ optimization, where the system states are subject to attacks which propagate over time. The structural constraint, which appears as inequalities involving the states and control inputs, will lead to sufficient conditions for the recovery of system matrices.

# Contents

# List of Figures

# List of Tables

# Acknowledgments

This dissertation is unimaginable without the vision and encouragement of my advisor Javad Lavaei, for whose advice I am grateful. I am fortunate to have excellent team members and friends. From them, I have learned not only excellent research works but also interesting facts about life in general. I appreciate the companionship of Erik Bertelli, Haoyang Cao, Junyu Cao, Yuhao Ding, Salar Fattahi, Dean Grosbard, Pedro Hespanhol, Anran Hu, Titouan Jehl, Hansheng Jiang, Ming Jin, Cedric Josz, Yusuke Kikuchi, Lihua Lei, Kevin Li, Meng Li, Darren Lin, Heyuan Liu, Alfonso Lobos, Igor Molybog, Julie Mulvaney-Kemp, Matt Olfat, Yi Ouyang, Cristobal Pais, SangWoo Park, Georgios Patsakis, Xu Rao, Mark Velednitsky, Renyuan Xu, Nan Yang, Ali Yekkehkhany, Donghao Yin, Haixiang Zhang, Richard Zhang, and Yu Zhang. SangWoo started the Martingales soccer team. The games with the team have taught me not only sportsmanship but also teamwork. Igor organized many trips and I was always surprised at his ability to obtain permits. Mark's game night has been a lot of fun. During my time at Berkeley, I have benefited immensely from the interactions with professors. In particular, I appreciate Professor Ilan Adler for his advice, Professor Venkat Anantharam for his wonderful materials and suggestions, Anil Aswani for his critical thinking on learning and optimization, Alper Atamtürk for his topics on integer programming, Adityanand Guntuboyina for the demonstration of technical excellence and attention to detail, Robert Leachman for his insight into supply chains, Barna Saha for her survey of Approximation Algorithms, and Lisa Pruitt for non-technical ideas commonly missed in an engineering program. The instructors of the Yongmudo Club upended my idea of mentoring and I am grateful for their coaching. During my internship, I was fortunate to be mentored by Andrew Sturges and Richard Chen, who have shown me how to collaborate and work effectively. Finally, I would like to thank my mother, only when I have to deal with children do I realize what it takes to be a good teacher.

# Chapter 1

# Introduction

As the cheapening computing power seeps into our daily lives, critical decisions in large-scale dynamical systems are bit by bit taken over by algorithms which are becoming increasingly complex. It is imperative that we validate the the decision making process, which often takes the form of an iterative optimization algorithm. Control theory is a field where we study problem structure and its interplay with optimization algorithms. The dissertation studies three problems in the theory of structured linear systems. The first two concerns the problem of optimal control where we show that

- The stability constraint of the decisions greatly complicates the geometric structure of the optimal control problem. In particular, when we require that the control policy stabilizes the system, we may have to sacrifice connectivity of the feasible set, which frustrates the attempt of a local search algorithm to find the best control policy.
- A homotopy scheme can overcome the issues of connectivity in the policy domain. This is achieved by constructing a series of artificial systems, eventually leading to one where the connectivity problem disappears and the locally optimal control policy is globally optimal.

For the last problem, we study the identification of an unknown linear dynamical systems under injections of a attacker who can potentially tune his input based on current state and control measurements. We find that

- When the attacker can only attack infrequently, we can shape the joint distribution of input and state pairs such that an accurate model of the linear dynamical system can be recovered in a single trajectory of state measurements.

The following three chapters will state our problem setting and conclusions in more detail. Chapter 2 and Chapter 3 study continuous-time linear systems, while Chapter 4 studies a discrete-time system.

# Chapter 2

# Connectivity Properties of the Set of Stabilizing Static Decentralized Controllers

The optimal decentralized control (ODC) problem is known to be NP-hard [13]. The NP-hardness is reflected in the properties of its feasible set. We study the complexity of the ODC problem through an analysis of the set of stabilizing static decentralized controllers, and show that there is no polynomial upper bound on its number of connected components. In particular, it is proved that this number is exponential in the order of the system for a class of problems. Since every point in each of these components is the unique solution of the ODC problem for some quadratic objective functional, the results of this work imply that, without prior knowledge for initialization, local search algorithms cannot solve the ODC problem to global optimality for all decentralized control structures. In an effort to understand the connection between the geometric properties of the feasible set of the ODC problem and the control structure, we further identify decentralized structures that admit tractable connectivity properties, using a combination of the Routh-Hurwitz criterion and Lyapunov stability theory.

## 2.1 Introduction

Classical state-space solutions to optimal centralized control problems do not scale well as the dimension increases [26]. Moreover, structural constraints such as locality and delay are ubiquitous in real-world controllers. The optimal decentralized control problem (ODC) has been proposed in the literature to bridge this gap. The model has found wide applications in electric power systems and robotics [65, 19, 89, 61]. On the one hand, ODC can have nonlinear optimal solutions even for linear systems and is NP-hard in the worst case [93, 14]. On the other hand, the existence of dynamic structured feedback laws is completely captured by the notion of fixed modes [80], and several works have discovered structural conditions on the system and/or the controller under which the ODC problem admits tractable solutions. The conditions include spatially invariance [5], partially nestedness [78], positiveness [75], and quadratic invariance [56]. More recently, the System Level Approach [90] has convexified structural constraints at the expense of working with a series of impulse response matrices. Promising approximation [30, 2, 59] and convex relaxation techniques [84, 16, 36, 17] also exist in the literature.

A recent line of research, initiated in the machine learning community, suggests using nonlinear programming methods based on local search for the optimal control problems [35]. These methods have been applied to instances of ODC to obtain approximate solutions [91] and to promote sparsity in controllers [57]. Local search methods are well-studied for convex problems, and they normally come with optimality guarantees [16]. However, when the problem is non-convex, these methods may converge to a saddle point or to a local minimum [8]. Local search algorithms are effective: (i) when they are initialized at a point close enough to the optimal solution, or (ii) when there is no spurious local optimum and it is possible to escape saddle points [43, 52, 95, 51]. These conditions are not evidently verifiable for ODC and the question whether local search is effective for ODC remains unanswered.

This chapter shows that the chances of success for the global convergence of local search methods applied to a general ODC problem are theoretically slim. Specifically, we prove that the feasible set of the ODC problem in the static case, which includes all structured static controllers that stabilize the system, can be not only non-convex but also disconnected where the number of connected components grows exponentially in the order of the system. Since any point in the feasible set is the unique globally optimal solution of ODC for some quadratic objective functional, this result implies that no reformulation of the problem with a smooth change of variables could convexify the problem. Moreover, if one seeks to solve a hard instance of the ODC problem through local search, the algorithm needs to be initialized an exponential number of times unless some prior information about the location of the solution is available in order to start in the correct connected component. This result contrasts with the recent findings in [35] and qualifies the applicability of local search methods in optimal control problems.

Although the number of connected components is shown to be exponential in this work, we also demonstrate that favorably structured systems can have a single connected component. In particular, it is proved that the set of static stabilizing controllers is connected for damped systems no matter what the control structure is. Moreover, a bound on the number of connected components is provided in the scalar case. For block structured systems with a sufficient number of free elements, we develop a series of equivalence relations that describe the exact number of connected components of structured stable matrices.

This work is related to several papers in the literature. The set of stabilizing controllers has been studied from many angles. The work [67] parametrizes the set of stable state-feedback controllers under no structural constraints. The paper [66] studies the connectivity of stable linear systems and concludes that single-input single-output systems of order $n$ have at most $n + 1$ connected components, while stable multi-input multi-output systems have only one connected component. The work [6] investigates what types of sparse patterns can sustain stable dynamics using graph theory. For systems with a few parameters, the number of stability regions can be bounded by the number of root-invariant regions using the D-decomposition method [46, 45]. However, the connectivity of decentralized stabilizing controllers, especially for multi-input multi-output systems, lacks a systematic study.

The remainder of this chapter is organized as follows. Notations and problem formulations are given in Section 2.2. We derive elementary connectivity properties of the set of stabilizing controllers and bound the number of connected components for scalar controllers in Section 2.3. Section 2.4 examines a subclass of decentralized control problems for which the number of connected components is exponential, and discusses the implications of this result on the number of locally optimal solutions of ODC. Section 2.5 extends the result to a broad class of controllers with a tri-diagonal-containing structure and shows that the set of stabilizing controllers with a bounded norm has an exponential number of connected components. Section 2.6 proves that highly damped systems admit a connected set of decentralized controllers. The section further discusses how this property could be used to approximate the solution of the ODC problem. Section 2.7 describes the connectivity properties of structured stable matrices with zero blocks. Concluding remarks are drawn in Section 2.8.

## 2.2 Problem Formulation

Consider the linear time-invariant system

$$\begin{aligned}\dot{x}(t) &= Ax(t) + Bu(t),\\ y(t) &= Cx(t),\end{aligned}$$

where $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$ and $C \in \mathbb{R}^{p\times n}$ are real matrices of compatible sizes. The vector $x(t)$ is the state of the system and $y(t)$ is the output. We focus on the static case, where the control input $u(t)$ is to be determined via a static output-feedback law $u(t) = Ky(t)$ with the gain $K \in \mathbb{R}^{m\times p}$ such that some measure of performance is optimized. Since the analysis to follow is on the feasible set, the initial state (being deterministic or stochastic) and the objective function (being quadratic or some other function of the system's signals) are unimportant. With no loss of generality, we assume that the initial state $x(0) = x_0$ is normally distributed with zero mean and unit variance. The quadratic performance measure is defined by

$$J_\lambda(K) = \mathbb{E}\int_0^\infty e^{-\lambda t}\left[x^\top(t)Qx(t) + 2x^\top(t)Du(t) + u^\top(t)Ru(t)\right]dt \tag{2.1}$$

where the matrix $L = \left[\begin{smallmatrix} Q & D\\ D^\top & R\end{smallmatrix}\right]$ is positive semi-definite and $R$ is positive definite. We use the notations $L \succeq 0$ and $R \succ 0$ to denote positive semi-definiteness and positive definiteness, respectively. The discount factor $\lambda \geq 0$. The expectation is taken over $x_0$. The closed-loop system is

$$\dot{x}(t) = (A + BKC)x(t).$$

A matrix is stable, or equivalently Hurwitz, if all its eigenvalues lie in the open left half plane. $K$ is said to stabilize the system if $A + BKC$ is stable. All the matrices considered in this work are real-valued unless otherwise noted. The objective is to study the set of structured stabilizing controllers

$$\mathcal{K}_{\mathcal{S}} = \{K : A + BKC \text{ is stable}, K \in \mathcal{S}\}, \tag{2.2}$$

where $\mathcal{S} \subseteq \mathbb{R}^{m \times p}$ is a linear subspace of matrices, often specified by fixing certain entries of the matrix to zero. Decentralized and distributed controllers could be specified by the set $\mathcal{S}$ with a prescribed sparsity pattern. The set of sparse stable matrices

$$\mathcal{A}_{\mathcal{T}} = \{A : A \text{ stable and } A \in \mathcal{T}\} \tag{2.3}$$

is a special case of (2.2), where $\mathcal{T} \subseteq \mathbb{R}^{n \times n}$ is a linear subspace of matrices. When $\mathcal{T}$ is a linear subspace of sparse matrices, we represent $\mathcal{T}$ with a sparsity pattern where $*$ denotes the positions of entries that can be non-zero. As an example, the set of tri-diagonal matrices can be represented by the following sparsity pattern:

$$\begin{bmatrix} * & * & 0 & \cdots & \cdots & 0 \\ * & * & * & \ddots & & \vdots \\ 0 & * & * & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & \ddots & * \\ 0 & \cdots & \cdots & 0 & * & * \end{bmatrix}.$$

Let $I_{\mathcal{T}} \in \mathcal{T}$ denote the indicator of the sparsity pattern of $\mathcal{T}$ so that $I_{\mathcal{T}}$ has an entry 1 at all positions of $\mathcal{T}$ that can be nonzero and 0 otherwise. The connectivity properties of $\mathcal{K}_{\mathcal{S}}$ and $\mathcal{A}_{\mathcal{T}}$ will be studied under the Euclidean topology. We use $\partial\mathcal{K}_{\mathcal{S}}$ to denote the boundary of the set $\mathcal{K}_{\mathcal{S}}$. The notation $\operatorname{diag}(a_1, \ldots, a_n)$ denotes the $n$-by-$n$ diagonal matrix with diagonal entries $a_1, \ldots, a_n$. We write $\operatorname{tr}(A)$ for the trace of the matrix $A$ and $\|A\|_2$ for the 2-norm of $A$. The notation $\mathbb{E}[X|Y]$ denotes the expectation of the random variable $X$ conditioned on the random variable $Y$.

Geometrically, the set of stable matrices is an open non-convex cone with the origin removed. The sets $\mathcal{K}_{\mathcal{S}}$ and $\mathcal{A}_{\mathcal{T}}$ are obtained by slicing this open cone of stable matrices along an affine subspace and a linear subspace, respectively. The slicing affects the number of connected components for each of these sets and thereby reflects the tractability of the optimal decentralized control problem.

## 2.3 Connectivity Properties in Special Cases

In this section, we prove global geometric properties of the stabilizing set $\mathcal{K}_{\mathcal{S}}$ for certain choices of $B, C$ and $\mathcal{S}$ using elementary arguments.

The stability of matrices can be characterized in different ways. Lyapunov's characterization [28, §4.1] states that a matrix $M$ is stable if and only if there is a solution $P \succ 0$ to the equation $MP + PM^\top + I = 0$. The Routh-Hurwitz criterion [7, §11.17] states that a matrix is stable if and only if the coefficients of its characteristic polynomial satisfy a set of polynomial inequalities. These basic techniques allow us to study the stabilizing set $\mathcal{K}$ when there are no structural constraints and full state measurements.

**Lemma 1.** *Assume that $\mathcal{S} = \mathbb{R}^{m\times p}$ and $C = I$. The set $\mathcal{K}_\mathcal{S}$ is connected, but generally non-convex.*

*Proof.* Observe that $\mathcal{K}_\mathcal{S}$ is the continuous image of the set

$$\mathcal{H} = \{(R, P) : AP + BR + PA^\top + R^\top B^\top = -I, P \succ 0\}$$

through the map $(R, P) \to RP^{-1}$. Moreover, $\mathcal{H}$ is connected since it is the intersection of a linear space and a convex cone. The map is well-defined as $P$ is positive definite; it is also surjective from the Lyapunov's characterization: whenever $A + BK$ is stable, there is a matrix $P \succ 0$ such that $(A + BK)P + P(A + BK)^\top = -I$ and the tuple $(R, P)$ can be mapped to the desired $K$ under the formula $KP = R$.

To show that $\mathcal{K}_\mathcal{S}$ is generally non-convex, consider the second-order system

$$A = \begin{bmatrix} 0 & 1 \\ -a_0 & -a_1 \end{bmatrix}, B = \begin{bmatrix} 0 & b_0 \\ 1 & b_1 \end{bmatrix}, K = \begin{bmatrix} k_{11} & k_{12} \\ k_{21} & k_{22} \end{bmatrix}$$

where $A$ and the first column of $B$ are in the canonical form to ensure controllability. The closed-loop matrix is equal to

$$A + BK = \begin{bmatrix} b_0 k_{21} & 1 + b_0 k_{22} \\ -a_0 + k_{11} + b_1 k_{21} & -a_1 + k_{12} + b_1 k_{22} \end{bmatrix}.$$

To analyze the stability, we use the Routh-Hurwitz criterion and write

$$\mathcal{K}_\mathcal{S} = \{K : \operatorname{tr}(A + BK) < 0, \det(A + BK) > 0\}.$$

Notice that $\mathcal{K}_\mathcal{S}$ is not convex in general since its intersection with the lower dimensional subspace $k_{21} = 0$ is given by

$$\left\{ K = \begin{bmatrix} k_{11} & k_{12} \\ k_{21} & k_{22} \end{bmatrix} : -a_1 + k_{12} + b_1 k_{22} < 0, (1 + b_0 k_{22})(-a_0 + k_{11}) < 0 \right\},$$

which turns out to be the union of two disjoint polyhedrons if $b_0 \neq 0$ (due to the product in the second condition). □

An implication of lemma 1 is that the feasible set of the linear-quadratic optimal centralized control problem is connected, which justifies the success of the local search algorithm proven in [35] for centralized controllers. Another insightful, but impractical, scenario is the case with $B = C = I$ and a mostly arbitrary $\mathcal{S}$. This is studied below.

**Lemma 2.** *Assume that $B = C = I$ and that $\mathcal{S}$ contains $-I$. Then, the set $\mathcal{K}_{\mathcal{S}}$ is connected.*

*Proof.* Since $\mathcal{S}$ is a linear subspace, we have $-\lambda I \in \mathcal{S}$ for every $\lambda \in \mathbb{R}$. Given two arbitrary matrices $K_1, K_2 \in \mathcal{K}_{\mathcal{S}}$, consider the following connected path from $A + K_1$ to $A + K_2$:

$$\begin{aligned} A + K_1 &\overset{\text{increase } \lambda}{\rightarrow} A + K_1 - \lambda I \\ &\overset{K_1 \rightarrow K_2}{\rightarrow} A + K_2 - \lambda I \\ &\overset{\text{decrease } \lambda}{\rightarrow} A + K_2, \end{aligned}$$

where

- $\lambda \geq 0$ is first increased to a large value;
- we move from $A + K_1 - \lambda I$ to $A + K_2 - \lambda I$ via an arbitrary continuous path between $K_1$ and $K_2$ in $\mathcal{S}$;
- $\lambda$ is decreased eventually.

The parameter $\lambda$ can be made so large that all matrices on the path from $A + K_1 - \lambda I$ to $A + K_2 - \lambda I$ could be regarded as a small (on the order of $K_2 - K_1$) perturbation of the large matrix $A + K_1 - \lambda I$. Such small perturbation preserves the stability condition of $A + K_1 - \lambda I$. The proof is completed by noting that the designed path, which connects $K_1$ and $K_2$, involves only controllers in $\mathcal{S}$ and passes through only stabilizing matrices continuously. □

If the measurement matrix $C$ is not the identity matrix, the set could become disconnected even in the simplest case $K = k \in \mathbb{R}$. This is demonstrated in the example below. To differentiate vectors from matrices, we rewrite $B$ as $b$ and $C$ as $c^\top$, where $b$ and $c$ are column vectors in $\mathbb{R}^n$.

**Example 1.** *Assume that $(A, b)$ is controllable and $c \neq 0$, where $A \in \mathbb{R}^{3\times 3}$. Then, the set $\mathcal{K}$ can have at most two connected components. To prove this statement, with no loss of generality we write the system in the controllable canonical form, i.e.,*

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -a_0 & -a_1 & -a_2 \end{bmatrix}, \ b = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \ c^\top = \begin{bmatrix} c_0 & c_1 & c_2 \end{bmatrix}.$$

*The Routh-Hurwitz criterion characterizes stability with the set of inequalities*

$$\begin{aligned} a_0 - kc_0 &> 0, \\ a_1 - kc_1 &> 0, \\ a_2 - kc_2 &> 0, \\ (a_0 - kc_0) &< (a_2 - kc_2)(a_1 - kc_1). \end{aligned}$$

*Consider the quadratic function $f(k) = (a_2 - kc_2)(a_1 - kc_1)$, which can have at most two branches that lie above the line $a_0 - kc_0$. The intersection of these branches with the interval defined by the first three linear inequalities leads to at most 2 connected components. An example with exactly two components can be produced by the parameters*

$$(a_0, a_1, a_2) = (-5, -1, 1), \quad (c_0, c_1, c_2) = (0.85, 0.2, 0.2).$$

*fig. 2.1 verifies the above result by plotting the maximum real part of the closed-loop eigenvalues versus $k$.*

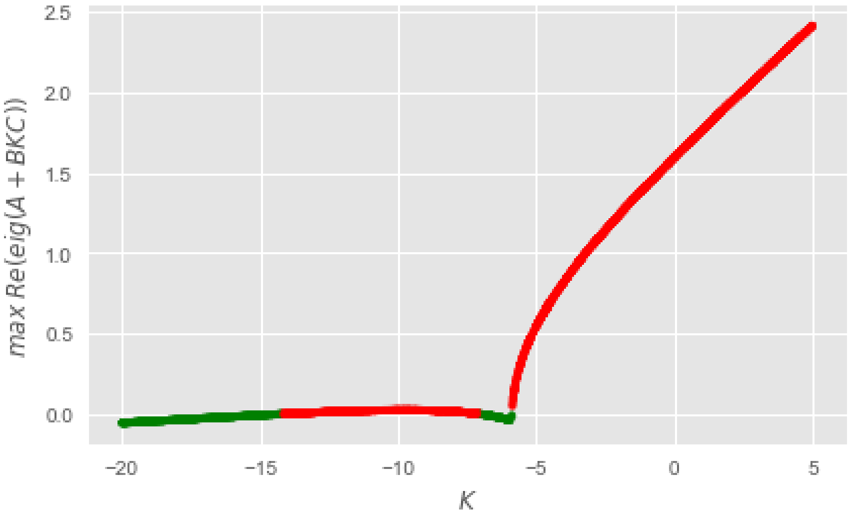


Figure 2.1: As discussed in Example 1, the set of stabilizing controllers can have two connected components for a third-order system. Observe that there are two intervals for $k$ that produce eigenvalues in the left-half complex plane.

It can be inferred from example 1 that the coordinates of the set of stabilizing controllers are "one-sided". This is not surprising since when $A + BKC$ is stable, it holds that $\mathrm{tr}(A + BKC) < 0$. We elaborate on this result in lemma 3.

**Lemma 3.** *Consider the case $m = p = 1$. Suppose that $(A, b)$ is controllable and $c \neq 0$. Then, the scalar set $\mathcal{K}_\mathcal{S}$ cannot extend to infinity on both sides.*

*Proof.* As before, with no loss of generality consider the canonical form

$$A = \begin{bmatrix} 0 & & I \\ -a_0 & \cdots & -a_{n-1} \end{bmatrix}, b = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, c^\top = [c_0, \ldots, c_{n-1}].$$

The matrix $A + bkc^\top$ has the characteristic polynomial

$$(a_0 - c_0k) + (a_1 - c_1k)x + \dots, +(a_{n-1} - c_{n-1}k)x^{n-1} + x^n = 0.$$

It follows from the Routh-Hurwitz criterion that the coefficients of this polynomial must be positive. Since $c \neq 0$, there is some entry $c_{i_0} \neq 0$ and, as a result, $k$ is prevented from extending to infinity on one side due to the inequality $a_{i_0} - c_{i_0}k > 0$. □

In what follows, we will bound the number of connected components for scalar controllers. Compared with [45, Theorem 1], our bound is tighter under the assumption of controllability. We denote by $\lceil \xi \rceil$ the smallest integer greater than or equal to the scalar $\xi$.

**Theorem 1.** *Consider the case $m = p = 1$. Suppose that $(A, b)$ is controllable and $c \neq 0$. The scalar set $\mathcal{K}_\mathcal{S}$ can have at most $\lceil \frac{n}{2} \rceil$ connected components.*

*Proof.* If there is no stabilizing controller in $\mathcal{S}$, then $\mathcal{K}_\mathcal{S} = \emptyset$; otherwise one can first stabilize $A$ with some controller $k_0$ and then analyze the set of shifted controllers $k - k_0$. As a result, without loss of generality one can assume that $A$ is stable. We call a controller $k$ *critical* when it is on the boundary of the set stabilizing controllers, implying the presence of a closed-loop eigenvalue on the imaginary axis. If necessary, we replace $A$ with $A - \epsilon I$ for a small $\epsilon > 0$ so that the number of connected components remains the same and that the intervals of $\mathcal{K}_\mathcal{S}$ share no boundary points. Consider the solution to the equation

$$\begin{aligned} 0 &= \det(\mathbf{j}wI - A - kbc^\top) \\ &= \det(\mathbf{j}wI - A)\det(1 - kc^\top(\mathbf{j}wI - A)^{-1}b) \end{aligned} \tag{2.4}$$

(the symbol $\mathbf{j}$ denotes the imaginary unit). Since $A$ is stable, the first term in the second line of (2.4) is not zero and therefore the second term must be zero. Taking its real and imaginary part yields

$$1 - k \times \mathrm{Re}\{c^\top(\mathbf{j}wI - A)^{-1}b\} = 0, \tag{2.5}$$

$$\mathrm{Im}\{c^\top(\mathbf{j}wI - A)^{-1}b\} = 0. \tag{2.6}$$

Equation (2.6) is of the form $\mathrm{Im}\left\{\frac{f(\mathbf{j}w)}{g(\mathbf{j}w)}\right\} = 0$ with $g(\mathbf{j}w) = \det(\mathbf{j}wI - A) \neq 0$; equivalently, one can write $\mathrm{Im}\{f(\mathbf{j}w)\overline{g(\mathbf{j}w)}\} = 0$ where $f(\mathbf{j}w)$ is a polynomial of degree at most $n-1$, $g(\mathbf{j}w) = \det(\mathbf{j}wI - A)$ is a polynomial of degree $n$, and overline denotes the complex conjugate. $\mathrm{Im}\{f(\mathbf{j}w)\overline{g(\mathbf{j}w)}\}$ is a polynomial of degree $2n-1$ in $w$ with only odd degree terms; it can have at most $2n - 1$ real roots that are symmetric around 0. Because $\mathrm{Re}\{f(\mathbf{j}w)\overline{g(\mathbf{j}w)}\}$ has only even degree terms, at most $n$ distinct pairs of the symmetric roots of eq. (2.6) can be plugged into (2.5). This leads to at most $n$ critical values for the scalar $k$ and divides the real line into at most $n+1$ intervals of interlacing stable-unstable controller regions. At most $\lceil \frac{n+1}{2} \rceil$ of them are stable. Note that when $n + 1$ is odd, lemma 3 rules out one interval that extends to infinity. As a result, the upper bound can be sharpened to $\lfloor \frac{n+1}{2} \rfloor = \lceil \frac{n}{2} \rceil$. □

Theorem 1 states that the number of connected components would grow with the dimension of the system even in the special case $m = p = 1$. Our bound is *tight* when $n = 3$ in light of example 1.

## 2.4 Exponential Subclass

One of the main results of this chapter is stated below.

**Theorem 2.** *There is no polynomial function with respect to the order of the system that can serve as an upper-bound on the number of connected components of the set of decentralized stabilizing controllers.*

To prove the theorem, it suffices to show the existence of a subclass of decentralized control problems whose set of stabilizing controllers has an exponential number of connected components. Our proof requires a lemma that characterizes the stability of tri-diagonal matrices whose diagonal elements are mostly purely imaginary complex numbers. Define the inertia $\mathrm{In}(G)$ of an $n \times n$ matrix $G$ as the triplet $\mathrm{In}(G) = (\alpha(G), \beta(G), \gamma(G))$, where $\alpha(G)$, $\beta(G)$ and $\gamma(G)$ count the eigenvalues of $G$ with positive, negative and zero real parts, respectively.

**Lemma 4** (From [92]). *Consider the tri-diagonal matrix*

$$G = \begin{bmatrix} f_1 + \mathbf{j}g_1 & f_2 & 0 & \cdots & \cdots & 0 \\ -h_2 & \mathbf{j}g_2 & f_3 & \ddots & & \vdots \\ 0 & -h_3 & \mathbf{j}g_3 & f_4 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & -h_{n-1} & \mathbf{j}g_{n-1} & f_n \\ 0 & \cdots & \cdots & 0 & -h_n & \mathbf{j}g_n \end{bmatrix},$$

*where $f_i$, $g_i$ and $h_i$ are real for $i = 1, ..., n$, $f_1 \neq 0$, and $f_i h_i \neq 0$ for $i = 2, \ldots, n$. Then,*

$$In(G) = In(D),$$

*where*

$$D = \mathrm{diag}(f_1, f_1 f_2 h_2, f_1 f_2 f_3 h_2 h_3, \ldots, f_1 \cdots f_n h_2 \cdots h_n).$$

A corollary of lemma 4 for the stability of real tri-diagonal matrices is given below.

**Corollary 1.** *Given the tri-diagonal real matrix $A$ of the form*

$$A = \begin{bmatrix} f_1 & f_2 & 0 & \cdots & \cdots & 0 \\ -h_2 & 0 & f_3 & 0 & & \vdots \\ 0 & -h_3 & 0 & f_4 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & -h_{n-1} & 0 & f_n \\ 0 & \cdots & \cdots & 0 & -h_n & 0 \end{bmatrix}, \tag{2.7}$$

*it holds that*

- *If $f_1 < 0$ and $f_i h_i > 0$ for all $i \in \{2, \dots, n\}$, then $A$ is stable.*
- *If $f_i h_i < 0$ for some index $i \in \{2, \dots, n\}$, then $A$ is unstable.*

**Remark 1.** *Sparse stable matrices theory [6] states that the graph associated with the sparsity pattern of the matrix in* (2.7) *is a chain and has nested Hamiltonian sub-graphs. The graph is sufficient to sustain stable dynamics. Moreover, the sparse matrix subspace is minimally stable because: (i) if $f_1$ is set to zero, then the trace of the matrix becomes zero and therefore at least one eigenvalue should be unstable, (ii) if any non-diagonal element is set to zero, then the matrix decomposes into a block triangular form where the lower diagonal block has a zero trace, leading to instability.*

Due to remark 1, corollary 1 gives necessary and sufficient conditions for the stability of a class of matrices, which can be used to analyze both connected components and separating hyper-surfaces. In what follows, we will first show the possibility of $2^{n-1}$ connected components in the case with a non-identity $C$ and then develop a similar result for $C = I$.

**Theorem 3.** *Let $A \in \mathbb{R}^{n\times n}$ be in the form of* (2.7)*, and set $B \in \mathbb{R}^{n\times(2n-2)}$, $C \in \mathbb{R}^{(2n-2)\times n}$*

*and $K \in \mathbb{R}^{(2n-2)\times(2n-2)}$ to*

$$B = \left[\begin{array}{cccc|cccc} 0 & \cdots & \cdots & 0 & +1 & 0 & \cdots & 0 \\ -1 & \ddots & & \vdots & 0 & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \vdots & \vdots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & 0 & \vdots & & \ddots & +1 \\ 0 & \cdots & 0 & -1 & 0 & \cdots & \cdots & 0 \end{array}\right],$$

$$C = \left[\begin{array}{ccccc} 1 & 0 & \cdots & \cdots & 0 \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 & 0 \\ \hline 0 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & 1 \end{array}\right],$$

$$K = \operatorname{diag}(k_2, \ldots, k_n, k_2, \ldots, k_n).$$

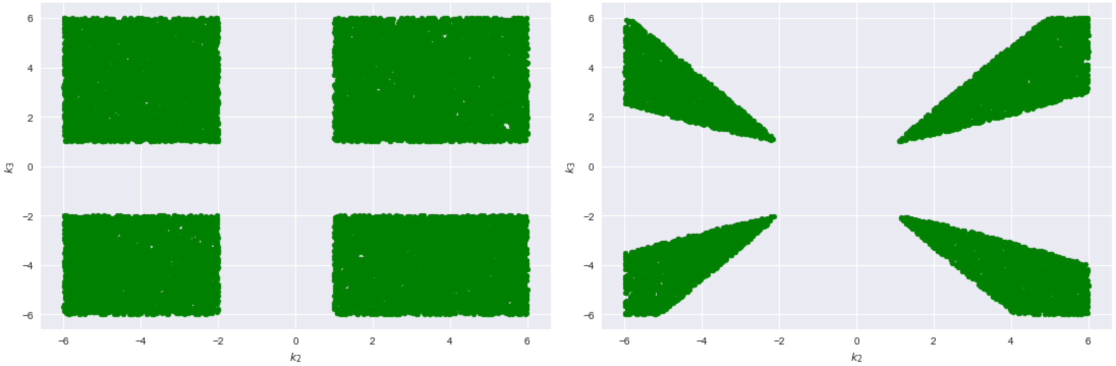
*Suppose that $f_1 < 0$ and $f_i \neq h_i$ for $i = 2, \ldots, n$. Then, the set $\mathcal{K}$ has at least $2^{n-1}$ connected components.*

*Proof.* The closed-loop matrix $A + BKC$ can be expressed as

$$\left[\begin{array}{cccccc} f_1 & f_2 + k_2 & 0 & \cdots & \cdots & 0 \\ -h_2 - k_2 & 0 & f_3 + k_3 & \ddots & & \vdots \\ 0 & -h_3 - k_3 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & 0 & f_n + k_n \\ 0 & \cdots & \cdots & 0 & -h_n - k_n & 0 \end{array}\right].$$

It results from corollary 1 and remark 1 that the closed-loop stability is equivalent to the conditions $(h_i + k_i)(f_i + k_i) > 0$ for $i = 2, \ldots, n$. Equivalently, either $k_i < \min(-h_i, -f_i)$ or $k_i > \max(-h_i, -f_i)$ holds for $i = 2, \ldots, n$. Therefore, the region of stabilizing $K$, parametrized in $(k_2, \ldots, k_n) \in \mathbb{R}^{n-1}$, is separated by $n - 1$ hyperplanes $k_i = -(f_i + h_i)/2$ for $i = 2, \ldots, n$. Since there are stable regions on both sides of each of those hyperplanes, the overall number of connected components becomes at least $2^{n-1}$. □

The result of theorem 3 is demonstrated in the left plot of fig. 2.2 for $n = 3$. Note that the "one-sided" result of lemma 3 does not hold here since $K$ is not a scalar.



(a) $\epsilon = 0$

(b) $\epsilon = 0.2$

Figure 2.2: We randomly sample $K$ and check the closed-loop stability for an instance of the system in theorem 3. The controller is parametrized in terms of $(k_2, k_3)$ where $n = 3$, with $f_i = -1$ and $h_i = 2$ for $i = 1, 2, 3$. The projection of the set $\mathcal{K}$ onto the 2-dimensional space corresponding to $(k_2, k_3)$ is shown in green. The left figure shows that there are $2^{n-1} = 4$ connected components, where each coordinate takes values in $(-\infty, -2)$ or $(1, \infty)$ to be stable. The right figure shows the connected components when the number 0.2 is added to each diagonal entry of $A$.

**Remark 2.** *Note that eigenvalues are continuous functions of the entries of a matrix and that the connected components studied in the proof of theorem 3 are separated by a positive margin. Therefore, one may speculate that a small perturbation of $A$ will not change the number of connected components. This is not the case in general since the eigenvalues of $A+BKC$ can become arbitrarily close to the imaginary axis when $\|K\|$ is large, as illustrated in fig. 2.3. However, one part of every connected component is resistant to perturbations. For example, with $\epsilon > 0$, the set $\{K : (A + \epsilon I) + BKC \text{ stable }\}$ is a subset of $\{K : A + BKC \text{ stable }\}$, the former contains only those controllers that make the closed-loop eigenvalues at least $\epsilon$ away from the imaginary axis. The number $\epsilon$ can be set so small that at least one point from each component remains stable. In other words, a new matrix $A$ obtained by adding $\epsilon$ to the diagonal entries of the matrix in* (2.7) *gives rise of an exponential number of connected components where the number cannot change with a very small perturbation of its elements. This is illustrated in the right plot of fig. 2.2.*

The subclass of problems studied in theorem 3 may be unsatisfactory as it requires that the free elements of $K$ repeat themselves and that $C \neq I$. The next theorem addresses these issues.

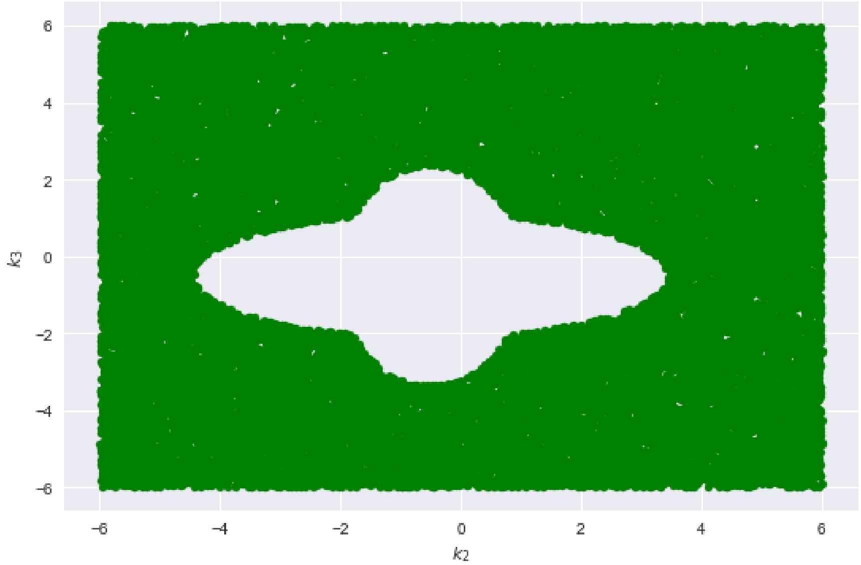


Figure 2.3: If the diagonal entries of $A$ are reduced by 0.2, then the set $\mathcal{K}$ becomes connected. The projection of the set $\mathcal{K}$ onto the 2-dimensional space corresponding to $(k_2, k_3)$ is shown in green.

**Theorem 4.** *Let $A$ be in the form*

$$A = \begin{bmatrix} f_1+\epsilon & f_2 & 0 & \cdots & \cdots & 0 \\ -h_2 & \epsilon & f_3 & \ddots & & \vdots \\ 0 & -h_3 & \epsilon & f_4 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & -h_{n-1} & \epsilon & f_n \\ 0 & \cdots & \cdots & 0 & -h_n & \epsilon \end{bmatrix}, \tag{2.8}$$

*where $\epsilon \geq 0$, $f_1 < 0$, and $(-1)^i(f_i - h_{i+1}) > 0$ for $i = 2, \dots, n$. Consider $B \in \mathbb{R}^{n\times n}$, $C \in \mathbb{R}^{n\times n}$ and $K \in \mathbb{R}^{n\times n}$ to be*

$$B = \begin{bmatrix} 0 & 1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & 0 & 1 \\ & & -1 & 0 \end{bmatrix}, \quad C = I,$$
$$K = \operatorname{diag}(k_1, k_2, \dots, k_n).$$

*For a small enough $\epsilon$, the set $\mathcal{K}$ has at least $fib_n$ connected components, where $fib_0$ =*

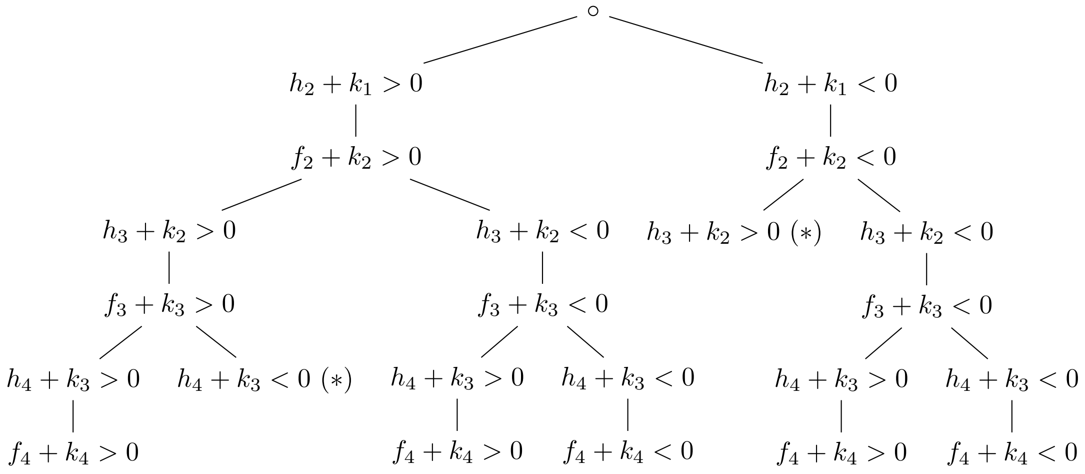


Figure 2.4: This tree shows the enumerating signs of the closed-loop matrix entries for $n = 4$. The branch marked with $(*)$ has contradictory inequalities.

*$1, fib_1 = 1, fib_{i+2} = fib_{i+1} + fib_i$ for $i = 0, 1, ...$ is the Fibonacci sequence, which is on the order of $\left(\frac{1+\sqrt{5}}{2}\right)^n$.*

*Proof.* First, assume that $\epsilon = 0$ and consider the closed-loop matrix $A + BKC$:

$$\begin{bmatrix} f_1 & f_2 + k_2 & 0 & \cdots & \cdots & 0 \\ -h_2 - k_1 & 0 & f_3 + k_3 & \ddots & & \vdots \\ 0 & -h_3 - k_2 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & 0 & f_n + k_n \\ 0 & \cdots & \cdots & 0 & -h_n - k_{n-1} & 0 \end{bmatrix}.$$

In light of corollary 1 and remark 1, the necessary and sufficient conditions for the closed-loop stability are $(h_i + k_{i-1})(f_i + k_i) > 0$ for $i = 2, ..., n$. As a result, if $h_2 + k_1 > 0$, then $f_2 + k_2 > 0$. Now, because $h_3 < f_2$, the term $h_3 + k_2$ can be positive or negative. If it is positive, then $f_3 + k_3$ must be positive, and we can move on to study the sign of $h_4 + k_3$. As we proceed, note that not all sign assignments for $h_i + k_{i-1}$ and $f_i + k_i$ are possible due to the assumptions on $f_i$ and $h_i$. The enumeration procedure is illustrated in fig. 2.4. Any path from the root to the bottom level leaf passes through a set of linear inequalities that together enclose an open polyhedron of stable regions. These stable regions are separated by the hyperplanes $h_{i+1} + k_i = 0$ for $i = 1, 2, \ldots, n - 1$ and $f_i + k_i = 0$ for $i = 2, 3, \ldots, n$.

Next, we count the number of branches. If $h_i + k_{i-1} > 0$ (or equivalently $f_i + k_i > 0$) appears $m_i$ times and $h_i + k_{i-1} < 0$ (or equivalently $f_i + k_i < 0$) appears $n_i$ times, assuming

$m_i \geq n_i$, the next level will have at most $(m_i + n_i) + \max(m_i, n_i) = 2m_i + n_i$ branches. This number is achievable if $f_i < h_{i+1}$, which means keeping all the children of the inequalities $f_i + k_i > 0$ and pruning one child from each inequality $f_i + k_i < 0$. Then, $m_{i+1} = m_i$, $n_{i+1} = m_i + n_i$, and $n_{i+1} \geq m_{i+1}$, which reverses the order of $m_i$ and $n_i$. It can be verified that the total number of connected regions $m_i + n_i$ satisfies the iteration of the Fibonacci sequence.

The connected regions are separated by the hyperplanes $k_i = -f_i$ or $k_i = -h_{i+1}$ with no margin. When $\epsilon > 0$, the connected components are strictly separated. More precisely, whenever $k_i = -f_i$ or $k_i = -h_{i+1}$, the matrix $A + BKC$ decomposes into a block triangular form where the lower diagonal block has a positive trace, which means that the matrix cannot be stable. When $\epsilon$ is small enough, the original connected regions described by linear inequalities do not shrink abruptly — in fact, at least one point from every polyhedron remains stable. As a result, these stable regions are the true connected components of the stabilizing controller set. □

To illustrate theorem 4, consider the matrix

$$A = \begin{bmatrix} -1+\epsilon & 2 & 0 & & & & \\ -2 & \epsilon & 1 & 0 & & & \\ 0 & -1 & \epsilon & 2 & 0 & & \\ & 0 & -2 & \epsilon & 1 & 0 & \\ & & 0 & -1 & \epsilon & 2 & 0 \\ & & & \ddots & \ddots & \ddots & \ddots & \ddots \end{bmatrix}. \tag{2.9}$$

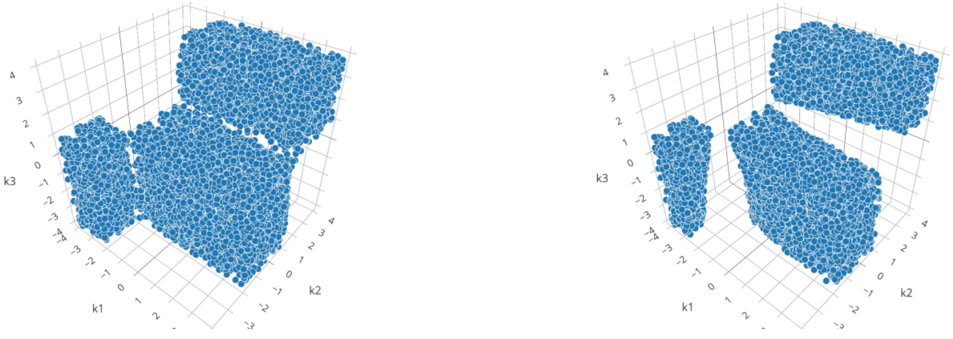
The corresponding set $\mathcal{K}$ obtained by sampling random matrices $K$ and checking the closed-loop stability is provided in fig. 2.5 for $n = 3$.

Our exponential examples are based on specific settings of the parameters $f_i$ and $h_i$ in the matrix $A$ that maximize the number of connected components. We next show that even if the parameters $f_i$ and $h_i$ are considered random, the expected number of connected components is still exponential.

**Theorem 5.** *Consider the matrices $A$, $B$, $C$, and $K$ defined in theorem 4, and let $f_i$ and $h_j$ be independent random variables whose distributions are standard normal for $i = 1, \ldots, n$ and $j = 2, \ldots, n$. If $\epsilon \geq 0$ is small enough, the expected number of connected component of $\mathcal{K}_{\mathcal{S}}$ is at least $\left(\frac{3}{2}\right)^{n-2}$.*

*Proof.* With the assumed distribution, $f_i < h_{i+1}$ and $f_i > h_{i+1}$ occur equally likely, while $f_i = h_{i+1}$ happens with zero probability. Our enumeration tree is random, and we count the number of leaves as follows. If $f_i + k_i > 0$ appears $m_i$ times and $f_i + k_i < 0$ appears $n_i$ times for $i \geq 2$, the next level has two possibilities:

**(i)** $f_i < h_{i+1}$, which keeps all the children of the inequalities $f_i + k_i > 0$ and prunes one child from each inequality $f_i + k_i < 0$. Therefore, $m_{i+1} = m_i$ and $n_{i+1} = m_i + n_i$.



(a) $\epsilon = 0$

(b) $\epsilon = 0.02$

Figure 2.5: We randomly sample $K$ and check the closed-loop stability for an instance of the system in theorem 4 with $n = 3$, the matrix $A$ given in (2.9), and $K = \operatorname{diag}(k_1, k_2, k_3)$. The projection of the set $\mathcal{K}$ onto the 3-dimensional space corresponding to $(k_1, k_2, k_3)$ is shown in blue.

**(ii)** $f_i > h_{i+1}$, which keeps all the children of the inequalities $f_i + k_i < 0$ and prunes one child from each inequality $f_i + k_i > 0$. Therefore, $m_{i+1} = m_i + n_i$ and $n_{i+1} = n_i$.

Combining the two cases, we can calculate the expected number of children $m_{i+1} + n_{i+1}$ conditioned on $m_i$ and $n_i$ in the previous level:

$$\begin{aligned}\mathbb{E}[m_{i+1} + n_{i+1}|m_i, n_i] &= \mathbb{E}[m_{i+1} + n_{i+1}|m_i, n_i, f_{i+1} < h_{i+2}]\mathbb{P}(f_{i+1} < h_{i+2}) \\ &\quad + \mathbb{E}[m_{i+1} + n_{i+1}|m_i, n_i, f_{i+1} > h_{i+2}]\mathbb{P}(f_{i+1} > h_{i+2}) \\ &= (2m_i + n_i)\frac{1}{2} + (2n_i + m_i)\frac{1}{2} = \frac{3}{2}(m_i + n_i).\end{aligned}$$

With the initial conditions $\mathbb{E}[m_2 + n_2|f_1 > 0] = 0$ and $\mathbb{E}[m_2 + n_2|f_1 < 0] = 2$, we have $\mathbb{E}[m_2 + n_2] = 1$. Using induction, it can be concluded that $\mathbb{E}[m_n + n_n] = \left(\frac{3}{2}\right)^{n-2}$. □

By adopting a randomized setting, we are able to analyze the change of connected components when one element $k_{i_0}$ is fixed to zero for some index $i_0 \in \{1, 2, \ldots, n-1\}$. The proof is based on a careful counting of branches and is provided in the Appendix.

**Proposition 1.** *With the same setting as in theorem 5, assume that* $K = \operatorname{diag}(k_1, \ldots, k_n)$ *and* $k_{i_0}$ *is fixed to zero for some index* $i_0 \in \{1, \ldots, n\}$*. Then, the expected number of connected components of* $\mathcal{K}_\mathcal{S}$ *for a small enough* $\epsilon$ *is at least*

$$\begin{cases} \frac{1}{6}(\frac{3}{2})^{n-2}, & \text{if } 2 \le i_0 \le n-1. \\ \frac{1}{2}(\frac{3}{2})^{n-2}, & \text{if } i_0 = 1 \text{ or } i_0 = n. \end{cases}$$

The above results on connectivity reflect not only the computational complexity of the original ODC problem with the hard constraint $K \in \mathcal{K}_{\mathcal{S}}$, but also the complexity of a modified ODC formulation with soft constraints. We explain this implication below. Consider an arbitrary continuous function $h : \mathbb{R}^{m \times p} \to \mathbb{R}$ that satisfies $h(K) = 0$ for all $K \in \mathcal{K}_{\mathcal{S}}$ and $h(K) > 0$ for all $K \in \mathbb{R}^{m \times p} \setminus \mathcal{K}_{\mathcal{S}}$. $h(K)$ serves as a penalty function that can be used to replace the hard constraints of ODC with soft constraints. The penalized form of ODC is given by

$$\min_{K} \quad J_0(K) + c \cdot h(K) \tag{2.10}$$

where $J_0(K)$ is defined in eq. (2.1) and $c$ is a large constant. The above optimization is unconstrained and can be solved using standard numerical algorithms for nonlinear optimization. Indeed, it is common in optimization to convert constrained problems to unconstrained ones via penalty or barrier functions since most efficient numerical algorithms for non-convex optimization are designed for unconstrained problems. The reason for such reformulation is that the constraints do not need to be satisfied in each iteration of a numerical algorithm, and their satisfaction is only required asymptotically when many iterations are taken. In what follows, we study how numerical algorithms perform on the unconstrained formulation eq. (2.10).

**Lemma 5.** *Suppose that $C$ has full row rank and $\left[\begin{smallmatrix} Q & D \\ D^\top & R \end{smallmatrix}\right]$ is positive definite. There are instances of the ODC problem for which the penalized formulation* (2.10) *has an exponential number of local minima if $c$ is sufficiently large.*

*Proof.* Consider any instance of the class of ODC problems provided in theorem 4 for which the feasible set of the problem has an exponential number of connected components. Due to the coercive property proven in Lemma 8 in the Appendix, each connected component in $\mathcal{K}_{\mathcal{S}}$ must have a local minimum for the unpenalized objective $J_0(K)$. Let $\mathcal{O}$ denote the set of all local minima in any arbitrary connected component of the feasible set of ODC, and $\mathcal{O}(\epsilon) \subseteq \mathbb{R}^{m \times p}$ be the set of all points in the feasible set of eq. (2.10) that are at most $\epsilon$ away from $\mathcal{O}$, for any given $\epsilon > 0$. If eq. (2.10) is numerically solved using gradient descent with an initial point in $\mathcal{O}(\epsilon)$, it follows from the proof in [58, §13.1] that the algorithm will converge to a local minimum that is in the interior of $\mathcal{O}(\epsilon)$ and approaches $\mathcal{O}$ as $c$ goes to infinity. This implies that eq. (2.10) has at least one local minimum corresponding to the set $\mathcal{O}$. Therefore, eq. (2.10) has an exponential number of local minima. □

Lemma 5 implies that common first-order and second-order numerical algorithms that work on unconstrained formulations and are guaranteed to converge to a stationary point may end up producing an exponential number of different solutions depending on their initialization.

## 2.5 Bounded Connectivity Number

The results of the preceding section were developed for systems with a very specific structure. We show in this section that for a large class of systems that contain a tri-diagonal structure, there exists a configuration of the matrices $(A, B)$ such that the set of static stabilizing controllers with a bounded norm has an exponential number of connected components. The restriction to a bounded control gain is natural since very high gain controllers cannot be implemented in practice due to the sensitivity of the closed-loop system to noise and disturbance.

Given a linear subspace of sparse matrices[1] $\mathcal{T}$, we say that $\mathcal{T}$ is *tri-diagonal-containing* if it contains all tri-diagonal matrices, i.e.,

$$\mathcal{T} \supseteq \{A : A_{ij} = 0 \text{ for all } |i - j| \geq 2\}.$$

We say that $(A, B)$ is *compatible* with $\mathcal{T}$ if both $A$ and $B$'s sparsity patterns coincide with $I_{\mathcal{T}}$. Since $\mathcal{T}$ is a linear subspace, $A + BK \in \mathcal{T}$ for every diagonal matrix $K$. Given a set $\mathcal{K}$, let $\#\mathcal{K}$ denote the number of connected components of $\mathcal{K}$. Given system matrices $(A, B)$ and a radius $r \geq 0$, we define the set of bounded stabilizing controllers $\mathcal{K}^r(A, B)$ as

$$\mathcal{K}^r(A, B) = \{K : A + BK \text{ stable, } K \text{ diagonal, } \|K\| \leq r\},$$

where $\|\cdot\|$ denotes an arbitrary matrix norm. Note that $\mathcal{K}^\infty(A, B)$ coincides with the set $\mathcal{K}_{\mathcal{S}}$ defined in eq. (2.2) with $\mathcal{S}$ being the set of diagonal matrices. We define the *bounded connectivity number*, which we denote by $c(A, B)$, as follows:

$$c(A, B) = \sup_{r \geq 0} \#\mathcal{K}^r(A, B).$$

The bounded connectivity number quantifies the number of connected components of the set of stabilizing decentralized controllers with a bounded norm in the worst case.

**Theorem 6.** *Given any tri-diagonal-containing sparse matrix subspace $\mathcal{T}$, there exist system matrices $(A, B)$ compatible with $\mathcal{T}$ such that the bounded connectivity number $c(A, B)$ is exponential in the order of the system.*

*Proof.* To prove that $c(A, B)$ is exponential in the order of the system, it suffices to find a radius $r$ and system matrices $(A, B)$ such that $K^r(A, B)$ has an exponential number of connected components and that $(A, B)$ has the same sparsity pattern as $\mathcal{T}$. We start with the matrices $(A, B)$ given in Theorem 4 with an $\epsilon > 0$, which may not be compatible with $\mathcal{T}$. Since $\mathcal{K}^\infty(A, B)$ has an exponential number of connected components, by continuity there exists an $r > 0$ such that $\mathcal{K}^r(A, B)$ has an exponential number of connected components[2].

[1] Recall in Section 2.2 that a linear subspace of sparse matrices is specified by positions of nonzero entries and $I_{\mathcal{T}}$ is the indicator matrix of the non-zero positions.

[2] If there is a connected component of $\mathcal{K}^\infty(A, B)$, it will intersect with a ball $\{K : \|K\| \leq r\}$ where $r$ is large enough, and the intersection will appear as one or more connected components in $\mathcal{K}^r(A, B)$.

Moreover, since $\epsilon > 0$, the connected components of $\mathcal{K}^r(A, B)$ are strictly separated in the sense that every component of $\mathcal{K}^r(A, B)$ is contained in a component of $\mathcal{K}^r(A - \frac{\epsilon}{2}I, B)$, and when $K \in \partial\mathcal{K}^r(A - \frac{\epsilon}{2}I, B)$, the eigenvalues of the closed-loop matrix $A + BK$ is at least $\frac{\epsilon}{2}$ away from the imaginary axis. Since eigenvalues of a matrix are continuous functions of the entries of the matrix and $K$ is bounded, we claim that for all small $\delta > 0$, the set $\mathcal{K}^r(A + \delta I_{\mathcal{T}}, B + \delta I_{\mathcal{T}})$ is also exponential, because (1) by continuity when $\delta > 0$ is small, there exists a controller in each connected component of $\mathcal{K}^r(A, B)$ that remains stabilizing in $\mathcal{K}^r(A + \delta I_{\mathcal{T}}, B + \delta I_{\mathcal{T}})$ and (2) no two connected components of $\mathcal{K}^r(A, B)$ in this bounded region can merge. We elaborate on the second point below. Let $N$ denote the number of connected components of $\mathcal{K}^r(A, B)$. We select one controller from each connected component of $\mathcal{K}^r(A, B)$ and denote them by $K_1, \ldots, K_N$. By continuity, when $\delta$ is small, they remain stabilizing for the system $(A + \delta I_{\mathcal{T}}, B + \delta I_{\mathcal{T}})$. Consider the quantity

$$a(A, B) = \min_{\substack{1 \leq i,j \leq N \\ i \neq j}} \min_{p_{ij} \in P_{ij}} \max_{K \in p_{ij}} \operatorname{spabs}(A + BK) \tag{2.11}$$

where spabs$(\cdot)$ denotes the spectral abscissa (maximum real part of the eigenvalues). The set $P_{ij}$ contains all paths $p_{ij}$ from $K_i$ to $K_j$ such that every controller $K \in P_{ij}$ satisfies $\|K\| \leq r$. We use min instead of inf because the minimum is achievable[3]. We also have $a(A, B) > \frac{\epsilon}{2}$ because all paths $p_{ij} \in P_{ij}$ with $i \neq j$ must intersect with a controller $K \in \partial\mathcal{K}^r(A - \frac{\epsilon}{2}I, B)$, at which point $\operatorname{spabs}(A + BK) > \frac{\epsilon}{2}$. Since the continuous function spabs$(\cdot)$ is absolutely continuous in a compact region, for all small $\delta > 0$, we have $|\operatorname{spabs}(A + BK) - \operatorname{spabs}(A + \delta I_{\mathcal{T}} + (B + \delta I_{\mathcal{T}})K)| < \frac{\epsilon}{4}$ for all $K$ with $\|K\| \leq r$. As a result, $a(A + \delta I_{\mathcal{T}}, B + \delta I_{\mathcal{T}}) > 0$, i.e., $K_1, \ldots, K_N$ belong to different connected components of $\mathcal{K}^r(A + \delta I_{\mathcal{T}}, B + \delta I_{\mathcal{T}})$. The proof is concluded by noting that $\delta$ can be selected so that $(A + \delta I_{\mathcal{T}}, B + \delta I_{\mathcal{T}})$ has the same sparsity pattern as $\mathcal{T}$. □

To understand the implication of Theorem 6, consider a multi-agent system, where each agent has a single state. As long as each agent interacts with its previous and next neighbors, no matter how many more interactions exist in the system, the ODC problem has an exponential number of local solutions for certain system parameters.

## 2.6 Highly Damped Systems

All previous results suggest that the diagonal entries of $A$ being positive contribute to the complexity of the feasible set $\mathcal{K}$. theorem 7 below shows that the diagonal entries of $A$ being negative is a desirable structure in the sense that if $A$ is highly dampened, the feasible set is connected independent of control structures.

[3] Even though the minimization of eq. (2.11) is over an infinite set $P_{ij}$, we can replace it with the minimization over the bounded part of a lower level-set of spabs$(A + BK)$, where the lower level-set is large enough so that $K_i$ and $K_j$ are connected.

**Theorem 7.** *Given arbitrary matrices $A$, $B$ and $C$ of compatible dimensions and a linear subspace of matrices $\mathcal{S}$, the set*

$$\mathcal{K}_{\mathcal{S},\lambda} = \{K : A - \lambda I + BKC \textit{ is stable }, K \in \mathcal{S}\}$$

*is connected when $\lambda > 0$ is large enough.*

*Proof.* First note that the Routh-Hurwitz criterion describes $\mathcal{K}_{\mathcal{S},\lambda}$ by polynomial inequalities in the entries of $A - \lambda I + BKC$, the set $\mathcal{K}_{\mathcal{S},\lambda}$ is semi-algebraic with a finite number of connected components given the order of the system [15]. Consider a number $\mu$ and let $\lambda$ be a parameter that increases from $\mu$ toward $\infty$. Since $\lambda \geq \mu$, we have $\mathcal{K}_{\mathcal{S},\lambda} \supseteq \mathcal{K}_{\mathcal{S},\mu}$, and therefore $\mathcal{K}_{\mathcal{S},\lambda}$ contains all components of $\mathcal{K}_{\mathcal{S},\mu}$ but could possibly connect them or add new components. The addition of new components with the increase of $\lambda$ could occur only a finite number of times. We explain the claim below.

We have noted that set $\mathcal{K}_{\mathcal{S},\lambda}$ is semi-algebraic and is a slice of the set

$$\mathcal{W} = \{(K, \lambda) : A - \lambda I + BKC \text{ is stable }, K \in \mathcal{S}\},$$

which is also semi-algebraic (described by a finite number of polynomial inequalities) and has a finite number of connected components. When a connected component starts to appear for a certain $\lambda_0$, this means that along the direction $(0, 1)$, the set $\mathcal{W}$ has a point of contact, say $K_0$, with a hyperplane orthogonal to $(0, 1)$. If we consider the linear function $(K, \lambda) \to \lambda$ over the set $\mathcal{W}$, $K_0$ is a local minimum of the function and $\lambda_0$ is a critical value. By the semi-algebraic Sard's theorem [15, Theorem 9.6.2], the set of critical values of a linear function over a semi-algebraic set is finite. This proves the claim that as $\lambda$ increase, new components in $\mathcal{K}_{\mathcal{S},\lambda}$ occur for a finite number of times.

To connect all those components, we first increase $\lambda$ until no new connected component appears, then select a controller from each connected component, and cover all those controllers with a ball $\mathcal{B} \subseteq \mathcal{S}$. By making $\lambda$ so large that all controllers in $\mathcal{B}$ become stabilizing, we glue all of the connected components. □

The interpretation of the result of theorem 7 is that if the open-loop matrix of the system can be written as $A - \lambda I$ for a large $\lambda$, then the feasible set of ODC is connected. This corresponds to highly damped systems.

**Remark 3.** *It is noted in [53] that if we consider the discounted cost*

$$J_{2\lambda}(K) = \mathbb{E} \int_0^\infty e^{-2\lambda t}(x^\top Qx + 2x^\top Du + u^\top Ru)dt,$$

*or equivalently make a change of variables $\hat{x}(t) = e^{-\lambda t}x(t)$ and $\hat{u}(t) = e^{-\lambda t}u(t)$, then the closed-loop dynamics become equal to $\dot{\hat{x}}(t) = (A - \lambda I + BKC)\hat{x}(t)$. Therefore, it follows from theorem 7 that the feasible set of the ODC problem is connected for discounted costs with a large discount factor.*

**Remark 4.** *It is known in the context of inverse optimal control [53] that any static state-feedback gain $K$ is the unique minimizer of some quadratic performance measure* (2.1) *for all initial states. One such measure is*

$$\int_0^\infty (u(t) - Kx(t))^\top R\,(u(t) - Kx(t))\,dt.$$

*where $R$ is a positive definite matrix. As a result, every point in any connected component is an optimal solution to some ODC problem. Since there is an exponential number of connected components in certain cases, random initialization is unlikely to successfully locate the optimal component unless prior information is available or the system is favorably structured. Local search algorithms, therefore, fail for general ODC problems.*

A by-product of Theorem 7 is a new controller design strategy, which is based on approximating the ODC problem with another one whose feasible set is connected. This new problem is obtained by damping the system's dynamics. Indeed, we have shown in [38] that minimizing $J_\lambda(K)$ with a large $\lambda$ is more tractable than solving the original ODC problem since the separate connected components will be glued together via damping (as proved in Theorem 7). In the following, we study the cost of this approximation by bounding the ratio of the two objectives.

**Lemma 6.** *Suppose that $\mathbb{E}x_0x_0^\top = I$ and $C = I$. Let $K^+$ be the solution of ODC with the objective function $J_\lambda(K)$ and assume that $K^+$ stabilizes $(A, B)$. Let $W(K^+) = (A+BK^+)+(A+BK^+)^\top$. We have the following upper bound*

$$\frac{J_0(K^+)}{J_\lambda(K^+)} \leq \begin{cases} \frac{\nu_{\min}(W(K^+))-\lambda}{\nu_{\max}(W(K^+))}, & \textit{if } \nu_{\max}(W(K^+)) < 0 \\ \frac{\nu_{\max}(W(K^+))-\lambda}{\nu_{\min}(W(K^+))}, & \textit{if } \nu_{\min}(W(K^+)) > 0 \end{cases}$$

*and lower bound*

$$\frac{J_0(K^+)}{J_\lambda(K^+)} \geq \begin{cases} \frac{\nu_{\max}(W(K^+))-\lambda}{\nu_{\min}(W(K^+))}, & \textit{if } \nu_{\max}(W(K^+)) < \lambda \\ \frac{\nu_{\min}(W(K^+))-\lambda}{\nu_{\max}(W(K^+))}, & \textit{if } \nu_{\min}(W(K^+)) > \lambda \end{cases},$$

*where $\nu_{\min}(\cdot)$ and $\nu_{\max}(\cdot)$ denote the smallest and largest eigenvalues of a matrix, respectively.*

The proof of Lemma 6 is provided in the appendix. We illustrate lemma 6 with a numerical simulation in Figure 2.6. The system matrices are of the form eq. (2.9), which are specified below:

$$A = \begin{bmatrix} -1 & 0.5 \\ -0.5 & 0 \end{bmatrix}, B = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, C = I, K = \operatorname{diag}(k_1, k_2), Q = 5I, R = I, D = 0.$$

Using extensive search, it can be shown that the system has two locally optimal controllers and their undamped costs $J_0(K)$ are as follows:

$$K_1^* \approx \operatorname{diag}(0.7178, 0.6643), \quad J_0(K_1^*) \approx 12.88,$$
$$K_2^* \approx \operatorname{diag}(-1.5384, -1.4369), \quad J_0(K_2^*) \approx 18.08.$$

Starting from the initial stabilizing controller $K_0 = \operatorname{diag}(-2, -2)$, we run gradient descent twice to minimize the cost $J_0(K)$ and its approximate function $J_1(K)$. The step sizes are selected by the Amijo rule as in [38] so that stability is preserved for all iterations. The iterations are stopped when the norm of the gradient is less than $10^{-6}$. When minimizing $J_0(K)$, the iterations converge to $K_2^*$. When minimizing $J_1(K)$, the iterations converge to $K^+ \approx \operatorname{diag}(0.4420, 0.3836)$. We calculate the damped cost $J_1(K^+) \approx 5.98$ and the undamped cost $J_0(K^+) \approx 13.44$. The local search solution to the approximate ODC is better than the solution to the original ODC. With

$$W(K^+) = (A + BK^+) + (A + BK^+)^\top \approx \begin{bmatrix} -3.0000 & -0.0584 \\ -0.0584 & -1.0000 \end{bmatrix},$$

we calculate $\nu_{\max}(W(K^+)) \approx -1.00$ and $\nu_{\min}(W(K^+)) \approx -3.00$. The conclusion of lemma 6 is verified:

$$\begin{aligned} \frac{J_0(K^+)}{J_1(K^+)} &\approx 2.25 < 4.00 \approx \frac{\nu_{\min}(W(K^+)) - 1}{\nu_{\max}(W(K^+))}, \\ \frac{J_0(K^+)}{J_1(K^+)} &\approx 2.25 > 0.67 \approx \frac{\nu_{\max}(W(K^+)) - 1}{\nu_{\min}(W(K^+))}. \end{aligned}$$

## 2.7 Stable Matrices with Block Patterns

In this section, we analyze the connectivity of the set of sparse stable matrices $\mathcal{A}_\mathcal{T}$, defined in (2.3). It follows from lemma 2 that only in matrices with constrained diagonal entries do nontrivial connectivity properties emerge, and we study sparse stable matrices with zero blocks in the diagonal entries.

### Two-by-two block

Below is the main theorem.

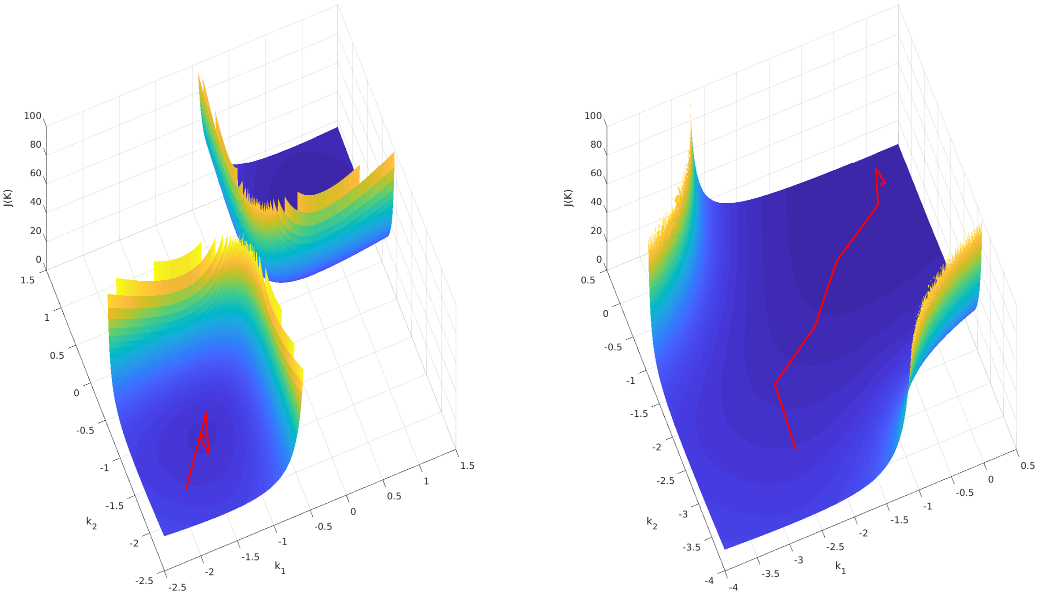
**Theorem 8.** *Consider the matrix subspace*

$$\mathcal{T} = \left\{ \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & 0_{(n-r)\times(n-r)} \end{bmatrix} \middle| A_{21} \in \mathcal{Z}, A_{11} \in \mathbb{R}^{r\times r}, A_{12} \in \mathbb{R}^{r\times(n-r)} \right\},$$

*where $\mathcal{Z}$ is any subspace of matrices in $\mathbb{R}^{(n-r)\times r}$. Then, the sets $\mathcal{A}_\mathcal{T}$ and*

$$\{A_{21} : A_{21} \text{ has full row rank}, A_{21} \in \mathcal{Z}\}$$

*have the same number of connected components.*



(a) minimize $J_0$

(b) minimize $J_1$

Figure 2.6: Cost surface and trajectory of gradient descent in the undamped regime and the damped regime. In the undamped regime, gradient descent is trapped in the initial component. In the damped regime, it almost reaches the globally optimal stabilizing controller.

*Proof.* For clarity the proof is first stated without the constraint $A_{21} \in \mathcal{Z}$; this incurs no loss of generality. $A$ is stable if and only if there is a matrix $P = \begin{bmatrix} P_{11} & P_{12} \\ P_{12}^\top & P_{22} \end{bmatrix} \succ 0$ partitioned accordingly that satisfies the Lyapunov equation

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & 0 \end{bmatrix} \begin{bmatrix} P_{11} & P_{12} \\ P_{12}^\top & P_{22} \end{bmatrix} + \begin{bmatrix} P_{11} & P_{12} \\ P_{12}^\top & P_{22} \end{bmatrix} \begin{bmatrix} A_{11}^\top & A_{21}^\top \\ A_{12}^\top & 0 \end{bmatrix} = \begin{bmatrix} -I & 0 \\ 0 & -I \end{bmatrix}. \tag{2.12}$$

Note that $P$ is unique and depends continuously on $A$ whenever $A$ is stable [28, §4.1]. We solve the partitioned equation

$$A_{11}P_{11} + A_{12}P_{12}^\top + P_{11}A_{11}^\top + P_{12}A_{12}^\top = -I \tag{2.13}$$

$$A_{11}P_{12} + A_{12}P_{22} + P_{11}A_{21}^\top = 0 \tag{2.14}$$

$$A_{21}P_{12} + P_{12}^\top A_{21}^\top = -I. \tag{2.15}$$

Since $P_{22} \succ 0$, eq. (2.14) uniquely determines the unconstrained block

$$A_{12} = -(A_{11}P_{12} + P_{11}A_{21}^\top)P_{22}^{-1}.$$

Substituting it back to (2.13) yields

$$A_{11}P_{11} + P_{11}A_{11}^\top - (A_{11}P_{12} + P_{11}A_{21}^\top)P_{22}^{-1}P_{12}^\top - P_{12}P_{22}^{-T}(A_{21}P_{11} + P_{12}^\top A_{11}^\top) = -I,$$

or equivalently

$$\begin{aligned} A_{11}(P_{11} - P_{12}P_{22}^{-1}P_{12}^\top) + (P_{11} - P_{12}P_{22}^{-1}P_{12}^\top)A_{11}^\top = & \\ - I + P_{11}A_{21}^\top P_{22}^{-1}P_{12}^\top + P_{12}P_{22}^{-T}A_{21}P_{11}. & \end{aligned} \tag{2.16}$$

The equation above can be simplified using the Schur complement $\tilde{P}_{11} = P_{11} - P_{12}P_{22}^{-1}P_{12}^\top$, which is an arbitrary positive definite matrix. One can write

$$\begin{aligned} A_{11}\tilde{P}_{11} + \tilde{P}_{11}A_{11}^\top = -I + \tilde{P}_{11}A_{21}^\top P_{22}^{-1}P_{12}^\top + P_{12}P_{22}^{-T}A_{21}\tilde{P}_{11} + P_{12}P_{22}^{-1}P_{12}^\top A_{21}^\top P_{22}^{-1}P_{12}^\top & \\ + P_{12}P_{22}^{-T}A_{21}P_{12}P_{22}^{-1}P_{12}^\top. & \end{aligned}$$

In light of (2.15), this is equivalent to

$$A_{11}\tilde{P}_{11} + \tilde{P}_{11}A_{11}^\top = -I + \tilde{P}_{11}A_{21}^\top P_{22}^{-1}P_{12}^\top + P_{12}P_{22}^{-T}A_{21}\tilde{P}_{11} - P_{12}P_{22}^{-2}P_{12}^\top. \tag{2.17}$$

Given $A_{21}$, $P_{12}$, $\tilde{P}_{11} \succ 0$, and $P_{22} \succ 0$, the eigenvalues of $\tilde{P}_{11}$ do not sum to zero. Therefore, (2.17) can be regarded as a Lyapunov equation where the unknown block $A_{11}$ has a unique symmetric solution $A_{11} = A_{11}^\top$; all other solutions $A_{11}$ lie in a linear subspace that contains this symmetric solution. The symmetric solution, moreover, depends continuously on $\tilde{P}_{11}$ as long as $\tilde{P}_{11}$ remains in the positive semi-definite cone, which is connected. As a result, not only are all $A_{11}$ connected to a symmetric $A_{11}$, all symmetric $A_{11}$ given $\tilde{P}_{11}$ are connected to the symmetric solution $A_{11}$ given $\tilde{P}_{11} = I$, which we denote by $\phi(A_{12}, P_{12}, P_{22})$:

$$\phi(A_{12}, P_{12}, P_{22}) = \frac{1}{2}\left(-I + A_{21}^\top P_{22}^{-1}P_{12}^\top + P_{12}P_{22}^{-T}A_{21} - P_{12}P_{22}^{-2}P_{12}^\top\right).$$

The above argument retracts the solutions of (2.13)-(2.15) while maintaining the topological property of connectivity. Using $\sim$ to denote the equivalence of connected components, we state the retraction procedure

$$\begin{aligned} \mathcal{A}_\mathcal{T} &\sim \left\{ \left( \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & 0 \end{bmatrix}, \begin{bmatrix} P_{11} & P_{12} \\ P_{12}^\top & P_{22} \end{bmatrix} \right) : (2.12), \begin{bmatrix} P_{11} & P_{12} \\ P_{12}^\top & P_{22} \end{bmatrix} \succ 0 \right\} && (2.18) \\ &\sim \{(A_{11}, A_{21}, P_{11}, P_{12}, P_{22}) : (2.15), (2.16), P_{11} \succ P_{12}P_{22}^{-1}P_{12}^\top, P_{22} \succ 0\} && (2.19) \\ &\sim \{\Big(A_{11}, A_{21}, \tilde{P}_{11}, P_{12}, P_{22}\Big) : (2.15), (2.17), \tilde{P}_{11} \succ 0, P_{22} \succ 0\} && (2.20) \\ &\sim \{(A_{11}, A_{21}, P_{12}, P_{22}) : (2.15), A_{11} = \phi(A_{12}, P_{12}, P_{22}), P_{22} \succ 0\} && (2.21) \\ &\sim \{(A_{21}, P_{12}, P_{22}) : (2.15), P_{22} \succ 0\} && (2.22) \\ &\sim \{(A_{21}, P_{12}) : (2.15)\}. && (2.23) \end{aligned}$$

The first equivalence (2.18) follows from the fact that for any stable matrix $A$, the formula

$$P = \int_0^\infty e^{A\tau} e^{A^\top \tau} d\tau,$$

gives the unique solution to the Lyapunov equation and the solution depends continuously on the matrix $A$. (2.19) follows from the unique solution of $A_{12}$ and the characterization of partitioned positive definite matrices with Schur complements:

$$\begin{bmatrix} P_{11} & P_{12} \\ P_{12}^\top & P_{22} \end{bmatrix} \succ 0 \iff P_{11} \succ P_{12} P_{22}^{-1} P_{12}^\top \text{ and } P_{22} \succ 0.$$

(2.20) follows from the simplification of Lyapunov equation, and the one-one correspondence between $\tilde{P}_{11}$ and $P_{11}$ given $(P_{12}, P_{22})$. (2.21) follows from the retraction of the solutions to (2.17); (2.22) follows from the continuity of function $\phi$, and finally (2.23) throws away the free variable $P_{22}$ because it does not appear in the relationship between $A_{21}$ and $P_{12}$.

(2.23) can be further simplified. We first show that (2.15) has a solution if and only if $A_{21}$ has full rank. If there is a vector $x \in \mathbb{R}^s$ such that $x^\top A_{21} = 0$, pre-multiply and post-multiply (2.15) by $x$ yields

$$0 = x^\top (A_{21} P_{12} + P_{12}^\top A_{21}^\top) x = -x^\top x,$$

or equivalently, $x = 0$. Therefore, $A_{21}$ has full row rank and similarly, $P_{12}$ has full column rank. On the other hand, given any full row rank matrix $A_{21}$, (2.15) has a full rank solution $P_{12} = -1/2 A_{21}^+$, where $A_{21}^+$ is the Moore-Penrose inverse. This completes the proof for the first equivalence in

$$\begin{aligned} \{(A_{21}, P_{12}) : (2.15)\} &\sim \{(A_{21}, P_{12}) : (2.15), A_{21} \text{ has full row rank}\} \\ &\sim \{\left(A_{21}, -1/2 A_{21}^+\right) : A_{21} \text{ has full row rank}\} \\ &\sim \{A_{21} : A_{21} \text{ has full row rank}\}. \end{aligned}$$

The second equivalence follows from the fact that, given $A_{21}$ has full row rank, a solution $P_{12} = -1/2 A_{21}^+$ to (2.15) always exists and all solutions lie in a subspace that can be retracted to that solution. The final equivalence comes from dropping the redundant second coordinate, since the Moore-Penrose inverse is continuous over full rank matrices.

The above proof imposes no restriction on $A_{21}$; it holds even if $A_{21}$ is restricted to a subspace $\mathcal{Z}$. □

In the special case where $\mathcal{Z}$ is the whole space and $A_{21}$ has more columns than rows, the set is connected.

**Corollary 2.** *Assume that $\mathcal{Z} = \mathbb{R}^{(n-r)\times r}$, where $2r > n$. Then, the set $\mathcal{A}_\mathcal{T}$ is connected.*

*Proof.* From theorem 8, if suffices to show the connectivity of

$$\left\{A_{21} \in \mathbb{R}^{(n-r)\times r} : A_{21} \text{ has full row rank}\right\}.$$

This set is the image of the continuous map $(U, D, V) \to UDV$ from the connected set $\mathcal{U} \times \mathcal{D} \times \mathcal{V}$, where

$$\begin{aligned}
\mathcal{U} &= \left\{U \in \mathbb{R}^{(n-r)\times(n-r)} : U \text{ is a orthogonal matrix with determinant } 1\right\}\\
\mathcal{D} &= \left\{D \in \mathbb{R}^{(n-r)\times r} : D_{ii} > 0 \text{ for } i = 1, \dots, r \text{ and all other entries are } 0\right\}\\
\mathcal{V} &= \left\{V \in \mathbb{R}^{r\times r} : V \text{ is a orthogonal matrix with determinant } 1\right\}
\end{aligned}$$

$\mathcal{U}$ and $\mathcal{V}$ are connected because the set of orthogonal matrices with positive determinant is connected. The map is surjective, because every full rank matrix $A_{21}$ has a singular value decomposition $A_{21} = UDV$, where $D_{ii} > 0$ for $i = 1, \dots, r$. If $\det(U) = -1$, we can flip the sign of the first column of $U$ and the first row of $V$ to ensure that $\det(U) = 1$ while preserving the product. If $\det(V) = -1$, we can flip the sign of the last row of $V$, and since $n - r < r$, the last row does not affect the product $UDV$. □

**Corollary 3.** *Suppose* $2r \geq n$ *and* $\mathcal{Z} = \{A_{21} \in \mathbb{R}^{(n-r)\times r} : A_{ij} = 0$ *for* $j \neq i\}$. *Then, the set* $\mathcal{A}_{\mathcal{T}}$ *has* $2^{n-r}$ *connected components.*

*Proof.* We invoke theorem 8. For a diagonal matrix to have full rank, all its diagonal entries must be nonzero, and therefore, every diagonal entry of $A_{21}$ can be either positive or negative. Those $(n - r)$ diagonal entries give rise to $2^{n-r}$ connected components. □

## More Complicated Block Patterns

We generalize the results in the previous section to the case where the space of matrices $\mathcal{T}$ has a block structure as in

$$\mathcal{T} = \left\{ \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & 0_{r\times r} & 0_{r\times(n-2r)} \\ 0_{(n-2r)\times r} & A_{32} & 0_{(n-2r)\times(n-2r)} \end{bmatrix} \middle| A_{21} \in \mathcal{Z}_1, A_{32} \in \mathcal{Z}_2; A_{11}, A_{12} \in \mathbb{R}^{r\times r}; A_{13} \in \mathbb{R}^{r\times(n-2r)} \right\}, \tag{2.24}$$

where $\mathcal{Z}_1 \subseteq \mathbb{R}^{r\times r}$ and $\mathcal{Z}_2 \subseteq \mathbb{R}^{(n-2r)\times r}$ are arbitrary subsets of matrices.

**Theorem 9.** *The set* $\mathcal{A}_{\mathcal{T}}$ *with* $\mathcal{T}$ *defined in* (2.24) *has the same number of connected components as the set*

$$\{(A_{21}, A_{32}) : A_{21} \in \mathcal{Z}_1, A_{32} \in \mathcal{Z}_2, A_{21} \textit{ and } A_{32} \textit{ have full row rank}\}.$$

We provide the proof in the Appendix. The result of theorem 9 is verified for $n = 3$ in Figure 2.7, where 4 connected components are found. In order to strictly separate the components, we plot the samples of sparse stable matrices whose eigenvalues are away from the imaginary axis by a fixed margin.

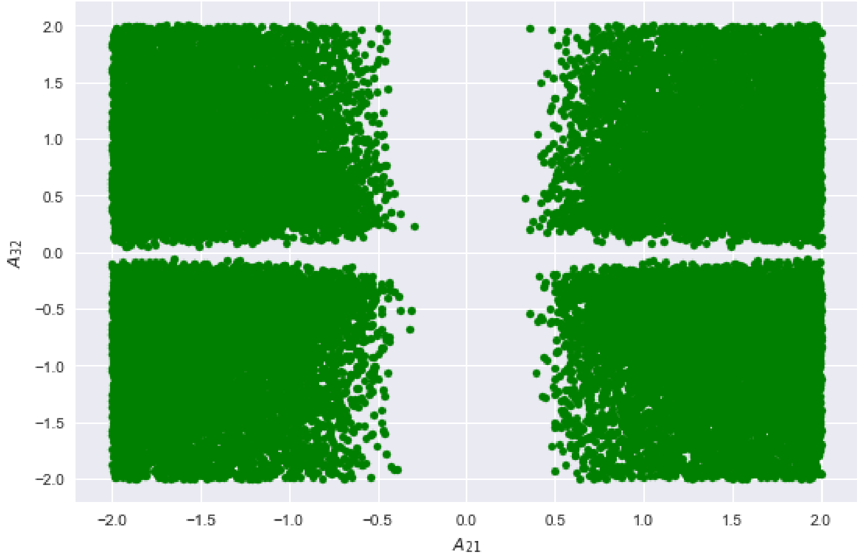


Figure 2.7: Verifying the result of theorem 9 in the case $n = 3$ and $r = 1$, we plot the projection of $A$ onto $(A_{21}, A_{32})$. The entries of the matrix $A$ are sampled uniformly over $[-2, 2]$. The green points marked those matrices $A$ such that $0.2I + A$ is stable.

**Remark 5.** *The result of theorem 9 can be generalized to n-by-n block matrices if the blocks are square and the first row and the lower diagonal blocks of $A$ are nonzero. The square block assumption on the sub-diagonals of $A$ ensures that, for any full rank sub-diagonals, the first row of $A$ and the upper-triangular entries of $P$ can always be solved from the Lyapunov equation. Specially, in case of scalar blocks, the set of stable matrices with the following pattern has $2^{n-1}$ connected components:*

$$\begin{bmatrix} * & * & \cdots & \cdots & * \\ * & 0 & \cdots & \cdots & 0 \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & * & 0 \end{bmatrix}$$

*This relaxes the condition $2r \leq n$ of corollary 3.*

The sparsity pattern discussed in remark 5 seems to suggest that the sparsity of the matrix space directly contributes to the number of connected components. The connection between sparsity and connectivity is complicated in that the number of connected components may remain exponential even when half of the matrix entries are free (such matrices are often regarded as dense).

**Theorem 10.** *The set $\mathcal{A}_{\mathcal{T}}$ has $2^{n-1}$ connected components, where $\mathcal{T}$ is the subset of matrices with the sparsity pattern:*

$$\begin{bmatrix} * & * & * & \cdots & \cdots & * \\ * & 0 & * & \cdots & \cdots & * \\ 0 & * & 0 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \ddots & * \\ 0 & \cdots & \cdots & 0 & * & 0 \end{bmatrix}$$

The theorem can be proved in a same manner as theorem 9 with a different reduction order. The proof is provided in the Appendix.

## 2.8 Conclusion

In this chapter, we studied the connectivity properties of the set of static stabilizing decentralized controllers. We demonstrated through a subclass of problems that the NP-hardness of optimal decentralized control could be attributed to a large number of connected components. In particular, we proved that the number of connected components for chain subsystems would follow a Fibonacci sequence. Even if the elements of the system matrix are random, the expected number of connected components is still exponential. A further implication of our study is that, for any tri-diagonal-containing structure, there exists a system with that structure and certain parameters for which the bounded connectivity number is exponential. The fact that the structure of the decentralized control problem can cause intractability leads to our study of specific system and controller properties that have connectivity guarantees. We bound the number of connected components for the scalar control case. We showed that connectivity would not be an issue for highly damped systems independent of the control structures. In case the system matrix has a certain block structure, we fully characterized the number of connected components. Our results qualified the applicability of local search algorithms to optimal decentralized control problems and emphasized structural considerations.

One future research direction is the analysis of the connectivity properties of dynamic controllers. Dynamic controllers have more flexibility in the choice of parameters and therefore we expect better connectivity properties to hold. On the constructive side, it is important to identify system or control structural properties that guarantee the connectivity of the feasible set. The connectivity result, combined with an analysis of the absence of saddle points, will shed light on the possibility of applying local search algorithms to decentralized control problems.

# Acknowledgment

The authors are grateful to Salar Fattahi for his comments and feedback. The authors are indebted to professor Mehran Mesbahi and Jingjing Bu for fruitful discussions about the bound in Theorem 1. The valuable comments from the anonymous reviewers have led to improvement of the original paper [37], on which this chapter is based. A follow-up paper [12] by researchers in our group has expanded the technique of Section 2.7.

## 2.9 Proofs

### Proof of Proposition 1

*Proof.* We adopt the same notation of $m_i$ and $n_i$ in theorem 5. Let $m'_{i+1}$ and $n'_{i+1}$ denote the number of appearances of $h_{i+1} + k_i > 0$ and $h_{i+1} + k_i < 0$, respectively. In theorem 5, $m'_{i+1} = m_{i+1}$ and $n'_{i+1} = n_{i+1}$. The situation is different when some $k_{i_0}$ is set to zero. We first consider the case $2 \leq i_0 \leq n - 1$.

The random variable $m_i + n_i$ evolves from $i = 1$ to $i = i_0 - 1$ in the same manner as theorem 5. Therefore, given $m_{i_0-1}$ copies of the inequality $f_{i_0-1} + k_{i_0-1} > 0$ and $n_{i_0-1}$ copies of the inequality $f_{i_0-1} + k_{i_0-1} < 0$, conditioned on $m_{i_0-1}$ and $n_{i_0-1}$, we have

$$(m'_{i_0}, n'_{i_0}) = \begin{cases} (m_{i_0-1}, m_{i_0-1} + n_{i_0-1}), & \text{with probability } \frac{1}{2} \\ (m_{i_0-1} + n_{i_0-1}, n_{i_0-1}), & \text{with probability } \frac{1}{2} \end{cases}.$$

Since $k_{i_0}$ is fixed to zero, when $f_{i_0} > 0$, all inequalities $f_{i_0} + k_{i_0} < 0$ are pruned, and when $f_{i_0} < 0$, all inequalities $f_{i_0} + k_{i_0} > 0$ are pruned. Therefore, conditioned on $m'_{i_0}$ and $n'_{i_0}$,

$$(m_{i_0}, n_{i_0}) = \begin{cases} (m'_{i_0}, 0), & \text{with probability } \frac{1}{2} \\ (0, n'_{i_0}), & \text{with probability } \frac{1}{2} \end{cases}.$$

Count similarly $m'_{i_0+1}$ and $n'_{i_0+1}$, we account for the loss of freedom in $h_{i_0+1} + k_{i_0}$:

$$(m'_{i_0+1}, n'_{i_0+1}) = \begin{cases} (m_{i_0}, 0), & \text{with probability } \frac{1}{2} \\ (0, n_{i_0}), & \text{with probability } \frac{1}{2} \end{cases}.$$

After this, the evolution of $(m_i, n_i)$ from $i$ to $i + 1$ is the same as *theorem* 5. It holds that $m_{i_0+1} = m'_{i_0+1}$ and $n_{i_0+1} = n'_{i_0+1}$. In sum,

$$\begin{aligned} \mathbb{E}[m_{i_0+1} + n_{i_0+1} | m_{i_0-1}, n_{i_0-1}] &= \mathbb{E}[m'_{i_0+1} + n'_{i_0+1} | m_{i_0-1}, n_{i_0-1}] \\ &= \frac{1}{2}\mathbb{E}[m_{i_0} + n_{i_0} | m_{i_0-1}, n_{i_0-1}] \\ &= \frac{1}{4}\mathbb{E}[m'_{i_0} + n'_{i_0} | m_{i_0-1}, n_{i_0-1}] \\ &= \frac{3}{8}(m_{i_0-1} + n_{i_0-1}). \end{aligned}$$

Hence, after fixing $k_{i_0} = 0$, the number of children is smaller by a factor of $\frac{1}{6}$ compared with theorem 5.

When $i_0 = 1$, $h_2 + k_1$ appears only once in the tree, and the expected number is cut by one half, because after fixing $k_1 = 0$, either $h_2 > 0$ or $h_2 < 0$ is kept. In the same vein, when $i_0 = n$, only half of the leaves are kept. □

## Proof of Theorem 9

*Proof.* Similar to theorem 8, we first ignore the constraints $A_{21} \in \mathscr{Z}_1$ and $A_{32} \in \mathscr{Z}_2$. $A$ is stable if and only if there is a matrix $P \succ 0$ partitioned accordingly that satisfies the Lyapunov equation

$$\begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & 0 & 0 \\ 0 & A_{32} & 0 \end{bmatrix} \begin{bmatrix} P_{11} & P_{12} & P_{13} \\ P_{21} & P_{22} & P_{23} \\ P_{31} & P_{32} & P_{33} \end{bmatrix} + \begin{bmatrix} P_{11} & P_{12} & P_{13} \\ P_{21} & P_{22} & P_{23} \\ P_{31} & P_{32} & P_{33} \end{bmatrix} \begin{bmatrix} A_{11}^\top & A_{21}^\top & 0 \\ A_{12}^\top & 0 & A_{32}^\top \\ A_{13}^\top & 0 & 0 \end{bmatrix} = -I. \quad (2.25)$$

The solution $P$ is unique whenever $A$ is stable.

We first show that

$$A_{21} \text{ and } A_{32} \text{ have full row rank.} \quad (2.26)$$

Consider the $(2,2)$ and $(3,3)$ blocks of eq. (2.25):

$$A_{21}P_{12} + P_{21}A_{21}^\top = -I \quad (2.27)$$
$$A_{32}P_{23} + P_{32}A_{32}^\top = -I. \quad (2.28)$$

If $x^\top A_{32} = 0$, conjugate (2.28) with $x$ to obtain

$$0 = x^\top (A_{32}P_{23} + P_{32}A_{32}^\top)x = -x^\top x,$$

or equivalently, $x = 0$, which means that $A_{32}$ has full row rank. Similarly, $A_{21}$ has full row rank.

Next we consider the $(1,3)$ and $(2,3)$ blocks of eq. (2.25):

$$A_{11}P_{13} + A_{12}P_{23} + A_{13}P_{33} + P_{12}A_{32}^\top = 0 \quad (2.29)$$
$$A_{21}P_{13} + P_{22}A_{32}^\top = 0. \quad (2.30)$$

Because $P_{33}$ is invertible, $A_{13}$ can be uniquely determined from (2.29). Because $A_{21}$ is full row rank and square, $P_{13}$ can be uniquely determined from (2.30). The equation corresponding to the remaining blocks after eliminating $A_{13}$ can be extracted by pre-multiply (2.25) by

$$W = \begin{bmatrix} I & 0 & -P_{13}P_{33}^{-1} \\ 0 & I & -P_{23}P_{33}^{-1} \end{bmatrix},$$

and post-multiply (2.25) by $W^\top$, which yields

$$\begin{bmatrix} A_{11} & A_{12} - P_{13}P_{33}^{-1}A_{32} \\ A_{21} & -P_{23}P_{33}^{-1}A_{32} \end{bmatrix} \begin{bmatrix} \bar{P}_{11} & \bar{P}_{12} \\ \bar{P}_{21} & \bar{P}_{22} \end{bmatrix} + \begin{bmatrix} \bar{P}_{11} & \bar{P}_{12} \\ \bar{P}_{21} & \bar{P}_{22} \end{bmatrix} \begin{bmatrix} A_{11}^\top & A_{21}^\top \\ A_{12}^\top - A_{32}^\top P_{33}^{-1}P_{32} & -A_{32}^\top P_{33}^{-1}P_{32} \end{bmatrix} \\ = \begin{bmatrix} -I - P_{13}P_{33}^{-2}P_{31} & -P_{13}P_{33}^{-2}P_{32} \\ -P_{23}P_{33}^{-2}P_{31} & -I - P_{23}P_{33}^{-2}P_{32} \end{bmatrix}. \quad (2.31)$$

where the partitioned Schur complement $\bar{P}_{ij}$ is equal to $P_{ij} - P_{i3}P_{33}^{-1}P_{3j}$ for $i, j = 1, 2$. The $(1, 2)$ and $(2, 2)$ blocks of (2.31) are

$$A_{11}\bar{P}_{12} + (A_{12} - P_{13}P_{33}^{-1}A_{32})\bar{P}_{22} + \bar{P}_{11}A_{21}^\top - \bar{P}_{12}A_{32}^\top P_{33}^{-1}P_{32} = -P_{13}P_{33}^{-2}P_{32} \quad (2.32)$$

$$A_{21}\bar{P}_{12} + \bar{P}_{21}A_{21}^\top = -I - P_{23}P_{33}^{-2}P_{32} + P_{23}P_{33}^{-1}A_{32}\bar{P}_{22} + \bar{P}_{22}A_{32}^\top P_{33}^{-1}P_{32}. \quad (2.33)$$

Since $\bar{P}_{22}$ is invertible, $A_{12}$ can be uniquely determined from (2.32). (2.33) is the same as (2.27) given (2.28) and (2.30). Eliminate $A_{12}$ similarly by conjugating (2.31) with $[\, I \;\; \bar{P}_{12}\bar{P}_{22}^{-1} \,]$, which yields

$$(A_{11} - \bar{P}_{12}\bar{P}_{22}^{-1}A_{21})\tilde{P}_{11} + \tilde{P}_{11}(A_{11}^\top - A_{21}^\top\bar{P}_{22}^{-1}\bar{P}_{21}) = *, \quad (2.34)$$

where $\tilde{P}_{11} = \bar{P}_{11} - \bar{P}_{12}\bar{P}_{22}^{-1}\bar{P}_{21}$, and the right hand side is a negative definite matrix determined by $P$. Since $\tilde{P}_{11}$ is positive definite, its eigenvalue do not sum up to zero; therefore, the solution $A_{11}$ always exists and can be shrunk to a symmetric solution that depends continuously on $P$, as explained in theorem 8. Using $\sim$ to denote the equivalence of connected components,

$$\mathcal{A}_\mathcal{T} \sim \left\{ \left( \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & 0 & 0 \\ 0 & A_{32} & 0 \end{bmatrix}, \begin{bmatrix} P_{11} & P_{12} & P_{13} \\ P_{21} & P_{22} & P_{23} \\ P_{31} & P_{32} & P_{33} \end{bmatrix} \right) : (2.25), \begin{bmatrix} P_{11} & P_{12} & P_{13} \\ P_{21} & P_{22} & P_{23} \\ P_{31} & P_{32} & P_{33} \end{bmatrix} \succ 0, (2.26) \right\} \quad (2.35)$$

$$\sim \left\{ \left( \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & 0 \end{bmatrix}, A_{32}, P_{23}, P_{33}, \begin{bmatrix} \bar{P}_{11} & \bar{P}_{12} \\ \bar{P}_{21} & \bar{P}_{22} \end{bmatrix} \right) : (2.28), (2.31), P_{33} \succ 0, \right. \\ \left. \begin{bmatrix} \bar{P}_{11} & \bar{P}_{12} \\ \bar{P}_{21} & \bar{P}_{22} \end{bmatrix} \succ 0, (2.26) \right\} \quad (2.36)$$

$$\sim \left\{ \left( A_{11}, A_{21}, A_{32}, P_{23}, P_{33}, \bar{P}_{12}, \bar{P}_{22}, \tilde{P}_{11} \right) : (2.28), (2.33), (2.34), \right. \\ \left. P_{33} \succ 0, \bar{P}_{22} \succ 0, \tilde{P}_{11} \succ 0, (2.26) \right\} \quad (2.37)$$

$$\sim \left\{ \left( A_{21}, A_{32}, P_{23}, P_{33}, \bar{P}_{12}, \bar{P}_{22} \right) : (2.28), (2.33), P_{33} \succ 0, \bar{P}_{22} \succ 0, (2.26) \right\} \quad (2.38)$$

$$\sim \left\{ \left( A_{21}, A_{32}, P_{33}, \bar{P}_{22} \right) : P_{33} \succ 0, \bar{P}_{22} \succ 0, (2.26) \right\} \quad (2.39)$$

$$\sim \{ (A_{21}, A_{32}) : (2.26) \}. \quad (2.40)$$

The first equivalence (2.35) is justified as in (2.18), with the additional condition that $A_{21}$ and $A_{32}$ must have full row rank. (2.36) follows from the unique continuous solution of $A_{13}$ and

$P_{13}$ in (2.29)-(2.30). (2.37) follows from the unique solution of $A_{12}$ in (2.32). (2.38) follows from the retraction of the solutions to (2.34). Since $A_{32}$ has full row rank, (2.28) is always solvable in $P_{23}$, and the solution subspace can be retracted to the pseudo-inverse solution $P_{23} = 1/2A_{32}^{+}$, which is a continuous function over the full-rank matrix $A_{32}$. The same argument applies to (2.33), where the solution $\bar{P}_{12}$ always exists and can be continuously retracted to the pseudo-inverse solution. This arrives at (2.39). (2.40) discards the redundant coordinates.

The proof above imposes no restriction on $A_{21}$ and $A_{32}$; it holds with any additional subspace constraint on them. □

## Proof of theorem 10

*Proof.* We show the proof for the case $n = 3$; the proof carries over to the general case. The idea is the same as theorem 9, with minor differences in the reduction order and in the justification for full-rank blocks. Consider the solution pair $(A, P)$ to the Lyapunov equation

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & 0 & a_{23} \\ 0 & a_{32} & 0 \end{bmatrix} \begin{bmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{bmatrix} + \begin{bmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{bmatrix} \begin{bmatrix} a_{11} & a_{21} & 0 \\ a_{12} & 0 & a_{32} \\ a_{13} & a_{23} & 0 \end{bmatrix} = -I. \tag{2.41}$$

where $P \succ 0$ is unique whenever $A = \left[\begin{smallmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & 0 & a_{23} \\ 0 & a_{32} & 0 \end{smallmatrix}\right]$ is stable. Consider the $(1,3)$, $(2,3)$ and $(3,3)$ blocks of eq. (2.41),

$$a_{11}p_{13} + a_{12}p_{23} + a_{13}p_{33} + p_{12}a_{32} = 0 \tag{2.42}$$

$$a_{21}p_{13} + a_{23}p_{33} + p_{22}a_{32} = 0 \tag{2.43}$$

$$a_{32}p_{23} + p_{32}a_{32} = -1. \tag{2.44}$$

Since $p_{33}$ is invertible, $a_{13}$ and $a_{23}$ are uniquely determined from (2.42) and (2.43). The equation in the remaining blocks after eliminating $a_{13}$ and $a_{23}$ can be extracted by pre-multiply (2.41) by

$$W = \begin{bmatrix} 1 & 0 & -p_{13}p_{33}^{-1} \\ 0 & 1 & -p_{23}p_{33}^{-1} \end{bmatrix}$$

and post-multiply (2.41) by $W^{\top}$:

$$\begin{bmatrix} a_{11} & a_{12} - p_{13}p_{33}^{-1}a_{32} \\ a_{21} & -p_{23}p_{33}^{-1}a_{32} \end{bmatrix} \begin{bmatrix} \bar{p}_{11} & \bar{p}_{12} \\ \bar{p}_{21} & \bar{p}_{22} \end{bmatrix} + \begin{bmatrix} \bar{p}_{11} & \bar{p}_{12} \\ \bar{p}_{21} & \bar{p}_{22} \end{bmatrix} \begin{bmatrix} a_{11} & a_{21} \\ a_{12} - a_{32}p_{33}^{-1}p_{32} & -a_{32}p_{33}^{-1}p_{32} \end{bmatrix}$$
$$= \begin{bmatrix} -1 - p_{13}p_{33}^{-2}p_{31} & -p_{13}p_{33}^{-2}p_{32} \\ -p_{23}p_{33}^{-2}p_{31} & -1 - p_{23}p_{33}^{-2}p_{32} \end{bmatrix}, \tag{2.45}$$

where the partitioned Schur complement $\bar{p}_{ij}$ is equal to $p_{ij} - p_{i3}p_{33}^{-1}p_{3j}$ for $i, j = 1, 2$. The $(1,2)$ and $(2,2)$ blocks of (2.45) are

$$a_{11}\bar{p}_{12} + (a_{12} - p_{13}p_{33}^{-1}a_{32})\bar{p}_{22} + \bar{p}_{11}a_{21} - \bar{p}_{12}a_{32}p_{33}^{-1}p_{32} = -p_{13}p_{33}^{-2}p_{32} \tag{2.46}$$

$$a_{21}\bar{p}_{12} + \bar{p}_{21}a_{21} = -1 - p_{23}p_{33}^{-2}p_{32} + p_{23}p_{33}^{-1}a_{32}\bar{p}_{22} + \bar{p}_{22}a_{32}p_{33}^{-1}p_{32}. \tag{2.47}$$

Similarly, since $\bar{p}_{22}$ is invertible, $a_{12}$ can uniquely solved from (2.46). Eliminating $a_{12}$ similarly by conjugating (2.45) with $\left[\begin{smallmatrix} 1 & \bar{p}_{12}\bar{p}_{22}^{-1} \end{smallmatrix}\right]$ gives

$$(a_{11} - \bar{p}_{12}\bar{p}_{22}^{-1}a_{21})\tilde{p}_{11} + \tilde{p}_{11}(a_{11} - a_{21}\bar{p}_{22}^{-1}\bar{p}_{21}) = * \tag{2.48}$$

where $\tilde{p}_{11} = \bar{p}_{11} - \bar{p}_{12}\bar{p}_{22}^{-1}\bar{p}_{21}$ and the right hand side is a negative definite matrix determined by $P$. Because $\tilde{p}_{11}$ is positive definite, its eigenvalues do not sum up to zero. As a result, the solution $a_{11}$ always exists and can be shrunk to a symmetric solution that depends continuously on $P$. We retract the solution set, where $\sim$ denotes the equivalence of connected components:

$$\begin{aligned}
\mathcal{A}_{\mathcal{T}} &\sim \left\{ \left( \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & 0 & a_{23} \\ 0 & a_{32} & 0 \end{bmatrix}, \begin{bmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{bmatrix} \right) : (2.41), \begin{bmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{bmatrix} \succ 0 \right\} \\
&\sim \left\{ \left( \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & 0 \end{bmatrix}, a_{32}, p_{13}, p_{23}, p_{33}, \begin{bmatrix} \bar{p}_{11} & \bar{p}_{12} \\ \bar{p}_{21} & \bar{p}_{22} \end{bmatrix} \right) : (2.44), (2.45), p_{33} \succ 0, \begin{bmatrix} \bar{p}_{11} & \bar{p}_{12} \\ \bar{p}_{21} & \bar{p}_{22} \end{bmatrix} \succ 0 \right\} \\
&\sim \{ (a_{11}, a_{21}, a_{32}, p_{13}, p_{23}, p_{33}, \bar{p}_{12}, \bar{p}_{22}, \tilde{p}_{11}) : (2.44), (2.47), (2.48), \\
&\qquad\qquad\qquad\qquad\qquad\qquad p_{33} \succ 0, \bar{p}_{22} \succ 0, \tilde{p}_{11} \succ 0 \} \\
&\sim \{ (a_{21}, a_{32}, p_{13}, p_{23}, p_{33}, \bar{p}_{12}, \bar{p}_{22}) : (2.44), (2.47), p_{33} \succ 0, \bar{p}_{22} \succ 0 \}\,.
\end{aligned}$$

The equivalence is justified similarly. We first add an additional the Lyapunov matrix $P$ and then repeatedly discard the upper-triangular entires of $A$, which are uniquely solved, while transforming the representation of $P$ with the Schur complement until we reach (2.48), which is always solvable in $a_{11}$. This discarding procedure produces a series of equations in the form of (2.47) and (2.44). Since scalar multiplication commutes, we substitute (2.44) to (2.47) and find that the right hand side of (2.47) is strictly less than zero, hence $a_{21} \neq 0$. In the same vein, (2.44) implies $a_{32} \neq 0$. We have proved that all lower sub-diagonal entries of $A$ cannot be zero. With nonzero $a_{21}$ and $a_{32}$, the remaining equations uniquely determine the sub-diagonal entries $(\bar{p}_{12}, p_{23})$, we arrive at the final series equivalences:

$$\begin{aligned}
\mathcal{A}_{\mathcal{T}} &\sim \{ (a_{21}, a_{32}, p_{13}, p_{23}, p_{33}, \bar{p}_{12}, \bar{p}_{22}) : (2.44), (2.47), p_{33} > 0, \bar{p}_{22} > 0, a_{32} \neq 0, a_{21} \neq 0 \} \\
&\sim \{ (a_{21}, a_{32}, p_{13}, p_{33}, \bar{p}_{22}) : p_{33} > 0, \bar{p}_{22} > 0, a_{32} \neq 0, a_{21} \neq 0 \} \\
&\sim \{ (a_{21}, a_{32}) : a_{32} \neq 0, a_{21} \neq 0 \}\,.
\end{aligned}$$

After discarding the redundant coordinates, we are left with $n-1$ nonzero conditions on the sub-diagonals of $A$, which give rise to $2^{n-1}$ connected components. □

## Proof of Lemma 6

The proof follows directly from the lemma below.

**Lemma 7.** *Suppose that* $\mathbb{E}x_0x_0^\top = I$*,* $C = I$ *and* $K$ *stabilizes both* $(A{-}\mu I, B)$ *and* $(A{-}\lambda I, B)$*. Define* $W(K) = (A + BK) + (A + BK)^\top$*. We have the following bound*

$$\frac{J_{2\mu}(K)}{J_{2\lambda}(K)} \leq \begin{cases} \frac{2\lambda - \nu_{\min}(W(K))}{2\mu - \nu_{\max}(W(K))}, & \textit{if } 2\mu > \nu_{\max}(W(K)) \\ \frac{2\lambda - \nu_{\max}(W(K))}{2\mu - \nu_{\min}(W(K))}, & \textit{if } 2\mu < \nu_{\min}(W(K)) \end{cases}.$$

*Proof.* The quadratic costs $J_{2\lambda}(K)$ and $J_{2\mu}(K)$ can be written as $\mathrm{tr}(P_\lambda(K))$ and $\mathrm{tr}(P_\mu(K))$, where

$$(A{-}\lambda I{+}BK)^\top P_\lambda(K) + P_\lambda(K)(A{-}\lambda I{+}BK) + K^\top RK + Q + DK + K^\top D^\top = 0 \tag{2.49a}$$

$$(A{-}\mu I{+}BK)^\top P_\mu(K) + P_\mu(K)(A{-}\mu I{+}BK) + K^\top RK + Q + DK + K^\top D^\top = 0. \tag{2.49b}$$

Taking the difference of eq. (2.49a) and eq. (2.49b) yields

$$(A{+}BK)^\top(P_\lambda(K){-}P_\mu(K)) + (P_\lambda(K){-}P_\mu(K))(A{+}BK) = 2\lambda P_\lambda(K) - 2\mu P_\mu(K). \tag{2.50}$$

Taking the trace of eq. (2.50), we obtain

$$\begin{aligned} &2\lambda\,\mathrm{tr}(P_\lambda(K)) - 2\mu\,\mathrm{tr}(P_\mu(K)) \\ &= \mathrm{tr}\left(((A + BK) + (A + BK)^\top)P_\lambda(K)\right) - \mathrm{tr}\left(((A + BK) + (A + BK)^\top)P_\mu(K)\right) \\ &\geq \nu_{\min}(W(K))\,\mathrm{tr}(P_\lambda(K)) - \nu_{\max}(W(K))\,\mathrm{tr}(P_\mu(K)), \end{aligned}$$

where the last step follows from the positive-semidefinite property of $P_\lambda(K)$ and $P_\mu(K)$. In the same vein,

$$2\lambda\,\mathrm{tr}(P_\lambda(K)) - 2\mu\,\mathrm{tr}(P_\mu(K)) \leq \nu_{\max}(W(K))\,\mathrm{tr}(P_\lambda(K)) - \nu_{\min}(W(K))\,\mathrm{tr}(P_\mu(K)).$$

Hence, if $2\mu > \nu_{\max}(W(K))$, we have

$$\mathrm{tr}(P_\mu(K)) \leq \frac{2\lambda - \nu_{\min}(W(K))}{2\mu - \nu_{\max}(W(K))}\,\mathrm{tr}(P_\lambda(K));$$

and if $2\mu < \nu_{\min}(W(K))$, we have

$$\mathrm{tr}(P_\mu(K)) \leq \frac{2\lambda - \nu_{\max}(W(K))}{2\mu - \nu_{\min}(W(K))}\,\mathrm{tr}(P_\lambda(K)).$$

□

## Proof of Coerciveness

We show that the ODC problem has a certain structure that disallows the locally optimal stabilizing $K$ to have arbitrarily large magnitude.

**Lemma 8.** *Consider the ODC problem with cost eq.* (2.1)*. Suppose that $C$ has full row rank, $L = \left[\begin{smallmatrix} Q & D \\ D^\top & R \end{smallmatrix}\right]$ is positive definite, $D_0 = \mathbb{E}x_0 x_0^\top$ is positive definite, and $K \in \mathcal{S}$ is stabilizing. Then, $J_0(K) \to \infty$ whenever $\|K\|_2 \to \infty$ or when $K$ approaches the boundary of the set of stabilizing controllers.*

*Proof.* We write

$$P(K) = \int_0^\infty e^{t(A+BKC)^\top} \hat{R}(K) e^{t(A+BKC)} \mathrm{d}t,$$

where

$$\hat{R}(K) = Q + DKC + C^\top K^\top D^\top + C^\top K^\top RKC.$$

When $K$ is stabilizing, $P(K)$ is well-defined. As $K$ approaches a finite $K_\dagger$ on the boundary of the set of stabilizing controllers, we show that $\|P(K)\|_2 \to \infty$. By assumption, the symmetric matrix $\hat{R}(K)$ in the integral is positive definite, because it can be written as

$$\hat{R}(K_\dagger) = \begin{bmatrix} I & C^\top K_\dagger^\top \end{bmatrix} L \begin{bmatrix} I \\ K_\dagger C^\top \end{bmatrix}.$$

Therefore, its minimum eigenvalue $\nu_{\min}(\hat{R}(K_\dagger)) > 0$, and when $K$ is close to $K_\dagger$, $\hat{R}(K) \succeq \frac{1}{2}\nu_{\min}(\hat{R}(K_\dagger))I$. We make the estimate

$$\begin{aligned} \mathrm{tr}(P(K)) &\geq \frac{1}{2}\nu_{\min}(\hat{R}(K_\dagger)) \int_0^\infty \mathrm{tr}\left(e^{t(A+BKC)^\top} e^{t(A+BKC)}\right) dt \\ &\geq \frac{1}{2}\nu_{\min}(\hat{R}(K_\dagger)) \int_0^\infty \|e^{t(A+BKC)}\|_2^2 dt \\ &= \frac{1}{2}\nu_{\min}(\hat{R}(K_\dagger)) \int_0^\infty e^{2t\cdot\mathrm{spabs}(A+BKC)} dt, \end{aligned}$$

where spabs$(\cdot)$ denotes the spectral abscissa (maximum real part of the eigenvalues). The estimate above shows that $\mathrm{tr}(P(K)) \to \infty$ as $K$ approaches $K_\dagger$ from the stabilizing set. Since $J_0(K) = \mathrm{tr}(P(K)D_0) \geq \mathrm{tr}(P(K))\nu_{\min}(D_0)$, $J_0(K)$ also approaches infinity.

In case $\|K\|_2 \to \infty$ from the stabilizing set, we use the fact that $P(K)$ is the unique solution to the equation

$$(A + BKC)^\top P + P(A + BKC) + \hat{R}(K) = 0.$$

Let $\sigma_{\min}(C)$ denote the smallest singular value of $C$, which is positive by assumption. From the triangle inequality,

$$\begin{aligned}\nu_{\min}(R)\sigma_{\min}(C)^2\|K\|_2^2 &\leq \|C^\top K^\top RKC\|_2 \\ &\leq 2\|A+BKC\|_2\|P(K)\|_2 + \|Q\|_2 + 2\|D\|_2\|K\|_2\|C\|_2 \\ &\leq 2(\|A\|_2 + \|B\|_2\|K\|_2\|C\|_2)\|P(K)\|_2 + \\ &\qquad\qquad \|Q\|_2 + 2\|D\|_2\|K\|_2\|C\|_2,\end{aligned}$$

Therefore,

$$\|P(K)\|_2 \geq \frac{\nu_{\min}(R)\sigma_{\min}(C)^2\|K\|_2^2 - \|Q\|_2 - 2\|D\|_2\|K\|_2\|C\|_2}{2(\|A\|_2 + \|B\|_2\|K\|_2\|C\|_2)}.$$

Hence, $\|P(K)\|_2 \to \infty$ as $\|K\|_2 \to \infty$ inside the stabilizing set. Similarly $J(K) = \mathrm{tr}(P(K)D_0) \geq \|P(K)\|_2\nu_{\min}(D)$ also approaches infinity. □

# Chapter 3

# Damping with Varying Regularization in Optimal Decentralized Control

This chapter studies a homotopy continuation method for the design of an optimal static or dynamic decentralized controller to minimize a quadratic cost functional. The proposed method involves a combination of the classical local search technique in the space of control policies, a gradual damping of the system dynamics, and a gradual variation of a parametrized cost functional. A series of optimal decentralized control (ODC) problems is generated via a continuous variation of parameters. Unlike the classical homotopy literature, which focuses on tracking a specific trajectory, we study the ensemble of critical controller trajectories and show how the properties of the ensemble can be leveraged to find a globally optimal solution of the ODC problem. After guaranteeing the continuity and asymptotic properties of the proposed method, we prove that with enough damping, there is no spurious locally optimal controller for a block-diagonal control structure. This leads to a sufficient condition under which an iterative algorithm can find a global solution to a class of optimal decentralized control problems. The "damping property" introduced in this analysis is shown to be unique for general system matrices. Empirical observations are presented for instances with an exponential number of locally optimal decentralized controllers, where the developed method could find the global solution even when initialized at a poor local solution.

## 3.1 Introduction

The optimal decentralized control (ODC) problem has found a wide range of applications in electric power systems and robotics [65, 19, 89, 61]. The problem differs from the classical centralized optimal control problem by an additional constraint on the control architecture, which breaks the separation principle and the classical solution formulae [26]. Furthermore, it renders the computational model intractable to solve in general [93]. To bridge the gap between model complexity and tractability, researchers have looked into convex reformulations under various assumptions [17, 78, 75, 56, 90]. Convex formulations have attractive theoretical properties, but they may not be exact and often cause the problem's dimension to explode. In practice, homotopy methods [18, 97, 1, 20, 73] and local-search algorithms [44, 62] are much more appealing and yet theoretically poorly understood.

The objective of this chapter is to study under what conditions ODC problems can be solved to global optimality using low-complexity numerical algorithms. To address this problem, the chapter attempts to delineate the boundary of tractable ODC instances through the lens of homotopy continuation methods. We propose a homotopy scheme that varies the cost and feasible region of the ODC problems. The attractive properties of the homotopy paths lead to sufficient conditions for the designed homotopy algorithm to find a globally optimal decentralized controller in the presence of many local solutions.

Our theoretical results on homotopy paths are related to the line of research on the landscape of non-convex optimization problems [51, 96, 55, 27, 52, 43]. The study of the landscape informs when local search methods are able to obtain high-quality solutions [48, 44, 35]. The ODC problem is distinguished from a general non-convex optimization problem studied in the machine learning literature due to a stability constraint. A recent investigation of the topological properties of ODC in [40] shows that the region of stabilizing controllers can be disconnected and that the number of locally optimal solutions to ODC can grow exponentially in the order of the system. This confirms that the landscape of ODC is highly complex.

The theoretical understanding of homotopy methods in control theory is limited, and no theoretical results are known for the ODC problem to explain when and what homotopy strategies are effective. A theoretical analysis of homotopy methods in the context of ODC is challenging, as illustrated by examples in [60] showing that the general homotopy setting can cause ill-behaviors such as stable-unstable interlaces and discontinuous solution paths. The theoretical challenges have not prevented homotopy methods from successful applications in the numerical solution of optimal control problems [68, 20, 73]. The damping technique in this chapter is similar to the idea in [18], where the author has proposed a homotopy map that connects a stable system to the original system to obtain a stabilizing controller and empirically documented its performance. The paper [97] has considered the $H_2$ reduced-order problem and proposed several homotopy maps and initialization strategies.

Compared with those earlier works, we analyze a specific type of continuation, namely, damping with varying regularization, and show how this method may escape unwanted local minima of the ODC problem. Moreover, diverging from the classical analysis of homotopy methods that focuses on tracking a specific trajectory, we study the ensemble of critical controller trajectories and how the tracking of those trajectories leads to the globally optimal controller.

The key contribution of this chapter is a theoretical analysis of the continuity and asymptotic properties of the trajectories of the locally optimal solutions with respect to the variation of the damping and regularization parameters. After formulating the problem and introducing the homotopy scheme in Section 3.2, we delineate the continuity and asymptotic properties of the proposed damping strategies in Section 3.3 and Section 3.4, respectively. Notably, the analysis leads to the result that if the system dynamics is dampened enough, as long as the condition number of the regularization matrices remains bounded, there is no spurious locally optimal controller, by which we mean all locally optimal controllers are globally optimal for the damped system. Furthermore, we show that this globally optimal controller in the damped system can be continuously connected to the globally optimal controller in the original system via a variation of the homotopy method, provided that the globally optimal decentralized controllers are unique in the damping process. Numerical experiments are detailed in Section 3.5, followed by concluding remarks in Section 3.6. Some of the proofs are relegated to the last chapter.

## 3.2 Homotopy for Optimal Decentralized Control

With no loss of generality, we study the optimal decentralized control problem (ODC) with a static controller and a quadratic cost. The reason is that the design of a dynamic controller can be reformulated as the design of a static controller for an augmented system, as discussed in Section 3.5.B. Consider the linear time-invariant system

$$\dot{x}(t) = Ax(t) + Bu(t),$$

where $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$ are real matrices of compatible sizes. The vector $x(t)$ is the state of the system with an unknown initialization $x(0) = x_0$, where $x_0$ is modeled as a random variable with zero mean and a positive definite covariance $\mathbb{E}[x(0)x(0)^\top] = D_0$ (where $\mathbb{E}[\cdot]$ denotes the expectation operator). The control input $u(t)$ is to be determined via a static stabilizing state-feedback law $u(t) = Kx(t)$ with the gain $K \in \mathbb{R}^{m\times n}$ such that the quadratic performance measure

$$\mathbb{E} \int_0^\infty [x(t)^T Qx(t) + u(t)^T Ru(t)]dt$$

is minimized for a positive semidefinite matrix $Q \succeq 0$ and positive definite matrix $R \succ 0$, where the expectation is taken over $x_0$. We model the decentralized nature of the controller via a structural constraint. Precisely, ODC optimizes over the set of structured stabilizing controllers

$$\mathcal{K}_S = \{K : A + BK \text{ is stable}, K \in \mathcal{S}\}, \tag{3.1}$$

where $\mathcal{S} \subseteq \mathbb{R}^{m\times n}$ is a linear subspace of matrices, often specified by fixing certain entries of the matrix to zero. The sparsity pattern can be equivalently described with an indicator matrix $I_\mathcal{S}$ whose $(i, j)$-entry is defined to be

$$[I_\mathcal{S}]_{ij} = \begin{cases} 1, & \text{if } K_{ij} \text{ is free} \\ 0, & \text{if } K_{ij} = 0. \end{cases}$$

The structural constraint $K \in \mathcal{S}$ is equivalent to $K \circ I_\mathcal{S} = K$, where $\circ$ denotes entry-wise multiplication.

We design a homotopy scheme by introducing a sequence of damped cost functions with a varying regularization, defined as[1]

$$\begin{aligned} J(K,\alpha) =& \mathbb{E}\int_0^\infty \left[e^{-2\alpha t}\left(\hat{x}^\top(t)Q\hat{x}(t) + \hat{u}^\top(t)R_\alpha\hat{u}(t)\right)\right] dt \\ s.t.\quad & \dot{\hat{x}}(t) = A\hat{x}(t) + B\hat{u}(t) \\ & \hat{u}(t) = K\hat{x}(t), \end{aligned} \tag{3.2}$$

where the varying regularization $R_\alpha \succ 0$ is positive definite for all $\alpha \geq 0$ with $R_0 = R$. The notation $R_\alpha$ denotes a function of $\alpha$. In this setting, $\alpha$ is a damping parameter that will be used to construct a series of ODC problems. Additional assumptions will be imposed on the varying regularization $R_\alpha$ in the following sections. The introduction of the discounting term $e^{-2\alpha t}$ is a common practice in infinite-horizon control problems to ensure that the cost is finite [9]. By a change of variables $x(t) = e^{-\alpha t}\hat{x}(t)$ and $u(t) = e^{-\alpha t}\hat{u}(t)$, the cost $J(K,\alpha)$ can be equivalently written as

$$\begin{aligned} J(K,\alpha) =& \mathbb{E}\int_0^\infty \left[x^\top(t)Qx(t) + u^\top(t)R_\alpha u(t)\right] dt \\ s.t.\quad & \dot{x}(t) = (A-\alpha I)x(t) + Bu(t) \\ & u(t) = Kx(t). \end{aligned} \tag{3.3}$$

The parameter $\alpha$ alters the matrix $A$ to $A - \alpha I$, which acts to decrease the real part of all eigenvalues of $A$. This is the reason that we refer to $\alpha$ as the damping parameter. The ODC problem under study is associated with $\alpha = 0$:

$$\begin{aligned} \min_K \quad & J(K,0) \\ s.t. \quad & K \text{ stabilizes } (A,B) \\ & K \in \mathcal{S}. \end{aligned}$$

Instead of directly solving the above optimization formulation of ODC, we leverage the flexibility of a varying parameter $\alpha$ and relax the notion of stability for some fictitious damped systems. We call $K$ a stabilizing solution to (3.3) if $K$ stabilizes the system $(A - \alpha I, B)$, in which case formulation (3.2) is also meaningful. Formally, we define ODC with damping and varying regularization as

$$\begin{aligned} \min_K \quad & J(K,\alpha) \\ s.t. \quad & K \text{ stabilizes } (A-\alpha I,B) \\ & K \in \mathcal{S}. \end{aligned} \tag{ODC($\alpha$)}$$

Our relaxed notion of stability coincides with the true notion of stability when $\alpha = 0$. We emphasize that the relaxation of stability for the damped system (where $\alpha > 0$) is an artificial

[1] Note that $J(K,\alpha)$ implicitly depends on the regularization parameter $R_\alpha$.

construction in our solution method; the goal is to obtain an optimal stabilizing controller for the undamped system with $\alpha = 0$. We use ODC$(\alpha, K_0)$ to denote the problem ODC$(\alpha)$ together with an initial stabilizing controller $K_0$ that is provided for local search methods.

The two equivalent formulations (3.2) and (3.3) motivate the notion of "damping property". We make a formal statement below.

**Lemma 9.** *The function $J(K, \alpha)$ satisfies a "damping property" in the sense that for every controller $K$ that stabilizes the system $(A - \alpha I, B)$, the following statements hold for all $\beta > \alpha$:*

- *$K$ stabilizes the system $(A - \beta I, B)$;*
- *$J(K, \beta) \leq J(K, \alpha)$ if $R_\beta \preceq R_\alpha$.*

*Proof.* It follows from the formulation ODC$(\alpha)$ that whenever $A - \alpha I + BK$ is stable and $\beta > \alpha$, the matrix $A - \beta I + BK = (A - \alpha I + BK) - (\beta - \alpha)I$ is also stable. Therefore, $J(K, \beta)$ is well-defined. Due to formulation (3.2), whenever $R_\beta \preceq R_\alpha$, we have $J(K, \beta) \leq J(K, \alpha)$. □

The matrix function $R_\alpha$ is said to be monotonically decreasing if $R_\beta \preceq R_\alpha$ for all $\beta > \alpha \geq 0$. Let $K^*(\alpha)$ denote the set of globally optimal solutions of ODC$(\alpha)$, by which we mean $J(K^*(\alpha), \alpha) \leq J(K, \alpha)$ for every $K \in \mathcal{S}$ that stabilizes $(A - \alpha I, B)$. Denote the set of critical controllers by $K^\dagger(\alpha)$, which contains all stabilizing controllers $K \in \mathcal{S}$ that satisfy the first-order optimality condition $\nabla J(K, \alpha) \circ I_{\mathcal{S}} = 0$ (where $\nabla$ denotes the gradient with respect to $K$). Since the ODC problem is smooth, the set of critical controllers contains the set of locally optimal controllers whose costs are the lowest in their neighborhoods. The first-order optimality condition can be expanded as follows (see [76] for details):

$$\begin{aligned}
&(A - \alpha I + BK)^\top P_\alpha(K) + \\
&\qquad P_\alpha(K)(A - \alpha I + BK) + K^\top R_\alpha K + Q = 0 && (3.4a)\\
&L_\alpha(K)(A - \alpha I + BK)^\top + \\
&\qquad\qquad (A - \alpha I + BK)L_\alpha(K) + D_0 = 0 && (3.4b)\\
&\left[(B^\top P_\alpha(K) + R_\alpha K)L_\alpha(K)\right] \circ I_S = 0 && (3.4c)\\
&K \circ I_{\mathcal{S}} = K. && (3.4d)
\end{aligned}$$

The matrices $P_\alpha(K)$ and $L_\alpha(K)$ are the closed-loop Gramians that depend on $\alpha$. The above conditions provide a closed-form expression for the cost

$$J(K, \alpha) = \mathrm{tr}(D_0 P_\alpha(K)), \tag{3.5}$$

where $\mathrm{tr}(\cdot)$ denotes the trace of a matrix. For every given $\alpha$, the equations (3.4a)-(3.4d) and (3.5) are algebraic, involving only polynomial functions of the unknown matrices $K$, $P_\alpha$ and $L_\alpha$. The matrices $P_\alpha$ and $L_\alpha$ are written as functions of $K$ because they are uniquely determined from (3.4a) and (3.4b) given a stabilizing controller $K$. When the context is clear, we drop the implicit dependence on $K$ in the notations $P_\alpha$ and $L_\alpha$.

The chapter studies the properties of $K^*(\alpha)$, $K^\dagger(\alpha)$, and $J(K,\alpha)$ for any controller $K$ belonging to $K^*(\alpha)$ or $K^\dagger(\alpha)$, and shows how these properties can be leveraged to find a global solution of the problem ODC($\alpha$). We refer to $J(K^\dagger(\alpha),\alpha)$ and $K^\dagger(\alpha)$, which are multi-valued functions of $\alpha$, as *critical cost trajectories* and *critical controller trajectories*, respectively. To motivate the study of $K^\dagger(\alpha)$, Figure 3.1 illustrates the evolution of many critical decentralized controllers for a particular system as $\alpha$ varies (see Section 3.5 for details on the experiment). Figure 3.1a plots selected trajectories of $J(K,\alpha)$ against $\alpha$, where $K \in K^\dagger(\alpha)$. Those trajectories are all connected to a stabilizing controller in $K^\dagger(0)$. The lowest curve corresponds to the cost of globally optimal controllers $J(K^*(\alpha),\alpha)$. Figure 3.1b plots the distance of selected controllers $K \in K^\dagger(\alpha)$ from a controller $K$ in $K^*(\alpha)$.

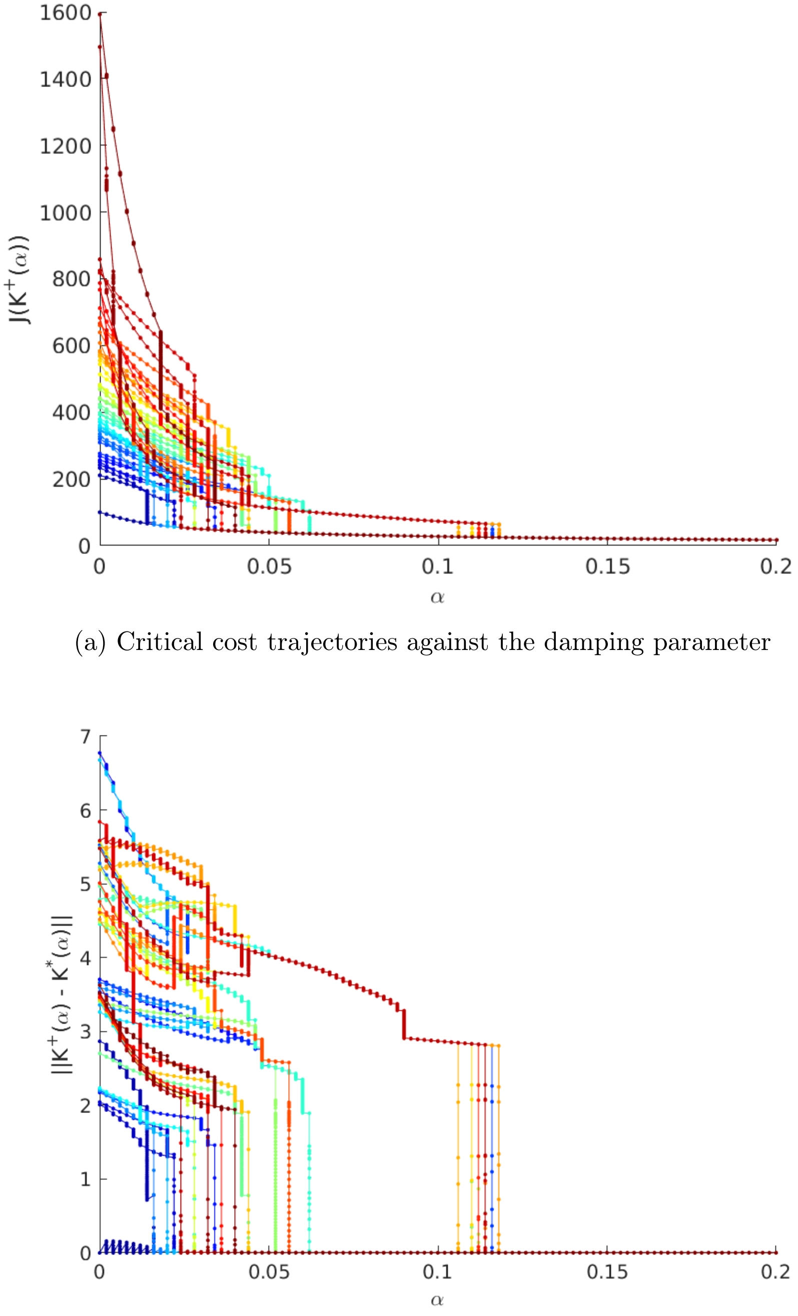
Figure 3.1 illustrates the property that a modest damping causes the locally optimal trajectories to “collapse” to each other. This attractive phenomenon suggests an effective technique for solving ODC by varying the damping parameter to relate the original ODC problem to a highly damped ODC problem. Two strategies based on the above idea are detailed in Algorithm 1 and Algorithm 2.

**Algorithm 1** The Forward-Backward Method

```
Input: $J(K,\alpha)$ and an initial controller $K_0 \in S$ that stabilizes the system $(A,B)$.
Output: A potentially improved controller in $K^\dagger(0)$.
Select a list of parameters $0=\alpha_0<\alpha_1,\dots,<\alpha_T$.
for $t \leftarrow 1,\dots,T$ do
    Obtain a $K_t \in K^\dagger(\alpha_t)$ by solving ODC($\alpha_t, K_{t-1}$) using local search.
end for
for $t \leftarrow T-1, T-2, \dots, 0$ do
    Obtain a $K_t \in K^\dagger(\alpha_t)$ by solving ODC($\alpha_t, K_{t+1}$) using local search.
end for
```

Algorithm 1 aims to find an optimal decentralized controller based on a given initial controller. One execution of the algorithm is plotted in Figure 3.2. Algorithm 2 starts with a large enough $\alpha$ for which $K=0$ is an initial stabilizing controller in the set $\mathcal{S}$ and iteratively solves for a better controller while reducing the damping parameter $\alpha$. The improvement at $\alpha=\alpha_t$ is achieved using local-search and the initialization $K_{t+1}$ from the previous step. Algorithm 1 is different from Algorithm 2 in that it starts with a potentially undesirable controller for $\alpha=0$ and gradually increases $\alpha$ to obtain an improved optimal controller for a highly-damped system and then applies a variant of Algorithm 2 to backtrack that controller to a globally optimal controller for $\alpha=0$. In what follows, we develop a theoretical analysis of the technique used in these algorithms.

The granularity of the space for $\alpha$, namely $\{\alpha_0,\alpha_1,\dots,\alpha_T\}$, does not affect the final solution as long as the discretization step is small enough so that the algorithm can approximately track the continuous paths. Admittedly, the literature of numerical continuation methods is rich with appealing predictor-corrector and piecewise-linear methods [1], and they can be applied in the tracking of $K^\dagger(\alpha)$ and $K^*(\alpha)$. Nevertheless, this chapter aims to analyze



(a) Critical cost trajectories against the damping parameter

(b) Distance between $K^{\dagger}(\alpha)$ and $K^{*}(\alpha)$

Figure 3.1: Samples of critical cost and critical controller trajectories of the system in equation (3.10) as the damping parameter $\alpha$ varies.

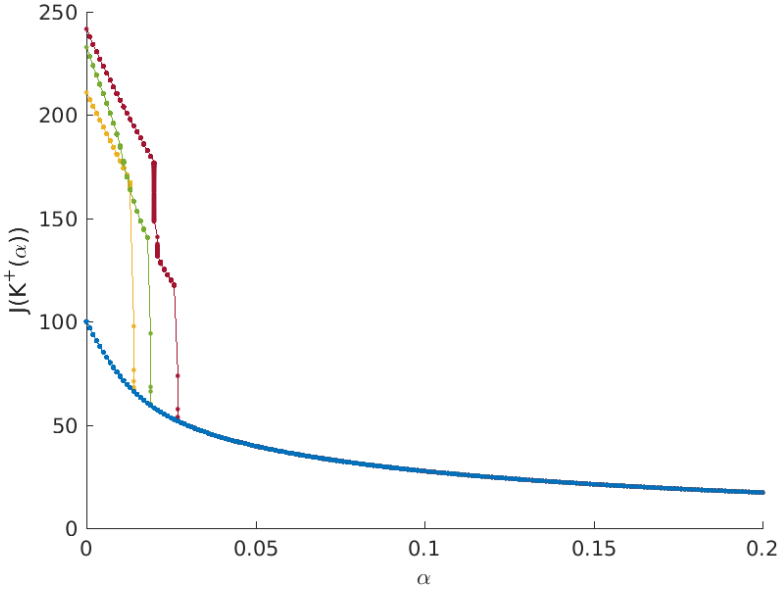


Figure 3.2: Selected cost trajectories of Algorithm 1 when applied to several critical controllers. The system is described in equation (3.10) and initialized at three sub-optimal controllers with costs greater than 200. After the damping parameter $\alpha$ is increased to around 0.2, the local-search algorithm starts to track the better blue curve. When $\alpha$ is gradually decreased to 0, the local-search algorithm tracks the blue curve and yields a controller whose cost is around 100.

**Algorithm 2** The Backward Method

```
**Input**: $J(K, \alpha)$
**Output**: A potentially stabilizing $K_0 \in K^\dagger(0)$.
Select a list of parameters $0 = \alpha_0 < \alpha_1, \dots, < \alpha_T$, where $\alpha_T$ is large enough such that $K_T = 0$ stabilizes the system $(A - \alpha_T I, B)$.
for $t \leftarrow T-1, T-2, \dots, 0$ do
    Obtain a $K_t \in K^\dagger(\alpha_t)$ by solving ODC$(\alpha_t, K_{t+1})$ using local search.
end for
```

the possibility of using local search to locate the globally optimal controller trajectory as opposed to following a specific trajectory closely.

**Remark 6.** *It is sometimes advantageous to use the pseudo-arclength or predictor-corrector method*[2] *[20] instead of monotonically increasing the homotopic parameter. This is the case when there are limit points on the homotopic curves for which continuing increasing the homotopy parameter cannot reach a solution nearby and it is necessary to make a turn.*

[2] These methods use unit direction tangent to the homotopy curve for prediction and then correct the deviation from the homotopy curve through some iterative schemes such as Newton-like methods.

*However, as Lemma 1 and Figure 1(a) demonstrate, our homotopy curves can be viewed with $J(K,\alpha)$ as a series of decreasing curves, for which no turn is necessary for the purpose of curve tracking. It is entirely possible, however, that in the backward process, decreasing $\alpha$ will reach a point where there are no nearby stabilizing controllers; this scenario can be detected because $J(K,\alpha)$ will approach infinity as $K$ reaches the boundary of the stabilizing controllers. In this case, an additional homotopy step such as Fixed-Point–Newton (FPN) Homotopy can be applied [73]. However, the theoretical properties of the FPN Homotopy are not well-understood.*

**Remark 7.** *Homotopy methods find solutions to a difficult problem by constructing a series of related problems. The idea of the chapter is to construct a homotopy path on which a tractable control problem exists; such a control problem has a stable $A$ matrix so that $K = 0$ stabilizes the system. More generally, one can apply homotopy methods in phases, where the first phase explores tractable problems and the second phase connects the tractable problems to the difficult problem of interest. The paper [72] recently investigated this idea for the problem of low-thrust trajectory optimization.*

Due to the NP-hardness of ODC, Algorithm 1 and Algorithm 2 cannot always produce a globally optimal or even a stabilizing decentralized controller, unless certain conditions are met. The conditions will be discussed later. In Section 3.3, we first prove the continuity of the trajectories, which is a prerequisite for tracking.

## 3.3 Continuity of Solution Trajectories

This section studies the continuity properties of the set-valued map $K^*(\alpha)$ and $K^\dagger(\alpha)$. The key notion of hemi-continuity captures the evolution of the parametrized optimization problems.

**Definition 1.** *The set-valued map $\Gamma : \mathcal{A} \to \mathcal{B}$ is said to be upper hemi-continuous at a point $z$ if for any open neighborhood $V$ of $\Gamma(z)$ there exists a neighborhood $U$ of $z$ such that $\Gamma(U) \subseteq V$.*

The related notion of lower hemi-continuity is provided in Section 3.7. A set-valued map is said to be continuous if it is both upper and lower hemi-continuous. A single-valued function is continuous if and only if it is upper hemi-continuous. We restate a version of Berge Maximum Theorem with a compactness assumption from [69].

**Lemma 10** (Berge Maximum Theorem [69])**.** *Let $\mathcal{A} \subseteq \mathbb{R}$ and $\mathcal{S} \subseteq \mathbb{R}^{m\times n}$. Assume that $J : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is jointly continuous and $\Gamma : \mathcal{A} \to \mathcal{S}$ is a compact-valued map*[3]*. Define*

$$K^*(\alpha) = \arg\min\{J(K,\alpha) | K \in \Gamma(\alpha)\}, \ \textit{for } \alpha \in \mathcal{A},$$
$$J(K^*(\alpha),\alpha) = \min\{J(K,\alpha) | K \in \Gamma(\alpha)\}, \ \textit{for } \alpha \in \mathcal{A}.$$

[3] Compact-valued means for every $x \in \mathcal{A}$, $\Gamma(x)$ is compact. The map's value can be set.

*If $\Gamma$ is continuous at some $\alpha \in \mathcal{A}$, then $J(K^*(\alpha), \alpha)$ is continuous at $\alpha$. Furthermore, $K^*$ is non-empty, compact-valued, closed, and upper hemi-continuous.*

The Berge Maximum Theorem does not readily apply to ODC since the set of stabilizing controllers is open and often unbounded. The difficulty can be overcome by restricting the relevant map to a lower level-set.

**Theorem 11.** *Assume that the matrix function $R_\alpha$ is continuous in $\alpha$ and that $K^*(0)$ is non-empty. Then, the set $K^*(\alpha)$ is non-empty for all $\alpha > 0$. Furthermore, $K^*(\alpha)$ is upper hemi-continuous and the optimal cost $J(K^*(\alpha), \alpha)$ is continuous. If $R_\alpha$ is monotonically decreasing, $J(K^*(\alpha), \alpha)$ is strictly decreasing in $\alpha$.*

*Proof.* When $K^*(0)$ is non-empty, there is an optimal decentralized controller for the undamped system. We can apply any controller in $K^*(0)$ to the damped system and conclude that

$$J(K^*(\alpha), \alpha) \leq J(K^*(0), \alpha) < \infty.$$

Note that since $K^*(\alpha)$ is the set of globally optimal controllers, $J(K^*(\alpha), \alpha)$ is a well-defined single-valued function of $\alpha$. The inequality above assumes the existence of a globally optimal controller for all values of the damping parameter $\alpha$. This is true because the lower-level set of $J(K, \alpha)$ is compact [87]. Precisely, define $\Gamma_M(\alpha)$ to be

$$\Gamma_M(\alpha) = \{K \in S : A - \alpha I + BK \text{ stable, } J(K, \alpha) \leq M\}. \tag{3.6}$$

The set-valued function $\Gamma_M$ is compact-valued for all fixed $\alpha$ given a fixed $M$. We select any $M > J(K^*(0), \alpha)$ and optimize $J(K, \alpha)$ instead over $K \in \Gamma_M(\alpha)$ without losing any globally optimal controller. The continuity of $\Gamma_M(\alpha)$ at $\alpha$ for almost all $M$ is proved in Section 3.7. Berge Maximum Theorem then yields the desired continuity of $K^*(\alpha)$ and $J(K^*(\alpha), \alpha)$. When $R_\alpha$ is monotonically decreasing, the "damping property" ensures that $J(K^*(\alpha), \alpha)$ is monotonically decreasing. □

The above argument can be extended to characterize all critical controllers. A caveat is the possible existence of critical controllers whose costs approach infinity in the damped problem. Such existence does not contradict the damping property — damping can introduce locally optimal controllers that are not stabilizing in the absence of damping.

**Theorem 12.** *Assume that the matrix function $R_\alpha$ is continuous in $\alpha$ and that $K^\dagger(0)$ is non-empty. Then, the set $K^\dagger(\alpha)$ is nonempty for all $\alpha > 0$. Furthermore, if for $\alpha_0 > 0$,*

$$\lim_{\epsilon \to 0^+} \sup_{\alpha \in [\alpha_0 - \epsilon, \alpha_0 + \epsilon]} \sup_{K \in K^\dagger(\alpha)} J(K, \alpha) < \infty, \tag{3.7}$$

*then, $K^\dagger(\alpha)$ is upper hemi-continuous at $\alpha_0$ and the optimal cost $J(K^\dagger(\alpha), \alpha)$ is upper hemi-continuous at $\alpha_0$.*

*Proof.* The fact that $K^{\dagger}(\alpha)$ is non-empty follows from the existence of globally optimal controllers in Theorem 11. Consider the parametrized optimization problem

$$\begin{aligned} \min \quad & \|\nabla J(K,\alpha)\| \\ s.t. \quad & K \in \Gamma_M(\alpha), \end{aligned} \tag{3.8}$$

where $\|\cdot\|$ denotes the 2-norm of a vector. The assumption (3.7) ensures the existence of a real number $M$ and $\epsilon > 0$ such that $M > J(K,\alpha)$ for all $K \in K^{\dagger}(\alpha)$ and $\alpha \in [\alpha_0 - \epsilon, \alpha_0 + \epsilon]$. This choice of $M$ guarantees that the formulation (3.8) does not cut off any critical controller. As will be proven in Section 3.7, $\Gamma_M(\alpha)$ is continuous at $\alpha_0$ for almost all $M$. Therefore, $M$ can be selected to make $\Gamma_M(\alpha)$ continuous at $\alpha_0$ without cutting off any critical controllers. Berge Maximum Theorem implies that $K^{\dagger}(\alpha)$ is upper hemi-continuous. Since $J(K,\alpha)$ is jointly continuous in $(K,\alpha)$, the map $J(K^{\dagger}(\alpha),\alpha)$ is upper hemi-continuous. □

## 3.4 Asymptotic Properties

In this section, we state asymptotic properties of the local solutions given by the set $K^{\dagger}(\alpha)$. They shed light on the trajectories in Figure 3.1. We use $\|\cdot\|$ to denote the 2-norm of a matrix and use $\lambda_{\min}(\cdot)$ to denote the minimum eigenvalue of a symmetric matrix. The matrix-valued function $R_\alpha : \mathbb{R} \to \mathbb{R}^{m\times m}$ is called semi-algebraic if its graph $\{(\alpha, W) \in \mathbb{R}\times\mathbb{R}^{m\times m} : R_\alpha = W\}$ can be represented by a finite set of polynomial equalities and inequalities.

**Assumption 1.** *The symmetric matrix function $R_\alpha$ is monotonically decreasing and is a semi-algebraic function of $\alpha$. Furthermore, there exist real constants $\Delta > \delta > 0$ such that $\Delta \geq \|R_\alpha\| \geq \lambda_{\min}(R_\alpha) \geq \delta$ for all $\alpha \geq 0$.*

Recall that the function $R_\alpha$ is introduced to make Algorithms 1 and 2 effectively solve $ODC(0)$. The above assumption provides guidance on how to select this function. Furthermore, we make the following technical assumption on the sparsity pattern of the decentralized controller to develop the results of this section.

**Assumption 2.** *The sparsity pattern $I_S$ is block-diagonal with square blocks and $R_\alpha$ has the same sparsity pattern as $I_S$ for all $\alpha$.*

The following theorem characterizes the evolution of critical controllers for a specific sparsity pattern. It also justifies the practice of random initialization around zero and the initialization strategy in Algorithm 2.

**Theorem 13.** *Under Assumption 1 and Assumption 2, we have $\sup_{K\in K^{\dagger}(\alpha)} \|K\| \to 0$ as $\alpha \to \infty$. Furthermore, $\sup_{K\in K^{\dagger}(\alpha)} J(K,\alpha) \to 0$ as $\alpha \to \infty$.*

*Proof.* Refer to Section 3.7. □

As $\alpha \to \infty$, not only do all critical controllers in $K^{\dagger}(\alpha)$ approach zero, the problem also becomes convex over bounded regions with enough damping.

**Theorem 14.** *Under Assumption 1 and Assumption 2, for any given $r > 0$, the Hessian matrix $\nabla^2 J(K, \alpha)$ is positive definite over the set $\|K\| \leq r$ for all large values of $\alpha$.*

*Proof.* Refer to Section 3.7. □

**Corollary 4.** *Under Assumption 1 and Assumption 2, there is no spurious locally optimal controller for large $\alpha$, meaning that $K^\dagger(\alpha) = K^*(\alpha)$ for all large values of $\alpha$.*

*Proof.* For any given $r > 0$, all controllers in the ball $\mathcal{B} = \{K : \|K\| \leq r\}$ are stabilizing when $\alpha$ is large. As a result, stability constraints can be relaxed over $\mathcal{B}$. Furthermore, it results from Theorem 13 that when $\alpha$ is large, all critical controllers will be inside $\mathcal{B}$. In light of Theorem 14, the objective function becomes convex over $\mathcal{B}$ for large values of $\alpha$. These observations imply that local and global solutions coincide. □

Corollary 4 implies that with a large damping parameter $\alpha$ and a well-conditioned $R_\alpha$, the damped problem is tractable. Now, the problem remains to connect the damped system to the undamped one. This requires the following assumption that excludes bifurcation — when a critical trajectory merges with the globally optimal trajectory, the homotopy method cannot distinguish one from the other.

**Assumption 3.** *The set of globally optimal controllers $K^*(\alpha)$ is a singleton set for all $\alpha \geq 0$. Furthermore, there is an $\epsilon > 0$ such that for all $\alpha > 0$ and any two distinct controllers $K_1 \in K^\dagger(\alpha)$ and $K_2 \in K^*(\alpha)$, it holds that $\|K_1 - K_2\| \geq \epsilon$.*

Even though Assumption 3 requires that $K^*(\alpha)$ be distinguished from $K^\dagger(\alpha)$ for all $\alpha > 0$, there is a value $\alpha_0$ such that this assumption automatically holds for all $\alpha > \alpha_0$. This immediately follows from Corollary 4 and the observation in Theorem 14 that the Hessian $\nabla^2 J(K, \alpha)$ is positive definite. The following lemma provides sufficient conditions under which Assumption 3 holds. We use spabs$(A)$ to denote the spectral abscissa (maximum real part of the eigenvalues), which is negative for a stable matrix $A$.

**Lemma 11.** *Assume that $A$ is stable and that the following conditions hold for all $\alpha \geq 0$:*

- $\lambda_{\min}(R_\alpha) \geq \delta > 0$
- *$\nabla^2 J(K, \alpha)$ is positive definite over the region $\{K : \|K\| \leq r\}$, where $r$ is large enough so that*

$$\begin{aligned} -spabs(A)\delta r^2 - n\|Q\|\kappa(D_0)\|B\|r \\ +spabs(A)\|Q\| - n\|Q\|\kappa(D_0)\|A\| > 0, \qquad (3.9a) \\ \delta r^2 - \|Q\|(1 + n\kappa(D_0)) \geq 0, \qquad (3.9b) \end{aligned}$$

*where $\kappa(D_0) = \|D_0\|/\lambda_{\min}(D_0)$. Then, Assumption 3 holds for all $\alpha \geq 0$.*

*Proof.* We first show that for all stabilizing $K$ with $\|K\| \geq r$, we have $J(K, \alpha) > J(0, \alpha)$. When $\|K\| \geq r$, note that $P_\alpha(K)$ is the unique solution to the equation (3.4a). From the triangle inequality,

$$\begin{aligned}\lambda_{\min}(R_\alpha)\|K\|_2^2 &\leq \|K^\top R_\alpha K\|_2 \\ &\leq 2\|A - \alpha I + BK\|_2\|P_\alpha(K)\|_2 + \|Q\|_2 \\ &\leq 2(\|A\|_2 + \alpha + \|B\|_2\|K\|_2)\|P_\alpha(K)\|_2 + \|Q\|_2,\end{aligned}$$

Therefore,

$$\|P_\alpha(K)\|_2 \geq \frac{\delta\|K\|_2^2 - \|Q\|_2}{2(\|A\|_2 + \alpha + \|B\|_2\|K\|_2)}.$$

Note that $\mathrm{tr}(P_\alpha(K)D_0) \geq \|P(K)\|_2\lambda_{\min}(D_0)$,

$$J(K, \alpha) \geq \lambda_{\min}(D_0)\frac{\delta r^2 - \|Q\|_2}{2(\|A\|_2 + \alpha + \|B\|_2 r)},$$

where we use the fact that the right-hand side is monotonically increasing in $r$. Similarly, using the matrix exponential expression of $P_\alpha(0)$, we can estimate $J(0, \alpha)$ as follows:

$$\begin{aligned}\mathrm{tr}(P_\alpha(0)D_0) &\leq n\|Q\|\|D_0\|\int_0^\infty \|e^{t(A-\alpha I)}\|_2^2 dt \\ &= n\|Q\|\|D_0\|\int_0^\infty e^{2t\cdot(\mathrm{spabs}(A)-\alpha)}dt = \frac{n\|Q\|\|D_0\|}{2(\alpha - \mathrm{spabs}(A))}.\end{aligned}$$

The inequalities (3.9a) and (3.9b) ensure that $J(K, \alpha) > J(0, \alpha)$ for all $\alpha \geq 0$. As a result, if a matrix $K$ that satisfies $\|K\| \geq r$ belongs to $K^\dagger(\alpha)$, it cannot be globally optimal, and there must exist an $\epsilon > 0$ that bounds the distance between $K$ and $K^*(\alpha)$ because their costs are different. The proof is completed by noting that the assumption of the positive definiteness of the Hessian ensures that $K^*(\alpha)$ is a singleton set. □

To track the global trajectory using local search as a part of Algorithm 1 or 2, it is necessary for the local search to converge when initialized close to the trajectory. Furthermore, the discretization needs to be adapted to the local search algorithm. The conditions are specified in the following two definitions.

**Definition 2.** *A local search algorithm for ODC($K_0, \alpha$) is said to be locally $\delta$-stable if for any $K_0$ with $\sup_{K \in K^*(\alpha)} \|K_0 - K\| < \delta$, it converges to a point in $K^*(\alpha)$.*

Since ODC is a smooth optimization problem, as long as the radius $\delta > 0$ is selected in such a way that the region $\{K_0 : \sup_{K \in K^*(\alpha)} \|K_0 - K\| < \delta\}$ remains inside the gradient dynamics's region of attraction of $K^*(\alpha)$, gradient descent with a small step-size is locally $\delta$-stable.

**Definition 3.** *Given a $\delta$-stable local search algorithm for ODC, a discretization $0 = \alpha_0 < \alpha_1, \ldots, < \alpha_T$ is said to be $\delta$-adaptive if for any $i \in \{0, 1, \ldots, T-1\}$ and any two controllers $K_1 \in K^*(\alpha_i)$ and $K_2 \in K^*(\alpha_{i+1})$, it holds that $\|K_1 - K_2\| < \delta$.*

As long as $K^*(\dot{})$ is a continuous function, we can find a discretization that is $\delta$-adaptive by selecting a small increment.

**Corollary 5.** *Suppose that Assumptions 1, 2 and 3 are satisfied. Then, the trajectory $K^*(\alpha)$ is continuous. Moreover, if $\delta > 0$ is chosen small enough such that the local search method is locally $\delta$-stable and the discretization $0 = \alpha_0 < \alpha_1, \ldots, < \alpha_T$ is $\delta$-adaptive, Algorithm 1 and Algorithm 2 both find the globally optimal stabilizing controller in $K^*(0)$ for a large $\alpha_T$.*

*Proof.* It is shown in Theorem 11 that $K^*(\alpha)$ is upper hemi-continuous. Under Assumption 3, the set $K^*(\alpha)$ is a singleton, and the continuity of $K^*(\alpha)$ can be concluded because a single-valued function is continuous if and only if it is upper hemi-continuous. When a $\delta$-adaptive discretization $0 = \alpha_0 < \alpha_1, \ldots, < \alpha_T$ is selected with $\alpha_T$ sufficiently large for which the "no spurious property" of Corollary 4 holds, Algorithm 1 and Algorithm 2 are able to locate the continuous globally optimal trajectory $K^*(\alpha)$ at $\alpha = \alpha_T$. To obtain $K^*(0)$, we follow the continuous $K^*(\alpha)$ based on Algorithm 1 and Algorithm 2. Since the local search algorithm is $\delta$-stable and the discretization is $\delta$-adaptive, we inductively obtain a series of controllers $K_t$ for $t = T, T-1, \ldots, 0$, which all lie on the path $K^*(\alpha)$ for $\alpha \in [0, \alpha_T]$. □

The previous results all rely on the "damping property" in Lemma 9. It is worth mentioning that damping the system with $-I$ is almost the only continuation method for general system matrices "$A$" that is able to achieve the monotonic increase of stable sets. This will be formalized below.

**Theorem 15.** *When $n \geq 3$, for any $n$-by-$n$ real matrix $H$ that is not a multiple of $-I$, there exists a stable matrix $A$ for which $A + H$ is unstable.*

The proof is given in Section 3.7. This theorem justifies the use of $-\alpha I$ as the continuation parameter and is the reason that our setting avoids the undesirable behaviors of homotopy documented in [60]. However, matrices other than $-\alpha I$ may be used for a specific $A$ matrix; if $A$ has certain structures (such as upper-triangular), there are non-trivial matrices $H$ (such as a non-positive diagonal matrix) for which $A + tH$ is always stable when $t > 0$. If the algorithm designer does not aim to customize the homotopy method for every specific system matrix $A$, the above theorem supports the use of the universal variation matrix $-\alpha I$.

## Discrete-time Stochastic Systems

We detour briefly to discuss damping with varying regularization for discrete-time stochastic systems. This shall illustrate the difference between discrete- and continuous-time systems. Consider the stochastic system

$$x[t+1] = Ax[t] + Bu[t] + d[t]$$

under a static feedback policy $u[t] = Kx[t]$, where $K$ is to be designed such that the damped objective

$$J(K, \alpha) = \lim_{t\to\infty} \mathbb{E}\left[\alpha^{2t}\left(x[t]^\top Qx[t] + u[t]^\top R_\alpha u[t]\right)\right]$$

is minimized. The damping parameter $\alpha$ belongs to the interval $[0, 1]$. Assume that the random variables $d[t], t = 0, 1, 2, \dots,$ are independent and $d[t]$ has the covariance matrix $\Sigma_d$. We consider the following construction of homotopy for the discrete-time ODC problem

$$\begin{aligned} &\min_K J(K, \alpha) = \operatorname{tr}[(K^\top R_\alpha K + Q)P_\alpha(K)],\\ &s.t.\ (\alpha A+\alpha BK)P_\alpha(K)(\alpha A+\alpha BK)^\top - P_\alpha(K) + \Sigma_d = 0,\\ &\quad \alpha\|(A + BK)\| < 1. \end{aligned} \tag{d-ODC($\alpha$)}$$

Even though the formulation is not linear in $K$ or in $P_\alpha$, we develop asymptotic results under an additional bounded assumption, as stated below. The proof of the lemma is given in Section 3.7.

**Lemma 12.** *Suppose that $\lambda_{\min}(R_\alpha) \geq \epsilon > 0$ for all $\alpha \in [0, 1]$. Assume further that a locally optimal solution $K_\alpha$ to* (d-ODC($\alpha$)) *exists and is uniformly bounded for all $\alpha \in [0, 1]$. Then, as $\alpha \to 0$, it holds that $P_\alpha(K_\alpha) \to \Sigma_d$ and $K_\alpha \to 0$.*

Due to the above lemma, one can use an analogue of Algorithm 1 or Algorithm 2 to solve ODC in the discrete setting, but the damping parameter $\alpha$ should be discretized over the interval $[0, 1]$.

**Remark 8.** *It appears in d-ODC($\alpha$) that the homotopy map is constructed by rescaling the matrices $A$ and $B$ at the same time. In fact, the covariance matrices is also rescaled implicitly, as shown in the following computation:*

$$\begin{aligned} J(K, \alpha) &= \lim_{t\to\infty} \mathbb{E}\operatorname{tr}[(Q + K^\top R_\alpha K)x[t]x[t]^\top \alpha^{2t}]\\ &= \operatorname{tr}\Bigg[(Q + K^\top R_\alpha K) \times\\ &\lim_{t\to\infty}\sum_{\tau=0}^{t}(\alpha A+\alpha BK)^{t-\tau}\mathbb{E}[d[\tau]d[\tau]^\top]\alpha^{2\tau}(\alpha A+\alpha BK)^{\top(t-\tau)}\Bigg]. \end{aligned}$$

## 3.5 Numerical Experiments

In this section, we study ODC problems that have poor local minima and therefore the existing nonlinear programming techniques based on local-search cannot provably find the global solution without the knowledge of the location of the global solution in the space of the control policies. We further catalog various homotopy behaviors as the damping parameter $\alpha$ varies. The focus is on the evolution of critical trajectories, which can be tracked by any local search or path-following methods. The experiments are performed on small-sized systems, so local search with random initialization is unlikely to miss locally optimal solutions and therefore, more likely to give a complete picture of the ensemble of trajectories $J(K^\dagger(\alpha), \alpha)$ and $K^\dagger(\alpha)$. Despite the small system dimension, the existence of many locally optimal solutions and their convoluted trajectories demonstrate the power and the limit of using homotopy methods in ODC.

For the local search method, we use the projected gradient descent. At a controller $K^i$, we perform line search along the direction $\tilde{K}^i = -\nabla J(K) \circ I_{\mathcal{S}}$. The step size is determined with backtracking and Armijo rule, namely, we select $s^i$ as the largest number in $\{\bar{s}, \bar{s}\beta, \bar{s}\beta^2, ...\}$ such that $K^i + s^i\tilde{K}^i$ is stabilizing while

$$J(K^i + s^i\tilde{K}^i) < J(K^i) + \gamma s^i \langle \nabla J(K^i), \tilde{K}^i \rangle.$$

We select the parameters $\gamma = 0.001$, $\beta = 0.5$, and $\bar{s} = 1$. We terminate the iteration when the norm of the gradient is less than $10^{-2}$.

## Systems with a Large Number of Local Minima

We first consider the examples from [40], where the feasible set is highly disconnected and admits many local minima. The system matrices are

$$A = \begin{bmatrix} -1 & 2 & 0 & & \\ -2 & 0 & 1 & 0 & \\ 0 & -1 & 0 & 2 & \ddots \\ & 0 & -2 & 0 & \ddots \\ & & \ddots & \ddots & \ddots \end{bmatrix}, B = \begin{bmatrix} 0 & 1 & 0 & & \\ -1 & 0 & 1 & 0 & \\ 0 & -1 & 0 & 1 & \ddots \\ & 0 & -1 & 0 & \ddots \\ & & \ddots & \ddots & \ddots \end{bmatrix}, \tag{3.10}$$

$$D_0 = I, \quad I_{\mathcal{S}} = I, \quad Q = I, \quad R_\alpha = I.$$

When the dimension $n$ is equal to 9, it is known that the set of stabilizing decentralized controllers has at least 55 connected components, each of them containing at least one locally optimal controller. We track 50 of those locally optimal solutions. The damping parameter $\alpha$ is gradually increased from 0 to 0.2 with a 0.002 increment. The trajectories of locally optimal solutions are tracked by solving the newly damped system with the previous local optimal solution as the initialization, in the same spirit of Algorithm 1. The evolution of the optimal cost and the distance from the best known optimal controller are plotted in Figure 3.1. Notice that all sub-optimal local trajectories terminate after a modest damping $\alpha \approx 0.12$. After that, the minimization algorithm always tracks a single trajectory. This illustrates the prediction of Corollary 4. Especially, if we start tracking a sub-optimal controller trajectory from $\alpha = 0$, we will be on a better trajectory when $\alpha \approx 0.2$. At that time, if we gradually decrease $\alpha$ to zero, we will obtain a stabilizing controller with a lower cost.

## Dynamic Controllers

So far, the design of a static controller has been discussed. However, the previous results all apply to the design of a dynamic controller via a reformulation. To explain the idea, consider the continuous-time setting with no loss of generality. Assume that the goal is to design a set of sub-controllers of pre-specified degrees as opposed to a static decentralized controller. The motivation is that a static stabilizing controller may not exist, but when the degrees of the local controllers are chosen high enough, a stabilizing decentralized controller always exists under the assumption of no unstable fixed modes [80]. More precisely, let the system consist of a number of interacting subsystems, where each subsystem should have a local controller that is allowed to communicated with some other local controllers based on a user-defined communication strategy. Moreover, each unknown local controller has a user-defined degree. The overall decentralized controller can be written as

$$\begin{aligned} \dot{x}_c(t) &= A_c x_c(t) + B_c x(t) \\ u(t) &= C_c x_c(t) + D_c x(t), \end{aligned}$$

where $x_c$ is the aggregate state of the decentralized controller. The matrices $A_c, B_c, C_c$ and $D_c$ are indeed block matrices with certain sparsity patterns due to the pre-specified distributed architecture of the controller. As opposed to $K$, we need to find the globally optimal values for the structured tuple $(A_c, B_c, C_c, D_c)$. The closed-loop system can be written as

$$\begin{bmatrix} \dot{x} \\ \dot{x}_c \end{bmatrix} = \left( \begin{bmatrix} A & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} B & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} D_c & C_c \\ B_c & A_c \end{bmatrix} \right) \begin{bmatrix} x \\ x_c \end{bmatrix}.$$

Define $\tilde{A} = \begin{bmatrix} A & 0 \\ 0 & 0 \end{bmatrix}$, $\tilde{B} = \begin{bmatrix} B & 0 \\ 0 & I \end{bmatrix}$, $\tilde{K} = \begin{bmatrix} D_c & C_c \\ B_c & A_c \end{bmatrix}$ and $\tilde{S}$ as the set of matrices $\tilde{K}$ with the correct sparsity pattern. The design of the dynamic decentralized controller can be reformulated as the design of a static gain $\tilde{K} \in \tilde{\mathcal{S}}$ for the system $(\tilde{A}, \tilde{B})$. Using the homotopy idea on $\tilde{A}$, one can deploy Algorithm 1 or 2 to solve the problem and all of the previous mathematical results hold true.

As an example, consider the design of a decentralized controller for the system given in (3.10), where each local controller has degree 1 and $A_c$ is a diagonal matrix. The cost trajectories are plotted in Figure 3.3. It can be observed that the homotopy behavior is similar to the static case, and the globally optimal controller can be found via Algorithm 1 or 2.

## Experiments on Small Random Systems

With the same initialization and optimization procedure, we perform the experiments on 3-by-3 system matrices $A$ and $B$ randomly generated from the normal distribution with zero mean and unit variance. For 92 out of 100 samples, we are not able to find more than one locally optimal trajectory. Some examples with more than one local trajectories are

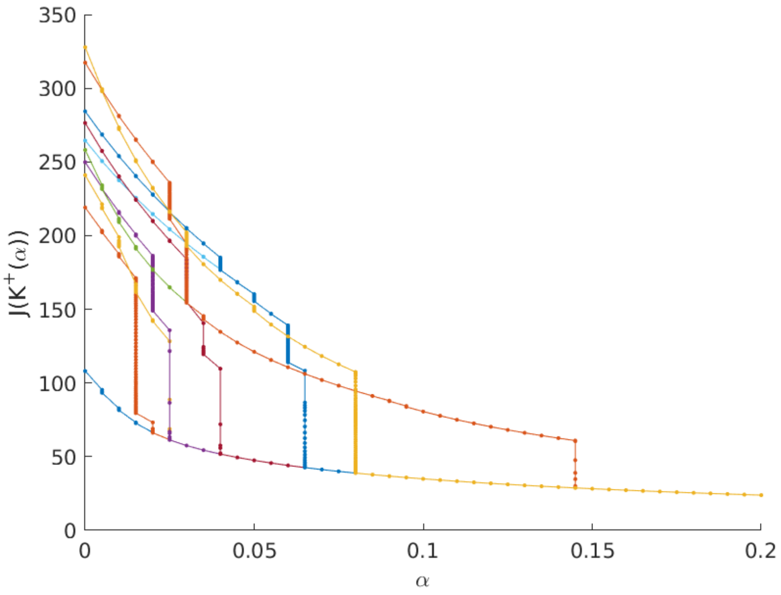


Figure 3.3: Cost trajectories of degree-1 dynamic controller design.

provided in Figures 3.4, 3.5, and 3.6. The top plot in each figure shows the costs of locally optimal controllers. The bottom plot shows the distance of each locally optimal controller to the controller with the lowest cost. Note that the order of the cost trajectories may be preserved (Figure 3.4) or may be disrupted (Figure 3.5 and Figure 3.6). In Figure 3.5, at the intersection of the two curves, there are two distinct global solutions and therefore, Algorithm 1 may fail to obtain the globally optimal decentralized controller. More than one trajectory may have the lowest cost as the damping increases (Figure 3.6), but with high damping, there is only one trajectory that has the lowest cost. If Algorithm 1 is applied with initialization on the purple curve, whose cost is around 180, after the damping parameter $\alpha$ is increased to around 2, the purple curve merges with the orange curve. When the damping parameter $\alpha$ is reduced to $\alpha = 0$, Algorithm 1 will return to the orange curve with cost around 80, which is a sub-optimal decentralized controller. This illustrates the necessity of assuming the uniqueness of the globally optimal controller in Corollary 5.

## 3.6 Conclusion

This chapter studied the optimal decentralized control problem with a large number of locally optimal solutions. To be able to find a globally optimal control policy, we proposed a homotopy method that gradually changes the control problem. We investigated the trajectories of the locally and globally optimal solutions to the optimal decentralized control problem as the damping parameter and the regularization of the decentralized control problem varied. Asymptotic and continuity properties of trajectories were proved, which were

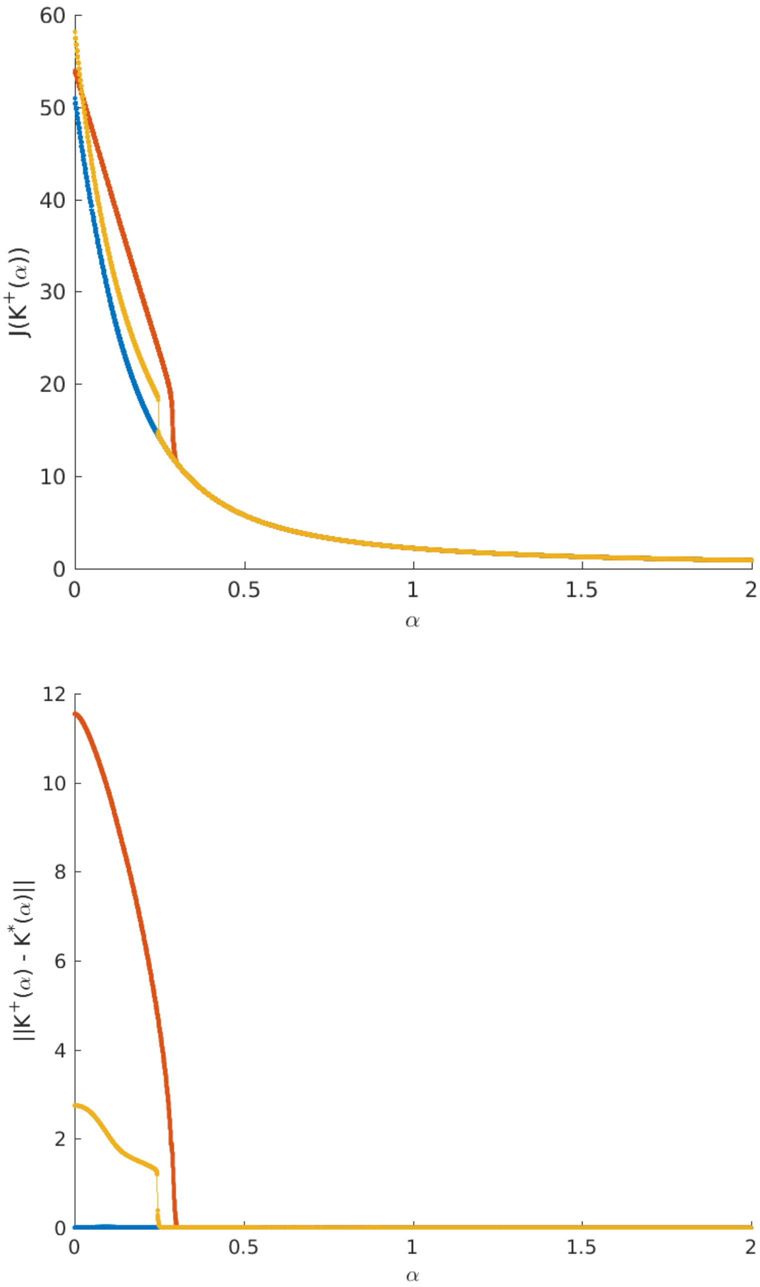


Figure 3.4: Trajectories of a randomly generated system where the order of locally optimal controllers is preserved as the damping parameter $\alpha$ changes.

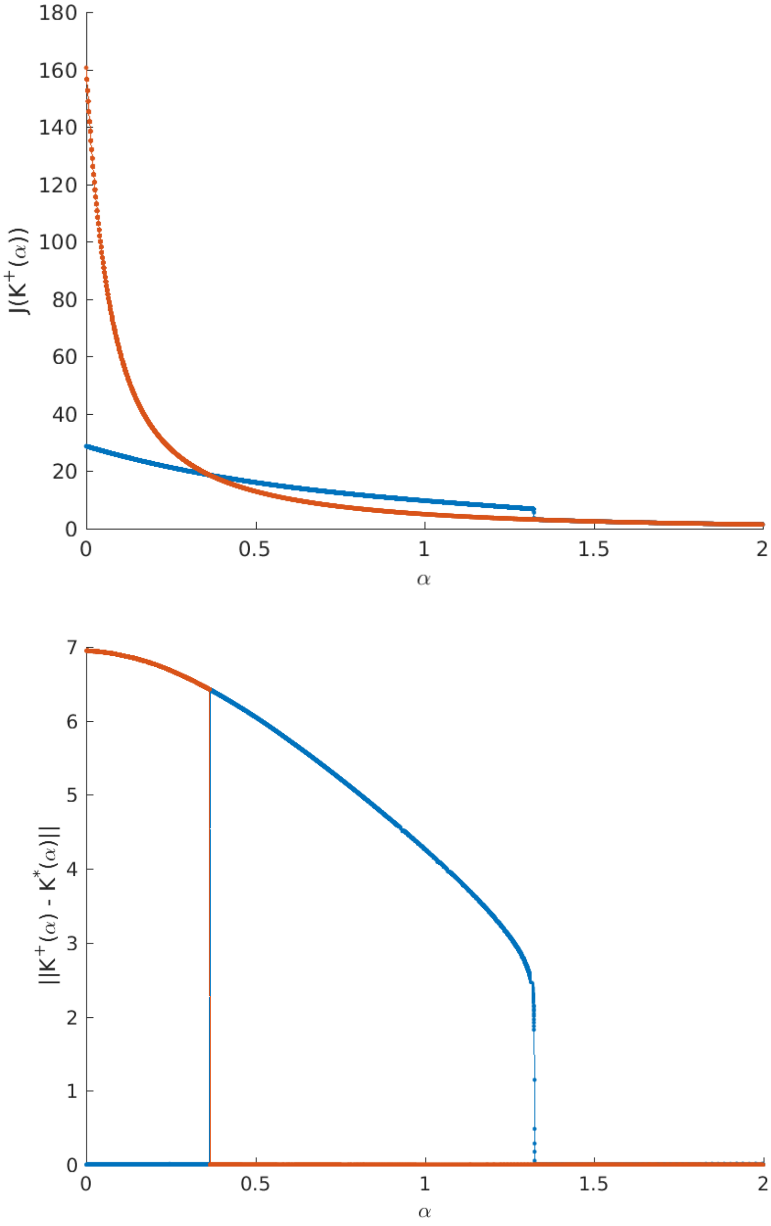


Figure 3.5: Trajectories of a randomly generated system where the order of locally optimal controllers is disrupted as the damping parameter $\alpha$ changes.

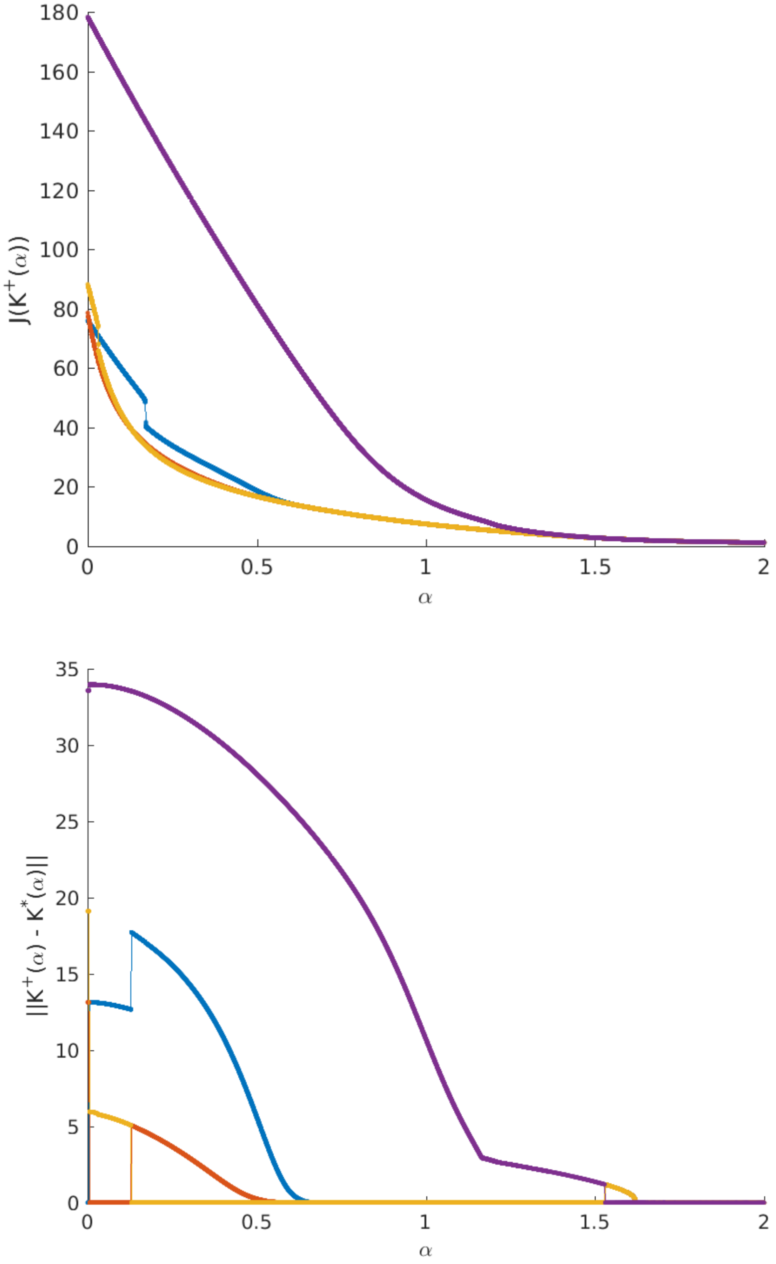


Figure 3.6: Trajectories of a randomly generated system with a complicated behavior.

based on the notion of “damping property”. A sufficient condition was developed together with an algorithm based on local search for finding the global solution of the optimal decentralized control problem. The complicated behavior of numerical continuation methods was illustrated with numerical examples with many local minima.

## 3.7 Proofs

The section collects the remaining proofs of the results of the previous sections.

### Continuity

Lemma 13 and Lemma 14 given below prove the continuity of the lower level-set map $\Gamma_M$ defined in (3.6). The continuity of $\Gamma_M$ is the prerequisite for applying the Berge Maximum Theorem. The reader is referred to [69] for an accessible treatment of relevant definitions.

Recall the notion of upper hemi-continuity of a set-valued map $\Gamma : A \to B$ in Definition 1. If $B$ is compact, upper hemi-continuity is equivalent to the graph of $\Gamma$ being closed, that is, if $a_n \to a^*$ and $b_n \in \Gamma(a_n) \to b^*$, then $b^* \in \Gamma(a^*)$. Lemma 13 resolves the upper hemi-continuity of $\Gamma_M$.

**Lemma 13.** *Assume that $R_\alpha$ is continuous in $\alpha$ and that for a given $M > 0$, $\Gamma_M(\alpha)$ is not empty for all $\alpha \geq 0$. Then, $\Gamma_M(\alpha)$ is an upper hemi-continuous set-valued map.*

*Proof.* From [87], $\Gamma_M(\alpha)$ is compact for all $\alpha$. To characterize the continuity of $\Gamma$ at a point $\alpha^* \geq 0$, it suffices to assume that the range of $\Gamma_M$ is compact and, therefore, the sequence characterization of upper hemi-continuity applies. Suppose that $\alpha_i \to \alpha^*$ and select a sequence of $K_i \in \Gamma_M(\alpha_i)$ that converges to $K^*$. The continuity of $J(K, \alpha)$ implies that $J(K^*, \alpha^*) \leq M$. The fact that the cost is bounded implies that $A - \alpha^* I + BK^*$ is stable. Since the matrix subspace $\mathcal{S}$ is a closed set, the limit point $K^*$ belongs to $\mathcal{S}$. We have verified all conditions for $K^* \in \Gamma_M(\alpha^*)$, and therefore $\Gamma_M$ is upper hemi-continuous. □

A complementary notion of upper hemi-continuity is lower hemi-continuity, which is stated below.

**Definition 4.** *The set-valued map $\Gamma : A \to B$ is said to be lower hemi-continuous at a point $z$ if for any open neighborhood $V$ intersecting $\Gamma(z)$ there exists a neighborhood $U$ of $z$ such that $\Gamma(x)$ intersects $V$ for all $x \in U$.*

Equivalently, for all $a_m \to a \in A$ and $b \in \Gamma(a)$, there exists a subsequence $a_{m_k}$ of $a_m$ and a corresponding $b_k \in \Gamma(a_{m_k})$, such that $b_k \to b$. The map $\Gamma_M$ is lower hemi-continuous for almost all $M$.

**Lemma 14.** *For any given $\alpha^* \geq 0$, $\Gamma_M(\alpha)$ is lower hemi-continuous at $\alpha^*$ except when $M \in \{J(K, \alpha^*) : K \in K^\dagger(\alpha^*)\}$, which is a finite set of critical costs.*

*Proof.* To prove by contradiction, consider a sequence $\alpha_i \to \alpha^*$ and a matrix $K^* \in \Gamma_M(\alpha^*)$, for which there exists no subsequence of $\alpha_i$ and $K_i \in \Gamma_M(\alpha_i)$ such that $K_i \to K^*$. We claim

- $J(K^*, \alpha^*) = M$ — otherwise by the continuity of $J$, $J(K^*, \alpha_i) < M$ for large $i$ and, since the set of stabilizing controllers is open, $K^* \in \Gamma_M(\alpha_i)$ for large $i$, which is a contradiction;

- $K^*$ must be a local minimum of $J(K, \alpha^*)$ — otherwise there exists a sequence $K_j \to K^*$ with $J(K_j, \alpha^*) < M$ and, by the continuity of $J$, there exists a sequence of large enough indices $n_j, j = 1, 2, \ldots$, such that $J(K_j, \alpha_{n_j}) < M$; the sequence $K_j \in \Gamma_M(\alpha_{n_j})$ converges to $K^*$.

The argument above implies that $M$ is the cost of some locally optimal controller at $\alpha^*$. Because the ODC problem is smooth, the locally optimal controller is a critical controller. Given $\alpha^*$, $J(K, \alpha^*)$ is expressed in (3.5) as a linear function in terms of $P_{\alpha^*}(K)$ over a semi-algebraic set given by (3.4a)-(3.4d), and the value $M$, being the cost of a locally optimal controller, is a *critical value* of $J(K, \alpha^*)$. By the semi-algebraic Sard's theorem [15, Theorem 9.6.2], the set of critical values of a linear function over a semi-algebraic set is finite. □

We are ready to proceed with the proof of the asymptotic properties.

## Asymptotic Properties

*Proof of Theorem 13.* Recall the expression of the objective function (3.2) together with the first-order necessary conditions (3.4a)-(3.4d) and the equation (3.5). Consider the semi-algebraic set

$$N = \{(K, P, L, \alpha) : (K, P, L) \text{ solves (3.4a)-(3.4d) given } \alpha,\\ P \succ 0, L \succ 0, \alpha \geq 0\}$$

and the map $f : N \to \mathbb{R}^2$ defined by

$$f(K, P, L, \alpha) = (J_\alpha(K), \alpha).$$

Due to (3.5), $f$ can be expressed as a linear function over the set $N$. Let $C_f$ denote the set of critical values of $f$. From semi-algebraic Sard's theorem [15, Theorem 9.6.2], the set $C_f \subseteq \mathbb{R}^2$ is a semi-algebraic set of dimension at most 1, and is therefore a finite union of semi-algebraic curves in $\mathbb{R}^2$. Because $C_f$ contains only finite many curves, by Bezout's theorem, they intersect at finitely many points. Furthermore, note that the set $C_f$ contains the critical cost trajectories $J(K^\dagger(\alpha), \alpha)$ in the sense that

$$C_f \supseteq \{(J(K, \alpha), K \in K^\dagger(\alpha), \alpha \geq 0\},$$

because $K^\dagger(\alpha)$ are critical points of $J(K, \alpha)$ and, therefore, critical points of $f$. The set $C_f$ may contain additional curves due to the vanishing of the Jacobian of $f$ along the second coordinate, but this does not affect the fact that the curves in $\{(J(K,\alpha), K \in K^\dagger(\alpha), \alpha \geq 0\}$ intersect at finitely many points, because the intersections of curves in $\{(J(K,\alpha), K \in K^\dagger(\alpha), \alpha \geq 0\}$ remain intersections in $C_f$. As a result, there exists an $\alpha_0 > 0$ such that for all $\alpha > \alpha_0$, the curves in $\{(J(K,\alpha), K \in K^\dagger(\alpha), \alpha \geq \alpha_0\}$ do not intersect. We claim that the following inequality holds for all $\beta > \alpha > \alpha_0$:

$$\max_{K \in K^\dagger(\beta)} J(K,\beta) \leq \max_{K \in K^\dagger(\alpha)} J(K,\beta).$$

The reason is that the maximization over both $K^\dagger(\beta)$ and $K^\dagger(\alpha)$ yields points on the same curve in $\{(J(K,\alpha), K \in K^\dagger(\alpha), \alpha \geq \alpha_0\}$, and the one in $K^\dagger(\beta)$ has a lower cost than the one in $K^\dagger(\alpha)$ from the damping property. The right-hand side of the above inequality optimizes over a fixed, finite set of controllers and approaches zero as $\beta \to \infty$ due to the representation (3.2), the bound on norm of $R_\alpha$ in Assumption 1 and the dominated convergence theorem. The left-hand side, therefore, also converges to zero as $\beta \to \infty$. From (3.5) and the assumption that $D_0$ is positive definite, we have $\|P_\beta(K)\| \to 0$ for all $K \in K^\dagger(\beta)$ as $\beta \to \infty$.

The assumption on sparsity allows the expression of the critical controllers in (3.4c) as

$$K = -R_\alpha^{-1}((B^\top P_\alpha(K) L_\alpha(K)) \circ I_S)(L_\alpha(K) \circ I_S)^{-1}. \tag{3.11}$$

Especially, we bound $\|BK\| \leq e_\alpha(K) \cdot \lambda_{\min}(L_\alpha(K))^{-1}$, where $e_\alpha(K) = \|BR_\alpha^{-1}\| \cdot \|B^\top P_\alpha(K) L_\alpha(K)\|$. The term $\|BR_\alpha^{-1}\|$ is bounded due to the assumption that the minimum eigenvalue of $R_\alpha$ is bounded away from zero. Pre- and post-multiplying (3.4b) by the unit eigenvector $v$ of the smallest eigenvalue of $L_\alpha(K)$ yields

$$\lambda_{\min}(L_\alpha(K))(2\alpha - 2v^\top(A+BK)v) = v^\top D_0 v. \tag{3.12}$$

Therefore,

$$\begin{aligned}
\lambda_{\min}(L_\alpha(K)) &\geq \frac{\lambda_{\min}(D_0)}{2\alpha + 2\|A+BK\|} \\
&\geq \frac{\lambda_{\min}(D_0)}{2\alpha + 2\|A\| + 2\|BK\|} \\
&\geq \frac{\lambda_{\min}(D_0)}{2\alpha + 2\|A\| + 2e_\alpha(K)\lambda_{\min}(L_\alpha(K))^{-1}},
\end{aligned}$$

which simplifies to

$$\lambda_{\min}(L_\alpha(K)) \geq \frac{\lambda_{\min}(D_0) - 2e_\alpha(K)}{(2\alpha + 2\|A\|)}. \tag{3.13}$$

Take the trace of (3.4b), consider the estimate

$$\begin{aligned}
2n\|A\|\|L_\alpha\| + \operatorname{tr}(D_0) &\geq 2\|A\| \operatorname{tr}(L_\alpha) + \operatorname{tr}(D_0) \\
&\geq 2\alpha \operatorname{tr}(L_\alpha) + 2 \operatorname{tr}[BR_\alpha^{-1}((B^\top P_\alpha L_\alpha) \circ I_S)(L_\alpha \circ I_S)^{-1} L_\alpha] \\
&\geq 2\alpha \operatorname{tr}(L_\alpha) - 2e_\alpha(K) \operatorname{tr}[(L_\alpha \circ I_S)^{-1} L_\alpha] \\
&= 2\alpha \operatorname{tr}(L_\alpha) - 2e_\alpha(K) n \\
&\geq 2\alpha \|L_\alpha\| - 2n\|BR_\alpha^{-1}\|\|B^\top\|\|P_\alpha\|\|L_\alpha\|,
\end{aligned} \tag{3.14}$$

where for clarity we drop the implicit dependence on $K$ in $L_\alpha$ and $P_\alpha$. The second and the third inequalities use the bound $|\operatorname{tr}(AL)| \le \|A\| \operatorname{tr}(L)$ for a positive definite matrix $L$ and any matrix $A$. The next equality in the above sequence follows from the assumption that $I_{\mathcal{S}}$ is block diagonal. The estimate (3.14), combined with the previous argument that $\|P_\alpha\| \to 0$, implies that $\|L_\alpha\| \to 0$ and thereby, $e_\alpha(K) \to 0$. The inequality (3.14) further suggests

$$\|L_\alpha\| \le \frac{\operatorname{tr}(D_0)}{2\alpha - 2n\|A\| - 2n\|BR_\alpha^{-1}\|\|B^\top\|\|P_\alpha\|}, \tag{3.15}$$

for a small enough $P_\alpha$. Combining (3.13) and (3.15) leads to

$$\begin{aligned}\|K\| &\le \|R_\alpha^{-1}\| \cdot \|(B^\top P_\alpha L_\alpha) \circ I_S\| \cdot \|(L_\alpha \circ I_S)^{-1}\| \\ &\le \|R_\alpha^{-1}\| \cdot \|B^\top\| \cdot \|P_\alpha\| \cdot \|L_\alpha\| \cdot \lambda_{\min}(L_\alpha)^{-1} \\ &\le \|R_\alpha^{-1}\| \cdot \|B^\top\| \cdot \|P_\alpha\| \\ &\quad \times \frac{\operatorname{tr}(D_0)}{2\alpha - 2n\|A\| - 2n\|BR_\alpha^{-1}\|\|B^\top\|\|\|P_\alpha\|} \\ &\quad \times \frac{(2\alpha + 2\|A\|)}{\lambda_{\min}(D_0) - 2e_\alpha(K)},\end{aligned}$$

which converges to 0 as $\alpha \to \infty$. □

We use $\otimes$ to denote the Kronecker project of two matrices and *vec* to denote the vectorized operation that stacks the columns of a matrix together into a vector. We make use of the vectorized Hessian formula in the following lemma, which will be used in the proof of Theorem 14.

**Lemma 15** (Borrowed from [76]). *Define $j_\alpha : \mathbb{R}^{m \cdot n} \to \mathbb{R}$ by $j_\alpha(vec(K)) = J(K, \alpha)$. The Hessian of $j_\alpha$ is given by the formula*

$$H_\alpha(K) = 2\left\{(L_\alpha(K) \otimes R_\alpha) + G_\alpha(K)^\top + G_\alpha(K)\right\}, \tag{3.16}$$

*where*

$$\begin{aligned}G_\alpha(K) =& [I \otimes (B^\top P_\alpha(K) + R_\alpha K)] \times \\ & [I \otimes (A - \alpha I + BK) + (A - \alpha I + BK) \otimes I]^{-1} \\ & (I_{n,n} + P(n,n))[L_\alpha(K) \otimes B]\end{aligned}$$

*and $P(n,n)$ is an $n^2 \times n^2$ permutation matrix.*

*Proof of Theorem 14.* We first show that $H_\alpha(K)$ in Lemma 15 is positive definite for any fixed $K$ when $\alpha$ is large. Recall the definition of $L_\alpha$ and $P_\alpha$ in (3.4a)-(3.4b) and apply the triangle inequality:

$$\begin{aligned}&2\alpha\|L_\alpha(K)\| \le \|D_0\| + 2\|A + BK\|\|L_\alpha(K)\|, \\ &2\alpha\|P_\alpha(K)\| \le \|Q\| + 2\|A + BK\|\|P_\alpha(K)\| + \|R_\alpha\|\|K\|^2.\end{aligned}$$

Due to Assumption 1, the term $\|R_\alpha\| \geq \delta > 0$ for all $\alpha \geq 0$. The above inequalities imply that $\|P_\alpha(K)\|/\|R_\alpha\| \to 0$ and $\|L_\alpha(K)\| \to 0$ as $\alpha \to \infty$. We now bound the minimum eigenvalue of $L_\alpha(K)$. Let $v$ be the unit eigenvector of $L_\alpha(K)$ corresponding to $\lambda_{\min}(L_\alpha(K))$; pre- and post-multiplying (3.4b) by $v$, we obtain

$$\begin{aligned}\lambda_{\min}(L_\alpha(K)) &\geq \frac{v^\top D_0 v}{2\alpha - 2v^\top (A+BK)v} \\ &\geq \frac{\lambda_{\min}(D_0)}{2\alpha + 2\|A+BK\|}.\end{aligned} \tag{3.17}$$

The first Hessian term $L_\alpha(K) \otimes R_\alpha$ in (3.16) can be lower bounded by (3.17). Due to Assumption 1, $\lambda_{\min}(R_\alpha)/\|R_\alpha\| \geq \delta/\Delta$ for all $\alpha \geq 0$. Therefore,

$$\begin{aligned}\lambda_{\min}\left(L_\alpha(K) \otimes R_\alpha\right) &= \lambda_{\min}(L_\alpha(K)) \cdot \lambda_{\min}(R_\alpha) \\ &\geq \frac{\lambda_{\min}(D_0)}{2\alpha + 2\|A+BK\|} \cdot \frac{\delta}{\Delta} \cdot \|R_\alpha\|.\end{aligned}$$

We bound the norm of the second and the third Hessian terms $\|G_\alpha(K)\|$ as follows:

$$\begin{aligned}\|G_\alpha(K)\| &\leq \|I \otimes (B^\top P_\alpha(K) + R_\alpha K)\| \\ &\quad \times \| \left[I \otimes (A - \alpha I + BK) + (A - \alpha I + BK) \otimes I\right]^{-1} \| \\ &\quad \times \| \left[I_{n,n} + P(n,n)\right] \left[L_\alpha(K) \otimes B\right] \| \\ &\lesssim \|R_\alpha\|(1 + \|P_\alpha\|/\|R_\alpha\|) \times \\ &\quad (-\lambda_{\max}\left(I \otimes (A - \alpha I + BK) + (A - \alpha I + BK) \otimes I\right))^{-1} \times \\ &\quad \|L_\alpha(K)\| \\ &\lesssim \|R_\alpha\|(2\alpha)^{-1}\|L_\alpha(K)\|,\end{aligned}$$

where $\lesssim$ hides constants that do not depend on $\alpha$. Comparing the two estimates above, for all large $\alpha$, we find that the first term $L_\alpha(K) \otimes R_\alpha$ in (3.16) dominates the following term $G_\alpha(K)^\top + G_\alpha(K)$ for a bounded $K$. Therefore, the Hessian $H_\alpha(K)$ is positive definite over bounded $K$ when $\alpha$ is large. Note that $H_\alpha(K)$ is the Hessian of the objective function when the controller is centralized. The conclusion carries over the decentralized controller because the Hessian for the decentralized controller is a principal sub-matrix of the Hessian for the centralized controller. □

*Proof of Lemma 12.* We use the Einstein notation where subscript variables that appear twice in a monomial are summed over and the subscripts that appear once are free over the corresponding set of indices. We use the lower-case letters to denote the entries of the corresponding upper-case letter matrices and write $A = (a_{ij}), B = (b_{ij}), K_\alpha = (k_{ij}), \Sigma_d = (\sigma_{ij}), P_\alpha = (p_{ij}), R_\alpha = (r_{ij}), Q = (q_{ij})$. The optimal solution $K_\alpha$ satisfies the first-order necessary condition to be derived below:

$$\begin{aligned} 0 = \frac{\partial J}{\partial k_{ij}} &= \frac{\partial[(k_{ba}r_{bc}k_{cd} + q_{ad})p_{ad}]}{\partial k_{ij}} \\ &= (r_{ic}k_{cd})p_{jd} + (k_{ba}r_{bi})p_{aj} + (k_{ba}r_{bc}k_{cd} + q_{ad})\frac{\partial p_{ad}}{\partial k_{ij}}. \end{aligned} \tag{3.18}$$

The constraints in (d-ODC($\alpha$)) may be written as

$$\alpha^2(a_{ab} + b_{ac}k_{cb})p_{bd}(a_{ed} + b_{ef}k_{fd}) - p_{ae} + \sigma_{ae} = 0 \tag{3.19}$$

Taking its partial derivatives with respect to $k_{ij}$ yields

$$\begin{aligned} &2\alpha^2 b_{ai}p_{jd}(a_{ed} + b_{ef}k_{fd}) + \\ &\quad \alpha^2(a_{ab} + b_{ac}k_{cb})\frac{\partial p_{bd}}{\partial k_{ij}}(a_{ed} + b_{ef}k_{fd}) - \frac{\partial p_{ae}}{\partial k_{ij}} = 0 \end{aligned} \tag{3.20}$$

By assumption, the entries of the controller $k_{ij}$ are bounded as $\alpha \to 0$. Hence, (3.19) implies that $P_\alpha(K_\alpha) \to \Sigma_d$ as $\alpha \to 0$ and is consequently bounded. This, combined with (3.20), implies that the partial derivatives of $P_\alpha(K)$ with respect to $K$ vanish as $\alpha \to 0$. This implies that the first two terms in (3.18), which are both $R_\alpha K_\alpha P_\alpha(K)^\top$ in matrix form, converge to zero. Since $P_\alpha(K)$ and $R_\alpha$ are invertible, $K_\alpha \to 0$ as $\alpha \to 0$. □

## Unique Stable Direction

To prove Theorem 15, define the set of *stable directions* as

$$\mathcal{H} = \{H : A + tH \text{ is stable for all stable } A \text{ and } t \geq 0\}, \tag{3.21}$$

where $A$ and $H$ are $n$-by-$n$ real matrices.

**Lemma 16.** *All matrices in $\mathcal{H}$ are similar to a diagonal matrix with non-positive diagonal entries. Especially, they cannot have complex eigenvalues.*

*Proof.* When $t$ is large, $A + tH$ is a small perturbation of $tH$. Thus, the eigenvalues of $H$ must be in the closed left half-plane. With a suitable similar transformation, assume that $H$ is in the real Jordan form. We first consider the case when the dimension $n = 2$, and we emphasize the dimension in the subscript in $H_2$ and $A_2$. To prove for contradiction, assume that $H_2$ is not diagonalizable. The non-diagonal real Jordan form of $H_2$ has the following possibilities:

- $H_2 = \begin{bmatrix} h & 1 \\ 0 & h \end{bmatrix}$, where $H_2$ has a real eigenvalue $h < 0$ of multiplicity 2: Let $A_2 = \begin{bmatrix} 4h & -2 \\ 10h^2 & -3h \end{bmatrix}$, which is stable because $\operatorname{tr}(A_2) = h < 0$ and $\det(A_2) = 8h^2 > 0$. We have $A_2 + tH_2 = \begin{bmatrix} ht + 4hby & t-2 \\ 10h^2 & ht-3h \end{bmatrix}$, whose stability criteria $\mathrm{tr}(A_2 + tH_2) < 0$ and $\det(A_2 + tH_2) > 0$ amount to

$$2ht + h < 0 \text{ and } h^2(t^2 - 9t + 8) > 0,$$

or equivalently $t \in (-1/2, 1) \cup (8, +\infty)$. In particular, when $t = 2$, the matrix $A_2 + tH_2$ is not stable.

- $H_2 = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$: Consider the stable matrix $A_2 = \begin{bmatrix} -1 & 0 \\ 1 & -1 \end{bmatrix}$, for which $A_2 + tH_2$ is not stable when $t = 2$.

- $H_2 = \begin{bmatrix} 0 & f \\ -f & 0 \end{bmatrix}$, where $f > 0$: By selecting $A_2 = \begin{bmatrix} -1 & -4 \\ 1 & -1 \end{bmatrix}$, the matrix $A_2 + \frac{2}{f} H_2 = \begin{bmatrix} -1 & -2 \\ -1 & -1 \end{bmatrix}$ is not stable.

- $H_2 = \begin{bmatrix} h & f \\ -f & h \end{bmatrix}$, where $h < 0$ and $f > 0$: By rescaling, assume that $f = 1$. Consider the matrix function

$$G(t) = \begin{bmatrix} 0 & \frac{1}{2} + (u+w)h \\ -\frac{1}{2} + (u-w)h & h \end{bmatrix} + t \begin{bmatrix} h & 1 \\ -1 & h \end{bmatrix}. \tag{3.22}$$

We have

$$\begin{aligned} \mathrm{tr}(G(t)) &= h + 2ht, \\ \det(G(t)) &= (1 + h^2)t^2 + (1 + h^2 + 2hw)t \\ &\quad + h^2(w^2 - u^2) + hw + \frac{1}{4}. \end{aligned}$$

Therefore,

$$\begin{aligned} \mathrm{tr}(G(-\frac{1}{2})) &= 0, \\ \frac{d}{dt} \mathrm{tr}\, G(t) &= 2h, \\ \det(G(-\frac{1}{2})) &= h^2(-\frac{1}{4} - u^2 + w^2), \\ \frac{d}{dt} \det G(t) \bigg|_{t=-\frac{1}{2}} &= 2hw. \end{aligned}$$

Hence, as long as

$$w > 0 \text{ and } -\frac{1}{4} - u^2 + w^2 > 0, \tag{3.23}$$

for a small enough $\epsilon > 0$, the matrix $A_2 = G(-\frac{1}{2} + \epsilon)$ is a stable matrix and there is a matrix $G(t)$ with $t > -\frac{1}{2}$ whose trace is negative and whose determinant is smaller than $\det(A_2)$. Consider the minimal value of $\det G(t)$, which is obtained at $-\frac{1}{2} - \frac{hw}{1+h^2}$:

$$\det G\left(-\frac{1}{2}-\frac{hw}{1+h^2}\right)=h^2\left(-\frac{1}{4}-u^2+\frac{h^2}{1+h^2}w^2\right).$$

As a result, when

$$-\frac{1}{4} - u^2 + \frac{h^2}{1+h^2}w^2 < 0, \tag{3.24}$$

the matrix $G(t)$ with $t = -\frac{1}{2} - \frac{hw}{1+h^2}$ is unstable. The parameters $u$ and $w$ that satisfy (3.23) and (3.24) always exist.

For the higher dimension $n > 2$, the real Jordan form of $H$ is a block upper-triangular matrix

$$H = \begin{bmatrix} H_2 & * \\ 0 & * \end{bmatrix},$$

where $H_2$ can take the four possibilities mentioned above ("$*$" denotes an arbitrary submatrix). We take the corresponding stable $A_2$ constructed above, which has the property that $A_2 + t_0 H_2$ is not stable for some $t_0 > 0$. Select a block diagonal matrix

$$A = \begin{bmatrix} A_2 & 0 \\ 0 & -I \end{bmatrix}.$$

Then, $A$ is stable, while $A + t_0 H = \begin{bmatrix} A_2 + t_0 H_2 & * \\ 0 & * \end{bmatrix}$ is unstable. □

We can strengthen the argument above and further characterize $\mathcal{H}$ in the case $n \geq 3$.

**Lemma 17.** *When $n \geq 3$, the set of stable directions $\mathcal{H}$ does not contain any matrices of rank $1, 2, \dots, n-2$.*

*Proof.* Due to Lemma 16, it suffices to consider a diagonal matrix $H$ with negative diagonal entries. Assuming that there is a matrix $H \in \mathcal{H}$ whose rank belongs to the set $\{1, 2, \dots, n-2\}$, we write

$$H = \begin{bmatrix} H_3 & 0 \\ 0 & * \end{bmatrix},$$

where $H_3 = \mathrm{diag}(-1, 0, 0)$. We will construct a stable 3-by-3 matrix $A_3$ such that $A_3 + t_0 H_3$ is unstable for some $t_0 > 0$, and then carry the instability to $A + t_0 H$ with the extended matrix

$$A = \begin{bmatrix} A_3 & 0 \\ 0 & -I \end{bmatrix}.$$

From [40], the set

$$T = \left\{ t : \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 5 & 1 & -1 \end{bmatrix} + t \begin{bmatrix} 0 \\ 0 \\ -1 \end{bmatrix} \begin{bmatrix} 0.85 \; 0.2 \; 0.2 \end{bmatrix} \text{ is stable} \right\}$$

has two disconnected components. Consider the Jordan decomposition of the matrix

$$\begin{bmatrix} 0 \\ 0 \\ -1 \end{bmatrix} \begin{bmatrix} 0.85 & 0.2 & 0.2 \end{bmatrix} = W \,\mathrm{diag}(-0.2, 0, 0) W^{-1},$$

where $W$ is some invertible matrix. This leads to the following matrix function:

$$G(t) = 5W^{-1} \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 5 & 1 & -1 \end{bmatrix} W + t \times \mathrm{diag}(-1, 0, 0).$$

After this similar transformation, the set $T$ can be written in terms of $G(t)$ as

$$T = \{ t : G(t) \text{ is stable} \}.$$

Since $T$ is disconnected, there exists some $t_1 < t_2$ such that $G(t_1)$ is stable while $G(t_2)$ is unstable with some eigenvalue in the open right half-plane. Setting $A_3 = G(t_1)$ and $t_0 = t_2 - t_1$ completes the proof. □

Since we can perturb the direction and make $H$ full-rank, the restriction on the rank of $H$ is not essential. The following lemma confirms this observation, and it completes the proof of Theorem 15.

**Lemma 18.** *When $n \geq 3$, it holds that $\mathcal{H} = \{-\lambda I, \lambda \geq 0\}$.*

*Proof.* From lemma 16, it suffices to consider the case where $H$ is diagonal with negative diagonal entries. Consider

$$H = \begin{bmatrix} H_3 & 0 \\ 0 & * \end{bmatrix},$$

where $H_3 = \mathrm{diag}(h_1, h_2, h_3)$. The diagonal entries $h_1, h_2$, and $h_3$ are non-positive and not all equal. We will construct a stable $A_3$ and a corresponding $t_0$ such that $A_3 + t_0 H_3$ is not stable, and extend to the general $A$ as in Lemma 17. The case with a rank-1 matrix $H_3$ has been considered in Lemma 17. In what follows, we prove the case for rank-2 and rank-3 matrices. Without loss of generality, we rescale $H_3$ and assume that $h_1 = -1$. Consider the following two standard forms for $H_3$:

- $H_3 = \text{diag}(-1, h_2, 0)$, where $h_2 < 0$. Consider the matrix function

$$G(t) = \begin{bmatrix} 0 & -1 & 0 \\ 0 & 0 & -h_2 \\ 2 & 1 & 0 \end{bmatrix} + tH_3 = \begin{bmatrix} -t & -1 & 0 \\ 0 & th_2 & -h_2 \\ 2 & 1 & 0 \end{bmatrix}.$$

The characteristic polynomial of $G(t)$, denoted by $\phi_{G(t)}(x)$, can be written as

$$\phi_{G(t)}(x) = x^3 + (t - th_2)x^2 + (h_2 - t^2h_2)x + (t-2)h_2.$$

The Routh-Hurwitz Criterion states that the stability of $G(t)$ is equivalent to the following system of inequalities:

$$\begin{gathered} t(1-h_2) > 0, \\ (t-2)h_2 > 0, \\ t(1-h_2)h_2(1-t^2) > (t-2)h_2. \end{gathered}$$

which can be simplified with $h_2 < 0$ to

$$0 < t < 2, \tag{3.25a}$$

$$(1-h_2)t^3 + th_2 - 2 > 0. \tag{3.25b}$$

When $t = \frac{3}{2}$, (3.25b) simplifies to the obvious expression $\frac{1}{8}(11 - 15h_2) > 0$; when $t = 3$, (3.25a) implies that $G(t)$ is not stable. Setting $A_3 = G(\frac{3}{2})$ and $t_0 = \frac{3}{2}$ completes the proof.

- $H_3 = \text{diag}(-1, h_2, h_3)$, where without loss of generally we assume that

$$-1 \leq h_2, h_3 < 0, \text{ and one of them is not } -1. \tag{3.26}$$

Consider the matrix

$$G(t) = \begin{bmatrix} 0 & -1 & 0 \\ 0 & 0 & h_2 \\ ah_3 & h_3 & 0 \end{bmatrix} + tH_3 = \begin{bmatrix} -t & -1 & 0 \\ 0 & th_2 & h_2 \\ ah_3 & h_3 & th_3 \end{bmatrix}.$$

The Routh-Hurwitz Criterion states that the stability of $G(t)$ is equivalent to the following system of inequalities:

$$t > 0, \tag{3.27a}$$

$$f_1(t) = a - t + t^3 > 0, \tag{3.27b}$$

$$\begin{aligned} f_2(t) = {} & -ah_2h_3 + th_2h_3(h_2+h_3)+ \\ & t^3(1-h_2)(1-h_3)(-h_2-h_3) > 0. \end{aligned} \tag{3.27c}$$

We claim that when

$$\sqrt{\frac{h_2h_3(h_2+h_3)^2}{(-h_2-h_3+h_2h_3)^3}} < a < \sqrt{\frac{4}{27}}, \tag{3.28}$$

the set of $t$ that satisfy the Routh-Hurwitz Criterion is disconnected. To prove this, we write the positive local minimum of $f_1(t)$ in (3.27b) as $t_1 = \sqrt{\frac{1}{3}}$ and write the positive local minimum of $f_2(t)$ in (3.27c) as $t_2 = \sqrt{\frac{h_2h_3}{3(1-h_1)(1-h_2)}}$. The condition (3.26) implies that $t_1 < t_2$ and the condition (3.28) implies that $f_1(t_1)$ and $f_2(t_2)$ are negative. Furthermore, consider $t_0 = a\frac{h_2+h_3-h_2h_3}{h_2+h_3}$, which is the root of $(1-h_2)(1-h_3)(-h_2-h_3)f_1(t) - f_2(t)$. It holds that $t_1 < t_0 < t_2$ and both $f_1(t_0)$ and $f_2(t_0)$ are positive. We conclude that when $t = t_0$, the matrix $G(t_0)$ is stable, and when $t$ is large, $G(t)$ is again stable. Yet, when $t = t_2 \in (t_0, \infty)$, the matrix $G(t_2)$ is not stable.

□

# Acknowledgment

The authors are grateful to Salar Fattahi and Cédric Josz for their constructive comments. The author thanks Yuhao Ding for sharing the implementation of local search methods. The chapter is based on the published paper [38].

# Chapter 4

# Learning of Dynamical Systems under Adversarial Attacks

This chapter complements the optimal control problem studied in the previous chapters, but instead focuses on the identification side of the problem. In particular, we are interested in the identification of a linear time-invariant dynamical system affected by sparse state disturbances modeling adversarial attacks or faults. Our identification scheme is based on $l_1$ minimization. We derive sufficient conditions for exact recovery in finite time when we have exact measurements. For noisy measurements, we provide an error bound on the estimation error. The conditions we provide are based on the null space property (NSP). We derive sufficient conditions for NSP and show that NSP holds in a particular attack model where the input is Gaussian and the adversary injects disturbances intermittently with a fixed policy based on the states and input measurements. On the other hand, we provide a lower bound to the estimation error, showing that the absence of NSP could lead to inconsistent estimates. Parts of the chapter is based on the published paper [39].

## 4.1 Introduction

The control of large-scale unknown dynamical systems, such as the power distribution networks, calls for an accurate model of the system. Recent interests in data-driven control and non-asymptotic analysis of statistical estimators provide a wealth of frameworks and tools applicable to the control of unknown dynamical systems [24, 31]. Although learning an accurate dynamical model is not necessary to achieve the control objectives, a state-space model has the advantage of being applicable to many control tasks and objectives. The issue is particularly salient in the operation of safety-critical systems, where a robust design of control laws is necessary [41].

This chapter focuses on the identification of a linear dynamical system where the states can be measured directly but are subject to unknown disturbance, accounting for adversarial attacks or faults. We prove that a type of identification scheme based on constrained lasso can perfectly recover the system matrices when the state disturbance is sparse and the measurement is perfect. The issue of robustness in identification has a long history. Dating back to Tukey [88] who made the observation that a small deviation from the model assumption could have dramatic effects on estimation and prediction, there have since been many attempts to robustify the M-estimators[1] and to use regularization to achieve robustness. The work [94] showed the equivalence of robust optimization and $l_1$-regularization for support vector machines and further attributed generalization ability to robustness against local disturbance. The more recent study [10] significantly extended the connection between robustification and regularization in regression problems.

In the system identification literature, there have been studies for the case of dense noise and the general non-smooth robust estimators [3, 4]. Those works proposed necessary and sufficient conditions for recovery that apply to all instances of robust estimation problems. The estimator of this chapter is a special instance of the general non-smooth sum-of-norms estimator studied in the above two papers, but we specialize the analysis to the case of system identification, which leads to insights on input design for a particular system matrix. Other related papers [32] and [33] studied the system identification problem subject to sparsity assumptions on the $A$ and $B$ matrices and derived improved sample complexity bounds. However, their models were based on Gaussian disturbance that is not applicable to adversarial analysis. The recent work [63] studied the identification problem using a conic relaxation, which linearizes the problem at the expense of increasing the problem dimension. More recently, [77] proved finite-time identification bounds for linear dynamical systems without control input. The identification method is based on ordinary least-squares, which succeeds under the important assumption of regular matrices. Concurrently, [70] proved non-asymptotic bounds for system identification with Markov parameters, which are estimated using least-squares and the Kalman-Ho algorithm. It is challenging to generalize those algorithms to the case when the samples are missing or when they are corrupted. The set-membership estimator can deal with missing samples and is consistent [49], but the disturbance is assumed to be bounded.

Other related lines of work in the control literature involve the identification of switched systems with noisy measurements [50, 71] and system identification in the presence of output attacks [81]. In contrast, we study the case with contaminated states, whose effect propagates over time. Other fruitful ideas include attack resilient state estimation [34, 22] (where the goal is to recover the system state) and Byzantine fault tolerance [86, 47] (where a collection of redundant agents can prevent an attack by faulty agents in the computation of an optimization problem).

To situate the work in the broader context, we discuss related works on robust regression. The paper [79] studied the related problem of outlier detection in linear regression. It proved the equivalence of adding a penalty to the least-squares loss function and using an alternative loss function to the least-squares loss. In particular, it noted that $l_1$ regularization is

[1] M-estimators optimize the sample average of loss.

equivalent to using the Huber loss[2] and that Huber loss may not be the best choice for guaranteeing robustness in many cases — a non-convex loss function may be more appropriate. However, unless in very specialized settings, the theoretical justifications of non-convex estimators are rare, and the computation of non-convex estimators is not well-understood [52, 64]. The work [11] solved the problem of regression with sparse disturbance via iterative hard thresholding. There has been a flurry of recent papers on robust training [25, 74, 85]. Nevertheless, the independence assumption between samples renders them inapplicable to system identification — the state measurements are dependent and cannot be re-ordered. Transforming the data samples to deal with missing data in linear regression does not directly translate to the system identification case due to the need to measure several trajectories or solve nonlinear optimization problems. It is undesirable to reset the system in practical applications. Furthermore, it is unclear how identification can be achieved robustly in an online fashion.

Section 4.2 considers a particular type of $l_1$ minimization problem as a solution to the identification of system matrices $(A, B)$. Our problem differs from the usual literature on $l_1$ minimization in that the system identification setting naturally has correlated inputs and outputs. In Section 4.3, we derive sufficient conditions for exact recovery in finite time when we have exact measurements. The noisy case is studied in Section 4.4, where we provide an error bound on the estimation error. The conditions are based on the null space property (NSP), which is hard to verify directly. We derive sufficient conditions for NSP in Section 4.5 and in Section 4.6 show that NSP holds in a particular attack model where the input is Gaussian and the adversary injects disturbances intermittently with a fixed policy based on the states and input measurements. Section 4.7 complements the upper bound by providing a lower bound of the estimation error where the state information is known to the attacker, showing that the absence of NSP leaves the door open for attacks that lead to inconsistent estimators.

## 4.2 Problem Formulation

Consider the linear time-invariant dynamical system over the time horizon $[0, T]$:

$$x_{t+1} = \bar{A}x_t + \bar{B}u_t + \bar{d}_t. \quad t = 0, 1, \dots, T-1,$$

where $\bar{A} \in \mathbb{R}^{n\times n}, \bar{B} \in \mathbb{R}^{n\times m}$ are unknown matrices in the state space model to be estimated and $\bar{d}_t$'s are unknown disturbances. Throughout the chapter the bar over the variables indicates the unknown ground truth. The goal is to find the matrices $\bar{A}$ and $\bar{B}$ from the state measurements $x_0, ..., x_T \in \mathbb{R}^n$ and input data $u_0, ..., u_{T-1} \in \mathbb{R}^m$. The disturbances $\bar{d}_0, ..., \bar{d}_{T-1}$ model both noises and anomalies in the system, such as attacks or actuator's faults. The initial state $x_0$ is assumed known and the remaining states $x_t, t > 0$ depend on the input and

[2] The Huber Loss is a piece-wise function defined by $H_b(x) = \begin{cases} \frac{1}{2}x^2, & \text{if } |x| \le b \\ b(|x| - \frac{1}{2}b), & \text{otherwise} \end{cases}.$

the disturbances. Without any assumptions on the disturbance, the identification problem is not well-defined due to the impossibility of separating $\bar{A}x_t + \bar{B}u_t$ from the disturbance $\bar{d}_t$. In particular, if $\bar{d}_t = A'x_t + B'u_t$, then the system evolves as if the system matrices are $(\bar{A} + A', \bar{B} + B')$. We will make certain sparsity assumption on the disturbance signal in the noiseless case, and generalize the result to the noisy case.

For clarity of notation, we introduce the matrix notation $X = [x_0, \ldots, x_{T-1}]$, $U = [u_0, \ldots, u_{T-1}]$, and $D = [d_0, \ldots, d_{T-1}]$. The last state $x_T$ will appear in our optimization problem but it is not a column in the matrix notation. The attack $D$ is assumed to be restricted to a set $\mathcal{D} \subseteq \mathbb{R}^{n\times T}$. The set $\mathcal{D}$ captures the user's belief of possible places of attack and its directions. The $i$-th largest singular value of a matrix $U$ is denoted by $\sigma_i(U)$, the minimal and maximum singular values by $\sigma_{\min}(U)$ and $\sigma_{\max}(U)$, respectively.

Define the sum of norm error $\|D\|_{2,col} := \sum_i \|d_i\|_2$, where the index is over the columns of $D$. The (column-wise) support of $D$ is defined as $\operatorname{supp}(D) = \{i \in \{0, \ldots, T-1\} : d_i \neq 0\}$. For any subset of indices $I \subseteq \{0, 1, \ldots, T-1\}$, the complement of $I$ is defined as $I^c = \{i \in \{0, \ldots, T-1\} : i \notin I\}$. For any matrix $U \in \mathbb{R}^{n\times T}$, the *projection* $\Pi_I U$ is a matrix of the same size as $U$ and its columns are zero except for those in $I$, i.e.,

$$(\Pi_I U)_i = \begin{cases} u_i, & \text{if } i \in I \\ 0, & \text{otherwise} \end{cases}.$$

In contrast, the subset matrix $U_I$ has size $n \times |I|$, selecting only the columns of $U$ in the index set $I$. We use $U_{\neq i}$ as a shorthand for $U_{\{0,\ldots,T-1\}\setminus\{i\}}$ and $U_{\notin I}$ as a shorthand for $U_{\{0,\ldots,T-1\}\setminus I}$. The range of $U$ is defined as $\operatorname{range} U = \{\sum_i \lambda_i u_i, \lambda_i \in \mathbb{R}\}$.

The Minkowski sum of sets $\mathcal{E}$ and $\mathcal{F}$ is denoted by $\mathcal{E} \oplus \mathcal{F} = \{e + f, e \in \mathcal{E}, f \in \mathcal{F}\}$. The sum with the inverse of the set is denoted by $\mathcal{E} \ominus \mathcal{F} = \{e - f, e \in \mathcal{E}, f \in \mathcal{F}\}$.

To recover the system matrices $A$ and $B$, we analyze the following $l_1$-optimization problem:

$$\min_{A,B,D\in\mathcal{D}} \sum_{i=0}^{T-1} \|d_i\|_2 \tag{4.1}$$

$$s.t \quad x_{i+1} = Ax_i + Bu_i + d_i, \quad i = 0, \ldots, T-1, \tag{4.2}$$

where the states $x_i, i \in \{0, \ldots, T\}$ are generated according to

$$x_{i+1} = \bar{A}x_i + \bar{B}u_i + \bar{d}_i, \quad i = 0, \ldots, T-1. \tag{4.3}$$

The initial state $x_0$ is known. The control inputs $u_i, i \in \{0, \ldots, T-1\}$ are to be designed but fixed in the optimization problem (4.1). Problem (4.1) differs from the classical $l_1$ minimization (basis pursuit) problem

$$\begin{aligned} \min_z \quad & \|z\|_1 \\ s.t. \quad & \Phi\bar{z} = \Phi z, \end{aligned}$$

in that

- We apply the $l_1$ norm at the group level to the disturbances $d_1, \dots, d_{T-1}$, because we only assume sparsity in the occurrence of the disturbance but not the disturbance itself.
- The disturbances matrix $D$ is limited to a set $\mathcal{D}$.
- We do not attempt to minimize the $l_1$ norm of all the unknown parameters. In particular, the system matrices $A$ and $B$ are not assumed to be sparse.
- The states $x_i, 0 < i \leq T-1$ appear both as inputs and as measurements in the constraints (4.2). They also depend on the exogenous input $u_i$ and the disturbances $d_i$.
- Because the states are correlated, we cannot independently rescale them, as is commonly done in the analysis of $l_1$ optimization problems.

## 4.3 The Noiseless Case

This section studies the noiseless case, where $\{\bar{d}_i, i \in \{0, 1, \dots, T-1\}\}$ are either 0 or come from the attacker who is attempting to disturb the running of the system. We aim to understand how to design the input of the system (in case that is an option) so that the identification of the excited system in the presence of adversarial disturbances is possible.

We use $S = \operatorname{supp}(\bar{D})$ to denote the time stamps of actual attacks. The set of disturbances $\mathcal{D}$ is assumed to be closed under the projection onto $S$.

**Assumption 4.** *The set of disturbances $\mathcal{D}$ is convex and contains $0$ in its interior. Furthermore, $\Pi_S(D) \in \mathcal{D}$ for all $D \in \mathcal{D}$.*

A key construction in the study of (4.1) is the Null Space Property[42], which is formalized below.

**Definition 5.** *Let $c > 0$, $S \subsetneq \{0, \dots, T-1\}$, and $\mathcal{R}$ be a subset of $\mathbb{R}^{n \times T}$. The matrix $\begin{bmatrix} X \\ U \end{bmatrix} \in \mathbb{R}^{(n+m) \times T}$ is said to satisfy the Null Space Property with constant c, index set S, and range set $\mathcal{R}$ ($(c, S, \mathcal{R})$-NSP), if for all matrix pair $A \in \mathbb{R}^{n \times n}, B \in \mathbb{R}^{n \times m}$ such that $-AX - BU \in \mathcal{R}$ and $(A, B)$ are not both zero, we have*

$$\|[A, B] \begin{bmatrix} X_S \\ U_S \end{bmatrix}\|_{2,col} < c \|[A, B] \begin{bmatrix} X_{S^c} \\ U_{S^c} \end{bmatrix}\|_{2,col}. \tag{4.4}$$

When the set $S$ or $\mathcal{R}$ is obvious in the context, we omit them and use $c$-NSP or $(c, S)$-NSP to highlight the parameters of interest. NSP was already mentioned in the original paper proving an exact recovery result with restricted isometry property [21] and was carefully studied in the paper [23]. The theorem below formalizes the standard result that roughly states that 1-NSP is sufficient for the exact recovery of all sparse disturbances.

**Theorem 16.** *Assume that* $\begin{bmatrix} X \\ U \end{bmatrix}$ *satisfies the* $(1, S, \mathcal{D} \ominus \mathcal{D})$*-NSP where* $S = \operatorname{supp}(\bar{D})$*, then* $(\bar{A}, \bar{B}, \bar{D})$ *is the unique solution to problem* (4.1)*.*

*Proof.* Let $(A, B,\ D)$ be any feasible solution to (4.1). We will show that if the matrices are not equal to the ground truth $(\bar{A}, \bar{B}, \bar{D})$, then they cannot be an optimal solution. The feasibility can be written as:

$$\begin{aligned} x_{i+1} &= \bar{A}x_i + \bar{B}u_i + \bar{d}_i, \quad i = 0, \dots, T-1, \\ x_{i+1} &= Ax_i + Bu_i + d_i, \quad i = 0, \dots, T-1. \end{aligned}$$

Taking the difference of the two equality, $(\bar{A} - A, \bar{B} - B, \bar{D} - D)$ lies in the null space in the sense that

$$0 = (\bar{A} - A)x_i + (\bar{B} - B)u_i + (\bar{d}_i - d_i), \quad i = 0, \dots, T-1.$$

which can be written in matrix form as

$$0 = (\bar{A} - A)X_S + (\bar{B} - B)U_S + (\bar{D}_S - D_S) \tag{4.5}$$

$$0 = (\bar{A} - A)X_{S^c} + (\bar{B} - B)U_{S^c} - D_{S^c}, \tag{4.6}$$

where $S = \operatorname{supp}(\bar{D})$. Note that $\bar{D} - D \in \mathcal{D} \ominus \mathcal{D}$. If $A - \bar{A} = 0$ and $B - \bar{B} = 0$, then $D = \bar{D}$. Suppose $D \neq \bar{D}$, then $A - \bar{A}$ and $B - \bar{B}$ are not both zero. We apply the null space property to derive the following inequalities.

$$\begin{aligned} &\|D\|_{2,col} = \|D_S\|_{2,col} + \|-D_{S^c}\|_{2,col} \\ &= \|D_S\|_{2,col} + \|(\bar{A} - A)X_{S^c} + (\bar{B} - B)U_{S^c}\|_{2,col} \\ &> \|D_S\|_{2,col} + \|(\bar{A} - A)X_S + (\bar{B} - B)U_S\|_{2,col} \quad \text{(1-NSP)} \\ &= \|D_S\|_{2,col} + \|(\bar{D} - D)_S\|_{2,col} \\ &\geq \|\bar{D}_S\|_{2,col} \quad \text{(triangle inequality)} \\ &= \|\bar{D}\|_{2,col} \text{ (sparsity of disturbance).} \end{aligned}$$

This means that $(A, B, D)$ is not an optimal solution to (4.1). □

**Remark 9.** *[3] showed that 1-NSP is necessary for the exact recovery for* all *instances of a certain class of robust regression problems. However, because* $x_i, 1 \leq i \leq T-1$ *appear on both sides of the constraints* (4.2)*, the system identification problems are determined by inputs* $U$ *and disturbances* $D$*, which is only a subset of all instances of the regression problems.*

**Remark 10.** *The proof of Theorem 16 can be directly applied to the case without control input, for which the* $(c, S, \mathcal{D} \ominus \mathcal{D})$*-NSP takes the form*

$$\|AX_S\|_{2,col} < c\|AX_{S^c}\|_{2,col} \tag{4.7}$$

*for all $A \neq 0 \in \mathbb{R}^{n\times n}$ such that $-AX \in \mathcal{D} \ominus \mathcal{D}$. NSP property with $c = 1$ ensures that $\bar{A}$ is the unique solution to the optimization problem*

$$\min_{A,D\in\mathcal{D}} \sum_{i=0}^{T-1} \|d_i\|_2 \tag{4.8}$$

$$s.t \quad x_{i+1} = Ax_i + d_i, \quad i = 0,\dots,T-1, \tag{4.9}$$

*where the states $x_i, i \in \{0,\dots,T\}$ are generated according to*

$$x_{i+1} = \bar{A}x_i + \bar{d}_i, \quad i = 0,\dots,T-1.$$

□

## 4.4 The Noisy Case

This section studies the noisy case, where $d_i, i \in \{0,\dots,T-1\}$ are not sparse but are a combination of sparse attack and noise. We use the set $S$ in this section to denote the location of attacks. $S$ is no longer the support of $\bar{D}$ in this section and $\bar{D}_{S^c}$, whose columns are noises at times without attack, may be nonzero.

The next theorem studies the error bound for estimating $A$ and $B$ matrices.

**Theorem 17.** *Assume that $T > (m+n)$ and that the matrix $\begin{bmatrix} X \\ U \end{bmatrix}$ has full row rank. Suppose $(X,U)$ satisfies the $(c, S, \mathcal{D}\ominus\mathcal{D})$-NSP with $c < 1$, then the solution $(\hat{A},\hat{B},\hat{D})$ to the optimization problem* (4.1) *satisfies*

$$\|[\hat{A}-\bar{A}, \hat{B}-\bar{B}]\|_F \leq 2\frac{1+c}{1-c}\frac{\|\bar{D}_{S^c}\|_{2,col}}{\sigma_{\min}\left(\begin{bmatrix} X \\ U \end{bmatrix}\right)}.$$

*Proof.* The optimality of the solution implies that $\|\hat{D}\|_{2,col} \leq \|\bar{D}\|_{2,col}$. The constraints implies

$$0 = (\bar{A}-\hat{A})X + (\bar{B}-\hat{B})U + (\bar{D}-\hat{D}).$$

$c$-NSP implies

$$\|\bar{D}_S - \hat{D}_S\|_{2,col} < c\|\bar{D}_{S^c} - \hat{D}_{S^c}\|_{2,col} \tag{4.10}$$

or

$$\|\bar{D} - \hat{D}\|_{2,col} < (1+c)\|\bar{D}_{S^c} - \hat{D}_{S^c}\|_{2,col} \tag{4.11}$$

We have the following bounds

$$\begin{aligned}
&\|\bar{D}\|_{2,col} \geq \|\hat{D}\|_{2,col}\\
&= \|\bar{D} + (\hat{D} - \bar{D})\|_{2,col}\\
&= \|\bar{D}_S + (\hat{D}_S - \bar{D}_S)\|_{2,col} + \|\bar{D}_{S^c} + (\hat{D}_{S^c} - \bar{D}_{S^c})\|_{2,col}\\
&\geq \|\bar{D}_S\|_{2,col} - \|\hat{D}_S - \bar{D}_S\|_{2,col} - \|\bar{D}_{S^c}\|_{2,col} + \|\hat{D}_{S^c} - \bar{D}_{S^c}\|_{2,col}\\
&\geq \|\bar{D}_S\|_{2,col} - \|\bar{D}_{S^c}\|_{2,col} + (1-c)\|\hat{D}_{S^c} - \bar{D}_{S^c}\|_{2,col}\\
&\geq \|\bar{D}_S\|_{2,col} - \|\bar{D}_{S^c}\|_{2,col} + \frac{1-c}{1+c}\|\hat{D} - \bar{D}\|_{2,col}
\end{aligned}$$

where we applied the triangle inequality and (4.10)-(4.11). Cancelling $\|\bar{D}_S\|_{2,col}$ on both sides, we obtain

$$\|\hat{D} - \bar{D}\|_{2,col} \leq 2\frac{1+c}{1-c}\|\bar{D}_{S^c}\|_{2,col}.$$

The bound above can be translated to the bound on $(A, B)$ through the matrix norm inequality (note that $T \geq (m+n)$).

$$\begin{aligned}
&\|[\hat{A} - \bar{A}, \hat{B} - \bar{B}]\|_F \sigma_{\min}\left(\begin{bmatrix} X \\ U \end{bmatrix}\right)\\
&\leq \|\hat{D} - \bar{D}\|_F \leq \|\hat{D} - \bar{D}\|_{2,col} \leq 2\frac{1+c}{1-c}\|\bar{D}_{S^c}\|_{2,col}.
\end{aligned}$$

□

**Remark 11.** *The term $2\|\bar{D}_{S^c}\|_{2,col}$ on the right-hand side of Theorem 17 can be improved to $2\|\bar{D}_{S^c}\|_{2,col} - \|\bar{D}\|_{2,col} + \|\hat{D}\|_{2,col}$ using similar techniques as those in basis pursuit, see for example [42, Theorem 4.14], where an equivalence to c-NSP for basis pursuit problems is proved. As in Remark 9, optimization problem 4.1 is a special case of basis pursuit where the inputs and measurements are correlated. The bound, including the constant $\frac{1+c}{1-c}$, could potentially be improved with the knowledge of the constraints, an example for the basis pursuit problem is provided in [29, Chapter 7].*

## 4.5 Satisfaction of NSP

Condition (4.4) can help us confirm, after observing the states and input sequence, whether we can recover the true dynamics exactly. Theorem 17 has shown that the NSP condition is useful in obtaining an error bound for the identification error. The following lemmas attempt to derive stronger conditions that are more tractable than $(c, S, \mathcal{D})$-NSP. They can be combined with the results of the previous two sections to understand the type of input design that is robust to attackers and when we can recover the system matrices through the $l_1$ optimization problem (4.1).

**Lemma 19.** *Assume that $T \geq (m+n)$ and*

$$\sqrt{|S|}\sigma_{\max}\begin{bmatrix}X_S\\U_S\end{bmatrix} < c \cdot \sigma_{\min}\begin{bmatrix}X_{S^c}\\U_{S^c}\end{bmatrix}, \tag{4.12}$$

*where $S = \operatorname{supp}(\bar{D})$ and $|S^c| \geq m+n$, then $\begin{bmatrix}X\\U\end{bmatrix}$ satisfies the $(c, S, \mathcal{R})$-NSP for any range set $\mathcal{R}$.*

*Proof.* For any matrices $A, B$ such that $-AX - BU \in \mathcal{R}$ and $(A, B)$ are not both zero, we can upper bound and lower bound the norms:

$$\begin{aligned}\|[A,B]\begin{bmatrix}X_S\\U_S\end{bmatrix}\|_{2,col} &\leq \sqrt{|S|}\|[A,B]\begin{bmatrix}X_S\\U_S\end{bmatrix}\|_F\\ &\leq \sqrt{|S|}\|[A,B]\|_F\sigma_{\max}\left(\begin{bmatrix}X_S\\U_S\end{bmatrix}\right)\\ \|[A,B]\begin{bmatrix}X_{S^c}\\U_{S^c}\end{bmatrix}\|_{2,col} &\geq \|[A,B]\begin{bmatrix}X_{S^c}\\U_{S^c}\end{bmatrix}\|_F\\ &\geq \|[A,B]\|_F\sigma_{\min}\left(\begin{bmatrix}X_{S^c}\\U_{S^c}\end{bmatrix}\right),\end{aligned}$$

where we use the relationship between $(2, col)$-norm and Frobenius norm. The last inequality uses the assumption $|S^c| \geq m+n$. The inequality (4.12) therefore implies 1-NSP. The last statement follows from Theorem 16 by setting $\mathcal{R} = \mathcal{D} \ominus \mathcal{D}$ and $c = 1$. □

**Definition 6.** *For a matrix $V = [v_0, \ldots, v_{T-1}]$. $V$ is said to be $s$-self-decomposable if for all indices $I \subseteq \{0, 1, \ldots, T-1\}$ of size $|I| = s$, we have*

$$V_i \in \operatorname{range}(V_{\notin I}) \textit{ for all } i \in I.$$

*The $s$-self-decomposable amplitude is defined as*

$$\xi_s(V) = \max_{\substack{I\subseteq\{0,\ldots,T-1\}\\|I|=s}} \min_{\substack{\Gamma_I\in\mathbb{R}^{(T-s)\times s}\\ \Gamma_I=[\gamma_i]_{i\in I}}} \left\{\sum_{k\in I}\|\gamma_k\|_\infty : V_I = V_{\notin I}\Gamma_I\right\}. \tag{4.13}$$

If $U$ is $s$-self-decomposable, by definition it is also $t$-self-decomposable for $t < s$. We are particularly interested in the cases when $s = 1$ and $s = |S|$.

**Lemma 20.** *Assuming that $\begin{bmatrix}X\\U\end{bmatrix}$ has full row rank and is $s$-self-decomposable where $s = |S|$, then it satisfies the $(c, S, \mathcal{R})$-NSP for $c > \xi_s(\begin{bmatrix}X\\U\end{bmatrix})$.*

*Proof.* Assuming that $\begin{bmatrix} X \\ U \end{bmatrix}$ is $s$-self-decomposable where $s = |S|$, then we can find a matrix $\Gamma^*_S = [\gamma^*_i]_{i \in S}$ that is the minimizer of the inner optimization problem (4.13).

$$\begin{aligned} \|[A, B] \begin{bmatrix} X_S \\ U_S \end{bmatrix}\|_{2,col} &= \|[A, B] \begin{bmatrix} X_{S^c} \\ U_{S^c} \end{bmatrix} \Gamma^*_S\|_{2,col} \\ &\leq \sum_{i \in S} \|\gamma^*_i\|_\infty \|[A, B] \begin{bmatrix} X_{S^c} \\ U_{S^c} \end{bmatrix}\|_{2,col} \\ &\leq \xi_s(\begin{bmatrix} X \\ U \end{bmatrix}) \|[A, B] \begin{bmatrix} X_{S^c} \\ U_{S^c} \end{bmatrix}\|_{2,col}. \end{aligned}$$

It remains to obtain a strict inequality when we relax $c > \xi_s(\begin{bmatrix} X \\ U \end{bmatrix})$. Suppose for contradiction that this does not yield a strict inequality, we must have

$$[A, B] \begin{bmatrix} X_{S^c} \\ U_{S^c} \end{bmatrix} = 0 \text{ and } [A, B] \begin{bmatrix} X_S \\ U_S \end{bmatrix} = 0.$$

Since $\begin{bmatrix} X \\ U \end{bmatrix}$ has full row rank, $[A, B] = 0$. □

**Lemma 21.** *Given $S = \mathrm{supp}(\bar{D})$ where $|S| > 1$, assume that $\begin{bmatrix} X \\ U \end{bmatrix}$ has full row rank and is 1-self-decomposable where*

$$\xi_1 = \xi_1(\begin{bmatrix} X \\ U \end{bmatrix}) \leq \frac{1}{|S| - 1}.$$

*Then, it satisfies the $(c, S, \mathcal{R})$-NSP for any subset $\mathcal{R}$ and*

$$c > \frac{|S|\xi_1}{1 - (|S| - 1)\xi_1}.$$

*Proof.* For any $s \in S$ we can find a vector $\gamma^*_s$ that is the minimizer of the inner optimization problem (4.13).

$$\begin{aligned} \|[A, B] \begin{bmatrix} X_s \\ U_s \end{bmatrix}\|_2 &= \|[A, B] \begin{bmatrix} x_s \\ u_s \end{bmatrix}\|_2 \\ &= \|[A, B] \begin{bmatrix} X_{\neq s} \\ U_{\neq s} \end{bmatrix} \gamma^*_s\|_2 \\ &\leq \|\gamma_s\|_\infty \|[A, B] \begin{bmatrix} X_{\neq s} \\ U_{\neq s} \end{bmatrix}\|_{2,col} \\ &\leq \xi_1 \|[A, B] \begin{bmatrix} X_{\neq s} \\ U_{\neq s} \end{bmatrix}\|_{2,col} \end{aligned}$$

Hence

$$\|[A,B]\begin{bmatrix}X_s\\U_s\end{bmatrix}\|_2 \leq \frac{\xi_1}{1+\xi_1}\|[A,B]\begin{bmatrix}X\\U\end{bmatrix}\|_{2,col}.$$

Summing over $s \in S$, we obtain

$$\|[A,B]\begin{bmatrix}X_S\\U_S\end{bmatrix}\|_{2,col} \leq |S|\frac{\xi_1}{1+\xi_1}\|[A,B]\begin{bmatrix}X\\U\end{bmatrix}\|_{2,col}.$$

Rearranging terms and the proof is concluded by noting that, as in the proof of Lemma 20, the full rank assumption implies that when we select $c > \frac{|S|\xi_1}{1-(|S|-1)\xi_1}$ the inequality is strict for $(A,B) \neq 0$. □

**Remark 12.** *Lemma 21 implies that when $\xi_1 < \frac{1}{2|S|-1}$, the ground truth $(\bar{A},\bar{B},\bar{D})$ is recoverable through* (4.1)*.*

We have yet not answered how to design the control input $u_i, i \in \{0,\ldots,T-1\}$ such that the NSP holds. This is best done after we consider a probabilistic model of the disturbances.

## 4.6 A Probabilistic Model

The results of the previous section are still applicable only when the state and input sequence has been observed — they have not yet given a concrete input design scheme that achieves exact recovery in the noiseless case, or asymptotic recovery in the noisy case. This section considers a particular type of random input. We will make the observation that, despite the attacker's attempt, we can apply the block martingale small ball condition [83] and obtain a probabilistic estimation on $\sigma_{\min}\left(\begin{bmatrix}X\\U\end{bmatrix}\right)$.

We first restate the block martingale small ball (BMSB) condition following [83]. Define the filtration $\mathcal{F}_t$ as

$$\mathcal{F}_t = \begin{cases}\sigma(x_0,u_0), & \text{if } t=0\\ \sigma\left(x_0,\ldots,x_t,u_0,\ldots,u_t,d_0,\ldots,d_{t-1}\right) & \text{if } t\geq 1\end{cases},$$

so that the vector-valued process $\begin{bmatrix}x_t\\u_t\end{bmatrix}, t \geq 0$ is $\{\mathcal{F}_t\}_{t\geq 0}$ adapted.

**Definition 7.** *Given a filtration $\{\mathcal{F}_t\}_{t\geq 0}$, and a vector-valued process $V_t \in \mathbb{R}^d, t \geq 0$. The process is said to satisfy $(k,\Gamma_{sb},p)$-BMSB for $\Gamma_{sb} \succ 0$ if*

$$\frac{1}{k}\sum_{i=1}^{k}\mathbb{P}\left(|\langle w,V_{j+i}\rangle|^2 \geq w^\top\Gamma_{sb}w|\mathcal{F}_j\right) \geq p, \text{ almost surely}$$

*for any fixed $w \in \mathbb{R}^d$ with $\|w\|_2 = 1$ and any $j \geq 0$.*

We make the following two assumptions about the input and attack model. They will ensure that the process is Gaussian.

**Assumption 5.** *The input sequences $u_t, t \geq 0$ are independent and identically distributed Gaussian $N(0, \sigma^2 I)$ random variables.*

**Assumption 6.** *The attack $d_t, d \geq 0$ satisfies the following condition*

- *The set of attack $S \subseteq \{0, \ldots, T-1\}$ is fixed.*
- *For $t \notin S$, $d_t$ is the noise and follows the distribution $N(0, \epsilon^2 I)$.*
- *For $t \in S$, $d_t = Px_t + Qu_t + e_t$, where $P$ and $Q$ are constant matrices of compatible size. They are not dependent on $\mathcal{F}_t$. The sequence of random noise $e_t$ follows the Gaussian distribution $N(0, \epsilon^2 I)$ and is independent of $\mathcal{F}_t$.*

We can bound the Gramian matrix, which is identified as a key measure of sample complexity in [83]. We formalize the result in the Lemma below. Define the following constants

$$\begin{aligned}
\alpha_{\min} &= \min(\sigma_{\min}(A+P), \sigma_{\min}(A)) \\
\alpha_{\max} &= \max(\sigma_{\max}(A+P), \sigma_{\max}(A)) \\
\beta_{\max} &= \max(\sigma_{\max}(B+Q), \sigma_{\max}(B)).
\end{aligned}$$

**Lemma 22.** *Let $\Gamma_t = \mathbb{E}\left[x_t x_t^\top\right]$, we have*

$$\begin{aligned}
\Gamma_t &\succeq \alpha_{\min}^2 \Gamma_{t-1} + \epsilon^2 I \\
\Gamma_t &\preceq \alpha_{\max}^2 \Gamma_{t-1} + \left(\epsilon^2 + \beta_{\max}^2\right) I.
\end{aligned}$$

*In particular, for $t \geq 1$,*

$$\begin{aligned}
\Gamma_t &\succeq \sum_{0 \leq i \leq t-1} \alpha_{\min}^{2i} \epsilon^2 I \\
\Gamma_t &\preceq \Gamma_t^{\max} = \alpha_{\max}^{2t} \Gamma_0 + \sum_{0 \leq i \leq t-1} \alpha_{\max}^{2i} \left(\epsilon^2 + \beta_{\max}^2\right) I.
\end{aligned}$$

*Proof.* Condition on $\mathcal{F}_{t-1}$; we have

$$\begin{aligned}
&\mathbb{E}[x_t x_t^\top | \mathcal{F}_{t-1}] \\
&= \mathbb{E}[(Ax_{t-1} + Bu_{t-1} + d_{t-1}) \\
&\quad \cdot (Ax_{t-1} + Bu_{t-1} + d_{t-1})^\top | \mathcal{F}_{t-1}].
\end{aligned}$$

We analyze two cases

- When $t-1 \in S$, the term becomes

$$\begin{aligned}&\mathbb{E}[((A+P)x_{t-1}+(B+Q)u_{t-1}+e_{t-1})\\&\quad\cdot((A+P)x_{t-1}+(B+Q)u_{t-1}+e_{t-1})^\top|\mathcal{F}_{t-1}].\end{aligned}$$

Taking the expectation on both sides, we obtain

$$\begin{aligned}\Gamma_t &= \mathbb{E}\left[x_t x_t^\top\right]\\&\overset{(b)}{=} (A+P)\Gamma_{t-1}(A+P)^\top + \sigma^2(B+Q)(B+Q)^\top + \epsilon^2 I.\end{aligned}$$

where (b) follows by noting that $u_{t-1}$ and $e_{t-1}$ are independent of $x_{t-1}$ and has mean zero.

- When $t-1 \notin S$, the term becomes

$$\begin{aligned}&\mathbb{E}[(Ax_{t-1}+Bu_{t-1}+e_{t-1})\\&\quad\cdot(Ax_{t-1}+Bu_{t-1}+e_{t-1})^\top|\mathcal{F}_{t-1}].\end{aligned}$$

Taking the expectation in a similar way

$$\Gamma_t = A\Gamma_{t-1}A^\top + \sigma^2 BB^\top + \epsilon^2 I.$$

In both cases, we can lower bound $\Gamma_t$ by leaving out the positive semi-definite term and using the minimal singular values of the multipliers.

$$\Gamma_t \succeq \alpha_{\min}^2\Gamma_{t-1} + \epsilon^2 I.$$

The upper bound follows similarly by bounding with the maximum singular values. The result of the lemma follows by induction. □

Let $\Gamma = \begin{bmatrix}\epsilon^2 I_n & \\ & \sigma^2 I_m\end{bmatrix}$. The next lemma confirms that indeed BMSB condition applies to our setting.

**Lemma 23.** *Under the Assumption 5-6. For any sequence of indices* $0 \le s_0 < s_1 < s_2, \ldots$, *the sub-process* $\begin{bmatrix}x_{s_t}\\u_{s_t}\end{bmatrix}, t \ge 0$ *satisfies the* $(k, \frac{1}{2}\Gamma, \frac{1}{12})$*-BMSB condition.*

*Proof.* For clarity of notation we will prove the result for $s_t = t$ and then note that the proof does not depend on the fact that $s_t = t$.

We will prove that the process is 3-Paley-Zygmund [82, Lemma 3.9] and conclude BMSB as a consequence, following a similar argument in [82]. Fix a vector $\begin{bmatrix}w\\v\end{bmatrix} \in \mathbb{R}^{n+m}$. At any fixed time $j \ge 0$ and $i \ge 1$, we write

$$x_{i+j}|F_j = A^i x_j + \sum_{0\le k\le i-1} A^{i-k-1}(Bu_{j+k}+d_{j+k})|F_j$$

We may substitute the expression of $d_{s+i}, 0 \leq i \leq t-1$ and find that the conditional distribution of $\langle w, x_{t+s}\rangle + \langle v, u_{t+s}\rangle | \mathcal{F}_s$ is Gaussian. Let $Y_i = (\langle w, x_{j+i}\rangle + \langle v, u_{j+i}\rangle)^2$ and $Z_{j+i} = \begin{bmatrix} x_{j+i} \\ u_{j+i} \end{bmatrix}$ for $i \geq 1$. We can calculate

$$\begin{aligned}
\mathbb{E}[Y_i|\mathcal{F}_j] &= [w, v]^\top \mathbb{E}[Z_{j+i} Z_{j+i}^\top | \mathcal{F}_j] \begin{bmatrix} w \\ v \end{bmatrix} \\
&\overset{(a)}{=} w^\top \mathbb{E}[x_{j+i} x_{j+i}^\top | \mathcal{F}_j] w + \sigma^2 v^\top v \\
&\overset{(b)}{\geq} \epsilon^2 w^\top w + \sigma^2 v^\top v
\end{aligned}$$

where (a) follows because $u_{j+i}$ is independent of $x_{j+i}$ and is Gaussian with variance $\sigma^2$; (b) follows from the lower bound of Gramian in Lemma 22. To evaluate the condition in BMSB, we note

$$\begin{aligned}
&\mathbb{P}\left( \frac{1}{k} \sum_{i=1}^{k} Y_i \geq \frac{1}{2} \epsilon^2 w^\top w + \sigma^2 v^\top v \middle| \mathcal{F}_j \right) \\
&\geq \mathbb{P}\left( \frac{1}{k} \sum_{i=1}^{k} Y_i \geq \frac{1}{2} \mathbb{E}[\frac{1}{k} \sum_{i=1}^{k} Y_i | \mathcal{F}_j] \middle| \mathcal{F}_j \right) \\
&\overset{(c)}{\geq} \frac{1}{4} \frac{[\mathbb{E}[\sum_{i=1}^{k} Y_i | \mathcal{F}_j]]^2}{\mathbb{E}\left[ (\sum_{i=1}^{k} Y_i)^2 | \mathcal{F}_j \right]},
\end{aligned} \tag{4.14}$$

where (c) uses the Paley-Zygmund inequality. Since $Y_i|\mathcal{F}_j$ takes the form of $Z^2$ where $Z$ is a Gaussian random variable, which satisfies the condition

$$\mathbb{E} Z^4 \leq 3(\mathbb{E} Z^2)^2, \tag{4.15}$$

we can upper bound the denominator as follows:

$$\begin{aligned}
\mathbb{E}\left[ (\sum_{i=1}^{k} Y_i)^2 | \mathcal{F}_j \right] &= \sum_{i,i'=1}^{k} \mathbb{E}\left[ Y_i Y_{i'} | \mathcal{F}_j \right] \\
&\overset{(d)}{\leq} \sum_{i,i'=1}^{k} \sqrt{\mathbb{E}[Y_i^2 | \mathcal{F}_j] \mathbb{E}[Y_{i'}^2 | \mathcal{F}_j]} \\
&\overset{(e)}{\leq} 3 \sum_{i,i'=1}^{k} \mathbb{E}[Y_i | \mathcal{F}_j] \mathbb{E}[Y_{i'} | \mathcal{F}_j],
\end{aligned}$$

where (c) uses the Cauchy inequality and (e) follows from (4.15). Combining this inequality with (4.14) concludes the proof of the BMSB condition. Finally, we note that we have only used the fact that the conditional distribution of $Y_i$ is the square of a Gaussian random variable, the inequality used in the proof does not depend on whether the selected indices are contiguous. The proof generates to any sub-process. □

BMSB-condition implies a non-asymptotic bound on the singular value of $\begin{bmatrix} X \\ U \end{bmatrix}$.

**Proposition 2.** *Under Assumptions 5 and 6. Let* $C(I) = \left(m\sigma^2|I| + \sum_{i\in I} \operatorname{tr}(\Gamma_i^{\max})\right)$*, where* $\Gamma_i^{\max}$ *is defined in Lemma 22. For any subset* $I \subseteq \{0, 1, \ldots, T-1\}$*, we have*

$$\mathbb{P}\left(\sigma_{\max}\left(\begin{bmatrix} X_I \\ U_I \end{bmatrix}\right) > \sqrt{\frac{C(I)}{\eta}}\right) \leq \eta \tag{4.16}$$

$$\begin{aligned}&\mathbb{P}\left(\sigma_{\min}\left(\begin{bmatrix} X_I \\ U_I \end{bmatrix}\right) < \min(\epsilon,\sigma)\sqrt{\frac{k\lfloor |I|/k\rfloor p^2}{16}}\right) \leq \eta \\ &+ \exp\left(-\frac{|I|p^2}{10k} + 2(m+n)\log(10/p)\right. \\ &\left.+\frac{1}{2}(m+n)\log\left(\frac{C(I)}{\min(\epsilon,\sigma)^2 \frac{k\lfloor |I|/k\rfloor p^2}{16}\eta^2}\right)\right)\end{aligned} \tag{4.17}$$

*Proof.* The proof is a direct consequence of the covering argument in [83, Section D]. We only highlight the major differences in our attack model and give an outline of the proof.

To prove (4.16), we use the Markov inequality.

$$\begin{aligned}&\mathbb{P}\left(\sigma_{\max}\left(\begin{bmatrix} X_I \\ U_I \end{bmatrix}\right) > \sqrt{\frac{C(I)}{\eta}}\right) \\ &\leq \frac{\eta}{C(I)}\mathbb{E}\left[\lambda_{\max}\left(\begin{bmatrix} X_I \\ U_I \end{bmatrix}\begin{bmatrix} X_I \\ U_I \end{bmatrix}^\top\right)\right] \\ &\leq \frac{\eta}{C(I)} \cdot \mathbb{E}\left[\operatorname{tr}\left(\begin{bmatrix} X_I \\ U_I \end{bmatrix}\begin{bmatrix} X_I \\ U_I \end{bmatrix}^\top\right)\right] \leq \eta\end{aligned}$$

To prove the bound (4.17), we use Lemma 23, which already shows that $(k, \Gamma_{sb}, p)$-BMSB condition is satisfied, where $\Gamma_{sb} = \frac{1}{2}\Gamma$ and $p = \frac{1}{12}$. [83, Proposition 2.5] shows that for any $w \in \mathbb{R}^n$, $v \in \mathbb{R}^m$, we have

$$\begin{aligned}&\mathbb{P}\left(\sum_{i\in I}\left(\langle w, x_i\rangle + \langle v, u_i\rangle\right)^2 \leq \frac{\epsilon^2 w^\top w + \sigma^2 v^\top v}{16} p^2 k\lfloor |I|/k\rfloor\right) \\ &\qquad\qquad \leq \exp\left(-\frac{\lfloor |I|/k\rfloor}{8}p^2\right)\end{aligned}$$

To obtain the bound for the smallest singular value, we use a covering argument per [83, Section D], which leads to the inequality

$$\begin{aligned}&\mathbb{P}\left(\left\{\sigma_{\min}\left(\begin{bmatrix}X_I\\U_I\end{bmatrix}\right) < \min(\epsilon,\sigma)\sqrt{\frac{k\lfloor |I|/k\rfloor p^2}{16}}\right\}\cap\right.\\ &\qquad\left.\left\{\sigma_{\max}\left(\begin{bmatrix}X_I\\U_I\end{bmatrix}\right)\le\frac{C(I)}{\eta}\right\}\right)\\ &\le\exp\left(-\frac{|I|p^2}{10k}+2(m+n)\log(10/p)+\right.\\ &\qquad\left.\frac{1}{2}(m+n)\log\left(\frac{C(I)}{\min(\epsilon,\sigma)^2\frac{k\lfloor |I|/k\rfloor p^2}{16}\eta^2}\right)\right)\end{aligned}$$

Noting (4.16), the union bound leads to (4.17). □

We are now able to provide a sufficient condition for the satisfaction of NSP in our attack model.

**Theorem 18.** *Assume that $\alpha_{\max} < 1$, for any $c, \eta > 0$, there exist constants $N$ and $h > 0$, such that when $|S|^2 < h|S^c|$ and $|S^c| > N$, $\begin{bmatrix}X\\U\end{bmatrix}$ is c-NSP with probability at least $1 - 3\eta$.*

*Proof.* When $\alpha_{\max} < 1$, Lemma 22 shows that $\mathrm{tr}(\Gamma_i^{\max})$ will be bounded, hence $C(I) = O(|I|)$. Applying Proposition 2 for $I = S$ and $I = S^c$, respectively, there exists constants $N, c'$, and $c'''$ that do not depend on $S$, such that when $|S^c| > N$,with probability as least $1 - 3\eta$, the following two conditions hold

$$\begin{aligned}\sigma_{\max}\left(\begin{bmatrix}X_S\\U_S\end{bmatrix}\right)&\le c'\sqrt{\frac{|S|}{\eta}}\\ \sigma_{\min}\left(\begin{bmatrix}X_{S^c}\\U_{S^c}\end{bmatrix}\right)&\ge c''\sqrt{|S^c|}\end{aligned}$$

Therefore, we can pick a small enough $h > 0$, such that when $|S|^2 < h|S^c|$,

$$\sqrt{|S|}\sigma_{\max}\begin{bmatrix}X_S\\U_S\end{bmatrix} < c\cdot\sigma_{\min}\begin{bmatrix}X_{S^c}\\U_{S^c}\end{bmatrix} \tag{4.18}$$

with probability at least $1 - 3\eta$. Lemma 19 then applies and we conclude that $\begin{bmatrix}X\\U\end{bmatrix}$ is c-NSP. □

## 4.7 Estimation Error Lower Bound

We have shown the satisfaction of NSP with the attack model in Section 4.6. Assuming NSP, we have provided an upper bound of the estimation error in Theorem 17. This section complements the story of error upper bound by providing a lower bound to the estimation error for the attack model in Section 4.6. The result suggests that the absence of NSP opens up the chance for the $l_1$ optimization problem to produce an inconsistent estimate for certain types of attack. For simplicity, we study the noiseless case with no control input, where the NSP property is formulated in (4.7).

**Proposition 3.** *Let $S = \mathrm{supp}\bar{D}$. Suppose there exist parameters $0 < \mu < \nu$ and a matrix $R$ such that*

$$\mu\|Rx_i\|^2 \leq \langle \bar{d}_i, Rx_i\rangle \text{ and } \|\bar{d}_i\| \leq \nu\|Rx_i\|, \forall i \in S.$$

*Let $\hat{A}$ be the solution to the optimization problem (4.8). Then we have*

$$\|\bar{A} - \hat{A}\|_F \geq \frac{\mu}{\sigma_{\max}(x_S)}\left[\frac{\mu}{2\nu+\mu}\sum_{i\in S}\|Rx_i\| - \sum_{i\in S^c}\|Rx_i\|\right].$$

*Proof.* The proof is based on a relaxation of the dual problem of (4.8). First we rewrite (4.8) in the second-order cone optimization form:

$$\begin{aligned} &\min_{A,t} \sum_i t_i \\ &s.t. \quad \begin{bmatrix} x_{i+1} - Ax_i \\ t_i \end{bmatrix} \in C, \forall i \in \{0, 1, \ldots, T-1\}. \end{aligned}$$

where $C$ is the second-order cone

$$C = \{(x^\top, t)^\top \in \mathbb{R}^{n+1}, \|x\|_2 \leq t\},$$

and the sum is over $i \in \{0, \ldots, T-1\}$. To find out the dual problem, we compute

$$\begin{aligned} &\max_{v_i\in\mathbb{R}^n, s_i\in\mathbb{R}} \min_{A,t} \sum_i t_i - v_i^\top(x_{i+1} - Ax_i) - s_i t_i \\ &s.t. \begin{bmatrix} v_i \\ s_i \end{bmatrix} \in C, \quad \text{for all } i \in \{0, 1, \ldots, T-1\}. \end{aligned}$$

Note that a finite minimum in the inner optimization problem requires $\sum_i x_i v_i^\top = 0$ and $s_i = 0$ for all $i \in \{0, 1, \ldots, T-1\}$. They imply

$$\sum_i v_i^\top x_{i+1} = \sum_i v_i^\top \bar{d}_i + \sum_i \mathrm{tr}(\bar{A}x_i v_i^\top) = \sum_i v_i^\top \bar{d}_i.$$

Therefore, the dual problem is simplified to

$$\begin{aligned}\max_{v_i \in \mathbb{R}^n} \quad & -\sum_i v_i^\top \bar{d}_i \\ s.t. \quad & \|v_i\| \leq 1, \quad \text{for all } i \in \{0, 1, \ldots, T-1\} \\ & \sum_i x_i v_i^\top = 0.\end{aligned}$$

Consider the following relaxation with a fixed matrix $R$, which upper bounds the dual objective

$$\begin{aligned}\max_{v_i} \quad & -\sum_i v_i^\top \bar{d}_i \\ s.t. \quad & \|v_i\| \leq 1, \quad \text{for all } i \in \{0, 1, \ldots, T-1\} \\ & \sum_i v_i^\top R x_i = 0.\end{aligned}$$

Pick any $\lambda > 0$, we have

$$\begin{aligned}\|\bar{D}_S\|_{2,col} - \|(\bar{A} - \hat{A})X_S\|_{2,col} &\leq \sum_{i \in S} \|\bar{A}x_i + d_i - \hat{A}x_i\| \\ &\leq \sum_i \|x_{i+1} - \hat{A}x_i\| \\ &\leq \max_{v_i} \quad -\sum_i v_i^\top \bar{d}_i + \lambda \sum_i v_i^\top R x_i \\ &\quad\ s.t. \quad \|v_i\| \leq 1 \\ &= \sum_i \|\bar{d}_i - \lambda R x_i\|_2,\end{aligned}$$

where we use the triangle inequality, relax the sum from $i \in S$ to $i \in \{0, 1, \ldots, T-1\}$, and use the relaxed dual problem with a fixed Lagrange multiplier as an upper bound to the primer objective. As a result,

$$\begin{aligned}\|(A - \hat{A})\|_F &\geq \frac{1}{\sigma_{\max}(x_S)} \left[ \sum_{i \in S} (\|\bar{d}_i\| - \|\bar{d}_i - \lambda R x_i\|) - \sum_{i \in S^c} \lambda \|R x_i\| \right] \\ &= \frac{1}{\sigma_{\max}(x_S)} \left[ \sum_{i \in S} \frac{(\|\bar{d}_i\|^2 - \|\bar{d}_i - \lambda R x_i\|^2)}{\|\bar{d}_i\| + \|\bar{d}_i - \lambda R x_i\|} - \sum_{i \in S^c} \lambda \|R x_i\| \right] \\ &\geq \frac{1}{\sigma_{\max}(x_S)} \left[ \sum_{i \in S} \frac{(2\lambda \langle \bar{d}_i, R x_i \rangle - \lambda^2 \|R x_i\|^2)}{2\|\bar{d}_i\| + \lambda \|R x_i\|} - \sum_{i \in S^c} \lambda \|R x_i\| \right]\end{aligned}$$

For a fixed $i \in S$, the quadratic term in the numerator finds its maximum at $\lambda^* = \frac{\langle \bar{d}_i, Rx_i \rangle}{\|Rx_i\|^2} > \mu$, where the inequality is due to the assumption $\langle \bar{d}_i, Rx_i \rangle \geq \mu \|Rx_i\|^2, \forall i \in S$. Selecting $\lambda = \mu$, then we have

$$\begin{aligned}
\|(A - \hat{A})\|_F &\geq \frac{\mu}{\sigma_{\max}(x_S)} \left[ \sum_{i \in S} \frac{\|Rx_i\|^2 \mu}{2\|\bar{d}_i\| + \mu \|Rx_i\|} - \sum_{i \in S^c} \|Rx_i\| \right] \\
&\geq \frac{\mu}{\sigma_{\max}(x_S)} \left[ \sum_{i \in S} \frac{\mu}{2\nu + \mu} \|Rx_i\| - \sum_{i \in S^c} \|Rx_i\| \right]
\end{aligned}$$

where the inequalities follow from our assumption and Cauchy's inequality $\nu \|Rx_i\| \geq \|\bar{d}_i\| \geq \mu \|Rx_i\|$. □

## 4.8 Numerical Experiments

This section provides numerical simulations to illustrate the efficiency of the identification approach. First, consider the autonomous case where $\bar{B} = 0$. Our baseline for comparison is the least-squares estimator

$$\min_A \sum_{t=0}^{T-1} \|x_{i+1} - Ax_i\|_2^2. \tag{4.19}$$

To obtain the system matrices, we consider the case $n = 5$. We use $N(0, \Sigma)$ to denote the multivariate Gaussian random variable with mean 0 and covariance $\Sigma$. We set the spectrum of $A$ to be $\Gamma = \text{diag}(0.9, 0.8, 0.7, 1.1, 0.1)$, and let $A = P\Gamma P^{-1}$, where $P$ is a random matrix whose entries are normally distributed with mean 0 and variance 1. Let $x_0$ be normally distributed with mean 0 and variance 1. Let the disturbance $d_t$ be non-zero 30% of the time. Moreover, for $t \in K$, let $d_t$ follow the distribution $N(0, 10I_5)$, where $I_5$ is the 5-by-5 identity matrix. As the horizon $T$ increases from 1 to 50, we compare the constrained Lasso estimator (4.8) and the least-squares estimator (4.19) in Figure 4.1. Due to the frequency and large magnitude of the disturbance, the least-squares estimator never converges to the true system matrix $\bar{A}$. In contrast, the lasso estimator quickly converges to the true system matrix, and after it converges, future disturbance has little effect on the estimation accuracy.

Figure 4.2 shows that the presence of noise makes perfect recovery impossible in finite time, but the sudden improvement of the performance of the estimator is still apparent.

For the second example, we consider the Tennessee Eastman challenge problem. We obtain the $A$ and $B$ matrices from a discretization of the continuous-time LTI model in [54]. The discretization uses zero-order hold with the sampling period being 0.25h. Since the continuous-time model has a large separation between fast and slow modes, the discretized $A$ matrix has four modes close to 0. The values of $A$ and $B$ are provided in (4.21) and (4.22).

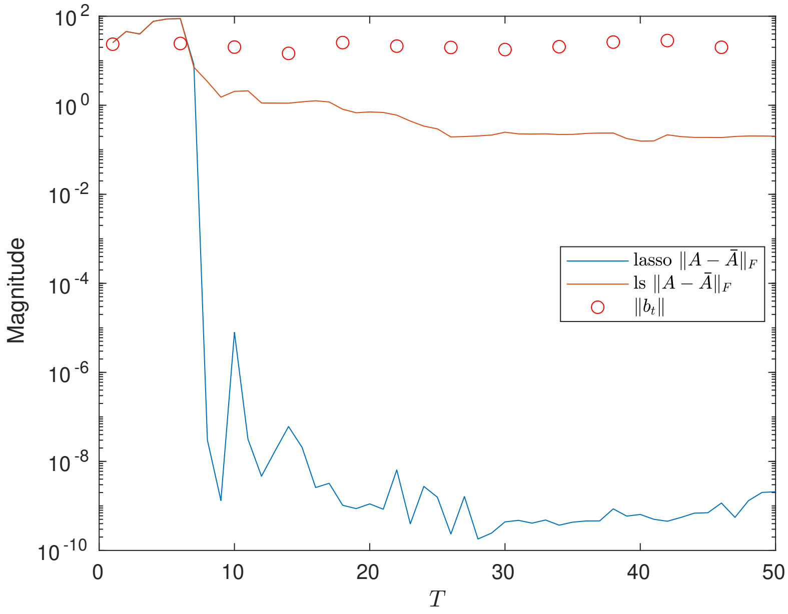


Figure 4.1: Comparing the constrained lasso estimator (4.8) and the least-squares (ls) estimator (4.19). The circles plot the magnitude of the disturbance $d_t$ when it is non-zero. The difference is measured in the Frobenius norm $\|\cdot\|_F$.

Our baseline for comparison is the least-squares estimator

$$\min_{A,B} \sum_{t=0}^{T-1} \|x_{i+1} - Ax_i - Bu_i\|_2^2. \tag{4.20}$$

Inspired by Theorem 18, the control inputs come from the distribution $N(0, I_4)$, and the initial state comes from $N(0, I_8)$. The disturbance is generated in the same fashion. Figure 4.3 shows that the constrained lasso estimator (4.1) vastly outperforms the least-squares estimator (4.20). Despite the fact that 30% of the states are disturbed, the identification of both $A$ and $B$ matrices is almost perfect.

## 4.9 Conclusion

This chapter studies an $l_1$-based identification scheme of a fully observable linear time invariant system affected by sparse state disturbances. We find that as long as the attack is not too frequent, even assuming that attack can take the form of a linear state and input feedback, an accurate state space representation can be obtained. The inequalities in the form of the null space property can provide conditions on the exact recovery of the model and give a bound of the estimation error. The lower bound on estimation error seems to suggest

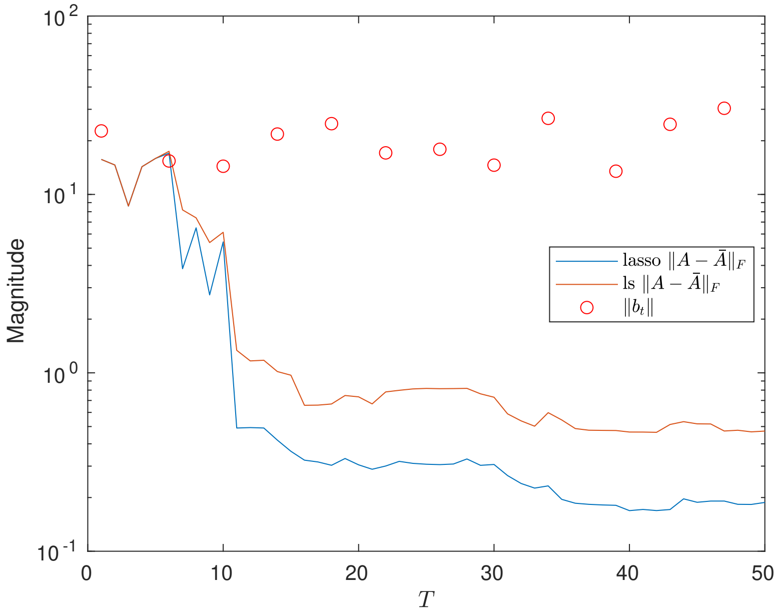


Figure 4.2: Comparing the constrained lasso estimator (4.8) and the least-squares (ls) estimator (4.19) with additional N(0,1) noise injected to the states. The circles plot the magnitude of the disturbance $d_t$.

$$A = \begin{bmatrix} 5.4893\times10^{-1} & 4.8137\times10^{-3} & -1.7226\times10^{-1} & -2.4752\times10^{-2} & 1.6520\times10^{-3} & 3.4343\times10^{-4} & -9.6398\times10^{-5} & 1.4510\times10^{-4} \\ 5.9242\times10^{-4} & 9.8284\times10^{-1} & 9.9585\times10^{-4} & -1.6428\times10^{-4} & 5.2225\times10^{-5} & 3.6788\times10^{-7} & -7.0184\times10^{-5} & 9.5650\times10^{-7} \\ -4.3298\times10^{-1} & 4.0718\times10^{-3} & 8.0876\times10^{-1} & -2.4586\times10^{-2} & 1.8725\times10^{-3} & -2.6758\times10^{-4} & -5.5680\times10^{-5} & 1.4413\times10^{-4} \\ 3.1393\times10^{-1} & -1.1807\times10^{-1} & 5.6784\times10^{-2} & 7.5675\times10^{-1} & 1.6457\times10^{-3} & 1.9424\times10^{-4} & -7.5567\times10^{-5} & -4.4716\times10^{-3} \\ 0 & 0 & 0 & 0 & 6.3656\times10^{-40} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 6.3656\times10^{-40} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 6.3656\times10^{-40} & 0 \\ 1.7555\times10^{-1} & -6.5758\times10^{-2} & 3.1911\times10^{-2} & 4.2687\times10^{-1} & 9.2087\times10^{-4} & 1.0861\times10^{-4} & -4.2300\times10^{-5} & -2.5223\times10^{-3} \end{bmatrix} \tag{4.21}$$

$$B = \begin{bmatrix} 0.2530 & 0.0412 & -0.0138 & -0.0111 \\ 0.0044 & 0.0000 & -0.0063 & -0.0001 \\ 0.2730 & -0.0138 & -0.0101 & -0.0111 \\ 0.0903 & 0.0104 & -0.0042 & 0.6455 \\ 1.0000 & 0 & 0 & 0 \\ 0 & 1.0000 & 0 & 0 \\ 0 & 0 & 1.0000 & 0 \\ 0.0499 & 0.0057 & -0.0023 & -1.0406 \end{bmatrix} \tag{4.22}$$

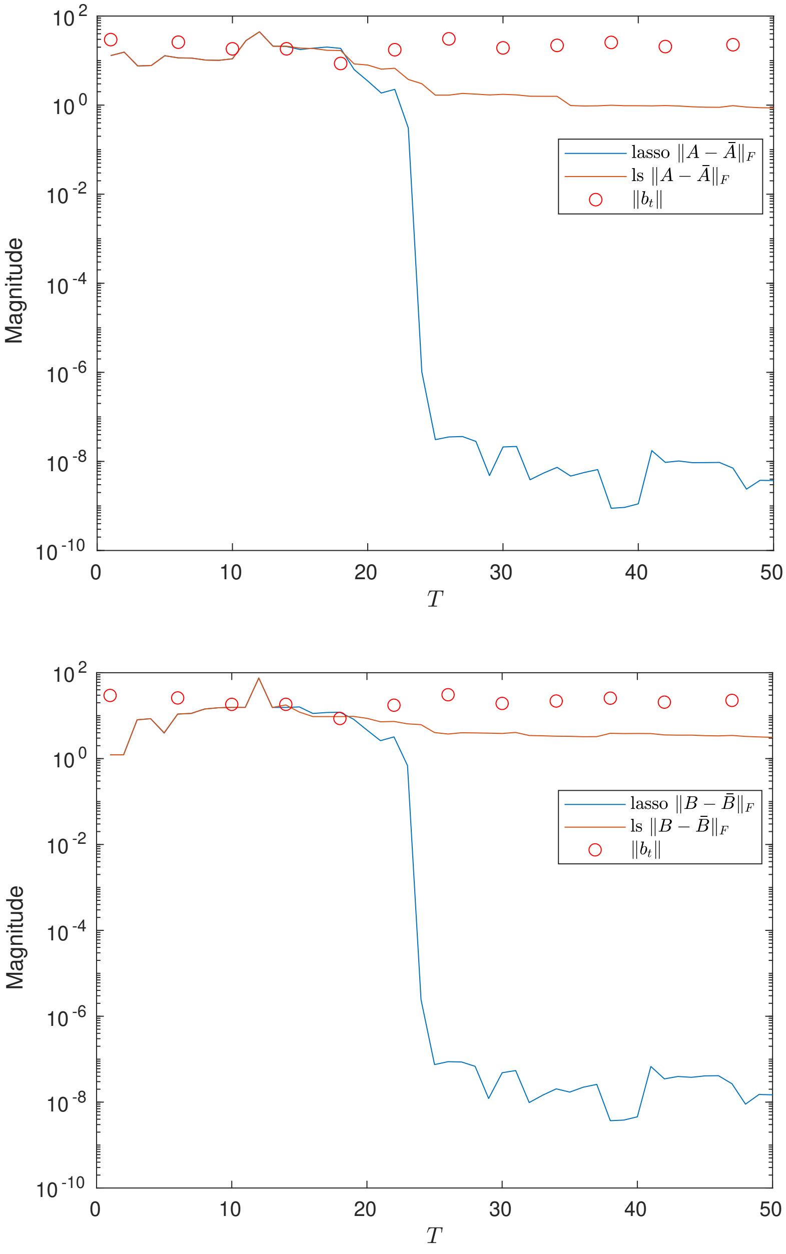


Figure 4.3: Comparing the constrained lasso estimator (4.1) and the least-squares (ls) estimator (4.20) for the Tennessee Eastman challenge problem. The circles plot the magnitude of the disturbance $d_t$ when it is non-zero. The difference is measured in the Frobenius norm $\| \cdot \|_F$.

that there are fundamental limits on this identification scheme. It would be interesting to study when consistency and error bounds hold for other models of attack. More generally, other identification schemes such as iterative re-weighted least squares and its variations are promising to analyze in the system identification context.

# Bibliography

[1] Eugene L. Allgower and Kurt Georg. *Introduction to Numerical Continuation Methods.* Society for Industrial and Applied Mathematics, 2003. ISBN: 978-0-89871-544-6 978-0-89871-915-4. DOI: 10.1137/1.9780898719154.

[2] R Arastoo et al. "Output Feedback Controller Sparsification via H2-Approximation". In: *IFAC Workshop on Distributed Estimation and Control in Networked Systems* 48.22 (2015), pp. 112–117. ISSN: 2405-8963. DOI: 10.1016/J.IFACOL.2015.10.316.

[3] Laurent Bako. "On a Class of Optimization-Based Robust Estimators". In: *IEEE Transactions on Automatic Control* 62.11 (Nov. 2017), pp. 5990–5997. ISSN: 0018-9286, 1558-2523. DOI: 10.1109/TAC.2017.2703308.

[4] Laurent Bako and Henrik Ohlsson. "Analysis of a Nonsmooth Optimization Approach to Robust Estimation". In: *Automatica* 66 (Apr. 2016), pp. 132–145. ISSN: 00051098. DOI: 10.1016/j.automatica.2015.12.024.

[5] B. Bamieh, F. Paganini, and M.A. Dahleh. "Distributed control of spatially invariant systems". In: *IEEE Transactions on Automatic Control* 47.7 (2002), pp. 1091–1107. ISSN: 0018-9286. DOI: 10.1109/TAC.2002.800646.

[6] M.-A. Belabbas. "Sparse stable systems". In: *Systems & Control Letters* 62.10 (2013), pp. 981–987. ISSN: 0167-6911. DOI: 10.1016/J.SYSCONLE.2013.07.004.

[7] Dennis S Bernstein. *Matrix mathematics: Theory, facts, and formulas with application to linear systems theory.* 2005.

[8] Dimitri P Bertsekas. *Nonlinear Programming. Athena Scientific, 2016.*

[9] Dimitri P Bertsekas. *Reinforcement Learning and Optimal Control.* Athena Scientific, 2019.

[10] Dimitris Bertsimas and Martin S. Copenhaver. "Characterization of the Equivalence of Robustification and Regularization in Linear and Matrix Regression". In: *European Journal of Operational Research* 270.3 (Nov. 2018), pp. 931–942. ISSN: 0377-2217. DOI: 10.1016/j.ejor.2017.03.051.

[11] Kush Bhatia et al. "Consistent Robust Regression". In: *Advances in Neural Information Processing Systems 30.* Ed. by I. Guyon et al. Curran Associates, Inc., 2017, pp. 2110–2119.

[12] Yingjie Bi and Javad Lavaei. “On the Connectivity Properties of Feasible Regions of Optimal Decentralized Control Problems”. In: *IEEE Transactions on Control of Network Systems* (2022).

[13] V. Blondel and J. Tsitsiklis. “NP-Hardness of Some Linear Control Design Problems”. In: *SIAM Journal on Control and Optimization* 35.6 (Nov. 1997), pp. 2118–2127. ISSN: 0363-0129. DOI: 10.1137/S0363012994272630.

[14] Vincent D Blondel and John N Tsitsiklis. “A survey of computational complexity results in systems and control”. In: *Automatica* 36.9 (2000), pp. 1249–1274. ISSN: 00051098. DOI: 10.1016/S0005-1098(00)00050-9.

[15] Jacek Bochnak, Michel Coste, and Marie-Françoise Roy. *Real algebraic geometry*. Vol. 36. Springer Science & Business Media, 2013.

[16] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge, 2004. arXiv: 1111.6189v1.

[17] Stephen P. Boyd et al. *Linear Matrix Inequalities in System and Control Theory*. Vol. 15. 1994, p. 193. ISBN: 9781611970777. DOI: 10.1109/TAC.1997.557595. arXiv: arXiv:1011.1669v3.

[18] J. R. Broussard and N. Halyo. “Active Flutter Control Using Discrete Optimal Constrained Dynamic Compensators”. In: *1983 American Control Conference*. 1983, pp. 1026–1034. DOI: 10.23919/ACC.1983.4788265.

[19] Francesco Bullo, Jorge Cortes, and Sonia Martinez. *Distributed control of robotic networks: a mathematical approach to motion coordination algorithms*. Vol. 27. Princeton University Press, 2009.

[20] J.-B. Caillau, O. Cots, and J. Gergaud. “Differential Continuation for Regular Optimal Control Problems”. In: *Optimization Methods and Software* 27.2 (Apr. 2012), pp. 177–196. ISSN: 1055-6788. DOI: 10.1080/10556788.2011.593625.

[21] Emmanuel J. Candès, Justin K. Romberg, and Terence Tao. “Stable Signal Recovery from Incomplete and Inaccurate Measurements”. In: *Communications on Pure and Applied Mathematics* 59.8 (2006), pp. 1207–1223. ISSN: 1097-0312. DOI: 10.1002/cpa.20124.

[22] Michelle S. Chong, Masashi Wakaiki, and João P. Hespanha. “Observability of Linear Systems under Adversarial Attacks”. In: *2015 American Control Conference (ACC)*. July 2015, pp. 2439–2444. DOI: 10.1109/ACC.2015.7171098.

[23] Albert Cohen, Wolfgang Dahmen, and Ronald DeVore. “Compressed Sensing and Best K-Term Approximation”. In: *Journal of the American Mathematical Society* 22.1 (2009), pp. 211–231. ISSN: 0894-0347, 1088-6834. DOI: 10.1090/S0894-0347-08-00610-3.

[24] Sarah Dean et al. “Regret bounds for robust adaptive control of the linear quadratic regulator”. In: *arXiv preprint arXiv:1805.09388* (2018).

[25] Ilias Diakonikolas et al. "Sever: A Robust Meta-Algorithm for Stochastic Optimization". In: *arXiv:1803.02815 [cs, stat]* (May 2019).

[26] J.C. Doyle et al. "State-space solutions to standard H-2 and H-infinity control problems". In: *IEEE Transactions on Automatic Control* 34.8 (1989), pp. 831–847. ISSN: 00189286. DOI: 10.1109/9.29425.

[27] Simon S Du et al. "Gradient Descent Can Take Exponential Time to Escape Saddle Points". In: *Advances in Neural Information Processing Systems 30*. 2017, pp. 1067–1077.

[28] Geir E Dullerud and Fernando Paganini. *A course in robust control theory: a convex approach*. Vol. 36. Springer Science & Business Media, 2013.

[29] Yonina C. Eldar and Gitta Kutyniok. *Compressed Sensing: Theory and Applications*. Cambridge university press, 2012.

[30] Salar Fattahi, Javad Lavaei, and Murat Arcak. "A Scalable Method for Designing Distributed Controllers for Systems with Unknown Initial States". In: *Proc. 56th IEEE Conference on Decision and Control* (2017).

[31] Salar Fattahi, Nikolai Matni, and Somayeh Sojoudi. "Efficient Learning of Distributed Linear-Quadratic Control Policies". In: *SIAM Journal on Control and Optimization* 58.5 (2020), pp. 2927–2951.

[32] Salar Fattahi and Somayeh Sojoudi. "Data-Driven Sparse System Identification". In: *2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. Oct. 2018, pp. 462–469. DOI: 10.1109/ALLERTON.2018.8635921.

[33] Salar Fattahi and Somayeh Sojoudi. "Sample complexity of sparse system identification problem". In: *accepted for publication in IEEE Transactions on Control of Network Systems* (2021).

[34] Hamza Fawzi, Paulo Tabuada, and Suhas Diggavi. "Secure Estimation and Control for Cyber-Physical Systems Under Adversarial Attacks". In: *IEEE Transactions on Automatic Control* 59.6 (June 2014), pp. 1454–1467. ISSN: 1558-2523. DOI: 10.1109/TAC.2014.2303233.

[35] Maryam Fazel et al. "Global Convergence of Policy Gradient Methods for Linearized Control Problems". In: *Proceedings of the 35th International Conference on Machine Learning*. 2018.

[36] Ghazal Fazelnia, Ramtin Madani, and Javad Lavaei. "Convex relaxation for optimal distributed control problem". In: *53rd IEEE Conference on Decision and Control*. 2014, pp. 896–903. ISBN: 978-1-4673-6090-6. DOI: 10.1109/CDC.2014.7039495.

[37] Han Feng and Javad Lavaei. "Connectivity properties of the set of stabilizing static decentralized controllers". In: *SIAM Journal on Control and Optimization* 58.5 (2020), pp. 2790–2820.

[38] Han Feng and Javad Lavaei. “Damping with Varying Regularization in Optimal Decentralized Control”. In: *IEEE Transactions on Control of Network Systems* (2021), pp. 1–1. ISSN: 2325-5870. DOI: `10.1109/TCNS.2021.3102008`.

[39] Han Feng and Javad Lavaei. “Learning of Dynamical Systems under Adversarial Attacks”. In: *2021 60th IEEE Conference on Decision and Control (CDC)*. Dec. 2021, pp. 3010–3017. DOI: `10.1109/CDC45484.2021.9683149`.

[40] Han Feng and Javad Lavaei. “On the Exponential Number of Connected Components for the Feasible Set of Optimal Decentralized Control Problems”. In: *Proceedings of the 2019 American Control Conference*. 2019, pp. 1430–1437.

[41] Jaime F. Fisac et al. “A General Safety Framework for Learning-Based Control in Uncertain Robotic Systems”. In: *IEEE Transactions on Automatic Control* 64.7 (July 2019), pp. 2737–2752. ISSN: 0018-9286, 1558-2523, 2334-3303. DOI: `10.1109/TAC.2018.2876389`.

[42] Simon Foucart and Holger Rauhut. *A Mathematical Introduction to Compressive Sensing*. Applied and Numerical Harmonic Analysis. New York, NY: Springer, 2013. ISBN: 978-0-8176-4947-0 978-0-8176-4948-7. DOI: `10.1007/978-0-8176-4948-7`.

[43] Rong Ge, Jason D. Lee, and Tengyu Ma. “Matrix Completion Has No Spurious Local Minimum”. In: *Advances in Neural Information Processing Systems (NIPS)* (2016).

[44] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. URL: `http://www.deeplearningbook.org`.

[45] E. N. Gryazina, B. T. Polyak, and A. A. Tremba. “D-Decomposition Technique State-of-the-Art”. en. In: *Automation and Remote Control* 69.12 (Dec. 2008), pp. 1991–2026. ISSN: 1608-3032. DOI: `10.1134/S0005117908120011`.

[46] Elena N. Gryazina and Boris T. Polyak. “Stability Regions in the Parameter Space: D-Decomposition Revisited”. en. In: *Automatica* 42.1 (Jan. 2006), pp. 13–26. ISSN: 0005-1098. DOI: `10.1016/j.automatica.2005.08.010`.

[47] Nirupam Gupta and Nitin H. Vaidya. “Fault-Tolerance in Distributed Optimization: The Case of Redundancy”. In: *Proceedings of the 39th Symposium on Principles of Distributed Computing*. PODC ’20. New York, NY, USA: Association for Computing Machinery, July 2020, pp. 365–374. ISBN: 978-1-4503-7582-5. DOI: `10.1145/3382734.3405748`.

[48] Moritz Hardt, Benjamin Recht, and Yoram Singer. “Train Faster, Generalize Better: Stability of Stochastic Gradient Descent”. In: *Proceedings of the 33rd International Conference on Machine Learning*. Vol. 48. 2015, pp. 1225–1234.

[49] Pedro Hespanhol and Anil Aswani. “Statistical Consistency of Set-Membership Estimator for Linear Systems”. In: *IEEE Control Systems Letters* 4.3 (July 2020), pp. 668–673. ISSN: 2475-1456. DOI: `10.1109/LCSYS.2020.2990998`.

[50] Sarah Hojjatinia, Constantino M. Lagoa, and Fabrizio Dabbene. "Identification of Switched Autoregressive Systems from Large Noisy Data Sets". In: *2019 American Control Conference (ACC)*. July 2019, pp. 4313–4319. DOI: 10.23919/ACC.2019.8814621.

[51] Chi Jin et al. "How to Escape Saddle Points Efficiently". In: *International Conference on Machine Learning*. 2017, pp. 1724–1732.

[52] Cedric Josz et al. "A Theory on the Absence of Spurious Solutions for Nonconvex and Nonsmooth Optimization". In: *Advances in Neural Information Processing Systems 31*. 2018, pp. 2441–2449.

[53] E. Kreindler and A. Jameson. "Optimality of linear control systems". In: *IEEE Transactions on Automatic Control* 17.3 (June 1972), pp. 349–351. ISSN: 0018-9286. DOI: 10.1109/TAC.1972.1099985.

[54] N. Lawrence Ricker. "Model Predictive Control of a Continuous, Nonlinear, Two-Phase Reactor". en. In: *Journal of Process Control* 3.2 (1993), pp. 109–123. ISSN: 0959-1524. DOI: 10.1016/0959-1524(93)80006-W.

[55] Jason D. Lee et al. "First-Order Methods Almost Always Avoid Strict Saddle Points". In: *Mathematical Programming* 176.1 (2019), pp. 311–337. ISSN: 1436-4646.

[56] Laurent Lessard and Sanjay Lall. "An algebraic approach to the control of decentralized systems". In: *IEEE Transactions on Control of Network Systems* 1.4 (2014), pp. 308–317. ISSN: 23255870. DOI: 10.1109/TCNS.2014.2357501. arXiv: 1309.5414.

[57] Fu Lin, Makan Fardad, and Mihailo R. Jovanovic. "Design of optimal sparse feedback gains via the alternating direction method of multipliers". In: *IEEE Transactions on Automatic Control* 58.9 (2013), pp. 2426–2431. ISSN: 00189286. DOI: 10.1109/TAC.2013.2257618. arXiv: 1111.6188.

[58] David G Luenberger and Yinyu Ye. *Linear and Nonlinear Programming*. Vol. 2. Springer, 1984.

[59] Nikolai Matni and John C Doyle. "A heuristic for sub-optimal H-2 decentralized control subject to delay in non-quadratically-invariant systems". In: *American Control Conference (ACC), 2013*. IEEE. 2013, pp. 5803–5808.

[60] Mathieu Mercadal. "Homotopy Approach to Optimal, Linear Quadratic, Fixed Architecture Compensation". In: *Journal of Guidance, Control, and Dynamics* 14.6 (1991), pp. 1224–1233. ISSN: 0731-5090. DOI: 10.2514/3.20778.

[61] Mehran Mesbahi and Magnus Egerstedt. *Graph theoretic methods in multiagent networks*. Princeton University Press, 2010.

[62] Hossein Mobahi and John W. Fisher Iii. "A Theoretical Analysis of Optimization by Gaussian Continuation". In: *Twenty-Ninth AAAI Conference on Artificial Intelligence*. 2015.

[63] I. Molybog, R. Madani, and J. Lavaei. "Conic Optimization for Robust Quadratic Regression: Deterministic Bounds and Statistical Analysis". In: *2018 IEEE Conference on Decision and Control (CDC)*. Dec. 2018, pp. 841–848. DOI: 10.1109/CDC.2018.8619037.

[64] Igor Molybog, Somayeh Sojoudi, and Javad Lavaei. "Role of sparsity and structure in the optimization landscape of non-convex matrix sensing". In: *Mathematical Programming* (2020), pp. 1–37.

[65] Daniel K. Molzahn et al. "A Survey of Distributed Optimization and Control Algorithms for Electric Power Systems". en. In: *IEEE Transactions on Smart Grid* 8.6 (Nov. 2017), pp. 2941–2962. ISSN: 1949-3053, 1949-3061. DOI: 10.1109/TSG.2017.2720471.

[66] Raimund J. Ober. "Topology of the set of asymptotically stable minimal systems". In: *International Journal of Control* 46.1 (1987), pp. 263–280. ISSN: 0020-7179. DOI: 10.1080/00207178708933897.

[67] A. Ohara and T. Kitamori. "Geometric structures of stable state feedback systems". In: *IEEE Transactions on Automatic Control* 38.10 (1993), pp. 1579–1583. ISSN: 00189286. DOI: 10.1109/9.241581.

[68] Toshiyuki Ohtsuka and Hironori Fujii. "Stabilized Continuation Method for Solving Optimal Control Problems". In: *Journal of Guidance, Control, and Dynamics* 17.5 (Sept. 1994), pp. 950–957. DOI: 10.2514/3.21295.

[69] Efe A. Ok. *Real Analysis with Economic Applications*. Princeton University Press, 2007.

[70] Samet Oymak and Necmiye Ozay. "Non-Asymptotic Identification of LTI Systems from a Single Trajectory". In: *2019 American Control Conference (ACC)*. July 2019, pp. 5655–5661. DOI: 10.23919/ACC.2019.8814438.

[71] Necmiye Ozay, Constantino Lagoa, and Mario Sznaier. "Robust Identification of Switched Affine Systems via Moments-Based Convex Optimization". In: *Proceedings of the 48h IEEE Conference on Decision and Control (CDC) Held Jointly with 2009 28th Chinese Control Conference*. Dec. 2009, pp. 4686–4691. DOI: 10.1109/CDC.2009.5399962.

[72] Binfeng Pan, Xun Pan, and Ping Lu. "Finding Best Solution in Low-Thrust Trajectory Optimization by Two-Phase Homotopy". In: *Journal of Spacecraft and Rockets* 56.1 (2019), pp. 283–291. ISSN: 0022-4650. DOI: 10.2514/1.A34144.

[73] Binfeng Pan et al. "Double-Homotopy Method for Solving Optimal Control Problems". In: *Journal of Guidance, Control, and Dynamics* 39.8 (2016), pp. 1706–1720. DOI: 10.2514/1.G001553.

[74] Adarsh Prasad et al. "Robust Estimation via Robust Gradient Estimation". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 82.3 (2020), pp. 601–627. ISSN: 1467-9868. DOI: 10.1111/rssb.12364.

[75] Anders Rantzer. "Scalable control of positive systems". In: *European Journal of Control*. Vol. 24. 2015, pp. 72–80. ISBN: 9781612848006. DOI: 10.1016/j.ejcon.2015.04.004. arXiv: 1203.0047.

[76] T. Rautert and E. W. Sachs. "Computational Design of Optimal Output Feedback Controllers". In: *SIAM Journal on Optimization* 7.3 (1997), pp. 837–852. ISSN: 1052-6234, 1095-7189. DOI: 10.1137/S1052623495290441.

[77] Tuhin Sarkar and Alexander Rakhlin. "Near Optimal Finite Time Identification of Arbitrary Linear Dynamical Systems". In: *International Conference on Machine Learning*. May 2019, pp. 5610–5618.

[78] Parikshit Shah and Pablo A. Parrilo. "H_2-optimal decentralized control over posets: A state-space solution for state-feedback". In: *IEEE Transactions on Automatic Control* 58.12 (2013), pp. 3084–3096. ISSN: 00189286. DOI: 10.1109/TAC.2013.2281881. arXiv: 1111.1498.

[79] Yiyuan She and Art B. Owen. "Outlier Detection Using Nonconvex Penalized Regression". In: *Journal of the American Statistical Association* 106.494 (June 2011), pp. 626–639. ISSN: 0162-1459. DOI: 10.1198/jasa.2011.tm10390.

[80] Shih-Ho Wang and E. Davison. "On the stabilization of decentralized control systems". In: *IEEE Transactions on Automatic Control* 18.5 (Oct. 1973), pp. 473–478. ISSN: 0018-9286. DOI: 10.1109/TAC.1973.1100362.

[81] Mehrdad Showkatbakhsh, Paulo Tabuada, and Suhas Diggavi. "System Identification in the Presence of Adversarial Outputs". In: *2016 IEEE 55th Conference on Decision and Control (CDC)*. Dec. 2016, pp. 7177–7182. DOI: 10.1109/CDC.2016.7799376.

[82] Max Simchowitz. "Statistical Complexity and Regret in Linear Control". PhD thesis. EECS Department, University of California, Berkeley, May 2021.

[83] Max Simchowitz et al. "Learning Without Mixing: Towards A Sharp Analysis of Linear System Identification". In: *Conference On Learning Theory*. PMLR, July 2018, pp. 439–473.

[84] Somayeh Sojoudi and Javad Lavaei. "On the exactness of semidefinite relaxation for nonlinear optimization over graphs: Part I". In: *Proceedings of the IEEE Conference on Decision and Control*. 2013, pp. 1043–1050. ISBN: 9781467357173. DOI: 10.1109/CDC.2013.6760020.

[85] Jacob Steinhardt, Pang Wei W Koh, and Percy S Liang. "Certified Defenses for Data Poisoning Attacks". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Curran Associates, Inc., 2017, pp. 3517–3529.

[86] Lili Su and Shahin Shahrampour. "Finite-Time Guarantees for Byzantine-Resilient Distributed State Estimation With Noisy Measurements". In: *IEEE Transactions on Automatic Control* 65.9 (Sept. 2020), pp. 3758–3771. ISSN: 1558-2523. DOI: 10.1109/TAC.2019.2951686.

[87] H. T. Toivonen. "A Globally Convergent Algorithm for the Optimal Constant Output Feedback Problem". In: *International Journal of Control* 41.6 (1985), pp. 1589–1599. ISSN: 0020-7179. DOI: 10.1080/00207178508961217.

[88] John W. Tukey. "The Future of Data Analysis". In: *The Annals of Mathematical Statistics* 33.1 (1962), pp. 1–67. ISSN: 0003-4851.

[89] Aswin N. Venkat et al. "Distributed MPC Strategies With Application to Power System Automatic Generation Control". In: *IEEE Transactions on Control Systems Technology* 16.6 (Nov. 2008), pp. 1192–1206. ISSN: 2374-0159. DOI: 10.1109/TCST.2008.919414.

[90] Yuh Shyang Wang, Nikolai Matni, and John C. Doyle. "System Level Parameterizations, constraints and synthesis". In: *Proceedings of the American Control Conference*. 2017, pp. 1308–1315. ISBN: 9781509059928. DOI: 10.23919/ACC.2017.7963133.

[91] C. Wenk and C. Knapp. "Parameter optimization in linear systems with arbitrarily constrained controller structure". In: *IEEE Transactions on Automatic Control* 25.3 (June 1980), pp. 496–500. ISSN: 0018-9286. DOI: 10.1109/TAC.1980.1102373.

[92] Harald K Wimmer. "An inertia theorem for tridiagonal matrices and a criterion of Wall on continued fractions". In: *Linear Algebra and its Applications* 9 (1974), pp. 41–44.

[93] H. S. Witsenhausen. "A counterexample in stochastic optimum control". In: *SIAM Journal on Control* 6.1 (Feb. 1968), pp. 131–147. ISSN: 0036-1402. DOI: 10.1137/0306011.

[94] Huan Xu, Constantine Caramanis, and Shie Mannor. "Robustness and Regularization of Support Vector Machines." In: *Journal of machine learning research* 10.7 (2009).

[95] Richard Y Zhang et al. "How Much Restricted Isometry is Needed In Nonconvex Matrix Recovery?" In: *Advances in Neural Information Processing Systems (NIPS)* (2018).

[96] Yuchen Zhang, Percy Liang, and Moses Charikar. "A Hitting Time Analysis of Stochastic Gradient Langevin Dynamics". In: *Conference on Learning Theory*. 2017, pp. 1980–2022.

[97] Dragan Zigic et al. "Homotopy Approaches to the $H\_2$ Reduced Order Model Problem". In: Computer Science Technical Reports TR 91-24 (1991). URL: http://eprints.cs.vt.edu/archive/00000269/.