DESIGN PRINCIPLES FOR LINEAR SYSTEMS: STABILITY AND OPTIMALITY

BY

ARTUR KIRKORYAN

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Mathematics
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2018

Urbana, Illinois

Doctoral Committee:

Professor Lee DeVille, Chair
Assistant Professor Mohamed-Ali Belabbas, Director of Research
Professor Yuliy Baryshnikov, Director of Research
Professor Vadim Zharnitsky

ABSTRACT

This thesis consists of two parts. Each of them deals with problems in the design of linear time-invariant systems with certain prescribed properties, such as stability and cost optimality.

The first part addresses theoretical questions arising in the design of autonomous decentralized systems. The network topology of such a system describes which agents are able to interact with each other.

We study the following problem: For a specified network topology, can one find a set of interaction laws that yield stable dynamics for the ensemble of agents? We restrict our analysis to systems with strictly linear dynamics. This problem can also be referred to as the structural stability problem, seen as the counterpart to the structural controllability problem.

In mathematical terms, we consider vector spaces of real square matrices for which every entry is either fixed at zero, or an arbitrary real number. We call them sparse matrix spaces, abbreviated SMS, and examine under what conditions they contain matrices for which all eigenvalues have strictly negative real parts. We call an SMS with this property stable.

We estimate the proportion of stable SMS when their size approaches infinity and when the locations of the free variables are chosen independently at random. Using graph theory techniques, we also develop polynomial-time algorithms for extension of a given stable SMS to a stable SMS with up to two additional nodes.

In the second part, we consider linear time-invariant systems with control. The well-known linear quadratic regulator (LQR) provides feedback controller that stabilizes the system while minimizing a quadratic cost function in the state of the system and the magnitude of the control. The optimal actuator design problem then consists of choosing an actuator that minimizes the cost incurred by an LQR.

While this procedure guarantees a low overall cost incurred, it only takes

into account the magnitude of the control signals the regulator sends to the actuator. Physical actuators are, however, also limited in their ability to follow rapid change in control signals. We show in this thesis how to design actuators so that the high-frequency content of the control signals is limited, while insuring stability and optimality of the resulting closed-loop system.

We also address optimal actuator design for linear systems with process noise. It is well-known that the control that minimizes a quadratic cost in the state and control for a system with linear dynamics corrupted by additive Gaussian noise is of feedback type and its design depends on the solution of an associated Riccati equation. We consider here the case where the noise is multiplicative, by which we mean that its intensity is dependent on the state. We show how to derive the actuator that minimizes a linear quadratic cost.

To my family, for their love and support.

ACKNOWLEDGMENTS

I am very grateful to my advisor, Prof. Mohamed-Ali Belabbas, for suggesting the topics for this thesis and supervising my research. His invaluable guidance allowed me to continue working through difficult times and made completion of this work possible.

I am also indebted to my co-advisor, Prof. Yuliy Baryshnikov, for our fruitful discussions and his support throughout my time at the University of Illinois.

Next, I appreciate my close friend Ivan Penev for our long discussions and his proofreading.

Finally, I want to thank my family, my girlfriend Andrea Hoyos, and friends who provided moral support, as well as all Professors in the university who enriched my knowledge and gave me the skills to tackle difficult problems.

TABLE OF CONTENTS

PART I

SPARSE MATRIX SPACES

CHAPTER 1

INTRODUCTION

Decentralized control deals with the design of controllers achieving a given task, e.g. stabilization of the system or optimal control, under constraints on what information about the system is available to the controller. By information available to a controller, we shall mean a subset of the variables used to describe the system. The study of decentralized control systems is motivated by the many problems that are characterized by an underlying network topology describing which interactions within a system are allowed: see e.g. [1, 2, 3, 4, 5, 6, 7, 8] and the references therein. Such problems include information transmission and distributed computation.

Despite its relatively long history, decentralized control remains a challenging area of control theory. In fact, some basic issues that underlie the subject are still mostly open. For example, consider the following: We call a vector space of matrices with entries that are either arbitrary real numbers or zeros a sparse matrix space (or SMS, a formal definition is given below). These vector spaces arise naturally in the study of linear, decentralized systems. In fact, we can associate to a such vector space a directed graph that describes the allowed interactions between the various parts of the system. With these considerations in mind, whether a matrix space contains a stable matrix is a natural property to study: indeed, the corresponding graphs can be thought of as describing the interactions that can sustain stable dynamics. In [9] are given necessary conditions and sufficient conditions for a SMS to be stable, as well as structural properties of a stable SMS. Since finding both necessary and sufficient conditions seems excessively hard to deal with, some restricted problems are considered instead. The case of SMS with symmetric structure is examined in [10], where necessary and sufficient conditions for stability are given. Creation of 1-node extensions from stable SMS is examined in [9].

Questions similar to the ones examined in this thesis also appear when studying the so called *signed patterns*. A signed pattern, as defined in [11],

is a set of all matrices for which the elements have some predefined signs. A signed pattern is called stable if it contains at least one Hurwitz matrix. Classifying all stable sign patterns is not yet complete and even though classifying stable sparse matrix spaces can be considered as only a special case of this undertaking, it is still a formidable task. Some sufficient and some necessary conditions for stability of signed patterns, as well as their equivalent counterparts for SMS, have been independently presented in [11, 9].

In this thesis, we estimate the amount of sparse matrix spaces that are stable, when each element of the SMS is a free variable independently with some fixed probability. We also build upon the work in [9] by considering node extensions of higher degrees, along with algorithms which test whether the extensions are stable.

The part is organized as follows: In Chapter 2, we introduce the required background material in Control Theory and Graph Theory. This includes basic notions about stability of a system, Hurwitz polynomials, directed graphs, cycles and their relations to permutations. In Chapter 3, we provide the main definitions concerning sparse matrix spaces, and discuss relations between their associated graphs and characteristic polynomials. In Chapter 4, we recall some results about stability of SMS from earlier papers, and discuss stability of random SMS when their size approaches infinity. In Chapter 5 we establish a necessary and sufficient condition for a SMS of $(n+1)\times(n+1)$ matrices to be stable given that it contains a Hurwitz SMS of $n \times n$ matrices. We call it a stable 1-extension of a stable SMS. Not every stable SMS can be obtained as a 1-extension, and we derive a sufficient condition for a 2-extension to be stable as well. We conclude this chapter by showing that there does not exist a finite set of extension rules that creates all stable SMS. In Chapter 6, we derive polynomial-time algorithms that implement the results of Chapter 5 to obtain stable 1- and 2-extensions. The algorithms are organized around two tasks. The first is to check for the existence of Hamiltonian decompositions. While Hamiltonian decompositions include cycles, which evoke hardness of underlying algorithms, we will see that a reduction to matching problems in bipartite graphs can be used to obtain fast algorithms. The other task is to check whether multivariable polynomials—in fact, coefficients of the characteristic polynomial of the symbolic adjacency matrix of a SMS—have common factors. Again, even though factoring polynomials is a hard task in general, the special form of the polynomials at hand

allows us to derive fast algorithms. The main new idea introduced is the notion of signature of a polynomial. We conclude and provide some directions for future work in the last chapter.

CHAPTER 2

PRELIMINARIES

In this chapter we provide some definitions and results from Linear Dynamical Systems and Graph Theory, which will be needed for the main chapters later.

2.1 Linear Dynamical Systems

We start with some basic notions from control theory, such as linear systems, stability, and Hurwitz polynomials. Proofs of the stated theorems can be found in any introductory control theory textbook and will not be presented below.

A linear dynamical system is given by a differential equation

$$\dot{x}(t) = A(t)x(t), \tag{2.1}$$

where $x(t)$ - the state - is a vector-valued function. When A is a constant matrix, the system is called linear time invariant (LTI).

One very important property of dynamical systems is stability.

Definition 2.1 (Stability)**.** *The system (2.1) is (globally) asymptotically stable if for any initial condition $x(0) = x_0$, the state $x(t)$ converges to 0 as the time t approaches infinity. If there exist constants $c < 0$ and $K > 0$, such that*

$$|x(t)| \leq Ke^{ct}|x_0|$$

for all $t \geq 0$ and all x_0, the system is called (globally) exponentially stable.

In the case of LTI systems (2.1), stability can be determined simply by examining the characteristic polynomial of the matrix A.

Definition 2.2. *A polynomial is called* Hurwitz *if all its roots have strictly negative real parts. A square matrix which has Hurwitz characteristic polynomial is called Hurwitz itself.*

Theorem 2.1. *The LTI system*

$$\dot{x}(t) = Ax(t)$$

is asymptotically stable if and only if the characteristic polynomial of the matrix A is Hurwitz. Furthermore, stable LTI systems are also exponentially stable.

There are different ways to check whether a given polynomial is Hurwitz or not, without explicitly computing its roots. One such way is by using the Hurwitz stability criterion, [12], presented below.

Theorem 2.2 (Hurwitz stability criterion)**.** *Let*

$$p(x) = a_0x^n + a_1x^{n-1} + ... + a_n$$

be a real polynomial. Consider the $n \times n$ matrix

$$H = \begin{pmatrix} a_1 & a_3 & a_5 & \dots & \dots & \dots & 0 & 0 & 0 \\ a_0 & a_2 & a_4 & & & & \vdots & \vdots & \vdots \\ 0 & a_1 & a_3 & & & & \vdots & \vdots & \vdots \\ \vdots & a_0 & a_2 & \ddots & & & 0 & \vdots & \vdots \\ \vdots & 0 & a_1 & & \ddots & & a_n & \vdots & \vdots \\ \vdots & \vdots & a_0 & & & \ddots & a_{n-1} & 0 & \vdots \\ \vdots & \vdots & 0 & & & & a_{n-2} & a_n & \vdots \\ \vdots & \vdots & \vdots & & & & a_{n-3} & a_{n-1} & 0 \\ 0 & 0 & 0 & \dots & \dots & \dots & a_{n-4} & a_{n-2} & a_n \end{pmatrix}.$$

Then the polynomial $p(x)$ is Hurwitz if and only if all leading principal minors of H are positive.

Even though the criterion above yields a straightforward way to determine whether given matrix is Hurwitz, it is computationally complex and difficult to use. In this thesis we will instead use a basic property of Hurwitz polynomials, given by the following lemma.

Lemma 2.1. *All coefficients of a real Hurwitz polynomial are non-zero and have identical signs.*

Proof. All roots of a real Hurwitz polynomial are either negative numbers or complex conjugates with negative real parts. Therefore the polynomial can be expressed as a scaled product of terms $x + a$ and $x^2 + bx + c$, where $a, b, c > 0$. ■

2.2 Graphs

Definition 2.3 (Undirected Graph)**.** *An undirected graph $G = (V, E)$ is a set of nodes V, along with a set of edges E, where every edge in E is a 2-element subsets of V.*

Definition 2.4 (Bipartite Graph)**.** *Bipartite graphs are undirected graphs for which the set of vertices V can be split into two subsets V_1 and V_2, such that no two vertices in V_1 and no two vertices in V_2 are connected with edges. Bipartite graphs are denoted as $G = (V_1, V_2, E)$.*

Definition 2.5 (Directed Graph)**.** *A set of nodes V along with a set of directed edges between them $E \subset V \times V$ is called a directed graph or also, digraph, and is denoted as $G = (V, E)$.*

Definition 2.6 (Subgraph)**.** *We say that the graph $G' = (V', E')$ is a subgraph of $G = (V, E)$ if V' is a subset of V, and E' is a subset of E.*

The number of nodes in V is called **cardinality** of the graph G and is denoted with $\|G\|$.

We recall that **path** of length k in a digraph G is a sequence of nodes $(u_1, u_2, ..., u_k)$, such that $(u_i, u_{i+1}) \in E$ for $1 \leq i < k$. We say that a subgraph $G' = (V', E')$ of G is *strongly connected* if for every $u_i, u_j \in V'$, $u_i \neq u_j$ there is a path in G' from u_i to u_j and from u_j to u_i. The maximal subgraphs which have this property are called "strongly connected components" of G.

A **cycle** of length k in G, or a **k-cycle**, is a closed path in G, that is a path $(u_1, ..., u_{k+1})$ of length $k + 1$, for which $u_{k+1} = u_1$. A **simple cycle** is a cycle for which all nodes are distinct, except for u_1 and u_{k+1}, i.e. $u_i \neq u_j$ for $1 \leq i \neq j \leq k$. In this thesis, *all the cycles considered are simple*, and we refer to them simply as cycles. Self-loops represent cycles of length 1.

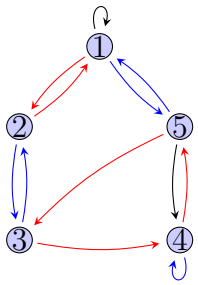

Figure 2.1: The graph depicted above admits several complete decompositions: one into the cycles (12) and (345), one into the cycles $(15), (23), (4)$ and one into the cycle (12345). The cycle (1) is a 1-decomposition and the cycle (23)(15) is a 4-decomposition of G. Finally, the cycles (1), (12), (1)(23) are nested 1-, 2- and 3-cycles.

We say that a set of cycles covers G if every node of G appears in at least one cycle. We say that two cycles are disjoint if they do not have any nodes in common. We call a **k-decomposition** of G a set of mutually-disjoint cycles in G whose union covers exactly k nodes. If $k = n$, i.e. the cycles cover the entire set V, we call it a **Hamiltonian decomposition**. We use the notation $(l_1 l_2 \dots F_k)$ to refer to the cycle $(v_{l_1}, v_{l_2}, \dots, v_{F_k}, v_{l_1})$ and write a k-decomposition as the formal product of its constituent cycles. For example (12)(3) refers to the 3-decomposition containing the cycle $(1, 2, 1)$ and the self-loop $(3, 3)$. In Fig. 2.1, the cycles (12) and (34) are disjoint, but the cycles (12) and (23) are not. We call a sequence of k-decompositions for $1 \leq k \leq n$ **nested** if the k-decomposition covers all nodes covered by the $k-1$-decomposition plus one additional node. We illustrate some of these notions in Figure 2.1.

There is a simple construction which associates a bipartite graph G' to every pair (G, f) consisting of a directed graph G on a set of nodes V and a bijection $f : V \to V'$ with $V \cap V' = \varnothing$. Namely, the bipartite graph is defined as $G' = (V, V', E')$, where $E' = \big\{\{v, f(w)\} \mid (v, w) \in E\big\}$.

2.3 Matchings

Finally, we recall few basic definitions related to matchings in bipartite graphs:

Definition 2.7 (Matchings and perfect matchings)**.** *Let $B = (V_1, V_2, E)$ be a bipartite graph.*

1. *A matching M of the bipartite graph B is a subset of E such that no edges in M are incident to the same node.*

2. *A matching M is said to be* perfect matching *if every node is adjacent to one edge in M.*

3. *A matching M is said to be* maximal matching *if no other matching contains M.*

4. *A matching M is said to be* maximum matching *if no other matching has higher cardinality than M.*

For example, the set of edges $(1, 3'), (2, 1'), (3, 2')$ constitutes a perfect matching in the bipartite graph depicted in Fig. 3.1-right.

In the classic theorem, [13], below, $N(X) = \{v \in G \mid \exists u \in X : (u, v) \in E\}$ will denote all neighbors of nodes in a subset $X \subset V$.

Theorem 2.3 (Hall's Marriage Theorem)**.** *Let $B = (V_1, V_2, E)$ be a bipartite graph. The graph B contains a perfect matching if and only if $|N(U_1)| \geq |U_1|$ for every $U_1 \subset V_1$ and $|N(U_2)| \geq |U_2|$ for every $U_2 \subset V_2$.*

Finding a maximal/maximum matching in a given graph is important and has many applications, most notably in computer science. There are various algorithms for doing this task.

We will make use of bipartite graphs in Chapter 4 and Chapter 6.

CHAPTER 3

SPARSE MATRIX SPACES

We start by introducing some vocabulary.

Definition 3.1. *We call a (real)* **sparse matrix space***, abbreviated SMS, a vector space of matrices with entries either arbitrary (real) or zero.*

Specifically, let $n > 0$ be an integer and let α be a set of pairs of integers between 1 and n, that is $\alpha \subset \{1, \ldots, n\} \times \{1, \ldots, n\}$ and denote by E_{ij} the $n \times n$ matrix with zero entries except for the ijth entry, which is equal to one. We define Σ_α to be the vector space of matrices of the form $A = \sum_{(i,j)\in\alpha} a_{ij} E_{ij}$, $a_{ij} \in \mathbb{R}$. For example, if $n = 3$ and $\alpha = \{(1,2), (1,3), (2,1), (2,2), (3,2)\}$, then Σ_α is the subspace of matrices of the form

$$A = \begin{bmatrix} 0 & * & * \\ * & * & 0 \\ 0 & * & 0 \end{bmatrix} \tag{3.1}$$

where $*$ are arbitrary real values.

A sparse matrix space Σ_α can be uniquely represented as a directed graph G with node set $V = \{1, 2, \ldots, n\}$ and edge set $E = \alpha$; we refer to G as the **graph associated with** Σ and vice-versa. For example, the graph associated to the SMS of Eq. (3.1) is depicted in Fig. 3.1-left.

Alternatively, Σ_α can be represented using a bipartite graph $B = \{V_1, V_2, E\}$ with node subsets $V_1 = \{1, 2, \ldots, n\}$, $V_2 = \{1', 2', \ldots, n'\}$ and edge set $E = \alpha$. The bipartite graph associated to the SMS of Eq. (3.1) is depicted in Fig. 3.1-right.

Given an SMS Σ, we refer to the matrix coefficients corresponding to indices in α (considered as functions on Σ_α) as the **free variables** of the SMS, or equivalently of the graph G associated with Σ. To emphasize that the free variables correspond to edges in the associated graph, we also refer

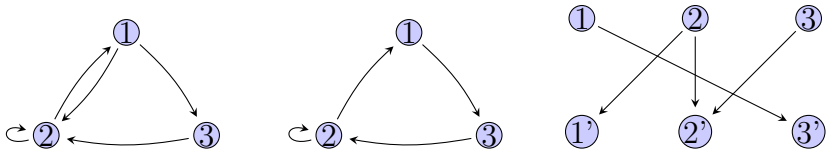

Figure 3.1: The graph on the left corresponds to the SMS of Eq. (3.1). It is Hurwitz, whereas the graph in the middle is not, even though both are strongly connected and have a node with a self-loop. Theorem 4.1 below allows to decide the stability of these graphs. The bipartite graph on the right gives another representation of the SMS given in Eq. (3.1).

to them as **edge-variables**. For example, the edge-variables of the SMS Σ in Fig.3.1-left are $(1,2),(1,3),(2,1),(2,2),(3,2)$. We call an **edge-product** a subset of edges of the graph or, with a slight abuse of notation, the product of the corresponding edge-variables. For example, $\alpha = \{(1,2),(2,1)\}$ is an edge-product, and so is $\alpha = a_{12}a_{21}$. This terminology, which allows to refer to the a_{ij} as entries of a matrix in Σ or edges in the corresponding graph G will prove useful below in proofs relying on both algebraic and graph theoretic concepts.

3.1 Permutations and digraph decompositions

We can establish a one-to-one correspondence between permutations of the set $\{1,\ldots,k\}$ and k-decompositions of digraphs— we explain this here and refer the reader to [9] for a more detailed exposition. Consider the set S_n of permutations (or equivalently, re-orderings) of the elements of $\mathcal{N} = \{1,2,\ldots,n\}$. We denote by $(l_1,l_2,\ldots,l_n)$ the permutation that sends i to l_i. There are $n!$ such permutations. Under the operation of composition of reorderings, the set of permutations can be made into a group, called the permutation group. A **permutation cycle** is a permutation that maps the elements of some subset $\mathcal{N}_1 \subset \mathcal{N}$ to each other in a cyclic fashion, while leaving the other elements fixed. For example, $(3,1,2,4)$ is a permutation cycle since it leaves 4 fixed, and maps the elements of $S = \{1,2,3\}$ to each other in a cyclic fashion, but the permutation $(2,1,4,3)$ is not a cycle.

We adopt the widely used convention of denoting a permutation cycle by $\mathbf{i} = (i_1 i_2 \ldots i_k)$, where the i_k are pairwise different, to indicate that the element in position i_1 is replaced by the element in position i_2, the element

in i_2 by the one in i_3 all the way to i_k by i_1 while the other elements are fixed. With this notation, the cycle $(3, 1, 2, 4)$ is written as $(132) = (321) = (231)$. We say that two permutation cycles **i** and **j** are *disjoint* if $i_l \neq j_m$ for all l, m. We call k the **order** of a cycle and we refer to cycles of order k as k-cycles. It is a fact from group theory that any permutation can be written as the composition of disjoint permutation cycles [14]. For example, the permutation $(2, 1, 4, 3)$ is the composition of (12) and (34) and is written as $(12)(34)$. It is easy to see that disjoint cycles commute (e.g. permuting $3, 4$ and then $1, 2$ produces the same result as permuting $1, 2$ first and then $3, 4$).

Now, the key observation is the following:

Lemma 3.1. *There is a one-to-one correspondence between permutations in S_n and n-decompositions in a complete graph with n nodes.*

For example, consider the complete decomposition $(12)(345)$ of the graph in Figure 2.1. It corresponds to the permutation $(2, 1, 4, 5, 3)$.

3.2 Characteristic polynomial and digraph decompositions

The proofs below will rely on the correspondence we establish here between terms of the characteristic polynomials of matrices in a SMS Σ and k-decompositions of its associated graph. Given the graph G on n nodes corresponding to Σ, we define its symbolic adjacency matrix A to be the $n \times n$ matrix with entries a_{ij} in position i, j if $(v_i, v_j) \in E$, and zero otherwise, where the symbols a_{ij} are formal variables. The matrix A is thought of as a generic matrix in the associated SMS Σ. Let $p_A(s) = \det(Is - A) = s^n + p_1 s^{n-1} + \ldots + p_n$ be the characteristic polynomial of A. The coefficients p_k are polynomials in the a_{ij} variables.

We denote by $\mathcal{I}$ an arbitrary subset of $\{1, 2, \ldots, n\}$ and write $|\mathcal{I}|$ for its cardinality. We denote by $A_{\mathcal{I}}$ the principal submatrix of A containing the rows and columns of A indexed by $\mathcal{I}$. It is well-known [15] that the coefficients

of $p_A(s)$ are given by

$$\begin{aligned} p_1 &= -\sum_{i=1}^{n} a_{ii}, \\ p_k &= (-1)^k \sum_{|\mathcal{I}|=k} \det(A_{\mathcal{I}}), \\ p_n &= (-1)^n \det(A), \end{aligned} \tag{3.2}$$

where the sums $\sum_{|\mathcal{I}|=k}$ are taken over the $\binom{n}{k}$ k-subsets of $\{1, 2, \ldots, n\}$. Thus, from the previous chapter and the expansion of the determinant as

$$\det(A_{\mathcal{I}}) = \sum_{\sigma \in S_k} (-1)^{\sigma} \prod_{l \in \mathcal{I}} a_{l\sigma(l)}, \tag{3.3}$$

where $(-1)^{\sigma}$ is the sign of the permutation σ [9], *we conclude that we can assign to each term in p_k a k-decomposition of G.* For example, for the graph G depicted in Figure 2.1, it is easy to see that $p_1 = -a_{11} - a_{44}$. Because this graph contains five 2-decompositions, namely $(12), (23), (34), (45), (15)$ and $(1)(4)$, we have that p_2 is the sum of five terms of degree 2: $a_{12}a_{21}$, $a_{23}a_{32}$, $\ldots$, $a_{11}a_{44}$. As a further example, the term corresponding to the 4-decomposition $(1)(345)$ is $a_{11}a_{34}a_{45}a_{53}$ and appears in p_4. We record here a few simple facts about the polynomials p_k (seen as polynomials in the free variables):

1. The p_k's are homogeneous polynomials.
2. The p_k's are *linear* in each of their variables (the entries a_{ij} of A).
3. The p_k's have coefficients only ± 1.

We now show that the polynomials satisfying the two items above enjoy the property that they have unique factorization over the reals and that there is no term cancellation when expanding the product of factors. We make this precise as follows: given p a polynomial in the variables $a_1, \ldots, a_n$, we denote by $\#p$ **the number of terms** with non-zero coefficients in p. We have the following result:

Lemma 3.2. *Let p be a polynomial in the variables $a_1, \ldots, a_n$ which satisfies properties 1 and 2 above. Then p can be factorized uniquely (up to constant*

factors) into a product of real homogeneous irreducible polynomials q_l, each of which is linear in the variables a_i. Furthermore,

$$\#p = \prod_l \#q_l. \tag{3.4}$$

Proof. The ring $R = \mathbb{R}[a_1, \ldots, a_n]$ of all polynomials in the variables a_i is a unique factorization domain, and moreover, the irreducible factors of any homogeneous element p of R are themselves homogeneous, of degrees whose sum is $\deg p$ [14].

We prove the remaining claims by induction on the number $m \leq n$ of variables a_i on which p depends non-trivially. For the case $m = 1$, p is a linear function which can not be factorized further and in this case Eq. (3.4) holds trivially.

Assume that Eq. (3.4) holds for polynomials satisfying property 2 above and depending non-trivially on at most $m - 1$ variables a_i. We show that it holds for polynomials p with m terms. Let q_l, $l = 1, \ldots, k$ be factors of p with

$$p = \prod_{l=1}^{k} q_l.$$

We can assume, perhaps after reordering the a_i, that p does not depend trivially on a_1. Indeed, if p were to depend trivially on every variable a_i, then $p \equiv 0$ and there is nothing to prove. We can express every factor q_l as $q_l = a_1^{n_l}\bar{q}_l + r_l$, where for every l we have $n_l \geq 0$, a_1 does not divide $\bar{q}_l$ and r_l is not divisible by $a_1^{n_l}$ unless it is zero. We thus obtain for p:

$$p = \prod_{l=1}^{k} (a_1^{n_l}\bar{q}_l + r_l) = a_1^{\sum n_l} q + r, \tag{3.5}$$

where $q = \prod_l \bar{q}_l$ and r is *not divisible* by $a_1^{\sum n_l}$, unless it is zero.

We conclude that, because p is linear in a_1, $\sum n_l = 1$. Assume without loss of generality that $n_1 = 1$, $n_l = 0$ for $l \neq 1$ and $r_l = 0$ for $l \neq 1$. Thus, we have

$$p = (a_1\bar{q}_1 + r_1)(\prod_{l=2}^{k} \bar{q}_l). \tag{3.6}$$

where we recall that the $\bar{q}_l$'s and r_1 are not divisible by a_1. Since the variable a_1 was randomly chosen (and relabeled), the same arguments apply to any

other variable on which p does not trivially depend. This implies that all polynomials q_l, as well as $\bar{q}_1$ and r_1 are linear in the variables $a_1, a_2, ..., a_n$.

For the last statement, we notice that from Eq. (3.6) we can conclude that

$$\#p = \#(\prod_{l=1}^{k} \bar{q}_l) + \#(r_1 \prod_{l=2}^{k} \bar{q}_l). \tag{3.7}$$

Furthermore, the numbers of variables a_i on which $\prod_{l=1}^{k} \bar{q}_l$ and $r_1 \prod_{l=2}^{k} \bar{q}_l$ depend non-trivially are both less than m. We use the induction hypothesis to obtain

$$\#(\prod_{l=1}^{k} \bar{q}_l) = \prod_{l=1}^{k} \#(\bar{q}_l) \text{ and } \#(r_1 \prod_{l=2}^{k} \bar{q}_l) = \#(r_1) \prod_{l=2}^{k} \#(\bar{q}_l). \tag{3.8}$$

Putting Eq. (3.7) and Eq. (3.8) together, we conclude that cancellations in the expansion of the product indeed do not occur. This also implies that all resulting monomials in the expansion have the same degree, which is possible only if the factors q_l of p are homogeneous. ■

Finally, we recall a result relating complete decompositions to sparse matrix spaces:

Lemma 3.3 ([9])**.** *The sparse matrix space* Σ *associated to a graph admitting a* n*-decomposition contains matrices that are* generically *non-singular.*

We recall that by *generic* is meant *everywhere except possibly on a subset of codimension at least one.*

CHAPTER 4

STABILITY OF SPARSE MATRIX SPACES

In this chapter we will examine under what conditions Sparse Matrix Spaces contain Hurwitz matrices. We will call such SMS "stable". We will also estimate the proportion of stable SMS when the free variables are randomly chosen.

Definition 4.1. *A Sparse Matrix Space is called "stable" if it contains a Hurwitz matrix. A graph corresponding to a stable SMS is called Hurwitz.*

4.1 Main Stability Conditions

A natural question to ask is how to determine whether given SMS Σ is stable or not. Some necessary and sufficient conditions for stability are given in [9] and presented below:

Theorem 4.1. *A Sparse Matrix Space $\Sigma \in \mathbb{R}^{n\times n}$ with corresponding directed graph G is stable:*

(a) if and only if each of the (strongly) connected components of G is stable;

(b) only if for every $k \in \{1, 2, ..., n\}$ there exists a k-decomposition of G.

(c) if G has a sequence of nested k-decompositions $G_1 \subset G_2 \subset ... \subset G_n$, $k = 1, 2, ..., n$;

4.2 Symmetric Sparse Matrix Spaces

In this section we briefly review some results on Symmetric Sparse Matrix Spaces, that is SMS for which the locations of the free variables are symmetric with respect to the main diagonal.

Definition 4.2. *A Sparse Matrix Space Σ is called "symmetric", if the element a_{ij} is a free variable if and only if the element a_{ji} is a free variable, for every $1 \leq i, j \leq n$.*

We note that in the case of symmetric SMS, for every edge (u, v) in its corresponding graph G, the graph G also contains the opposite edge (v, u). Therefore, to every symmetric SMS, we can attach an undirected graph, possibly containing self-loops. The notions of k-decomposition and Hamiltonian decomposition are naturally carried over to undirected graphs.

Using Theorem 4.1, we can completely classify stability of Symmetric SMS based on their graph structure.

Theorem 4.2 ([10], Theorem 6)**.** *Let G be a graph corresponding to a symmetric sparse matrix space. Then G is stable* if and only if*:*

1. *Every node in G is connected to a self-loop.*
2. *The graph G contains a Hamiltonian decomposition.*

In the case of symmetric Sparse Matrix Spaces, we are also able to estimate the proportion of stable spaces when the locations of the free variables are randomly chosen.

4.3 Random Symmetric Sparse Matrix Spaces

Let p and q be real numbers in the interval $[0, 1]$.

Definition 4.3. *A random symmetric SMS $\mathcal{M}^n_{p,q}$ is a random variable which takes values in the set of symmetric SMS of size n, such that every element on the main diagonal of the SMS is a free variable with probability q, and every element strictly below the main diagonal is a free variable with probability p.*

Definition 4.4. *A random undirected graph $\mathcal{G}^n_{p,q}$ is a random variable which takes values in the set of undirected graphs on vertices $V = \{1, 2, ..., n\}$, such that for every $u, v \in V, u > v$, the edge (u, v) belongs to E with probability p, and for every $u \in V$, the self-loop (u, u) belongs to E with probability q.*

Similarly to sparse matrix spaces and graphs, we have a natural correspondence between random SMS and random graphs.

Definition 4.5. *A property H of an undirected graph G is called monotone, if adding new edges to the graph preserves the property.*

Connectivity and existence of perfect matching are examples of monotone properties of graphs. Being a tree is an example of a non-monotone property.

As usual, with $\mathbb{P}(F)$ we will denote the probability of an event F.

Definition 4.6. *Let $X = \{X^n\}_{n=1}^{\infty}$ be a sequence of random variables. We say that almost every X^n exhibits a property H if and only if*

$$\lim_{n\to\infty} \mathbb{P}(X^n \textit{ exhibits } H) = 1.$$

The following theorem is a trivial generalization of [16], Theorem 2.1.

Theorem 4.3. *If H is a monotone property of a random graph $\mathcal{G}_{p,q}^n$ and $0 \le p_1 \le p_2 \le 1$, $0 \le q_1 \le q_2 \le 1$, then*

$$\mathbb{P}(\mathcal{G}_{p_1,q_1}^n \textit{ exhibits } H) \le \mathbb{P}(\mathcal{G}_{p_2,q_2}^n \textit{ exhibits } H).$$

We will be interested in the asymptotic properties of $\mathcal{M}_{p,q}^n$ when the size n grows to infinity.

Definition 4.7. *By $S_{p,q}^n, H_{p,q}^n, L_{p,q}^n$ we denote the following events:*

$S_{p,q}^n$ *a random symmetric sparse matrix space $\mathcal{M}_{p,q}^n$ is stable;*

$H_{p,q}^n$ *a random graph $\mathcal{G}_{p,q}^n$ contains a Hamiltonian decomposition;*

$L_{p,q}^n$ *a random graph $\mathcal{G}_{p,q}^n$ contains a self-loop.*

By $\bar{S}_{p,q}^n, \bar{H}_{p,q}^n, \bar{L}_{p,q}^n$ we denote the complements of these events.

Clearly, the properties corresponding to the events $S_{p,q}^n, H_{p,q}^n, L_{p,q}^n$ are monotone.

4.4 Stability of Random Symmetric Sparse Matrix Spaces

In this section, we will estimate the magnitudes of p and q for which most random symmetric SMS $\mathcal{M}_{p,q}^n$ are stable. Let

$$p = p(n) = \frac{\ln(n) + \omega_1}{n}, \qquad q = q(n) = \frac{\omega_2}{n},$$

where ω_1 and ω_2 are functions of n. We will assume that q is bounded away from 1, i.e. $q < 1 - \varepsilon$ for some $\varepsilon > 0$.

The following Lemma is a direct corollary of Hall's Marriage Theorem (Theorem 2.3).

Lemma 4.1. *Let G be an undirected graph without self-loops. Then G does not contain a Hamiltonian decomposition if and only if it contains an independent set $I = \{u_1, u_2, ..., u_k\}, I \subset V$, such that $|N(I)| = k - 1$ for some k.*

Proof. Along with the graph $G = (V, E), V = \{1, 2, ..., n\}$, we consider the corresponding bipartite graph $B = (V', V'', E^*)$, where $V' = \{1', 2', ..., n'\}$, $V'' = \{1'', 2'', ..., n''\}$, and $(i, j) \in E$ if and only if $(i', j') \in E^*$.

First, assume that there exists an independent set $I = \{u_1, u_2, ..., u_k\} \subset G$ such that $|N(I)| < k$. Then the same is true for the corresponding set $I' = \{u'_1, u'_2, ..., u'_k\} \subset V_1$ in the bipirtatite graph. Therefore, applying Hall's Theorem, we conclude that B does not contain a perfect matching, and therefore G does not contain a Hamiltonian decomposition.

Now assume that G does not contain a Hamiltonian decomposition, and thus B does not contain a perfect matching. Applying Hall's Theorem again, we conclude that there exists a subset $I' = \{u'_1, u'_2, ..., u'_k\} \subset V_1$, such that $N(I') < k$. Let $I' = I'_1 \cup I'_2, I'_1 \cap I'_2 = \emptyset$, where $I'_1 = \{u'_i \in I' | u''_i \in N(I')\}, I'_2 = \{u'_i \in I' | u''_i \notin N(I')\}$. The set I'_2 is non-empty, because otherwise $N(I') \geq k$, which is a contradiction. Since $N(I'_2) \cap I''_1 = \emptyset$, where $I''_1 = \{u''_i \in V'' | u'_i \in I'_1\}$, we have

$$|N(I'_2)| \leq |N(I')| - |I''_1| < k - |I'_1| = |I'_2|.$$

Therefore the corresponding set $I_2 \in V$ is independent and satisfies $|N(I_2)| < |I_2|$.

Now, choose the smallest independent set I, such that $|N(I)| < |I|$. If $|N(I)| < |I| - 1$, then we can remove any vertex v from I and get an independent set $J = I \setminus v$ for which $|N(J)| < |J|$. This is a contradiction, and therefore $|N(I)| = |I| - 1$. ■

The following theorem is given as Exercise 3.2 in [16].

Theorem 4.4 (Bollobás)**.** *Let $\omega_1 = c + o(1)$. Then the probability that the graph $\mathcal{G}^n_{p,0}$ contains an isolated vertex is equal to $1 - e^{-e^{-c}} + o(1)$.*

The proof of the next proposition follows ideas from [17].

Proposition 4.1. *Let $\omega_1 = c + o(1)$. Then the probability that $\mathcal{G}^n_{p,0}$ does not contain a Hamiltonian decomposition is equal to $1 - e^{-e^{-c}} + o(1)$.*

Proof. According to Lemma 4.1, $\mathcal{G}^n_{p,0}$ does not contain a Hamiltonian decomposition if and only if it does not contain an independent set of size $k \geq 1$ which is incident with exactly $k-1$ vertices.

Let F_k be the event that there exists an independent set I with k vertices, such that $|N(I)| = k-1$, and for every independent set J with $l < k$ vertices, $|N(J)| \neq l$. We have that $\bar{H}^n_{p,q} = \cup_{k=1}^{[(n+1)/2]} F_k$. Then, the probability that $\mathcal{G}^n_{p,q}$ does not contain a Hamiltonian decomposition is equal to

$$\mathbb{P}(\bar{H}^n_{p,q}) = \sum_{k=1}^{[(n+1)/2]} \mathbb{P}(F_k) = \mathbb{P}(F_1) + \sum_{k=2}^{[(n+1)/2]} \mathbb{P}(F_k).$$

Theorem 4.4 gives $\mathbb{P}(F_1) = 1 - e^{-e^{-c}} + o(1)$, so it remains to prove that $\sum_{k=2}^{[(n+1)/2]} \mathbb{P}(F_k) = o(1)$.

Now we evaluate F_k. We can choose the k vertices of the independent set $I \subset V$ in $\binom{n}{k}$ ways. Then, we can choose their $k-1$ neighbors in $\binom{n-k}{k-1}$ ways. If any vertex $v \in N(I)$ is adjacent with only one vertex $u \in I$, then $J = I \setminus v$ will be such that $|J| = k-1$ and $|N(J)| = k-2$, which is a contradiction. Therefore every vertex $v \in N(I)$ is adjacent to at least two vertices $u_1, u_2 \in I$. We have

$$\mathbb{P}(F_k) \leq \binom{n}{k}(1-p)^{\binom{k}{2}}\binom{n-k}{k-1}(1-p)^{(n-2k+1)k}\left(\binom{k}{2}p^2\right)^{k-1}.$$

First, we consider the terms F_k for which $n \neq 2k-1$. Using Stirling's approximation, we get

$$\begin{aligned}\mathbb{P}(F_k) &\leq \frac{n!}{k!(n-k)!}\frac{(n-k)!}{(k-1)!(n-2k+1)!}\frac{k^{k-1}(k-1)^{k-1}}{2^{k-1}}p^{2(k-1)}(1-p)^{(n-1.5k+0.5)k} \\ &\ll \frac{n^{n+0.5}k^{k-1}(k-1)^{k-1}}{k^{k+0.5}(k-1)^{k-0.5}(n-2k+1)^{n-2k+1.5}2^{k-1}}p^{2(k-1)}(1-p)^{(n-1.5k+0.5)k}.\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{P}(F_k) &\ll \frac{n^{n+0.5}}{k^2(n-2k+1)^{n-2k+1.5}2^{k-1}}\frac{C^k(\ln n)^{2(k-1)}}{n^{2(k-1)}}(1-p)^{(n-1.5k+0.5)k}\\ &\ll \left(1+\frac{2k-1}{n-2k+1}\right)^{n-2k+1.5}\frac{C^k n(\ln n)^{2(k-1)}}{k^2 2^{k-1}}(1-p)^{(n-1.5k+0.5)k}\\ &\ll \left(\left(1+\frac{2k-1}{n-2k+1}\right)^{\frac{10(n-2k+1)}{2k-1}} Cn^{\frac{1}{k}}(\ln n)^2(1-p)^{n-1.5k+0.5}\right)^k,\\ &\ll \left(Cn^{\frac{1}{k}}(\ln n)^2(1-p)^{n-1.5k+0.5}\right)^k,\end{aligned}$$

where with C we denote any constant which depends only on c. The last inequality follows from the fact that the function $f(x)=(1+\frac{1}{x})^x$ takes values between 1 and e.

Expanding $\ln(1-p)$ in Taylor series, we get

$$1-p=\exp(\ln(1-p))=\exp\left(-\frac{\ln(n)(1+o(1))}{n}\right)=n^{-\frac{1+o(1)}{n}}. \tag{4.1}$$

Therefore, for $2\le k<5$ we have

$$(1-p)^{(n-1.5k+0.5)}=n^{-(1-\frac{1.5k-0.5}{n})(1+o(1))}=n^{-1+o(1)},$$

and thus

$$\mathbb{P}(F_k)\ll\left(C\frac{n^{\frac{1}{2}}(\ln n)^2}{n^{1+o(1)}}\right)^k\ll\left(C\frac{(\ln n)^2}{n^{0.01}}\right)^k. \tag{4.2}$$

For $k\ge 5$, (4.1) implies

$$(1-p)^{(n-1.5k+0.5)}=n^{-(1-\frac{1.5k-0.5}{n})(1+o(1))}\ll n^{-0.25+o(1)},$$

and once again

$$\mathbb{P}(F_k)\ll\left(C\frac{n^{\frac{1}{5}}(\ln n)^2}{n^{0.25+o(1)}}\right)^k\ll\left(C\frac{(\ln n)^2}{n^{0.01}}\right)^k. \tag{4.3}$$

Finally, we consider F_k, such that n is odd and $n=2k-1$. If n is a

sufficiently large, we have

$$\begin{aligned}
\mathbb{P}(F_k) &\leq \frac{n!}{\left(\frac{n+1}{2}\right)!\left(\frac{n-1}{2}\right)!}(1-p)^{\binom{(n+1)/2}{2}}\binom{\frac{n+1}{2}}{2}^{\frac{n-1}{2}}p^{n-1} \\
&\ll \frac{n^{n+\frac{1}{2}}}{(n^2-1)^{\frac{1}{2}}(n+1)2^{\frac{n-5}{2}}}n^{-\frac{n^2-1}{8n}(1+o(1))}\left(\frac{C^{\frac{1}{2}}\ln n}{n}\right)^{n-1} \\
&\ll \left(n^{2-\frac{1}{n+1}}n^{-\frac{1}{4}\left(1-\frac{1}{n}\right)(1+o(1))}\left(\frac{C^{\frac{1}{2}}\ln n}{n}\right)^{2-\frac{4}{n+1}}\right)^{\frac{n+1}{2}} \\
&\ll \left(C\frac{(\ln n)^2}{n^{0.25+o(1)}}\right)^k \ll \left(C\frac{(\ln n)^2}{n^{0.01}}\right)^k, \qquad (4.4)
\end{aligned}$$

where $C = \left(1+\frac{1+c}{\ln 2}\right)^2$.

Combining (4.2), (4.3), and (4.4), we see that $\sum_{k=2}^{[(n+1)/2]}\mathbb{P}(F_k) = o(1)$, which concludes the proof. ■

Corollary 4.1. *Almost every $\mathcal{G}^n_{p,0}$ contains a Hamiltonian decomposition if and only if $\omega_1 \to \infty$. Almost every $\mathcal{G}^n_{0,q}$ contains a self-loop if and only if $\omega_2 \to \infty$.*

Proof. Notice that if $\omega_2 = c + o(1)$, then the probability that $\mathcal{G}^n_{0,q}$ does not contain a self-loop is equal to

$$\begin{aligned}
\mathbb{P}(\bar{L}^n_{0,q}) &= \lim_{n\to\infty}\left(1-\frac{c+o(1)}{n}\right)^n \\
&= \lim_{n\to\infty}\exp\left(n\ln\left(1-\frac{c+o(1)}{n}\right)\right) \\
&= \lim_{n\to\infty}\exp(-c+o(1)) \\
&= e^{-c}+o(1). \qquad (4.5)
\end{aligned}$$

Now the corollary follows from Proposition 4.1, (4.5), and from the fact that existence of a Hamiltonian decomposition, resp. of self-loops, are monotone graph properties. ■

The next Theorem gives a sharp threshold for p and q at which almost every random symmetric sparse matrix space $\mathcal{M}^n_{p,q}$ is stable.

Theorem 4.5. *Almost every $\mathcal{M}^n_{p,q}$ is stable if and only if $\omega_1, \omega_2 \to \infty$.*

Proof. If ω_2 does not tend to infinity, then Corollary 4.1 implies that there exists $P > 0$, such that for infinitely many n, $\mathbb{P}(\bar{S}^n_{p,q}) \geq \mathbb{P}(\bar{L}^n_{p,q}) > P$.

Let F_1 be the event that $\mathcal{G}^n_{p,q}$ contains an isolated vertex. If ω_1 does not tend to infinity, then Theorem 7.3 in [16] states that there exists $P > 0$, such that for infinitely many n, $\mathbb{P}(F_1) > P$. Therefore, for infinitely many n,

$$\mathbb{P}(\bar{S}^n_{p,q}) \geq \mathbb{P}(F_1)(1-q) > \varepsilon P.$$

If $\omega_1, \omega_2 \to \infty$, then from Corollary 4.1 follows that a.e. $\mathcal{G}^n_{p,q}$ contains a Hamiltonian decomposition and a self-loop. Furthermore, from Theorem 7.3 in [16] follows that a.e. $\mathcal{G}^n_{p,q}$ is connected. Therefore, using Theorem 4.2, we conclude that a.e. $\mathcal{M}^n_{p,q}$ is stable. ■

CHAPTER 5

SPARSE MATRIX SPACE EXTENSIONS

In this chapter, we will examine extensions of stable SMS, and determine under what conditions they are stable.

Definition 5.1. *Let G and G' are directed graphs, such that $G \subset G'$. We call G' a k-extension of G if $\|G'\| = \|G\| + k$.*

We will be interested in determining whether given k-extension of a stable graph is also stable. This question is a natural generalization of the problem of determining stability of arbitrary graphs and therefore is hard to solve completely. However, the 1- and 2-extension cases are manageable and here we give an almost complete analysis of them.

5.1 1-node Extensions

Proposition 5.1. *Let G and G' be graphs with n and $n+1$ vertices respectively, such that G is stable and G' is 1-extension of G. Then G' is stable if and only if it contains an $n+1$-decomposition.*

Proof. The necessity follows directly from Theorem 4.1 (b). The sufficiency can be deduced using the inductive step in Theorem 4.1 (c). ■

Now we can use Proposition 5.1 to create larger stable graphs from given smaller ones.

Corollary 5.1. *Let $G = (V, E)$ be a stable graph with n vertices. If the edge $(v_1, v_2) \in E$ belongs to some n-decomposition of G, then the 1-extension $G' = (V \cup \{v\}, E \cup \{(v_1, v)\} \cup \{(v, v_2)\})$ is also stable.*

Proof. Let the edge (v_1, v_2) belongs to the n-decomposition Γ of G. Then $\Gamma \setminus \{(v_1, v_2)\} \cup \{(v_1, v)\} \cup \{(v, v_2)\}$ is an $n+1$-decomposition of G'. ■

5.2 2-node Extensions

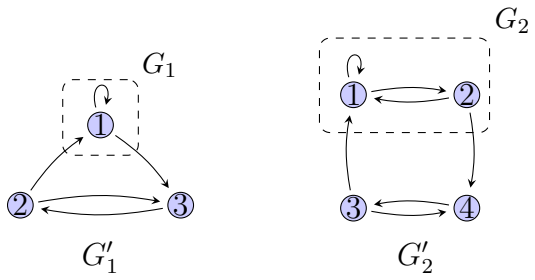

Figure 5.1: The graph G_1' (resp. G_2') is a Hurwitz two-node extension of the graph G_1 (resp. G_2). This cannot be deduced from Theorem 4.2.

We now address the design of Hurwitz graphs with $n+2$ nodes given a Hurwtiz graph with n nodes. One could of course use Proposition 5.1 twice for this task, but the resulting graph will necessarily contain a Hurwitz $n+1$ subgraph. As we will see in the following section, not all Hurwitz graphs can be obtained in this fashion. Hence, the method we provide adds two nodes in a way that does not necessarily reduce to repeated uses of Proposition 5.1.

Let us fix an SMS $\Sigma \subset \mathbb{R}^{n\times n}$ with corresponding graph $G = (V, E)$. Let Σ' be an SMS such that $\Sigma \subset \Sigma' \subset \mathbb{R}^{(n+2)\times(n+2)}$ with corresponding graph $G' = (V \cup \{v_{n+1}, v_{n+2}\}, E \cup E')$, where the edges E' are incident to either, or both of v_{n+1} and v_{n+2}. A generic matrix in Σ' is of the form

$$A' = \left[\begin{array}{c|cc} & a'_{1,n+1} & a'_{1,n+2} \\ & \vdots & \vdots \\ a_{ij} & a'_{n-2,n+1} & a'_{n-2,n+2} \\ & a'_{n-1,n+1} & a'_{n-1,n+2} \\ & a'_{n,n+1} & a'_{n,n+2} \\ \hline a'_{n+1,1} \cdots a'_{n+1,n} & a'_{n+1,n+1} & a'_{n+1,n+2} \\ a'_{n+2,1} \cdots a'_{n+2,n} & a'_{n+2,n+1} & a'_{n+2,n+2} \end{array}\right] \tag{5.1}$$

where the a'_{ij}'s represent the newly added variables, which are either free or zeros. Recall that the characteristic polynomial of a matrix A' as above can be written as $s^{n+2} + p_1 s^{n+1} + \ldots + p_{n+2}$, where p_{n+2} is the determinant of A' (up to a sign). The polynomial p_{n+2} is the sum of terms of degree $n+2$, each of which corresponds to a $n+2$-decomposition of G' by Section 3.2. Similarly, p_{n+1} is a polynomial in the variables a_{ij}, a'_{kl}, in which each term

corresponds to a $n+1$-decomposition of G'.

We can naturally decompose p_{n+1} into the sum of two terms by noticing that $n+1$-decompositions of G' need to cover *at least* one of the newly added nodes, v_{n+1} and v_{n+2}. Hence we write

$$p_{n+1} = p^1_{n+1} + p^2_{n+1}, \tag{5.2}$$

where

- p^1_{n+1} contains the terms corresponding to an $(n+1)$-decomposition that only cover v_{n+1} *or* v_{n+2}, but *not both*.
- p^2_{n+1} contains the terms corresponding to $(n+1)$-decompositions in G' that cover *both* the nodes v_{n+1} and v_{n+2}.

We have the following result:

Proposition 5.2. *Let G' be a two-node extension of the Hurwitz graph G which satisfies the necessary condition for stability (see Theorem 4.1). If p^1_{n+1} as defined above is not the zero polynomial, then G' is Hurwitz.*

Proof. As usual, we denote by Σ and Σ' the SMS associated to G and G' respectively. If p^1_{n+1} is non-zero, there are $(n+1)$-decompositions in Σ' that contain only v_{n+1} or v_{n+2}. We show that in that case, we can use Proposition 5.1 twice to prove the stability of Σ'. To wit, if there is an $(n+1)$-decomposition that only uses say node v_{n+1}, then the graph G_1 obtained by adding node v_{n+1} (and incident edges) to G satisfies the conditions of Theorem 4.2 and is thus Hurwitz. Now adding v_{n+2} to G_1 to obtain G', we see that G' satisfies the conditions of Theorem 4.2 and is thus Hurwitz. ■

We show in Prop. 6.1 below that we can check whether $p^1_{n+1} \not\equiv 0$ in polynomial time. We thus focus on the case $p^1_{n+1} \equiv 0$, which implies that $p^2_{n+1} \not\equiv 0$, since otherwise the extension fails to meet the necessary condition for stability. We cannot hope for a result akin to Theorem 4.2 in this case, as such a result would imply that conditioned on G being Hurwitz, any 2-node extensions G' that satisfies the necessary conditions is Hurwitz—a statement to which there are counter-examples. Therefore we need to make sure that the newly added edges are distributed in a way that allows us to have sufficient control over the roots of the characteristic polynomials of the matrices

in the corresponding G'. We introduce here a simple test to check that the edges are well-distributed (see Eqs. (5.1) and (5.2) for Definitions of a'_{ij} and p_{n+1}, p_{n+2})

Definition 5.2 (Edge distribution test)**.** *We say that a 2-node extension G' of a graph G passes the edge distribution test if $p_{n+1} \not\equiv 0, p_{n+2} \not\equiv 0$ and the rational function p_{n+1}/p_{n+2} is not a function of the a_{ij} only (that is, not all the a'_{kl} variables simplify in the ratio).*

We can now state the main result of this section:

Theorem 5.1. *If a 2-node extension G' of a Hurwitz graph G passes the edge distribution test, then G' is Hurwitz.*

We show in the second part of the thesis that we can check in polynomial time whether a two-node extension passes the edge-distribution test.

The following Lemma will be needed in the proof of Theorem 5.1.

Lemma 5.1. *Let G' be a 2-extension, passing the edge distribution test, of a Hurwitz graph G. With the notation of Definition 5.2, the ratio p_{n+1}/p_{n+2} is not a polynomial in the a'_{kl} variables.*

The above Lemma says that if a 2-extension passes the edge distribution test, the ratio p_{n+1}/p_{n+2} is a rational function of the variables a'_{kl} with coefficients in the field $\mathbb{R}(\{a_{ij}\}_{i,j\leq n})$ and a non-constant denominator.

Proof. We denote by $\bar{a}'$ (resp. $\bar{a}$) the vector containing all a'_{kl} variables (resp. a_{ij} variables). Assume, by contradiction, that p_{n+1}/p_{n+2} is a polynomial in the a'_{kl} variables, that is there exist polynomials $s(\bar{a}, \bar{a}')$ and $r(\bar{a})$ such that

$$p_{n+1}/p_{n+2} = s(\bar{a}, \bar{a}')/r(\bar{a}).$$

In general, we can thus write $p_{n+2} = q(\bar{a}, \bar{a}')r(\bar{a})$ and $p_{n+1} = q(\bar{a}, \bar{a}')s(\bar{a}, \bar{a}')$, where q is a polynomial in $\bar{a}$ and $\bar{a}'$ with coefficients in $\mathbb{R}$. By the correspondence between determinants and k-decompositions, every $n+2$-decomposition of G' is obtained by multiplying a term of q with a term of r. Similarly, every $n+1$-decomposition is obtained by multiplying a term of q with a term of s. Notice that the a'_{kl} variables correspond to edges incident to either node v_{n+1} or v_{n+2}. Furthermore, in any k-decomposition covering v_{n+1} and v_{n+2}, both of these nodes are incident to 2 edges each - one incoming and one

outgoing. Let us choose an arbitrary $n+2$-decomposition D_1 of G' and let a'_{kl} be the edge-variables used in the decomposition. Because of the remark above, all of these variables appear in one of the terms of q—call it α. If we assume that s does not have a trivial dependence on $\bar{a}'$, then it contains a term which depends on some a'_{kl}'s. Call that term β. Now we consider the edge-product $\alpha\beta$ and notice that it corresponds to a $n+1$-decomposition D_2 of G'. By construction, D_2 uses all edges in D_1 incident to the nodes v_{n+1} and v_{n+2} and at least one additional edge also incident to v_{n+1} or v_{n+2}. This implies that either node v_{n+1} or v_{n+2} has degree more than 2, which is a contradiction. ■

We now give the proof of the main theorem of this section.

Proof. (Theorem 5.1) For the sake of convenience, we shall consider the equivalent problem of proving that in the SMS corresponding to G', there exist matrices, all of whose eigenvalues have positive real parts (the negative of every such matrix is a stable matrix). Because of Proposition 5.2, it is sufficient to only consider the case $p^1_{n+1} \equiv 0$. Because G' satisfies the necessary condition for stability, we can conclude that $p_{n+1} = p^2_{n+1} \neq 0$.

Let $-A \in \Sigma$ be a Hurwitz matrix and define the matrix

$$A'_0 = \left[\begin{array}{ccccc|cc} & & & & & 0 & 0 \\ & & & & & \vdots & \vdots \\ & & A & & & 0 & 0 \\ & & & & & 0 & 0 \\ & & & & & 0 & 0 \\ \hline 0 & \cdots & 0 & 0 & & 0 & 0 \\ 0 & \cdots & 0 & 0 & & 0 & 0 \end{array}\right]. \tag{5.3}$$

We will show that there exists A^* in Σ' with n eigenvalues close to the eigenvalues of A, and hence with positive real parts, and such that the two other eigenvalues have positive real parts as well. The eigenvalues of A'_0 are $\lambda_1, \ldots, \lambda_n, \lambda_{n+1}, \lambda_{n+2}$ where λ_i, $1 \leq i \leq n$ are the eigenvalues of A, and $\lambda_{n+1} = \lambda_{n+2} = 0$. We use the notation

$$\sum_{j=1}^{n} \hat{\prod} \lambda_j := \sum_{j=1}^{n} \prod_{i=1, i\neq j}^{n} \lambda_i.$$

In particular, $p_{n+1} = \sum \hat{\prod} \lambda_i(A')$ for $A' \in \Sigma'$.

Because the eigenvalues λ_i of a matrix depend continuously on its entries, there exists $\varepsilon > 0$ such that for all A' in an ε-neighborhood of A'_0 in Σ' (for, say, the ∞-norm), the following three items hold:

1. $\min_{1\leq i\leq n} |\lambda_i(A')| > \max(|\lambda_{n+1}(A')|, |\lambda_{n+2}(A')|)$.
2. $\prod_{i=1}^{n} \lambda_i(A')$ is bounded away from zero.
3. $\sum_{i=1}^{n} \hat{\prod}\lambda_i(A')$ is bounded.

Indeed, property 1 follows from $\lambda_{n+1}(A'_0) = \lambda_{n+2}(A'_0) = 0$, $\lambda_i(A'_0) > 0$ for $i = 1, \ldots, n$ and continuity of the λ_i. Property 2 follows from $\prod_{i=1}^{n} \lambda_i(A'_0) \neq 0$ and Property 3 from the fact that $\sum_{i=1}^{n} \hat{\prod}\lambda_i(A'_0)$ is bounded. Let A' be a matrix in the ε-neighborhood of A'_0. Recall that the coefficients of the characteristic polynomial $\det(Is - A')$ of A' are given by formula 3.2. Since the roots of the characteristic polynomial are λ_i, we also have $\det(Is - A') = \prod_{i=1}^{n+2}(s - \lambda_i)$. Equating the coefficients of the terms in s^0 and s^1, we obtain

$$\begin{aligned} \det(A') = p_{n+2} &= \lambda_{n+1}\lambda_{n+2} \prod_{i=1}^{n} \lambda_i \\ \sum_{i=1}^{n+2} \det(A'_{[i,i]}) = p_{n+1} &= (\lambda_{n+1} + \lambda_{n+2}) \prod_{i=1}^{n} \lambda_i + \lambda_{n+1}\lambda_{n+2} \sum \hat{\prod} \lambda_i \end{aligned}$$

where $\det(A'_{[i,i]})$ is the principal minor obtained by removing the i-th row and the i-th column from A'. Since $p^1_{n+1} \equiv 0$, there are no $n+1$-decompositions in G' that contain only one of the nodes $n+1$ and $n+2$. From the relation between k-decompositions and principal minors from Section 3.2, we know that these $n+1$-decompositions correspond to principal minors of entries $(n+1, n+1)$ and $(n+2, n+2)$. Thus

$$\sum_{i=1}^{n+2} \det(A'_{[i,i]}) = \sum_{i=1}^{n} \det(A'_{[i,i]}).$$

From the above two relations, we obtain

$$\begin{aligned} \lambda_{n+1}\lambda_{n+2} &= \det(A') / \textstyle\prod_{i=1}^{n} \lambda_i \\ \lambda_{n+1} + \lambda_{n+2} &= \frac{\det(A')}{\prod_{i=1}^{n} \lambda_i} \left[\frac{\sum_{i=1}^{n} \det(A'_{[i,i]})}{\det(A')} - \frac{\sum \hat{\prod} \lambda_i}{\prod \lambda_i} \right] \end{aligned} \tag{5.4}$$

We first show that the eigenvalues $\lambda_{n+1}, \lambda_{n+2}$ of A' are either real or complex conjugate. Because conjugate numbers have the same norm, λ_{n+1} and λ_{n+2} cannot be complex conjugates of λ_i, $i = 1 \dots n$ by item 1. This implies, in turn, that $\det(A')/\prod_{i=1}^{n} \lambda_i(A') = \lambda_{n+1}\lambda_{n+2}$ is real, and the same is true for $\lambda_{n+1} + \lambda_{n+2}$.

From here on, we focus on showing that there exists A^* in the ε-neighborhood of A'_0 such that $\lambda_{n+1}\lambda_{n+2}$ and $(\lambda_{n+1} + \lambda_{n+2})$ are both strictly positive, and hence so are the real parts of λ_{n+1} and λ_{n+2}. We will do so by showing that

a) we can make the term $\frac{\sum_{i=1}^{n} \det(A'_{[i,i]})}{\det(A')}$ arbitrarily large

b) we can control the sign of $\det(A')/\prod_{i=1}^{n} \lambda_i$ without affecting $\frac{\sum_{i=1}^{n} \det(A'_{[i,i]})}{\det(A')}$.

The above two requirements, in view of (5.4) and properties 2 and 3 above, allow us to control the signs of λ_{n+1} and λ_{n+2}. We first focus on *a)*. We will make the ratio arbitrarily large by making its denominator arbitrarily close to zero and controlling the numerator. Because G' passes the edge-distribution test by Lemma 5.1, there is an edge $e^* \in E'$, with corresponding entry a^*_{kl} in A', such that $\frac{\sum_{i=1}^{n} \det(A'_{[i,i]})}{\det(A')}$ is a non-constant rational function of a^*_{kl}. Without loss of generality, we can assume that a^*_{kl} is in one of the last two rows of A'. We expand $\det(A')$ along the row containing a^*_{kl} to obtain the relation

$$\det(A') = a^*_{kl} q(A') + r(A'),$$

where $q(A')$ and $r(A')$ are polynomials in the other free variables (viz, besides a^*_{kl}) of degrees $n+1$ and $n+2$ respectively. Note that the root of $\det(A')$, seen as a linear function of a^*_{kl}, is at $-\frac{r(A')}{q(A')}$. Similarly, $\sum_{i=1}^{n} \det(A'_{[i,i]})$ can be expressed as

$$\sum_{i=1}^{n} \det(A'_{[i,i]}) = a^*_{kl} \bar{q}(A') + \bar{r}(A')$$

for appropriately defined polynomials $\bar{q}$ and $\bar{r}$. We claim that the following holds

$$\bar{q}(A') \frac{r(A')}{q(A')} \not\equiv \bar{r}(A'). \tag{5.5}$$

Indeed, assuming by contradiction that the previous relation is an identity, then

$$\det(A')/\sum_{i=1}^{n} \det(A'_{[i,i]}) = q(A')/\bar{q}(A')$$

does not depend on a^*_{kl}—a contradiction with the definition of a^*_{kl}.

Now choose A'_1 in the ε neighborhood of A'_0 such that $\bar{q}(A'_1)\frac{r(A'_1)}{q(A'_1)} \neq \bar{r}(A'_1)$ and $\mu > 0$ small enough so that

$$\left|\mu \frac{r(A'_1)}{q(A'_1)}\right| < \varepsilon \tag{5.6}$$

holds. Furthermore, set $A'_2 = A'_1 I_\mu$ where I_μ is the identity matrix with its last entry replaced by μ:

$$I_\mu = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & \cdots & & \mu \end{bmatrix}.$$

By choosing a^*_{kl} close to $-\frac{r(A'_2)}{q(A'_2)} = \mu\frac{r(A'_1)}{q(A'_1)}$, the ratio $\frac{\sum_{i=1}^n \det([A'_2]_{[i,i]})}{\det(A'_2)}$ can be made arbitrarily large. We thus fix a^*_{kl} so that the previous ratio is larger than $\sup \frac{\sum \hat{\prod}\lambda_i}{\prod \lambda_i}$, which is bounded by item 2 and 3. We denote by A'_3 the matrix obtained for that choice of a^*_{kl}.

For *b)* observe that on the one hand,

$$\det(A'_3 I_{-1}) / \prod_{i=1}^n \lambda_i(A'_3 I_{-1}) = -\det(A'_3) / \prod_{i=1}^n \lambda_i(A'_3)$$

since the numerator changes its sign if we invert the sign of the last row of A'_3, but the product $\prod_{i=1}^n \lambda_i(A'_3)$ does not as a consequence of item 2. On the other hand, we have that

$$\frac{\sum_{i=1}^n \det((A'_3 I_{-1})_{[i,i]})}{\det(A'_3 I_{-1})} = \frac{\sum_{i=1}^n \det([A'_3]_{[i,i]})}{\det(A'_3)}$$

because every minor in the summation depends on the last row of $A'_3 I_{-1}$. This ends the proof of points *a)* and *b)*. Putting the above together, we take A^* to be $A'_3 I_\delta$ where $\delta = 1$ if $\det(A'_3 I_{-1}) / \prod_{i=1}^n \lambda_i(A'_3 I_{-1})$ is positive and -1 otherwise. This concludes the proof. ■

We summarize the steps of the proof as it provides a method to obtain a Hurwitz matrix in the 2-extension Σ':

1. Pick a Hurwitz matrix $A \in \Sigma$ and create A'_0 as in (5.3).

2. Set $\varepsilon > 0$ such that items 1, 2 and 3 above hold.

3. Find an edge in G' (with corresponding variable a^*_{kl}) which meets meets Definition 5.2.

4. Find A'_1 in the ε neighborhood of A'_0 such that (5.5) holds (the relation holds for almost all matrices A' in the neighborhood).

5. Choose μ small enough so that (5.6) holds and set $A'_2 = A'_1 I_\mu$.

6. Obtain A'_3 by updating the entry of A'_2 corresponding to the edge a^*_{kl} to a value close enough to $\frac{r(A'_2)}{q(A'_2)}$ so that

$$\frac{\sum_{i=1}^n \det([A'_3]_{[i,i]})}{\det(A'_3)} > \sup \frac{\sum \hat{\prod} \lambda_i(A'_3)}{\prod \lambda_i(A'_3)}.$$

7. Set A^* to be $A'_3 I_\delta$ where $\delta = 1$ if $\det(A'_3 I_{-1}) / \prod_{i=1}^n \lambda_i(A'_3 I_{-1})$ is positive and -1 otherwise.

5.3 Higher k-extensions

We have demonstrated in the previous sections that one can obtain simple conditions which guarantee that 1- and 2-node extensions of Hurwitz graphs are Hurwitz. This begs the question of whether there exists a finite set of extension rules that would allow to create *all* Hurwitz graphs via node extensions. We show here that such hope is unfortunately vain. To wit, if the above conjecture was true, there would exists a finite k^* such that every Hurwitz graph on $n > k^*$ nodes admits a Hurwitz subgraph on $n - l$ nodes for some $0 < l \leq k^*$. We show in this section that, on the contrary, there exist Hurwitz graphs of arbitrary cardinality whose *sole Hurwitz subgraph* is the trivial Hurwitz graph (that is, the graph on one node with one self-loop). In order to characterize these graphs, we require a sufficient condition for a graph to be Hurwitz that does not follow from the conditions given in Th. 4.1. We give it in the next Proposition.

Proposition 5.3. *Let G be a digraph with n nodes. If there exists a sequence $e_1, \ldots, e_n$ of edges, and a permutation $(\sigma(1), \sigma(2), \ldots, \sigma(n))$ of $\{1, 2, \ldots, n\}$*

such that the edge e_i appears in at least one $\sigma(i)$-decomposition of G but not in any $\sigma(l)$-decompositions, $1 \leq l < i$, then G is Hurwitz.

Proof. Let Σ be the SMS associated to G, let $A \in \Sigma$ and denote by e_i the entry in A corresponding to the edge e_i, $1 \leq i \leq n$, and a_j, $1 \leq j \leq n^2 - n$ the other entries. Let $s^n + p_1 s^{n-1} + ... + p_{n-1}s + p_n$ be the characteristic polynomial of A. We think of p_k as polynomials in e_i, a_j. We show that for arbitrary real numbers $b_1, b_2, \ldots, b_n$ and $\epsilon > 0$, we can find values for the e_i's and the a_j's such that $|p_i - b_i| < \epsilon$, $i = 1, \ldots, n$. This previous statement clearly implies the claim of the proposition.

We set $\mathbf{e}_i = (e_1, \ldots, e_i)$, $\mathbf{p}_{\sigma(i)} = (p_{\sigma(1)}, \ldots, p_{\sigma(i)})$ and $\mathbf{a} = (a_1, a_2, ..., a_{n(n-1)})$.

Start with the first edge in the sequence: we know that the edge e_1 appears in at least one $\sigma(1)$-decomposition and no $\sigma(1)$-decomposition contains e_j for $j > 1$. Therefore, by (3.2) we can write $p_{\sigma(1)} = e_1 q_1(\mathbf{a}) + r_1(\mathbf{a})$, where $q_1 \neq 0$ and r_1 are polynomials in the variables a_j. Next, we consider e_i for $i = 2$. By the same argument, we see that $p_{\sigma(2)} = e_2 q_2(e_1, \mathbf{a}) + r_2(e_1, \mathbf{a})$, where $q_2 \neq 0$ and r_2 are polynomials in a_j and e_1. In general, we have

$$p_{\sigma(i)} = e_i q_i(\mathbf{e}_{i-1}, \mathbf{a}) + r_i(\mathbf{e}_{i-1}, \mathbf{a}).$$

Since the polynomials q_i are not zero, we can express e_i in terms of $p_{\sigma(i)}$, q_i and r_i as

$$e_i = \frac{p_{\sigma(i)} - r_i(\mathbf{e}_{i-1}, \mathbf{a})}{q_i(\mathbf{e}_{i-1}, \mathbf{a})}.$$

On the set where all $q_i(\mathbf{e}_{i-1}, \mathbf{a})$ are non-zero, we can regard $p_{\sigma(i)}, 1 \leq i \leq n$, and a_k, $1 \leq k \leq n^2 - n$, as independent variables. We replace $e_1, \ldots, e_{i-1}$ by their expressions in the equation of e_i.

We first see that we can express

$$e_1(\mathbf{p}_{\sigma(1)}, \mathbf{a}) = \frac{p_{\sigma(1)} - r_1(\mathbf{a})}{q_1(\mathbf{a})} =: \frac{f_1(\mathbf{p}_{\sigma(1)}, \mathbf{a})}{g_1(\mathbf{p}_{\sigma(1)}, \mathbf{a})}$$

for some relatively prime polynomials $f_1(x, y)$ and $g_1(x, y)$. We plug that expression into the one for e_2 to get

$$e_2(\mathbf{p}_{\sigma(2)}, \mathbf{a}) = \frac{p_{\sigma(2)} - r_2(e_1(\mathbf{p}_{\sigma(1)}, \mathbf{a}), \mathbf{a})}{q_2(e_1(\mathbf{p}_{\sigma(1)}, \mathbf{a}), \mathbf{a})} =: \frac{f_2(\mathbf{p}_{\sigma(2)}, \mathbf{a})}{g_2(\mathbf{p}_{\sigma(2)}, \mathbf{a})}.$$

In general, we have

$$e_i(\mathbf{p}_{\sigma(i)}, \mathbf{a}) = \frac{p_{\sigma(i)} - r_i(\mathbf{e}_{i-1}, \mathbf{a})}{q_i(\mathbf{e}_{i-1}, \mathbf{a})} =: \frac{f_i(\mathbf{p}_{\sigma(i)}, \mathbf{a})}{g_i(\mathbf{p}_{\sigma(i)}, \mathbf{a})}.$$

Going step by step over the process of substitution, we can see that all e_i are well-defined, i.e. the polynomials $g_i(x, y)$ are non-zero for all i.

Recall that our objective is to find values for e_i and a_k such that p_j are close to given numbers b_j. In order to do this, we will find appropriate values for a_{kl} and $p_j \approx b_j$, such that after making a substitution in the equations above, we will get proper (finite) values for e_i. Then the chosen a_k and the found e_i when plugged in the initial equations for $p_{\sigma(i)}$ will give $p_j \approx b_j$, which will solve the problem.

To find suitable a_k, first we consider the polynomials $g_i(x, y)$. Since none of them is identically zero, we can find some value $\alpha = (\alpha_1, \alpha_2, ..., \alpha_{n^2-n})$ for the vector variable y, such that $g_i(x, \alpha) \neq 0$ for $i = 1, ..., n$. Denote $\bar{g}_i(x) = g_i(x, \alpha)$ and consider the zero set of $\prod \bar{g}_i(x)$. It is of codimension at least one, so we conclude that there exist values β_j such that $|b_j - \beta_j| < \epsilon$ and $\bar{g}_i(\beta_{\sigma(1)}, ..., \beta_{\sigma(i)}) \neq 0$ for all i. Thus we find $a_k = \alpha_k$ and $e_i = \frac{f_i(\beta_{\sigma(1)}, ..., \beta_{\sigma(i)}, \alpha_1, ..., \alpha_{n^2-n})}{g_i(\beta_{\sigma(1)}, ..., \beta_{\sigma(i)}, \alpha_1, ..., \alpha_{n^2-n})}$, where all e_i are well defined. Clearly, these values satisfy the conditions and if we plug them in equations for $p_{\sigma(i)}$, we will get $p_j = \beta_j$. This concludes the proof. ■

We now show how to construct graphs which satisfy the conditions of Prop. 5.3 but so that none of their subgraphs satisfy the necessary conditions from Th. 4.1.

Theorem 5.2. *For any $n \geq 3$, there exists a Hurwitz graph G_n on n nodes such that all subgraphs of G_n with k nodes, $1 < k < n$, are not Hurwitz.*

Proof. We define the following sequence of graphs: G_n is a graph on n nodes, labeled $1, 2, \ldots, n$, with edges

1. $(1, k)$ for $k < n$,
2. $(k, k + 1)$ for $k < n$,
3. $(3, 2)$ and $(n, 1)$.

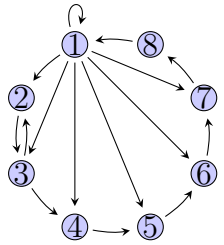

Figure 5.2: A Hurwitz graph on $n = 8$ nodes with the property that none of its subgraphs are Hurwitz, save for the trivial graph.

We depict G_8 in Fig. 5.2. Let n be fixed and consider the following sequence of edges in G_n:

$$e_1 = (1), e_2 = (2,3), e_3 = (1,n-1), e_4 = (1,n-2), \ldots, e_{n-1} = (1,3), e_n = (1,2).$$

We claim that edge e_i appears in at least one i-decomposition of G_n, but not in any l-decompositions for $l < i$ and thus G_n is Hurwitz by Prop. 5.3. To see that the claim holds, it is easier to start with edge e_n, which connects node 1 to node 2. From node 2, the only accessible node is 3, and from node 3 we can go to 4 or back to 2. The latter option yields the sequence $(1,2,3,2)$ which can not be a part of a cycle. The former option yields $(1,2,3,4)$. From any node $i > 3$, the only accessible node is $1 + (i \mod n)$. Hence the only cycle to which e_n belongs is $(123\cdots n)$ — an n-decomposition. We now take $e_{n-1} = (1,3)$. Using the same reasoning as above, the only cycle to which e_{n-1} belongs is $(1345\cdots n)$. The situation for e_{n-j}, $n-1 \geq j > 2$ is simpler to handle as the only cycle to which this edge belongs is $(1, j+1, j+2, \cdots, n)$ (this is again a consequence of the fact that from node $i > 3$, the only accessible node is $i+1$). Hence the only decompositions to which e_j belongs are $(1(j+1)\cdots n)$ and $(23)(1(j+1)\cdots n)$. Finally, $e_2 = (2,3)$ clearly can not belong a 1-decomposition (self-loop), which proves the claim.

We now show that every Hurwitz subgraph of G_n has either 1 or n nodes. To prove the claim, assume that G_k is a subgraph of G_n with $1 < k < n$ nodes and that G_k is Hurwitz. From the necessary conditions of Th. 4.1, we know that, first, there is an l-decomposition in G_k for $1 \leq l \leq k$ and, second, every node in G_k is strongly connected to node 1 (the only node with a self-loop). Notice that there is a unique 1-decompostion—the self-loop (1)—and a unique 2-decomposition— the cycle (23). Therefore, any Hurwitz subgraph

with $k \geq 2$ nodes must contain nodes 1, 2 and 3 by the first point above. Now observe that the only path from 3 to 1 is $345 \cdots n$. Hence, if any of the nodes $4, 5, \cdots, n$ is missing, 3 is not connected to 1. Thus we cannot spare any nodes in G_n and still satisfy the necessary conditions for stability as claimed. ■

CHAPTER 6

POLYNOMIAL TIME ALGORITHMS FOR NODE-EXTENSIONS

In this final chapter, we show that there exist deterministic, polynomial-time algorithms to verify whether the extensions of Hurwitz graphs discussed in the first part are Hurwitz. The main results are the following two Theorems, dealing with one-extensions and two-extensions respectively:

Theorem 6.1. *Let G' be a 1-node extension of a Hurwitz graph G. There is a polynomial time algorithm to decide whether G' is Hurwitz.*

and

Theorem 6.2. *Let G' be a 2-node extension of a Hurwitz graph G. There is a polynomial time algorithm to check whether G' passes the edge-distribution test and hence is Hurwitz.*

The remaining sections are devoted to proving Theorem 6.1 and Theorem 6.2. A basic tool is the relationship between n-decompositions of graphs and perfect matchings in an associated bipartite graph, which we present next.

6.1 Hamiltonian decompositions and bipartite matchings

Given a digraph $G = (V, E)$, we introduce the bipartite graph $G^2 = (V^2, E^2)$ with V^2 and E^2 defined as follows: if $V = \{1, 2, \ldots, n\}$, we set

$$V^2 = \{1, 2, \ldots, n, 1', 2', \ldots, n'\} \tag{6.1}$$

and

$$E^2 = \{(i, j') \text{ for } (i, j) \in E\}. \tag{6.2}$$

It is clear from its definition that the graph G^2 is a bipartite graph, with edges going from $V = \{v_1, \ldots, v_n\}$ to $V' = \{v_{1'}, \ldots, v_{n'}\}$. We have the following correspondence:

Lemma 6.1. *Hamiltonian decompositions of the directed graph $G = (V, E)$ are in one-to-one correspondence with perfect bipartite matchings of G^2.*

Proof. We first show that to a Hamiltonian decomposition of G corresponds a perfect matching in G^2. Denote by $C_1, \ldots, C_l$ disjoint cycles whose union covers V. Consider the list (a_j, b_j) of edges that appear in the cycles C_i. There are exactly n such edges and every node in G appears exactly twice in the list: once as the origin node of an edge (viz as a a_j) and once as the destination node of an edge (viz as a b_j). By definition of G^2, the edge (a_j, b_j) corresponds to the edge (a_j, b'_j) of E^2; let $M = \{(a_j, b'_j)\}$. By construction, every node in V and every node in V' is incident to exactly one edge of M and thus M is a perfect matching of G^2.

Now assume that $M = \{(a_j, b'_j)\}$ is a complete matching of G^2. Consider the set of edges $M' = \{(a_j, b_j) \mid (a_j, b'_j) \in M\} \subset E$. We claim that edges in M' yield a Hamiltonian decomposition of G. To see this, observe that because M is a complete matching, every node in G is the origin node of exactly one edge of M' and the destination node of exactly one edge of M'. Hence one can uniquely assign every node of G to a path made of edges in M'—namely the path obtained by following the unique edge leaving the node and iterating. Because every node has an incoming and outgoing edge, this path does not have any terminal or starting node and is thus a cycle. In addition, the fact every node has a unique incoming edge ensures the fact that every node is visited exactly once by a cycle. Hence M' yields disjoint cycles that visit every node of G exactly once, that is a Hamiltonian decomposition of G. ■

It has been known, at least since the time of Jacobi, that maximum matchings can be found in polynomial time, a common algorithm for this task being the *Hopcroft-Karp* algorithm [18]. Putting these facts together, we obtain the following algorithm:

Algorithm 1: finding an n-decomposition containing a specified edge.

Input: a digraph $G = (V, E)$ and an edge $e^* \in E$.

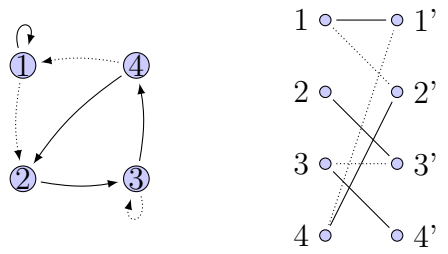

Figure 6.1: The bipartite graph to the right has two nodes for each node i of the graph on the left, labelled i and i'. A directed edge (k, l) of G correspond to an edge (k, l') in G^2. The plain edges show a 4-decomposition of G and the corresponding perfect matching of G^2.

Output: an n-decomposition of G containing the edge e if one exists, the empty set otherwise.

1. Construct the bipartite graph G^2 as described in Eq. (6.1) and (6.2). Set $b^* \in E^2$ to be the edge corresponding to e.

2. Discard all edges in G^2 adjacent to b^* and call the new graph $G^2(b^*)$.

3. Check whether the graph $G^2(b^*)$ contains a perfect matching using the Hopcroft-Karp algorithm. If it does not, then output the empty set. If G^2 contains a perfect matching, return the edges of E corresponding to it.

We prove the correctness of the above algorithm in the proposition below:

Proposition 6.1 (Polynomial-time algorithm for finding Hamiltonian decompositions)**.**

1. *Algorithm 1 determines in polynomial time whether a digraph G with n nodes contains an n-decomposition containing a specified edge e.*

2. *There is a polynomial time algorithm to decide whether a directed graph admits an n-decomposition.*

3. *There is a polynomial time algorithm to decide whether a directed graph admits an $n-1$-decomposition that contains a specified edge.*

Proof. Let $G = (V, E)$ be a directed graph with n nodes. For the first part, we first construct the bipartite graph $G^2 = ((V, V'), E^2)$ as described in (6.1)

and (6.2)—this can be done in polynomial time. We call b^* the edge in G^2 corresponding to e^* and set $G^2(b^*)$ be the subgraph of G^2 induced by all edges of G^2 that are *not adjacent* to b^*. We then run the maximum matching algorithm on $G^2(b^*)$ and denote its output by M. If M contains n edges, it is a perfect matching. Moreover, from Lemma 6.1 and the fact that b^* is the only edge in $G^2(b^*)$ incident with its vertices, it follows that M produces an n-decomposition containing the edge e. Reciprocally, if the graph G admits an n-decomposition containing e, then we can easily see that all of its edges have corresponding ones in the graph G^2 which form a perfect matching.

The second part can be proved by directly applying the Hopcroft-Karp algorithm to G^2.

Finally, for the last part, it suffices to run Algorithm 1 on all graphs obtained from G by removing a node which is not incident with e^*. ■

6.2 Signatures and factorization

Verifying whether a 2-extension passes the edge distribution test requires to check whether two multi-variable polynomials have factors in common. It is a well-known fact that such problems are hard to solve (in fact, of exponential complexity in the general case) and mostly intractable when the number of variables is large. Since a Hurwitz graph on n nodes has at least n edges [9], off-the-shelf methods of computational algebra [19] are unlikely to yield tractable algorithms. We show in this section that the relation between coefficients of the characteristic polynomials and k-decompositions can be brought to yield a polynomial-time algorithm.

Recall that we refer to a_{ij}, the ijth entry of a matrix $A \in \Sigma$, which corresponds to the edge (i, j) of the graph G, as an edge-variable and that we call an edge-product a monomial in the edge-variables, that is an expression of the form $\alpha = \prod a_{ij}$; we also treat α as a *set of edges*, and thus *we can take intersections and unions of edge-products.*

The **in-degree** of node v_l **with respect to an edge-product** α is the number of edges in α entering node v_l, i.e. it is the number of edge-variables $a_{\cdot l}$ in α. We define the out-degree similarly as the number edge-variables $a_{l\cdot} \in \alpha$. We denote in- and out-degree by $\deg^-(v_l, \alpha)$ and $\deg^+(v_l, \alpha)$ respectively. For example, the out-degree of node v_1 with respect to the edge-product

$a_{12}a_{23}a_{13}$ is two and the in-degree of node two is one.

We collect the in- and out-degrees of every node with respect to a given edge-product α in the vector $S(\alpha)$, which we call the **signature** of α and explicitly define as

$$S(\alpha) := \begin{bmatrix} (\deg^-(v_1,\alpha), \deg^+(v_1,\alpha)) \\ (\deg^-(v_2,\alpha), \deg^+(v_2,\alpha)) \\ \vdots \\ (\deg^-(v_n,\alpha), \deg^+(v_n,\alpha)) \end{bmatrix} =: \begin{bmatrix} S_1(\alpha) \\ S_2(\alpha) \\ \vdots \\ S_n(\alpha) \end{bmatrix}. \tag{6.3}$$

Note that the map $\alpha \to S(\alpha)$ is many-to-one in general. We can **partially order signatures** via component-wise comparisons: we say that $S \preceq T$ if $S_{ij} \leq T_{ij}$ for $1 \leq i \leq n$ and $1 \leq j \leq 2$. The following property of signatures is easily verified: for α and β edge-products we have

$$S(\alpha\beta) = S(\alpha) + S(\beta). \tag{6.4}$$

Lemma 6.2 (Signatures of k-decompositions)**.** *Let G be a digraph and α an edge-product in G. Then α is a k-decomposition in G if and only if $S(\alpha)$ contains exactly k rows equal to $(1,1)$ and $n-k$ rows equal to $(0,0)$.*

Proof. The result is a consequence of the fact that every node in a k-decomposition has in- and out-degree one. ■

In the next results, we relate signatures and factorization of the coefficients of the characteristic polynomial of a SMS.

Lemma 6.3. *Suppose that a polynomial q divides $p_n = (-1)^n det(A)$, and let $q = \alpha_1 \pm \alpha_2 \ldots \pm \alpha_k$ be the expansion of q into a sum of edge-product. Then all edge-products α_i have the same signature.*

We illustrate this fact below in Fig. 6.2.

Proof. The polynomial $p_n = \sum \pm\gamma_i$ is a signed sum of edge-products γ_i of degree n, each corresponding to an n-decomposition. Since every node is visited exactly once in a Hamiltonian decomposition, all edge-products γ_i have a signature with all entries one.

Let $q = \alpha_1 \pm \alpha_2 \ldots \pm \alpha_k$ and $r = \beta_1 \pm \ldots + \pm\beta_l$ be such that $p_n = qr$. Because p_n is homogeneous, so is q (resp. r) and thus all the α_i (resp. β_i)

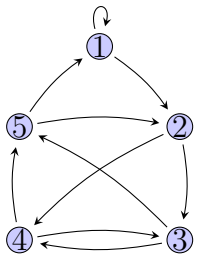

Figure 6.2: For the graph G above, we can factor p_5 as $p_5 = (-a_{11}a_{52} + a_{51}a_{12}).(a_{23}a_{34}a_{45} + a_{24}a_{35}a_{43})$. The edge products $a_{11}a_{52}$ and $a_{51}a_{12}$ in the first factor have the same signature, and so do the edge-products $a_{23}a_{34}a_{45}$ and $a_{24}a_{43}a_{35}$ appearing in the second factor.

have the same degree as polynomials. After expanding the product qr, we get edge-products of the type $\alpha_i\beta_j$ that have to correspond to Hamiltonian decompositions. We thus obtain using (6.4) and the fact that Hamiltonian decompositions have associated signatures of all ones, that

$$S(\alpha_i\beta_j) = S(\alpha_i) + S(\beta_j) = 1$$

for all pairs $i = 1, \ldots, k, j = 1, \ldots, l$. We conclude that $S(\alpha_i)$ is the same for all α_i. ■

From Lemma 6.3, we conclude that *we can associate a signature to a* factor q *of* p_n. In other words, the Lemma shows that we can make the following definition:

$$S(q) := S(\alpha) \text{ for } \alpha \text{ any term in the factor } q \text{ of } p_n.$$

Definition 6.1 (Signatures of the determinant p_n.)**.** *We call a* **signature of** p_n *any n-dimensional $Q = (Q_1, Q_2, ..., Q_n)$ such that there exists an irreducible factor of p_n with Q as signature. We denote by $\mathcal{S}(p_n)$, the* **set of signatures of** p_n.

We also refer to signatures of p_n as **signatures** of the graph. The following result relates common factors of the polynomials p_k. We let $(-1)^\gamma$ denote the sign of a permutation γ.

Proposition 6.2. *Let G be a digraph and Q a given integer vector. Denote by p_k the kth coefficient of the characteristic polynomial of G . The following statements are equivalent:*

1. *All k-decompositions of G contain a subgraph with corresponding signature Q.*

2. *We can factor p_k as $p_k = qr$ with $S(q) = Q$.*

3. *We can factor p_k as $p_k = qr$ with $q = \sum_{\alpha:S(\alpha)=Q}(-1)^{s(\alpha)}\alpha$.*

   The function $s(\alpha) \in \{0,1\}$ and depends on neither k nor Q.

Proof. We prove that $1 \Rightarrow 3 \Rightarrow 2 \Rightarrow 1$. That $3 \Rightarrow 2$ follows trivially from the definition of $S(q)$. To prove that $2 \Rightarrow 1$, we recall that every k-decomposition corresponds to a term in p_k. Since we assumed that $p_k = qr$ with $S(q) = Q$, all terms of p_k can be obtained by multiplying a term in q with corresponding signature Q with a term in r; this proves the statement.

We now show that $1 \Rightarrow 3$. Consider a term γ_1 of p_k. By assumption, we have $\gamma_1 = \alpha_1\delta_1$ for some edge-products δ_1 and α_1 where $S(\alpha_1) = Q$. We claim that if α_2 is another edge-product of G with signature Q, then the product $\gamma_2 := \alpha_2\delta_1$ is again a term of p_k. To see this, note that $S(\gamma_2) = S(\alpha_2) + S(\delta_1) = S(\gamma_1)$ and γ_2 is a term of p_k by Lemma 6.2. We can thus write

$$p_k = \sum_i \sum_{\alpha_j:S(\alpha_j)=Q} (-1)^{s(\delta_i,\alpha_j)}\delta_i\alpha_j \tag{6.5}$$

where $s(\delta_i, \alpha_j) \in \{0,1\}$ are functions indicating the sign with which the term $\delta_i\alpha_j$ appears in p_k.

Finally, we show that there exists functions $s_1(\delta)$, $s_2(\alpha)$, both with values in $\{0,1\}$, such that the following holds:

$$(-1)^{s(\delta_i,\alpha_j)} = (-1)^{s_1(\delta_i)+s_2(\alpha_j)}. \tag{6.6}$$

Note that it is sufficient to show that $(-1)^{s(\delta_i,\alpha_1)-s(\delta_i,\alpha_2)}$ only depends on α_1 and α_2; i.e. whether the signs of $\delta_i\alpha_1$ and $\delta_i\alpha_2$ in p_k are the same only depends on α_1 and α_2. In order to see this, recall that a term in p_k defines a permutation on k nodes of G—the nodes incident to the edges in the term. This permutation can be naturally extended to a permutation on $\{1, 2, ..., n\}$ by mapping the non-incident nodes to themselves via the identity. Moreover, the sign of a term in p_k is given by the sign of the permutation it defines, see Eq. (3.3). Therefore, two terms in the expansion of p_k, thinking of these terms as permutations, have the same sign if and only if they are related by

an even permutation. Since whether the parities of the permutations $\alpha_1\delta_i$ and $\alpha_2\delta_i$ are the same or not can be deduced by only examining α_1 and α_2, Eq. (6.6) is proven.

Putting Eqs. (6.5) and (6.6) together, we obtain that p_k can be factorized as

$$p_k = \left[\sum_{\alpha_j : S(\alpha_j)=Q} (-1)^{s_2(\alpha_j)}\alpha_j\right]\left[\sum_i (-1)^{s_1(\delta_i)}\delta_i\right],$$

which concludes the proof. ■

Corollary 6.1. *Let G be a digraph with determinant p_n. Let q be a factor of p_n with signature $S(q)$. If every $n-1$-decomposition contains a subgraph with signature equal to $S(q)$, then q is also a factor of p_{n-1}.*

Proof. Notice that since q is a factor of p_n, every n-decomposition of G contains a subgraph with corresponding signature equal to Q. Applying Prop. 6.2, we can write that

$$p_{n-1} = \left(\sum_{\alpha : S(\alpha)=Q} (-1)^{s(\alpha)}\alpha\right) r_1$$

and $p_n = \left(\sum_{\alpha:S(\alpha)=Q}(-1)^{s(\alpha)}\right) r_2$ for some edge-products r_1 and r_2. Thus $q = \pm\sum_{\alpha:S(\alpha)=Q}(-1)^{s(\alpha)}$ is a factor of p_{n-1} as well. ■

The previous Proposition and its Corollary show that signatures can be used to determine whether p_{n-1} and p_n have factors in common. The following result shows that by looking at a single n-decomposition of G, we have access to all the signatures of p_n. Recall that $\mathcal{S}(p_n)$ is the set of all signatures corresponding to factors of the determinant p_n of a graph G.

Proposition 6.3. *Let G be a digraph on n nodes which has at least one n-decomposition. Let E_h be the set of edges of an arbitrary n-decomposition of G. Let $a_{ij} \in E_h$ and $S(a_{ij})$ be its corresponding signature. Then there exists a unique signature $\bar{S}(a_{ij}) \in \mathcal{S}(p_n)$ satisfying $\bar{S}(a_{ij}) \succeq S(a_{ij})$. Moreover, the map*

$$\bar{S} : E_h \to \mathcal{S}(p_n) : a_{ij} \to \bar{S}(a_{ij})$$

is surjective.

Proof. First, recall that to each irreducible factor of p_n corresponds a signature. In fact, these irreducible factors are the sum of edge-products that each have the same signature by Lemma 6.3. Second, recall that p_n can be seen as a signed sum of n-decompositions. Now if we are given an arbitrary n-decomposition of G, with edge set E_h, putting the above two points together, we conclude that every edge-variable in E_h can be assigned to a unique signature, the one of the factor in which it appears. We now show that $\bar{S}$ is surjective. Choose an arbitrary signature $Q \in \mathcal{S}(p_n)$ and denote by q its corresponding factor. We claim that there is an edge-variable $a_{ij} \in E_h$ such that $\bar{S}(a_{ij}) = Q$. To see this, note that all n-decompositions of G are obtained by expanding the factorization of p_n. Therefore, any n-decomposition, incluing E_h, contains at least one edge-variable of q. By the first part of the Proposition, $\bar{S}(a_{ij}) = Q$, which shows that the map is surjective. ■

By the above proposition, we can assign to each edge of E_h a unique signature (and hence factor of p_n) and, moreover, every signature of p_n will appear in this assignment. Thus we can obtain from the edges of an *arbitrary* n-decomposition of G all signatures of p_n. This fact is very important in this context, as it is well-known that enumerating, let alone exhibiting, all possible n-decompositions of a graph is a hard problem, related to permanent computations [20].

6.3 Polynomial time algorithm for the signatures of the determinant of G

The main ingredient of our method to check whether a 2-extension is stable is the following algorithm, which computes in polynomial time *the signatures of all the factors of* p_n. As mentioned earlier, factorizing p_n is in general hard, and we cannot deduce a factor from its signature. But since we only care about *common factors* of p_n and p_{n-1}, regardless of the actual value of these factors, knowing the signature of factors is sufficient for our purpose. We now state the algorithm:

Algorithm 2: finding signatures of p_n

Input: a digraph G on n nodes with at least one n-decomposition.

Output: the signatures of p_n.

1. Pick an arbitrary n-decomposition using Proposition 6.1. Denote by E_h its set of edges. Proceed to step 2.

2. If E_h is empty: terminate the algorithm—all signatures of G have been found.
   Otherwise: pick an edge $(i,j) \in E_h$. Set $Q := (Q_1, Q_2, ..., Q_n) = S(a_{ij})$ as defined in (6.3). Proceed to step 3.

3. For every edge (k,l) in the graph such that

$$Q_k = (\cdot, 0) \text{ and } Q_l = (1, \cdot) \text{ or } Q_k = (\cdot, 1) \text{ and } Q_l = (0, \cdot),$$

   where the $\cdot$ denotes an arbitrary value, check if there exists a n-decomposition containing that edge using Proposition 6.1.

   (a) If such decomposition exists, then Q is not a signature and we update it as follows:

   - $Q_k \leftarrow Q_k + (1,0)$ if $Q_k = (\cdot, 0), Q_l = (1, \cdot)$.
   - $Q_l \leftarrow Q_l + (0,1)$ if $Q_k = (\cdot, 1), Q_l = (0, \cdot)$.

   Repeat step 3.

   (b) If such decomposition does not exist, then Q is a signature. Proceed to step 4.

4. Find the set α of edges in E_h for which $S(\alpha) = Q$. The set α can be found by running through all edges in E_h. Set $E_h = E_h - \alpha$ and go to step 2.

Proposition 6.4. *Algorithm 2 computes all the signatures of p_n in polynomial time.*

Proof. We first address the complexity of the algorithm. The first step is of polynomial complexity as shown in Proposition 6.1. Next, a single run of step 3 requires going through edges of the graph (at most n^2) and checking whether any of them satisfies the conditions listed. Since every iteration increases the number of non-zero entries of Q by 1, and this can be done at most $2n$ times for G, the number of such iterations is in $O(n)$. Therefore,

the complexity of finding a signature Q of p_n using step 3 is $O(n^3)$. The next step, namely finding all edges corresponding to Q as in step 4 has complexity $O(n^2)$. The number of signatures of p_n is at most n^2, so the algorithm runs step 2 at most this many times. Collecting all these observations, we see that the algorithm terminates after $O(n^5)$ time steps.

We now show that Algorithm 2 indeed produces all signatures of p_n, that is all elements of the set $\mathcal{S}(p_n)$. Let E_h be the edge set of an arbitrary n-decomposition of G and $a_{ij} \in E_h$. Let $\bar{S}(a_{ij})$ be the signature corresponding to a_{ij} as in Prop. 6.3. Starting with $S(a_{ij})$, the algorithm uses a greedy approach to build a sequence of increasing signatures, the variable Q in the algorithm, which upon completion of step 3, will hold the value of $\bar{S}(a_{ij})$, as will be shown below.

We denote by $Q(m)$ the value of Q at the mth update of this variable in step 3. The variable $Q(0)$ is initialized at $S(a_{ij})$. We first observe that according to the update rule of step 3, $Q(m+1) \succeq Q(m)$.

Next, we show that if $Q(m) \preceq \bar{S}(a_{ij})$ and the algorithm does not proceed to step 4, then $Q(m+1) \preceq \bar{S}(a_{ij})$. To see this, denote by q the irreducible factor of p_n in which a_{ij} appears and let (k,l) be the edge selected in the $m+1$ iteration of step 3. If a_{kl} and a_{ij} appear in different irreducible factors of p_n, then there is an edge-product $\alpha = qa_{kl}r$ in p_n. From Eq. (6.4), we obtain $S(\alpha) = S(q) + S(a_{kl}) + S(r)$ and, because $S(q) = \bar{S}(a_{ij})$, it follows that $S(\alpha)$ either has kth coordinate equal to $(\cdot, 2)$ or lth coordinate equal to $(2, \cdot)$. This implies that α can not correspond to a n-decomposition and we get a contradiction. Therefore the variable a_{kl} has to appear in the same term as a_{ij}. We conclude that the signature $Q(m+1) = Q(m) + S(a_{kl})$ satisfies the inequalities $S(a_{ij}) \prec Q(m) \prec Q(m+1) \preceq \bar{S}(a_{ij})$. Since the sequence $Q(m)$ increases with every iteration of step 3, the argument above also show that if $Q(m)$ becomes equal to $\bar{S}(a_{ij})$, then the algorithm proceeds to step 4.

Finally, we prove that if $Q(m)$ is such that the condition of step 3b holds, then $Q(m) = \bar{S}(a_{ij})$. To see this, assume that there are no edges satisfying either

$$Q_k(m) = (\cdot, 0) \text{ and } Q_l(m) = (1, \cdot) \text{ or } Q_k(m) = (\cdot, 1) \text{ and } Q_l(m) = (0, \cdot)$$

and which are contained in a n-decomposition. Note that this statement is

equivalent to saying that every n-decomposition contains a set of edges with a corresponding signature equal to $Q(m)$. Applying Proposition 6.2, we see that $p_n = qr$, where $q = \sum_{\alpha:S(\alpha)=Q(m)} \pm\alpha$. Since q is a factor of p_n and contains a_{ij}, its corresponding signature $Q(m)$ satisfies $Q(m) \succeq \bar{S}(a_{ij})$. By virtue of the inequality $Q(m) \prec \bar{S}(a_{ij})$ established in the previous paragraph, this shows that $Q(m+1) = \bar{S}(a_{ij})$.

Once we have $\bar{S}(a_{ij})$, we can find the subset of edges α in E_h, which contains a_{ij} and has signature $\bar{S}(a_{ij})$ — these are the edges a_{kl} for which $\bar{S}(a_{ij})_{l1} = 1$ and $\bar{S}(a_{ij})_{k2} = 1$. We then update $E_h \leftarrow E_h - \alpha$ and iterate. By Prop. 6.3, we obtain all signatures of p_n. ■

6.3.1 Polynomial time algorithm for checking common factors of p_{n+2} and p_{n+1}

Now we consider a 2-node extension G' of the graph G which contains both $n+2$ and $n+1$-decompositions. The last step of the proof of Theorem 6.2 is the following algorithm, which verifies whether G' passes the edge-distribution test.

Algorithm 3: checking common factors of p_{n+2} and p_{n+1}

Input: a digraph G on n nodes; a 2-extension G' of G with at least one $n+2$-decomposition and one $n+1$-decomposition; the set of signatures of G'.

Output: True if G' passes the edge-distribution test, False otherwise.

1. Denote by E_{ext} the set of edges in G' appearing in at least one $n+2$-decomposition and incident with node v_{n+1} or v_{n+2}.

2. If E_{ext} is empty: terminate the algorithm and return False. Otherwise: pick an edge $(i,j) \in E_{ext}$, find the signature $\bar{S}(a_{ij})$ in G' and denote by q the factor in p_{n+2} corresponding to it.

3. For every edge kl in the graph such that

   $$\bar{S}(a_{ij})_k = (\cdot, 0) \text{ and } \bar{S}(a_{ij})_l = (1, \cdot) \text{ or } \bar{S}(a_{ij})_k = (\cdot, 1) \text{ and } \bar{S}(a_{ij})_l = (0, \cdot),$$

   where the $\cdot$ denotes an arbitrary value, check if there exists an $n+1$-decomposition containing that edge using Theorem 6.1.

(a) If such decompositions exist for all considered edges (k, l), then q is not a factor of p_{n+1}. Terminate the algorithm and return True.

(b) If such decomposition does not exist for at least one considered edge (k, l), proceed to step 4.

4. Find the set α of edges (k, l) in E_{ext} for which $S(kl) \preceq \bar{S}(a_{ij})$. The set α can be found by running through all edges in E_{ext}. Set $E_{ext} = E_{ext} - \alpha$ and go to step 2.

Proposition 6.5. *Algorithm* 3 *checks whether a* 2*-extension* G' *of a digraph* G *passes the edge-distribution test in polynomial time.*

Proof. The proof is similar to the one of Proposition 6.4, so we just provide a sketch. First, it follows from the fact that the number of edges in G' is $O(n^2)$ and from Proposition 6.1 that the complexity of Algorithm 3 is polynomial. Second, we prove correctness of the algorithm. Assume that the algorithm returns False for a given extension G' of G. We claim that every factor of p_{n+2} containing an edge of E' is a factor of p_{n+1}. To see this, let q be a factor of p_{n+2} containing an edge-variable that is in E_{ext}.

Let a_{ij} be an edge-variable appearing in q as selected in step 2. Because the algorithm returned False, by step 3 all $n+1$-decompositions contain subgraphs with corresponding signatures equal to $S(q)$. Using Proposition 6.2 we see $p_{n+1} = qr$ for some r.

Reciprocally, assume that the extension does not pass the edge distribution test. This means that every factor of p_{n+2} containing an edge-variable corresponding to an element of E_{ext} also appears in the factorization of p_{n+1}. Choose an arbitrary edge $(i, j) \in E_{ext}$ and let $p_{n+2} = qr$, with a_{ij} appearing in q. We apply again Proposition 6.2 with $Q = \bar{S}(a_{ij}) = S(q)$ and conclude that every $n+1$-decomposition contains a subgraph with corresponding signature Q. This implies that for every iteration of the algorithm through step 3, the outcome is (b). Thus the number of edges in E_{ext} decreases monotonically until it reaches zero and the algorithm then returns False. ■

6.4 Proofs of Theorem 6.1 and Theorem 6.2

We now have all the ingredients necessary to prove Theorems 6.1 and 6.2.

Proof of Th. 6.1. Knowing that G is Hurwitz, Proposition 4.2 states that in order to determine whether G' is Hurwitz or not, we have to check only whether the latter contains a Hamiltonian decomposition. Proposition 6.1 offers a polynomial time algorithm to do this, which completes the proof. ■

Proof of Th. 6.2. First, using Theorem 6.1, we can check in polynomial time whether G' has $n+2$ and $n+1$ decompositions. If it is the case, then both p_{n+1} and p_{n+2} are non-zero polynomials. Applying Algorithm 2 we can find all signatures of p_{n+2}. Then using Algorithm 3 we can check whether the extension passes the edge-distribution test. ■

PART II

ACTUATOR DESIGN

CHAPTER 7

INTRODUCTION

In recent years there has been a resurgence of interest in the design of optimal actuators and sensors for linear systems. Driven by the rise of distributed control systems as models for large scale social or biological networks, or novel manufacturing and sensing methods that allow for more flexibility in the design/choice of actuators/sensors, there is an increased need for a better understanding of the way in which the performance of a system depends on the placement of its actuators.

Amongst the relevant recent work in the area, we mention [21], where it is shown that the optimal actuator/sensor design problem admits an essentially (up to symmetries) unique optimum when the magnitude of the actuator is small to moderate, and a provably convergent algorithm to find the optimal actuator is proposed. For related work when dealing only with control energy [22] (in contrast to linear-quadratic cost), we refer to [23]. In this case, the set of allowed actuators is a continuous set, corresponding physically to the placement of an actuator or sensor in the system (e.g., the placement of a camera in physical space). In this regard, we also mention [24, 25, 26]. Other types of problems require choosing a set of sensors c_i out of a *finite* family of available sensors. These have been investigated in various forms by several authors. In [27], the authors assign a cost to each sensor and show that optimally choosing a subset of sensors meeting cost constraints is an NP-hard problem, and furthermore exhibit a class of dynamics for which greedy algorithms yield a provably good approximation to the optimal selection. In [28], the authors look at a "relaxed" selection problem, where sensors are selected with a weight w_i to be optimized and propose a convex optimization algorithm. A different type of methods, based on L_1 optimization as a proxy for sensor selection has been investigated in [29]. Similar scenarios have also been investigated in the statistics literature in the field of experiment design, see [30] for a start to the relevant literature. Methods based

on greedy selection are also popular in the area, see [31] for an evaluation of the performance of such methods, and for an algorithmic approach to sensor/actuator selection in the structural setting (i.e., to guarantee structural controllability/observability).

In this part of the thesis we address two separate problems of optimal actuator design. Their solutions are obtained using the techniques employed in [21], which involve the minimization of a certain non-convex function on a compact manifold.

In Chapter 8 we examine the problem of designing a (dynamic) actuator/controller that minimizes a combination of linear-quadratic cost and the variations in the applied controls, as well.

Specifically, consider the linear time-invariant system

$$\dot{x} = Ax + Bu, \quad x(0) = x_0, \tag{7.1}$$

with the quadratic cost function

$$C = \int_0^\infty (x(t)^\top Q x(t) + u(t)^\top u(t))\,\mathrm{d}t. \tag{7.2}$$

It is well-known that the optimal control (i.e. the one minimizing the above cost) u_{opt} can be expressed in a feedback form [32]. The optimal actuator design problem is to design the system's actuator B to minimize the optimal cost.

In many practical situations, however, the physical actuators driven by $u(t)$ cannot vary their effort very fast (a DC motor, for example, may not be able to change its rotations per minute very fast due to physical constraints), or a sensor that cannot update its reading very quickly. To address this concern, we add the assumption that u is continuously differentiable and introduce the extra term $\dot{u}(t)^\top \dot{u}(t)$ in the integrand of the cost function (7.2). We allow the use of dynamic controls, i.e. $\dot{u}(t)$ is a function of $u(t)$ and $x(t)$.

We consider actuators B which satisfy $B^\top B = \gamma^2 I$, where $\gamma \neq 0$ and I is the identity matrix, and seek to find the ones that minimize the cost

$$\min_u \int_0^\infty (x^\top Q x + c u^\top u + \dot{u}^\top \dot{u})\mathrm{d}t, \tag{7.3}$$

which depends implicitly on B. Note that bounding the norm of B is neces-

sary as otherwise one can exchange control effort (measured as the magnitude of u) for actuator gain (measured as the norm of B), and artificially decrease the value of the cost. We show that generically for A, Q, and x_0, when B is small, there exists an essentially (up to symmetries) *unique* actuator which is locally optimal, and it is also globally optimal. Furthermore, this actuator can be found using a gradient algorithm over a suitable manifold. We illustrate the performance of the design in the last section of Chapter 8.

In Chapter 9, we focus on linear stochastic systems. Such models are widely used in engineering, biology and physics, due to the breadth of the situations they can describe [33, 34]. The most commonly used among them is the linear dynamics with additive Gaussian noise model, which can be described by the stochastic differential equation (SDE)

$$dx_t = Ax_t dt + bu dt + G dw_t,$$

where $x \in \mathbb{R}^n$, $A \in \mathbb{R}^{n\times n}$, $b \in \mathbb{R}^{n\times m}$ and $G \in \mathbb{R}^{n\times p}$ and w_t is a standard vector-valued Wiener process [35]. It is well known that the control minimizing the expected cost

$$\lim_{T\to\infty} \mathbb{E}\left(\frac{1}{T}\int_0^T (x^\top Q x + u^\top u) dt\right)$$

is of feedback type, and its explicit form is known. Now it is clear that the value of the optimal (with respect to u) cost will be dependent on the actuator b. The problem of finding the b that minimizes this cost is called the *optimal actuator placement*. This problem is in general difficult, and easily seen to be non-convex. We provide in this paper a solution to the related problem of optimal actuator placement problem for dynamics corrupted with *multiplicative noise*

$$dx_t = Ax_t\, dt + bu\, dt + G_1 x\, dw_t, \tag{7.4}$$

where $G_1 \in \mathbb{R}^{n\times n}$ and w_t is a Wiener process.

CHAPTER 8

ON THE OPTIMAL DESIGN OF LOW FREQUENCY ACTUATORS

8.1 Preliminaries

We now state the problem we are addressing precisely. Consider the control system

$$\dot{x} = Ax + \gamma Bu, \quad x(0) = x_0, \quad u(0) = 0, \tag{8.1}$$

where $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$ is such that $B^\top B = I_m$, and $\gamma > 0$. We consider also the associated cost functional

$$V(x_0) = \int_0^\infty (x^\top Q x + c u^\top u + \dot{u}^\top \dot{u}) \mathrm{d}t,$$

where $Q \in \mathbb{R}^{n\times n}$ is a symmetric positive definite (spd) matrix and c is a positive constant. We denote by

$$\mathcal{B} = \{B \in \mathbb{R}^{n\times m} \mid B^\top B = I_m\},$$

where I_m is the identity $(m \times m)$-matrix, be the Stiefel manifold of orthonormal m-frames in $\mathbb{R}^n$. Roughly speaking, our goal is to find the actuator $B \in \mathcal{B}$ which minimizes the cost functional V, for either a specific x_0 or in average (over x_0). To make this more precise, we first rewrite the problem by introducing the extended variable

$$\bar{x} = \begin{pmatrix} x \\ u \end{pmatrix},$$

for which we can write

$$\dot{\bar{x}} = M\bar{x} + \bar{B}v, \quad M = \begin{pmatrix} A & \gamma B \\ 0 & 0 \end{pmatrix}, \quad \bar{B} = \begin{pmatrix} 0 \\ I \end{pmatrix}, \tag{8.2}$$

$$\bar{V}(\bar{x}_0) = \int_0^\infty (\bar{x}^\top T\bar{x} + v^\top v)\mathrm{d}t, \quad T = \begin{pmatrix} Q & 0 \\ 0 & cI \end{pmatrix}, \tag{8.3}$$

where $v = \dot{u}$ is a continuous control. The optimal control v_{opt} which minimizes the cost $\bar{V}$ in the LQR problem (8.2), (8.3) is given by

$$v_{\text{opt}} = -\bar{B}^\top P\bar{x}_{\text{opt}}, \tag{8.4}$$

where P is the unique positive definite solution of the Riccati equation

$$M^\top P + PM - PEP + T = 0, \tag{8.5}$$

with $E = \bar{B}\bar{B}^\top$ (see Theorem 8.1 below), and $\bar{x}_{\text{opt}}$ is the solution to the system

$$\dot{\bar{x}}_{\text{opt}} = (M - EP)\bar{x}_{\text{opt}}, \quad \bar{x}_{\text{opt}}(0) = \begin{pmatrix} x_0 \\ 0 \end{pmatrix}. \tag{8.6}$$

The corresponding minimum cost is equal to

$$V_{\min}(x_0) = \mathrm{tr}(P\bar{L}),$$

where

$$\bar{L} = \bar{x}_0\bar{x}_0^\top = \begin{pmatrix} x_0x_0^\top & 0 \\ 0 & 0 \end{pmatrix} =: \begin{pmatrix} L & 0 \\ 0 & 0 \end{pmatrix} \tag{8.7}$$

is a positive semi-definite matrix.

If we write the matrix P in the form

$$P = \begin{pmatrix} P_1 & P_2 \\ P_2^\top & P_3 \end{pmatrix},$$

we can represent the system using Figure 8.1.

Our goal is the following: solve the minimization problem

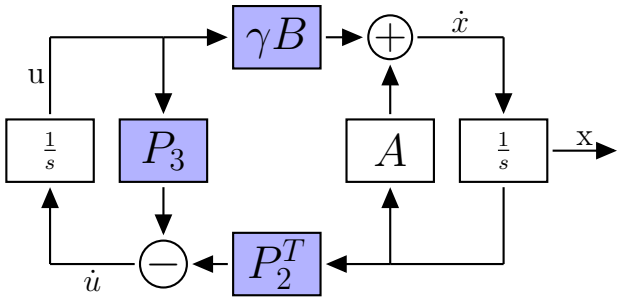

Figure 8.1: In this problem, the matrix A is fixed, and we design the blocks B, P_3 and $P_2^\top$ so as to minimize the cost (8.3).

$$B^* = \arg\min_{B \in \mathcal{B}} \operatorname{tr}(\bar{L}P),$$

where $\bar{L}$ and P are as defined above. It is well-known that if x_0 is sampled from an isotropic distribution centered at zero, then $\mathbb{E}_{x_0} V(x_0) \propto \operatorname{tr} P$ [21]. Hence taking $L = I_n$ provides the minimum in the average sense discussed above. In words, we find the actuator that minimizes the control effort V when paired with an optimal control.

For fixed A, Q, c, x_0, the value of the optimal cost as a function of B will be denoted by $F_\gamma(B)$, i.e.,

$$F_\gamma(B) := \operatorname{tr}(\bar{L}P).$$

It is useful to work in the space of matrices $BB^\top$, as B enters in the definition of P in this form through the Riccati equation (8.5). Hence, we introduce the set

$$\mathcal{G} = \{G \in \mathbb{R}^{n \times n} \mid G^2 = G = G^\top, \operatorname{rk} G = m\},$$

and the map

$$H : \mathcal{B} \to \mathcal{G} : B \mapsto BB^\top.$$

Each $G \in \mathcal{G}$ is a positive semi-definite matrix, representing the orthogonal projection operator onto the m-dimensional subspace of $\mathbb{R}^n$ spanned by the columns of G. The set $\mathcal{G}$ is an analytic submanifold of $\mathbb{R}^{n \times n}$ ([36], p. 275). Moreover, the analytic map H is surjective and its level sets are precisely the orbits of the action of the orthogonal group $\mathcal{O}(m) = \{\Theta \in \mathbb{R}^{m \times m} \mid \Theta^\top \Theta = I_m\}$ in $\mathcal{B}$ given by $(B, \Theta) \mapsto B\Theta$ for $B \in \mathcal{B}$, $\Theta \in \mathcal{O}(m)$. Thus, H

induces an analytic diffeomorphism of the Grassmann manifold $\mathcal{B}/\mathcal{O}(m)$ of all m-dimensional subspaces of $\mathbb{R}^n$ onto $\mathcal{G}$.

In Section 8.2, it is shown that there exists a unique analytic function $J : \mathcal{G} \times \mathbb{R} \to \mathbb{R}$, such that $J(H(B), \gamma) = F_\gamma(B)$ for all $B \in \mathcal{B}$. For fixed $\gamma \in \mathbb{R}$, we also use the notation $J_\gamma : \mathcal{G} \to \mathbb{R} : G \mapsto J(G, \gamma)$.

The main result of this chapter is the following theorem.

Theorem 8.1. *For A Hurwitz, and generically for Q positive definite, for $\gamma > 0$ small enough, the function $J_\gamma : \mathcal{G} \to \mathbb{R}$ has $\binom{n}{m}$ critical points, exactly one of which is a local minimum.*

As a consequence of the Theorem, we have

Corollary 8.1. *Under the assumptions of Theorem 8.1, the gradient flow of $F_\gamma(B)$ converges from a generic initial condition $B_0 \in \mathcal{B}$ to an optimal actuator $B^* \in \mathcal{B}$.*

We provide the gradient flow and illustrate its performance in simulations.

8.2 Gradient of the function F_γ

We are to optimize the function F_γ (or, equivalently, J_γ) over the manifolds $\mathcal{B}$ (resp. $\mathcal{G}$) using a gradient flow. To this end, we need to characterize the tangent spaces of said manifolds and introduce inner products. We do so in the next few paragraphs.

Tangent space and inner product Recall that $\mathfrak{so}(n) = \{\Omega \in \mathbb{R}^{n\times n} \mid \Omega = -\Omega^\top\}$ is the vector space of skew-symmetric matrices. We introduce the following operators:

Definition 8.1. *The linear operators $\rho_B : \mathfrak{so}(n) \to \mathbb{R}^{n\times m}$ and $\mathrm{ad}_G : \mathfrak{so}(n) \to \mathbb{R}^{n\times n}$ are defined for every $B \in \mathcal{B}$ and $G \in \mathcal{G}$ as:*

$$\rho_B(\Omega) = \Omega B$$
$$\mathrm{ad}_G(\Omega) = [G, \Omega],$$

where $[G, \Omega] := G\Omega - \Omega G$. The operator $\Pi_{\mathfrak{so}(n)}(V) = \frac{1}{2}(V - V^\top)$ denotes the projection of the matrix $V \in \mathbb{R}^{n\times n}$ onto $\mathfrak{so}(n)$.

It is well-known that the tangent spaces of $\mathcal{B}$ and $\mathcal{G}$ at B and G respectively are given by [37]

$$\begin{aligned}\mathrm{T}_B\mathcal{B} &= \{\rho_B(\Omega) \mid \Omega \in \mathfrak{so}(n)\},\\ \mathrm{T}_G\mathcal{G} &= \{\mathrm{ad}_G(\Omega) \mid \Omega \in \mathfrak{so}(n)\}.\end{aligned}$$

In order to compute the gradient of F_γ over $\mathcal{B}$, we will use the inner product $\tau(\cdot,\cdot) : T_B\mathcal{B} \times T_B\mathcal{B} \to \mathbb{R}$ which is defined as follows. Let

$$\bar{\rho} : \ker\,(\rho_B)^\perp \to \mathrm{T}_B\mathcal{B}$$

where $\ker\,(\rho_B)^\perp$ is the orthogonal complement of $\ker\,(\rho_B)$ inside $\mathfrak{so}(n)$ taken with respect to the Frobenius inner product $(\Omega_1, \Omega_2) := -\operatorname{tr}(\Omega_1\Omega_2)$. Now for $\dot{B}_i := \rho_B(\Omega_i)$, $i = 1, 2$, tangent vectors, the inner product $\tau(\dot{B}_1, \dot{B}_2)$ is defined as:

$$\tau(\dot{B}_1, \dot{B}_2) = -\operatorname{tr}(\bar{\rho}_B^{-1}(\dot{B}_1)\bar{\rho}_B^{-1}(\dot{B}_2)).$$

On the manifold $\mathcal{G}$, we will use the metric $\kappa(\cdot,\cdot)$ which is defined similarly. Let

$$\bar{ad}_G : \ker\,(\mathrm{ad}_G)^\perp \to \mathrm{T}_G\mathcal{G},$$

where $\ker\,(\mathrm{ad}_G)^\perp$ is the orthogonal complement of $\ker\,(\mathrm{ad}_G)$ inside $\mathfrak{so}(n)$ taken with respect to the Frobenius inner product. The inner product κ is defined for $\dot{G}_i := \mathrm{ad}_G\,\Omega_i$ as:

$$\kappa(\dot{G}_1, \dot{G}_2) = -\operatorname{tr}(\bar{\mathrm{ad}}_G^{-1}(\dot{G}_1)\bar{\mathrm{ad}}_G^{-1}(\dot{G}_2)).$$

The inner products τ and κ have often been used in optimization on manifold problems [38, 39].

We will need the following definition:

Definition 8.2 (First variation)**.** *Let $P : \mathcal{B} \to \mathbb{R}^{n\times n}$ be differentiable and $\dot{B} \in T_B\mathcal{B}$. We call the first variation of P at B the map*

$$\dot{P}_B : \mathrm{T}_B\mathcal{B} \to \mathbb{R}^{n\times n} : \dot{B} \to \lim_{\varepsilon\to 0} \frac{P(B + \varepsilon\dot{B}) - P(B)}{\varepsilon}.$$

Note that the first variation of a *real-valued* function is exactly the differential of this function. Finally, we recall that the gradient $\operatorname{grad} F$ of a function F on $\mathcal{B}$ with inner product $\tau(\cdot,\cdot)$ is defined as the unique solution of

$$\tau(\operatorname{grad} F(B), \dot{B}) = \dot{F}_B(\dot{B}), \text{ for all } \dot{B} \in T_B\mathcal{B},$$

where $\dot{F}$ is the first variation of F.

On the generalized Riccati equation The following result, which says that the cost function F depends nicely on B, is needed to obtain the gradient of F.

Proposition 8.1. *The Riccati equation* (8.5) *has a unique positive definite solution P for every $B \in \mathcal{B}$, $\gamma \in \mathbb{R}$ and A Hurwitz. Furthermore, the solution $P(B,\gamma) : \mathcal{B} \times \mathbb{R} \to \mathbb{R}^{n\times n}$ is analytic, and, for fixed $\gamma \in \mathbb{R}$ its first variation $\dot{P}$ at B is given by the convergent integral*

$$\dot{P}_B(\dot{B}) = \int_0^\infty e^{(M-EP)^\top t}(N^\top P + PN)e^{(M-EP)}\mathrm{d}t, \tag{8.8}$$

where

$$N = \begin{pmatrix} 0 & \gamma\dot{B} \\ 0 & 0 \end{pmatrix}.$$

Proof. The statement is standard, we thus only sketch the proof. Since A is Hurwitz, the uncontrollable modes of the pair (M,E) are stable and thus (M,E) is stabilizable for all $B \in \mathcal{B}$, $\gamma \in \mathbb{R}$. Since Q is positive definite and $c > 0$, the pair (M,T) is similarly shown to be detectable. Hence, (8.5) has a unique positive definite solution P and moreover the matrix $M - EP$ is Hurwitz ([32], Theorem 3.7). Furthermore, this solution depends analytically on $(B,\gamma) \in \mathcal{B} \times \mathbb{R}$ ([40], Lemma 1.1).

For the second part, we compute the first variation of Eq. (8.5) in an arbitrary direction $\dot{B}$ and get (we omit writing $\dot{B}$ as an argument of $\dot{P}, \dot{M}$ for clarity)

$$\begin{aligned} \dot{M}^\top P + \dot{P}M + P\dot{M} - \dot{P}EP - PE\dot{P} = 0, \\ \Leftrightarrow (M-EP)^\top\dot{P} + \dot{P}(M-EP) + N^\top P + PN = 0. \end{aligned}$$

This is a Lyapunov equation with unknown $\dot{P}$, so its solution can be written in the form (8.8). ■

Using the above Proposition, we can evaluate the differential of F at B as follows:

Proposition 8.2. *The differential of F_γ acts on a tangent vector $\dot{B} = \Omega B$ in $\mathrm{T}_B\mathcal{B}$ according to*

$$(\dot{F}_\gamma)_B(\dot{B}) = 2\,\mathrm{tr}(KPN),$$

where the matrix K is the unique positive semi-definite solution of the Lyapunov equation

$$(M - EP)K + K(M - EP)^\top + \bar{L} = 0. \tag{8.9}$$

Proof. We compute the directional derivative of $F_\gamma(B)$ for fixed γ with respect to $\dot{B} = \Omega B$. Multiplying both sides of (8.8) by $\bar{L}$ and applying the trace operator on both sides, we get

$$\dot{F}_\gamma = \mathrm{tr}\left(\int_0^\infty e^{(M-EP)^\top t}(N^\top P + PN)e^{(M-EP)t}\bar{L}\mathrm{d}t\right).$$

Using the relations $\mathrm{tr}(X) = \mathrm{tr}(X^\top)$ and $\mathrm{tr}(YZ) = \mathrm{tr}(ZY)$, we get

$$\dot{F}_\gamma = 2\,\mathrm{tr}(KPN),$$

where

$$K = \int_0^\infty e^{(M-EP)t}\bar{L}e^{(M-EP)^\top t}\mathrm{d}t.$$

■

The next two results show that we can set-up the optimization problem on the manifold $\mathcal{G}$ as well, the advantage of that formulation being that the critical points of the corresponding function on $\mathcal{G}$ (defined below) are isolated—in contrast, since $F_\gamma(B) = F_\gamma(\Theta B)$ for any matrix Θ such that $\Theta\Theta^\top = I$, the critical points of F are not isolated.

Lemma 8.1. *There exists a unique analytic function $J : \mathcal{G} \times \mathbb{R} \to \mathbb{R}$, such that $J(BB^\top, \gamma) = F_\gamma(B)$ for every $B \in \mathcal{B}$, $\gamma \in \mathbb{R}$.*

Proof. We construct the function J explicitly. Suppose that $\Theta \in \mathcal{O}(m)$, and set $\bar{\Theta} = \left(\begin{smallmatrix} I_n & 0 \\ 0 & \Theta \end{smallmatrix}\right) \in \mathcal{O}(n+m)$, where I_n the $n \times n$ identity matrix. The substitution $B \rightsquigarrow B\Theta$ transforms (8.5) into the equation

$$(\bar{\Theta}^\top M \bar{\Theta})^\top P + P(\bar{\Theta}^\top M \bar{\Theta}) - PEP + T = 0. \tag{8.10}$$

Since $\bar{\Theta}^\top E \bar{\Theta} = E$ and $\bar{\Theta}^\top T \bar{\Theta} = T$, it follows that the solutions $P(B, \gamma)$ of (8.5) and $P(B\Theta, \gamma)$ of (8.10) satisfy the relation

$$P(B\Theta, \gamma) = \bar{\Theta}^\top P(B, \gamma) \bar{\Theta}.$$

Note that we have $\bar{\Theta}^\top \bar{L} \bar{\Theta} = \bar{L}$. Hence, $\operatorname{tr}(P(B\Theta, \gamma)\bar{L}) = \operatorname{tr}(P(B, \gamma)\bar{L})$, and thus, the function $F : \mathcal{B} \times \mathbb{R} \to \mathbb{R}$, defined as $F(B, \gamma) = F_\gamma(B) = \operatorname{tr}(P(B, \gamma)\bar{L})$ is invariant under the right action of the group $\mathcal{O}(m)$ on $\mathcal{B} \times \mathbb{R}$ given by $(B, \gamma) \cdot \Theta = (B\Theta, \gamma)$ for $B \in \mathcal{B}$, $\gamma \in \mathbb{R}$, $\Theta \in \mathcal{O}(m)$. Since F is constant on the level sets of the surjective map $\mathcal{G} : \mathcal{B} \times \mathbb{R} \to \mathcal{G} \times \mathbb{R}$, where $\mathcal{G}(B, \gamma) = (H(B), \gamma)$, it induces a well defined function $J : \mathcal{G} \times \mathbb{R} \to \mathbb{R}$ such that $J \circ \mathcal{G} = F$. The analyticity of J follows from the facts that F, as well as the action of $\mathcal{O}(m)$ on $\mathcal{B} \times \mathbb{R}$, are analytic, and that $\mathcal{G}$ induces an analytic diffeomorphism of the orbit manifold $(\mathcal{B} \times \mathbb{R})/\mathcal{O}(m)$ onto $\mathcal{G} \times \mathbb{R}$ ([41], (16.10.4)). ■

Next, we will compute the gradients of F_γ and J_γ over $\mathcal{B}$ and $\mathcal{G}$ with respect to the natural inner products on these manifolds introduced above. We set

$$P = \begin{pmatrix} P_1 & P_2 \\ P_2^\top & P_3 \end{pmatrix}, \qquad K = \begin{pmatrix} K_1 & K_2 \\ K_2^\top & K_3 \end{pmatrix},$$

and get, from Prop. 8.2,

$$\dot{F}_\gamma(\dot{B}) = 2\gamma \operatorname{tr}((K_2^\top P_1 + K_3 P_2^\top)\dot{B}), \tag{8.11}$$

where $\dot{B} = \Omega B, \Omega \in \mathfrak{so}(n)$.

The following lemma relates the gradients of a function $f : \mathcal{G} \to \mathbb{R}$ to the gradient of the function $f \circ \mathcal{H} : \mathcal{B} \to \mathbb{R}$

Lemma 8.2. *Let $f : \mathcal{G} \to \mathbb{R}$ be differentiable and $B \in \mathcal{B}$. Let $\bar{f} : \mathcal{B} \to \mathbb{R} : B \to f(H(B))$. If $\Omega^* \in \mathfrak{so}(n)$ is such that*

$$\operatorname{grad} \bar{f}(B) = \Omega^* B,$$

then

$$\operatorname{grad} f(H(B)) = -[H(B), \Omega^*].$$

Proof. We show that the differential $dH : \mathrm{T}_B\mathcal{B} \to \mathrm{T}_{H(B)}\mathcal{G}$ sends the gradient $\operatorname{grad}(f \circ H)$ of $f \circ H$ to the gradient $(\operatorname{grad} f)$ of f.

Given $B \in \mathcal{B}$, let $\mathcal{H}_\mathcal{B} = \{\dot{\mathcal{B}} \in \mathrm{T}_\mathcal{B}\mathcal{B} \mid \dot{\mathcal{B}}' \in \ker(\mathrm{T}_\mathcal{B}\mathcal{H}) \Rightarrow \tau(\dot{\mathcal{B}}, \dot{\mathcal{B}}') = \prime\}$ be the horizontal subspace of $\mathrm{T}_B\mathcal{B}$ with respect to H. One can verify that for each $\dot{B} \in \mathcal{H}_\mathcal{B}$, we have $\tau(\dot{B}, \dot{B}) = \kappa(dH(\dot{B}), dH(\dot{B}))$, i.e. that H is a Riemannian submersion. Hence, $\operatorname{grad}(f \circ H)$ is a horizontal lifting of $\operatorname{grad} f$, i.e. $\operatorname{grad}(f \circ H)_B \in \mathcal{H}_\mathcal{B}$ and $dH(\operatorname{grad}(f \circ H)(B)) = (\operatorname{grad} f)(H(B))$ for all $B \in \mathcal{B}$, ([42], § 4). The second claim follows from the relation $dH(\Omega B) = [\Omega, H(B)]$ for $\Omega \in \mathfrak{so}(n)$. ∎

We now evaluate the gradient of F_γ:

Proposition 8.3. *The gradient of the function F_γ over $\mathcal{B}$ with respect to the metric τ is given by*

$$\operatorname{grad} F_\gamma(B) = -2\gamma \Pi_{\mathfrak{so}(n)}(B(K_2^\top P_1 + K_3 P_2^\top))B,$$

where K is given in Eq. (8.9)*, and P is given in Eq.* (8.5)*.*

Proof. For every $B \in \mathcal{B}$ and $\Omega \in \mathfrak{so}(n)$, we have

$$\begin{aligned}
\tau(\operatorname{grad} F_\gamma, \Omega B) &= -\operatorname{tr}(\bar{\rho}_B^{-1}(\operatorname{grad} F_\gamma)\bar{\rho}^{-1}(\Omega B)) \\
&= -\operatorname{tr}(\bar{\rho}_B^{-1}(\operatorname{grad} F_\gamma)\Omega) \\
&= 2\gamma \operatorname{tr}((K_2^\top P_1 + K_3 P_2^\top)\Omega B),
\end{aligned}$$

where the last row comes from the definition of the gradient and Eq. (8.11). Since the equality above holds true for arbitrary skew symmetric matrices Ω, we conclude that

$$\bar{\rho_B}^{-1}(\operatorname{grad} F_\gamma) = -2\gamma \Pi_{\mathfrak{so}(n)}(B(K_2^\top P_1 + K_3 P_2^\top)).$$

Applying ρ_B on both sides, we obtain

$$\operatorname{grad} F_\gamma = -2\gamma \Pi_{\mathfrak{so}(n)}(B(K_2^\top P_1 + K_3 P_2^\top))B$$

as announced. ■

Using Lemma 8.2, we obtain the gradient of J:

Proposition 8.4. *The gradient of the function J_γ over $\mathcal{G}$ with respect to the metric κ is given by*

$$\operatorname{grad} J_\gamma(G) = 2\gamma[G, \Pi_{\mathfrak{so}(n)}(B(K_2^\top P_1 + K_3 P_2^\top))],$$

where $B \in \mathcal{B}$ is any matrix for which $G = BB^\top$.

8.3 Analysis of J_γ.

In the case of a general γ, the analysis of the critical point of J_γ is difficult, and simulations show that the function can have many local minima. However, for γ small, the function is quite well-behaved, and we analyze it here. We quantify in the last section how small γ needs to be for the analysis to go through in practice.

When γ is small, the gradient of J_γ can be well-approximated by a well-behaved vector field (see Theorem 8.2 below), and the critical points of the original gradient and its approximation can be shown to be the same (we do so below). In order to obtain this approximation, we start with computing the first order expansion of the matrices M, P and K with respect to γ. First, we have

$$M = M^{(0)} + \gamma M^{(1)},$$

where

$$M^{(0)} = \begin{pmatrix} A & 0 \\ 0 & 0 \end{pmatrix}, \quad M^{(1)} = \begin{pmatrix} 0 & B \\ 0 & 0 \end{pmatrix}.$$

Lemma 8.3. *The matrix P in a neighborhood of $\gamma = 0$ satisfies:*

$$P = \begin{pmatrix} X & \gamma S \\ \gamma S^\top & \sqrt{c}I \end{pmatrix} + O(\gamma^2), \tag{8.12}$$

where X is the unique positive definite solution of the Lyapunov equation

$$A^\top X + XA + Q = 0 \tag{8.13}$$

and

$$S = -(A^\top - \sqrt{c}I)^{-1}XB. \tag{8.14}$$

Proof. Let $P = P^{(0)} + \gamma P^{(1)} + O(\gamma^2)$, where $P^{(0)}$ is a positive definite matrix and $P^{(1)}$ is a symmetric matrix.

First, we substitute the expansion of P in equation (8.5), and after grouping together the terms of degrees zero and one in γ, we get, respectively,

$$M^{(0)^\top}P^{(0)} + P^{(0)}M^{(0)} - P^{(0)}EP^{(0)} + T = 0, \tag{8.15}$$

and

$$M^{(1)^\top}P^{(0)} + P^{(0)}M^{(1)} + M^{(0)^\top}P^{(1)} + P^{(1)}M^{(0)} - P^{(1)}EP^{(0)} - P^{(0)}EP^{(1)} = 0. \tag{8.16}$$

Now we write

$$P^{(0)} = \begin{pmatrix} X_1 & X_2 \\ X_2^\top & X_3 \end{pmatrix},$$

and plugging it in equation (8.15), we get the system of equations

$$\begin{aligned} 0 &= A^\top X_1 + X_1 A - X_2 X_2^\top + Q, \\ 0 &= A^\top X_2 - X_2 X_3, \\ 0 &= -X_3^2 + cI. \end{aligned}$$

Since $P^{(0)}$ is positive definite, $X_3 = \sqrt{c}I$. Then, $(A^\top - \sqrt{c}I)X_2 = 0$, and since $A - \sqrt{c}I$ has full rank, we get $X_2 = 0$. Finally, we conclude that X_1 is the unique positive definite solution of the Lyapunov equaion

$$A^\top X_1 + X_1 A + Q = 0.$$

We thus have

$$P^{(0)} = \begin{pmatrix} X_1 & 0 \\ 0 & \sqrt{c}I \end{pmatrix},$$

we plug it in equation (8.16) and get:

$$W^\top + W + {M^{(0)}}^\top P^{(1)} + P^{(1)} M^{(0)} \\ - \sqrt{c} P^{(1)} E - \sqrt{c} E P^{(1)} = 0,$$

where

$$W = \begin{pmatrix} 0 & X_1 B \\ 0 & 0 \end{pmatrix}.$$

Similarly to the computation for $P^{(0)}$ above, we find that

$$P^{(1)} = \begin{pmatrix} 0 & S \\ S^\top & 0 \end{pmatrix},$$

where

$$S = -(A^\top - \sqrt{c}I)^{-1} X_1 B.$$

We conclude that $P^{(0)} + \gamma P^{(1)}$ is given by the matrix in the right hand side of (8.12), with $X = X_1$.

■

We can similarly obtain the expansion of K in a neighborhood of $\gamma = 0$:

Lemma 8.4. *The matrix K in a neighborhood of $\gamma = 0$ satisfies:*

$$K = \begin{pmatrix} Y & \gamma U \\ \gamma U^\top & 0 \end{pmatrix} + O(\gamma^2), \tag{8.17}$$

where

$$U = -(A - \sqrt{c}I)^{-1} Y (A^\top - \sqrt{c}I)^{-1} X B, \tag{8.18}$$

X is defined by (8.13)*, and Y is the unique positive semi-definite solution of the Lyapunov equation*

$$AY + YA^\top + L = 0 \tag{8.19}$$

Proof. Now we perform similar computations for the matrix K. Let

$$K = K^{(0)} + \gamma K^{(1)} + O(\gamma^2),$$

where $K^{(0)}$ is a positive semi-definite matrix, and $K^{(1)}$ is a symmetric matrix. Using Lemma 8.3 and equation (8.9)

$$(M - EP)K + K(M - EP)^\top + \bar{L} = 0,$$

we get:

$$\begin{aligned}(M^{(0)} - \sqrt{c}E)K^{(0)} + K^{(0)}(M^{(0)} - \sqrt{c}E)^\top + \bar{L} = 0,\\ (M^{(0)} - \sqrt{c}E)K^{(1)} + K^{(1)}(M^{(0)} - \sqrt{c}E)^\top\\ + \begin{pmatrix} 0 & B \\ -S^\top & 0 \end{pmatrix} K^{(0)} + K^{(0)} \begin{pmatrix} 0 & -S \\ B^\top & 0 \end{pmatrix} = 0,\end{aligned}$$

where S is given by (8.14).

We set

$$K^{(0)} = \begin{pmatrix} Y_1 & Y_2 \\ Y_2^\top & Y_3 \end{pmatrix}$$

and plug it in the first equation to get the following system of equations:

$$\begin{aligned} AY_1 + Y_1A^\top + L = 0,\\ (A - \sqrt{c}I)Y_2 = 0,\\ -2\sqrt{c}Y_3 = 0.\end{aligned}$$

Therefore,

$$K^{(0)} = \begin{pmatrix} Y_1 & 0 \\ 0 & 0 \end{pmatrix},$$

where Y_1 is the unique positive semi-definite solution of the Lyapunov equations

$$AY_1 + Y_1A^\top + L = 0.$$

Subsequently, we find

$$K^{(1)} = \begin{pmatrix} 0 & (A - \sqrt{c}I)^{-1} Y_1 S \\ S^\top Y_1 (A^\top - \sqrt{c}I)^{-1} & 0 \end{pmatrix}.$$

Finally, we conclude that $K^{(0)} + \gamma K^{(1)}$ is given by the matrix in the right hand side of (8.17), with $Y = Y_1$. ■

Now, for $\gamma \neq 0$ let us define the function

$$J_\gamma^*(G) = \frac{1}{\gamma^2}(J_\gamma(G) - J_0(G)),$$

and

$$J_0^*(G) = \lim_{\gamma \to 0} J_\gamma^*(G).$$

From Lemma. 8.1, $J_\gamma^*(G)$ is analytic. Since $F_\gamma(B) = F_\gamma(-B)$ for $\gamma \in \mathbb{R}, B \in \mathcal{B}$, the function $J_\gamma(G) - J_0(G)$ is even, analytic, and vanishes at $\gamma = 0$. Therefore, it can be written as the product of γ^2 and an analytic function. Furthermore, since $J_0 : \mathcal{G} \to \mathbb{R}$ is constant, we have $\operatorname{grad} J_\gamma^* = \frac{1}{\gamma^2} \operatorname{grad} J_\gamma^*$, and therefore, for $\gamma \neq 0$, the functions J_γ and J_γ^* have the same critical points in $\mathcal{G}$.

Theorem 8.2. *In a neighborhood of $\gamma = 0$ we have*

$$\operatorname{grad} J_\gamma(G) = \gamma^2 [G, [G, Z]] + O(\gamma^3), \tag{8.20}$$

where

$$Z = -X(A - \sqrt{c}I)^{-1} Y (A^\top - \sqrt{c}I)^{-1} X, \tag{8.21}$$

and X and Y are defined by (8.13) *and* (8.19)*, respectively.*

Proof. Using Lemma 8.3 and Lemma 8.4, we find

$$\begin{aligned} \operatorname{tr}(KPN) &= \gamma^2 \operatorname{tr}(U^\top X \dot{B}) + O(\gamma^3) \\ &= \gamma^2 \operatorname{tr}(-B^\top X (A - \sqrt{c}I)^{-1} Y (A^\top - \sqrt{c}I)^{-1} X \dot{B}) + O(\gamma^3) \\ &= \gamma^2 \operatorname{tr}(GZ\Omega) + O(\gamma^3), \end{aligned} \tag{8.22}$$

where $G = BB^\top$ and Z is given by (8.21). Similarly to the computations

we made in Section 8.2, we find that the gradient of J_γ with respect to the metric κ on $\mathcal{G}$ satisfies

$$\operatorname{grad} J_\gamma(G) = 2\gamma^2[G, \Pi_{\mathfrak{so}(n)}(GZ)] + O(\gamma^3). \tag{8.23}$$

Since G and Z are symmetric matrices, we have

$$\Pi_{\mathfrak{so}(n)}(GZ) = \frac{1}{2}[G, Z],$$

which concludes the proof. ■

We can now characterize the critical points of J as follows:

Corollary 8.2. *The critical points of J_0^* satisfy the equation*

$$[G, Z] = 0.$$

Proof. Using Theorem 8.2, we see that the critical points of J_0^* satisfy the equation $[G, [G, Z]] = 0$. Let $G = \Theta \bar{G} \Theta^\top$, where $\bar{G} = \operatorname{diag}(d_1, d_2, \ldots, d_n)$ is a diagonal matrix. If $Z = \Theta \bar{Z} \Theta^\top$, where $\bar{Z} = (z_{i,j})$ is a symmetric matrix, we get $[\bar{G}, [\bar{G}, \bar{Z}]] = 0$. After expanding the commutators, we see that the entry of the matrix $[\bar{G}, \bar{Z}]$ at position (i, j) is equal to $(d_i - d_j)z_{i,j}$, and the entry of the matrix $[\bar{G}, [\bar{G}, \bar{Z}]]$ at position (i, j) is equal to $(d_i - d_j)^2 z_{i,j}$. Therefore, $[\bar{G}, [\bar{G}, \bar{Z}]] = 0$ implies that $[\bar{G}, \bar{Z}] = 0$, which is equivalent to $[G, Z] = 0$. ■

8.4 Signature of the critical points of J_0^*

Recall that the signature of a symmetric bilinear form represented as a matrix Q is the triple (p_+, p_-, p_0), where p_+ (resp. p_-, resp. p_0) is the number of positive (resp. negative, resp. zero) eigenvalues of Q. We evaluate in this section the signature of the Hessian of J at its critical points. From this information, we can derive the number of local minima of J, since a local minima has signature $(n, 0, 0)$.

We first determine the number of critical points for J. Since symmetric matrices commute if and only if they have the same eigenspaces, Corollary 8.2 implies that the critical points G of J_0^* have the same eigenspace as the matrix Z.

Proposition 8.5. *Generically for A, Q, the matrix Z of Eq.* (8.21) *has n distinct eigenvectors.*

Proof. Let Sym_n denote the vector space of all symmetric matrices in $\mathbb{R}^{n\times n}$ and Pos_n denote its open subset consisting of all positive definite matrices. Given a matrix $G \in \mathbb{R}^{n\times n}$, let $\mathcal{L}_G : \mathrm{Sym}_n \to \mathrm{Sym}_n$ denote the Lyapunov operator, defined as $\mathcal{L}_G(X) = G^\top X + XG$ for $X \in \mathrm{Sym}_n$. Consider the polynomial map $\mathcal{Z} : \mathrm{Pos}_n \times \mathbb{R}^n \to \mathrm{Sym}_n$, defined as $\mathcal{Z}(Q, x) = -XVX$, where $X = -\mathcal{L}_A^{-1}(Q)$, $V = (A - \sqrt{c}I_n)^{-1}Y(A^\top - \sqrt{c}I_n)^{-1}$, and $Y = -\mathcal{L}_{A^\top}^{-1}(xx^\top)$ (cf. (8.7), (8.13), (8.19) and (8.21)). If $x_0 \in \mathbb{R}^n$ is such that the pair (A, x_0) is controllable, then $Y_0 = -\mathcal{L}_{A^\top}^{-1}(x_0x_0^\top)$ is a positive definite matrix ([43], Theorem 4), whence $V_0 = (A - \sqrt{c}I_n)^{-1}Y_0(A^\top - \sqrt{c}I_n)^{-1}$ is positive definite, as well. Then, for the partial differential of $\mathcal{Z}$ with respect to Q at point of the form (Q_0, x_0), where Q_0 is any positive definite matrix, we have $d\mathcal{Z}_{(Q_0,x_0)}(\dot{Q}, 0) = -\dot{X}V_0X_0 - X_0V_0\dot{X} = \mathcal{L}_{-V_0X_0}(\dot{X})$, where $\dot{Q} \in \mathrm{Sym}_n$, $X_0 = -\mathcal{L}_A^{-1}(Q_0)$, and $\dot{X} = -\mathcal{L}_A^{-1}(\dot{Q})$. Since the matrix $R = V_0^{\frac{1}{2}}X_0V_0^{\frac{1}{2}}$ is positive definite, the matrix $-V_0X_0 = V_0^{\frac{1}{2}}(-R)V_0^{-\frac{1}{2}}$ is Hurwitz, and hence the linear operator $\dot{Q} \mapsto d\mathcal{Z}_{(Q_0,x_0)}(\dot{Q}, 0) = \mathcal{L}_{-V_0X_0}(-\mathcal{L}_A^{-1}(\dot{Q}))$ is an isomorphism of Sym_n onto itself. Now, a corollary of the inverse function theorem implies that there exists a neighborhood $\mathfrak{U}$ of (Q_0, x_0) in $\mathrm{Pos}_n \times \mathbb{R}^n$, such that $\mathcal{Z}(\mathfrak{U})$ is open in Sym_n. Hence, there exists $(Q_1, x_1) \in \mathfrak{U}$, such that $\mathcal{Z}(Q_1, x_1)$ has distinct eigenvalues. Therefore, the discriminant of the characteristic polynomial of the matrix $\mathcal{Z}(Q, x)$ is a non-zero polynomial function on $\mathrm{Pos}_n \times \mathbb{R}^n$, whence we conclude that the set of all (Q, x), for which $\mathcal{Z}(Q, x)$ has distinct eigenvalues, i.e. the set where this function does not vanish, is open and dense in $\mathrm{Pos}_n \times \mathbb{R}^n$, and its complement has Lebesgue measure 0. Similarly, the set of all Hurwitz matrices A, for which there exists $x_0 \in \mathbb{R}^n$ such that the pair (A, x_0) is controllable, i.e. $\det(x_0, Ax_0, \ldots, A^{n-1}x_0) \neq 0$, is open and dense in the set of all Hurwitz matrices in $\mathbb{R}^{n\times n}$ and its complement has Lebesgue measure 0. ∎

Theorem 8.3. *Generically for A, Q, the function J_0^* has $\binom{n}{m}$ critical points.*

Proof. Proposition 8.5 states that generically Z has n different eigenvectors. Since Z and G commute, they must have the same eigenspaces. There are exactly $\binom{n}{m}$ matrices G on the Grassmannian $\mathcal{G}$ which satisfy this property. If $\Theta \in \mathcal{O}(n)$ is such that $Z = \Theta\bar{Z}\Theta^\top$ for some diagonal matrix $\bar{Z}$, they have

the form $G_i = \Theta \bar{G}_i \Theta^\top$, where $\bar{G}_i$ is any diagonal matrix, m of whose diagonal entries are equal to 1, the remaining diagonal entries being equal to 0. ■

The Hessian $\mathrm{H}J_0^*$ of J_0^* is the bilinear form

$$\mathrm{H}J_0^*(X, Y) := X(Y(J_0^*)) - \mathrm{d}J_0^*(\nabla_X Y), \tag{8.24}$$

where X, Y are arbitrary vector fields [44].

Let Ω_X and Ω_Y be fixed matrices in $\mathfrak{so}(n)$, and consider the vector fields X and Y on $\mathcal{G}$, defined as $X_G = [G, \Omega_X]$ and $Y_G = [G, \Omega_Y]$ for $G \in \mathcal{G}$. From (8.22) we obtain

$$Y_G(J_0^*) = \mathrm{tr}(GZ\Omega_Y),$$

and therefore,

$$X_G(Y(J_0^*)) = \mathrm{tr}([G, \Omega_X]Z\Omega_Y). \tag{8.25}$$

At the critical points G of J_0^*, the second term in equation (8.24) vanishes, and we have

$$(\mathrm{H}J_0^*)_G(X_G, Y_G) = X_G(Y(J_0^*)).$$

Now suppose that the matrices A, Q and L are such that the matrix Z, defined in (8.21), has distinct eigenvalues (Proposition 8.5), and let G be a critical point of J_0^*. Since G and Z are symmetric matrices, with $G^2 = G$ and $[G, Z] = 0$ (Corollary 8.2), there exists a matrix $\Theta \in \mathcal{O}(n)$ such that $G = \Theta \bar{G} \Theta^\top$ and $Z = \Theta \bar{Z} \Theta^\top$, where

$$\bar{G} = \mathrm{diag}(d_1, d_2, \ldots, d_n), \quad \bar{Z} = \mathrm{diag}(z_1, z_2, \ldots, z_n),$$

so that $d_i \in \{0, 1\}$ for $i = 1, \ldots, n$, and $z_1 > z_2 > \cdots > z_n$.

Theorem 8.4. *Suppose that the matrix Z of Eq. (8.21) has distinct eigenvalues, and let G be a critical point of J_0^*. Let $\alpha = \sum_{i=1}^n i d_i$, where $\bar{G} = \mathrm{diag}(d_1, \ldots, d_n)$ is the diagonal matrix defined in the preceding paragraph. Then the signature of the Hessian of J_0^* at G is $(p_+, p_-, 0)$, where*

$$p_+ = \alpha - \tfrac{1}{2} m(m+1), \tag{8.26}$$

$$p_- = nm - \tfrac{1}{2} m(m-1) - \alpha. \tag{8.27}$$

In particular, J_0^* has a unique local minimum, attained at the matrix for which $d_i = 1$ if and only if $i > n - m$.

Proof. Let $\Theta \in \mathcal{O}(n)$ and $\bar{Z}$ be defined as in the preceding paragraph. For $\Omega_X, \Omega_Y \in \mathfrak{so}(n)$, and the vector fields X, Y, defined as above, we can rewrite (8.25) as:

$$\begin{aligned} X_G(Y(J_0^*)) &= \operatorname{tr}([\Theta\bar{G}\Theta^\top, \Omega_X]\Theta\bar{Z}\Theta^\top\Omega_Y) \\ &= \operatorname{tr}([\bar{G}, \Theta^\top\Omega_X\Theta]\bar{Z}\Theta^\top\Omega_Y\Theta). \end{aligned}$$

Let $E_{i,j} = (e_{k,l})_{k,l\leq n}, i \neq j$, where $e_{i,j} = 1, e_{j,i} = -1$, and $e_{k,l} = 0$ otherwise. Also, let $\Pi = \Pi(G)$ be the set of all pairs of indices (i, j) such that $d_i = 1$ and $d_j = 0$. The matrices $[G, \Theta E_{i,j}\Theta^\top]$, for which $(i, j) \in \Pi$ form a basis for $\mathrm{T}_G\mathcal{G}$, ([21], Lemma 3.3). If we choose $\Omega_X = \Theta E_{i,j}\Theta^\top$ and $\Omega_Y = \Theta E_{k,l}\Theta^\top$, where $(i, j), (k, l) \in \Pi$, simple calculations show that the Hessian of J_0^* at the critical point G satisfies

$$(\mathrm{H}J_0^*)_G(X_G, Y_G) = \operatorname{tr}([\bar{G}, E_{i,j}]\bar{Z}E_{k,l}) = (z_k - z_l)\delta_{i,k}\delta_{j,l},$$

where $\delta_{i,j}$ is the Kronecker symbol. Thus, we see that the Hessian $(\mathrm{H}J_0^*)_G$ is in diagonal form with respect to that basis. In particular, $(\mathrm{H}J_0^*)_G$ is non-degenerate. The same argument as in the proof of Lemma 3.12 in [21] shows that its signature is given by (8.26) and (8.27). ■

Combining Theorems 8.3 and 8.4 with the results from Sections 8.2 and 8.3, next we prove Theorem 8.1.

Proof of Theorem 8.1. Suppose the matrices A, Q, and L are such that the matrix Z has distinct eigenvalues. This is satisfied generically, as per Proposition 8.5. Then, according to Theorem 8.3, the function J_0^* has $\binom{n}{m}$ critical points. Furthermore, Theorem 8.4 states that all the critical points are non-degenerate, and exactly one of them is a local minimum. Since the manifold $\mathcal{G}$ is compact, and the function J_γ^* is analytic on $\mathcal{G} \times \mathbb{R}$, there exists $\varepsilon > 0$, such that if $|\gamma| < \varepsilon$, then J_γ^* also has $\binom{n}{m}$ critical points which have the same signatures as the critical points of J_0^*. Finally, since the functions J_γ and J_γ^* have identical critical points, the statement of the theorem follows. ■

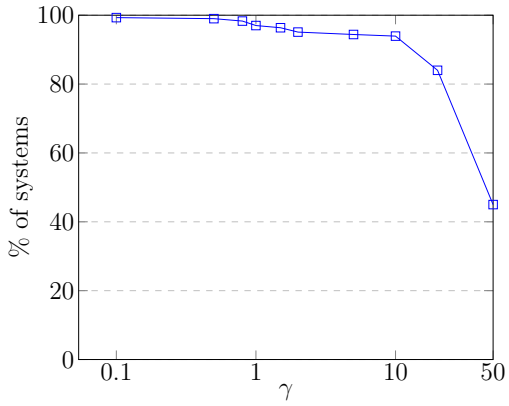

Figure 8.2: Empirical proportion of systems for which there is a unique optimal actuator as a function of γ.

8.5 Simulations and discussion

We implemented in Matlab the gradient flow of Prop. 8.3 to verify its convergence. For the simulation results presented here, we set $n = 4$ and $m = 1$. We choose $Q = I_4/\sqrt{4}$, $c = 1$, and sample random Hurwitz matrices (via rejection sampling) $A \in \mathbb{R}^{4\times 4}$. For each A, we run the gradient flow several (100) times from randomly chosen initial states and verify whether the flow converges to the same actuator, thus verifying empirically that F_γ has an essentially unique local minimum. We plot in Fig 8.2, as a function of γ, the percentage of systems (i.e., matrices A) for which there is a unique locally optimal actuator. Of course, when F_γ has more than one local minimum, the gradient will still converge to a local minimum, but it may not be the global minimum.

We conjecture that the number of local minima is always upper bounded by n, irrespective of the parameter γ. Another open problem of interest is to study the large γ asymptotic, i.e. systems with very high gain actuators.

CHAPTER 9

OPTIMAL ACTUATOR DESIGN FOR LINEAR SYSTEMS WITH MULTIPLICATIVE NOISE.

9.1 Preliminaries

9.1.1 Terminology and notation

First we will recall some basic definitions and results that are needed in the paper. A matrix M is called *Hurwitz*, if all of its eigenvalues have negative real parts. It is not hard to see that a matrix M is Hurwitz if and only if $\exp(Mt)$ approaches 0 as t approaches infinity. Hence Hurwitz matrices describe stable linear dynamics in continuous time. We denote by $[A, B] := AB - BA$ the commutator of matrices A and B. We also write

$$[B, \Omega] =: \mathrm{ad}_B\, \Omega = B\Omega - \Omega B.$$

We let Sym_n the set of real symmetric $n \times n$ matrices. For $A, G \in \mathbb{R}^{n\times n}$, we set

$$\mathcal{L}_{A,G} : \mathrm{Sym}_n \to \mathrm{Sym}_n : X \mapsto A^\top X + XA + G^\top XG.$$

9.1.2 Problem statement and background

Background and preliminary results. We consider the LTI control system

$$\dot{x} = Ax + bu, \tag{9.1}$$

where $A \in \mathbb{R}^{n\times n}, b \in \mathbb{R}^{n\times m}$, and introduce the quadratic cost

$$C(u, x_0) = \int_0^\infty (x^\top(t)Qx(t) + u^\top(t, x)u(t, x))dt, \tag{9.2}$$

where $Q > 0$ is given positive definite matrix. We recall that the pair (A, b)

is called stabilizable if the uncontrollable modes of (A, b) are stable. It is known that if (A, b) is stabilizable [45], then the control u that minimizes the above cost, which we denote as $u_{\min}(x_0)$, is given by $u_{\min}(x) = -b^\top Px$, where the matrix P is the unique positive-definite solution of the algebraic Riccati equation

$$A^\top P + PA - Pbb^\top P + Q = 0.$$

Furthermore, we can show that $C(u_{\min}(x_0), x_0) = \mathrm{tr}(PL)$ with $L = x_0 x_0^\top$.

We care in this paper about actuator design, and hence b is considered to be a free parameter. Note that if the matrix A is stable, then for any $b \in \mathbb{R}^{n\times m}$, the pair (A, b) is stabilizable, and hence the optimal cost is well-defined over $\mathbb{R}^{n\times m}$. Thus, for a stable matrix A, a positive-definite matrix Q, and an initial state x_0 given, we can ask the question:

How should we design the matrix b, such that the optimal cost $C(u_{\min}(x_0), x_0)$ is as small as possible?

First, we must place restriction on the matrices b. Indeed, it is not too difficult to see intuitively that if $\|b\|$ increases, all other things equal, then $C(u_{\min}(x_0), x_0)$ decreases. A proof of this fact is essentially reduced to results about the monotonicity of the Ricatti equation such as the ones in [46]. We thus constraint the norm of b by considering the set so that

$$b^\top b = \gamma^2 I.$$

This also adds the requirement that the actuators are orthogonal to each other, an assumption we will discuss below. Now noting that the cost C depends on the product $bb^\top$, we can rephrase the problem as follows. Let γ be a real parameter, and A, Q, L be such that A is Hurwitz, and Q, L are positive-definite. Minimize the function $J_\gamma(B) = \mathrm{tr}(LP)$, where $P = P(B)$ is the solution of

$$A^\top P + PA - \gamma^2 PBP + Q = 0,$$

over the set

$$\Gamma := \{B = bb^\top | b^\top b = I\}. \tag{9.3}$$

We can furthermore remove the dependence of the optimal design from the initial state x_0 by averaging over an "isotropic" initial state as follows: assuming the initial state is distributed according to a rotationally invariant

distribution (about the origin), such as a multivariate normal distribution center at the origin, then

$$\mathbb{E}C(u_{\min}(x_0)x_0)) = k \operatorname{tr} P,$$

for some positive constant k and where $\mathbb{E}$ is the expectation operator. This is the deterministic actuator placement problem.

9.1.3 Statement of the results

We explore in this paper the actuator placement problem for control systems which are corrupted by additive and multiplicative noise. To be more precise, consider the control system described by the stochastic differential equation (7.4). We introduce the cost

$$C = \lim_{T\to\infty} \mathbb{E}\left(\frac{1}{T}\int_0^\top (x^\top Q x + u^\top u)dt\right), \tag{9.4}$$

where $Q > 0$ is given positive-definite matrix. It can be shown that when $G_2 = 0$, the optimal control $u_{\min}$ in steady state is given again by the equations (9.2). Hence the addition of additive noise does not change the methods to solve the problem, nor the properties of the solution set in a meaningful way.

We will hence focus on the multiplicative noise case

$$dx = Ax\,dt + bu\,dt + Gx\,dw, \tag{9.5}$$

with associated cost as in Eq. (9.4).

Throughout the paper, we will assume that the matrices A and G satisfy the following technical condition:

$$\left|\int_0^\infty e^{tA^\top} G^\top G e^{tA} dt\right| < 1. \tag{9.6}$$

Equivalently, we require that the unique positive semi-definite solution X of

the Lyapunov equation

$$A^\top X + XA + G^\top G = 0$$

is a convergent matrix, i.e. all of its eigenvalues have norm less than 1. Under these assumptions, the control minimizing the cost in this case can be seen to be $u_{\min} = -b^\top Px$, where P is the unique positive-definite solution [47] of

$$A^\top P + PA + Q + G^\top PG - PBP = 0. \tag{9.7}$$

The minimum expected cost is equal to $C_{\min} = \mathrm{tr}(PL), L = x_0 x_0^\top$. Thus, the problem we will be solving is:

Problem 9.1. *Let $\gamma \in \mathbb{R}$ and A, Q, L be given matrices, such that A is Hurwitz, and Q is positive-definite, and L is positive semi-definite of rank 1. Minimize the function $J_\gamma(B) = tr(LP)$, where $P(B)$ is the solution of*

$$A^\top P + PA + Q + G^\top PG - \gamma^2 Pbb^\top P = 0,$$

over the set $\Gamma = \{B = bb^\top | b^\top b = I\}$.

We prove the following result:

Theorem 9.1. *Generically for A, G, Q, for $\gamma > 0$ small enough, the function $J_\gamma(B)$ has $\binom{n}{m}$ critical points over the manifold Γ, exactly one of which is local minimum. Furthermore, the differential equation*

$$\dot{B} = -\gamma^2 [B, [B, M]], \quad B(0) = B_0 \in \Gamma$$

where $M := PRP$, and P, R satisfy

$$\begin{aligned} A^\top P + PA + Q + G^\top PG - \gamma^2 PBP = 0, \\ (A - \gamma BP)R + R(A - \gamma BP)^\top + GRG^\top - L = 0. \end{aligned}$$

converges to the global minimizer of $J_\gamma(B)$ from almost all initial state B_0.

This result in essence extends the results of [21] to the case of multiplicative noise, and show that one can also obtain an optimal design in this case, since the gradient flow of J, derived in this paper, will converge to the optimal design from a generic initial state.

We briefly sketch the proof. First, we will compute the gradient ∇J_γ of the function J_γ, with respect to an appropriately defined metric on the space Γ. Then we will show that as γ approaches 0, after well-chosen normalization, the function J_γ has a proper limit J_0^*. We will find the points at which ∇J_0^* vanishes, will show that their number is $\binom{n}{m}$, and that all of them are non-degenerate. We will compute the Hessian of J_0^* and thus find the signatures of the critical points. Since the number of critical points and their signatures are constant in the vicinity of 0, the theorem will follow.

9.2 Proof of the main result

9.2.1 Preliminary results

We now derive some preliminary results which may be of independent interest, and will be needed to prove the main result. They pertain to positive definite solutions of Lyapunov equations and the dependence of the Riccati equation with respect to its defining parameters.

The first result deals with the "generalized" Lyapunov equation

$$AX + XA^\top + G^\top XG + Q = 0,$$

which is a mix of the "discrete-time" Lyapunov equation $AXA^\top - X + Q = 0$ and "continuous-time" Lyapunov equation $AX + XA^\top + Q = 0$. It is also referred to as a Lyapunov Equation of mixed type [48]. In [48], [47], the following lemma is proved:

Lemma 9.1. *Let $A \in \mathbb{R}^{n\times n}$, $G \in \mathbb{R}^{n\times n}$, where A is a Hurwitz matrix. We consider the generalized Lyapunov equation*

$$A^\top X + XA + G^\top XG + Q = 0. \tag{9.8}$$

The following statements are equivalent:

- *Equation* (9.8) *has a positive semi-definite solution $X \geq 0$ for some positive definite matrix $Q \in \mathbb{R}^{n\times n}$.*

- *The eigen-values of $\mathcal{L}_{A,G}$ have negative real parts.*

If any of the statements above are satisfied, equation (9.8) *has has a unique symmetric (positive definite) solution X for any symmetric (positive definite) matrix Q. In this case, the solution X can be represented as the converging sum*

$$X = \sum_{i=0}^{\infty} \mathcal{T}^i \left(\int_0^{\infty} e^{A^\top t} Q e^{At} dt \right),$$

where

$$\mathcal{T}(X) = \int_0^{\infty} e^{A^\top t} G^\top X G e^{At} dt.$$

The second preliminary result is to show that the positive definite solution of the Riccati equation (9.7) depends analytically on its parameters (under some assumptions to be listed). This result is an extension of [49], and the proof follows the same lines. We thus only sketch it.

Lemma 9.2. *Let $A, G, Q \in \mathbb{R}^{n\times n}$ be so that A is Hurwitz, Q is positive definite, and inequality* (9.6) *is satisfied. We introduce the function*

$$X : \Gamma \times \mathbb{R} \to \mathbb{R}^{n\times n} : (B, \gamma) \mapsto X(B, \gamma)$$

where $X(B, \gamma)$ is the unique positive definite solution of the Riccati equation

$$A^\top X + XA + Q + G^\top XG - \gamma^2 XBX = 0. \tag{9.9}$$

Then the map X is analytic.

Proof. As already mentioned, the proof follows Delchamps' approach and consists of using the inverse function theorem on an appropriately defined map. Namely, consider the map

$$\phi(B, \gamma, X) = A^\top X + XA + Q + G^\top XG - \gamma^2 XBX.$$

Its differential with respect to X is given by

$$d\phi = dX(A - \gamma^2 BX) + (A^\top - \gamma^2 XB)dX + G^\top dXG.$$

Introduce the map $M : \mathrm{Sym}_n \to \mathrm{Sym}_n$ defined as

$$M_{(B,\gamma)} : T \to T(A - \gamma^2 BX) + (A^\top - \gamma^2 XB)T + G^\top TG.$$

Note that equation (9.9) can be rewritten, adding and subtracting $\gamma^2 XBX$, as

$$\begin{aligned} X(A - \gamma^2 BX) + (A^\top - \gamma^2 XB)X + G^\top XG \\ +(Q + \gamma^2 XBX) = 0. \end{aligned}$$

Therefore, $X(\gamma, B)$—defined as the unique psd solution of Eq. (9.9)—is also a solution of the equation

$$M_{(B,\gamma)}(T) + (Q + \gamma^2 XBX) = 0,$$

i.e., setting $T = X$ solves the above equation. Thus, we can apply Lemma 9.1 and conclude that there exists a unique symmetric solution T to the equation $M_{(B,\gamma)}(T) = S$ for symmetric S. We conclude that $M_{(B,\gamma)}(T) : \mathrm{Sym}_n \to \mathrm{Sym}_n$ is surjective. Now, from the implicit function theorem applied to $\phi(B, \gamma, X)$, we conclude that every solution X of (9.9) for a given (B, γ) can be extended uniquely in a small enough neighborhood of (B, γ). Since the Riccati equation has a unique positive definite solution [50] for every $\gamma \in \mathbb{R}, B \in \Sigma$, the claim of the lemma follows. ■

9.2.2 Gradient of J_γ and its critical points

We now evaluate the gradient of the function J_γ defined over Γ. Recall that on a Riemannian manifold, the gradient ∇J is defined with respect to an inner product $\langle \cdot, \cdot \rangle$ on Γ—we will introduce an inner product below—as the unique solution of

$$D_\Delta J := \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} J(B + \varepsilon\Delta) = \langle \nabla J, \Delta \rangle, \forall\, \Delta \in T_B\Gamma, \tag{9.10}$$

where we also introduce the notation $D_\Delta J$ for the directional derivative of J along Δ. In words, the variation of the function along the direction Δ is equal to the inner product of the gradient ∇J with Δ. Because J is defined on Γ, we need to first find the set of allowed variations around B, or the

tangent space of Γ at B. We note that every $B \in \Gamma$ is so that rank $B = m$ and $B^2 = B$, and thus is an orthogonal projection to the subspace spanned by the columns of $b \in \mathbb{R}^{n\times m}$, where $bb^\top = B$. Reciprocally, to each m-dimensional subspace of $\mathbb{R}^n$, we can assign a unique orthogonal projection matrix B onto that subspace. Hence elements in Γ are in one-to-one correspondence with m-dimensional subspaces of $\mathbb{R}^n$, i.e. with elements of the Grassmanian [51] of m-planes in $\mathbb{R}^n$. It is furthermore well-known that Γ is a differentiable manifold, and admits a well-defined tangent space at any $B \in \Gamma$ denoted by $T_B\Gamma$. It is given by

$$T_B\Gamma = \{[B, \Omega] | \Omega \in \text{skew}(n)\}, \tag{9.11}$$

where we recall that $[A, B] := AB - BA$ is the commutator or Lie bracket of A and B, and $\text{skew}(n) \subset \mathbb{R}^{n\times n}$ is the set of skew-symmetric matrices, i.e. $A \in \text{skew}(n)$ if $A = -A^\top$.

An inner product on $T_B\Gamma$: We now introduce the inner product on $T\Gamma$ we will work with. We keep the introduction short, since the same inner product was used in [52, 39, 53]. We emphasize that the choice of inner product does not change the main results, but makes the analysis simpler.

Since every tangent vector $\Delta \in T_B\Gamma$ is of the form $\Delta = [B, \Omega]$ for some $\Omega \in \text{skew}(n)$, a seemingly good choice $\langle \cdot, \cdot \rangle$ would be

$$\langle \Delta_1, \Delta_2 \rangle_B = -\operatorname{tr}(\text{ad}_B^{-1}(\Delta_1)\, \text{ad}_B^{-1}(\Delta_2)) = -\operatorname{tr}(\Omega_1\Omega_2),$$

where $\Delta_1 = [B, \Omega_1]$ and $\Delta_2 = [B, \Omega_2]$.

However, the choice of Ω_1 and Ω_2 is not unique, i.e., $\text{ad}_B : \text{skew}_n \to T_B\Gamma$ is not invertible. We thus define $\bar{\text{ad}}_B(\cdot)$ as:

$$\bar{\text{ad}}_B : \text{skew}(n)/\ker\,(\text{ad}_B) \to T_B\Gamma,$$

where we regard $\text{skew}(n)/\ker\,(\text{ad}_B)$ as the orthogonal of $\ker\ \text{ad}_B$ in $\text{skew}(n)$ for the well-defined inner product in $\text{skew}(n)$ given by $\operatorname{tr}(\Omega_1\Omega_2^\top)$. Now $\bar{\text{ad}}_B$ is invertible for every B by construction, and we can define the operator

$$\langle \Delta_1, \Delta_2 \rangle_B = -\operatorname{tr}(\bar{\text{ad}}_B^{-1}(\Delta_1)\bar{\text{ad}}_B^{-1}(\Delta_2)). \tag{9.12}$$

One can show that it is a well-defined inner product on Γ.

We now evaluate the left-hand side of Eq. (9.10), i.e. we compute the derivative of $J_\gamma(B)$, denoted by $D_\Delta J$ in the direction Δ. This derivative is well-defined from Lemma 9.2. From Eq. (9.11), it suffices to consider $\Delta = [B, \Omega]$ for $\Omega \in \text{skew}(n)$. We sometimes write $D_\Omega J$ for $D_{[B,\Omega]} J$. Now assume Ω fixed and note that from that because the Riccati equation has a unique positive definite solution for all $B \in \Gamma$, the function $P(B)$ is well-defined as the solution of (9.7).

We introduce the short-hand notation $\dot{B} := [B, \Omega]$ and $\dot{P} := D_\Delta P$, for P defined as the positive definite of (9.7), and for $\Delta = [B, \Omega]$. Differentiating (9.7) in the direction Δ, we obtain

$$A^\top \dot{P} + \dot{P} A + G^\top \dot{P} G - \gamma^2 \dot{P} B P - \gamma^2 P \dot{B} P - \gamma^2 P B \dot{P} = 0.$$

Gathering the terms multiplying $\dot{P}$ and $\dot{B}$, we obtain

$$(A - \gamma^2 BP)^\top \dot{P} + \dot{P}(A - \gamma^2 BP) + G^\top \dot{P} G - \gamma^2 P \dot{B} P = 0$$

We can regard the equality above as a generalized Lyapunov equation in $\dot{P}$, similar to the one studied in Lemma 9.1.

Lemma 9.3. *Under the assumptions of Lemma 9.1, the derivative of J in the direction $\Delta = [B, \Omega]$ is given by*

$$D_\Omega(J) = -\gamma^2 \operatorname{tr}([M, B]\Omega)$$

where $M := P R_i P$ and

$$R_i := \int_0^\infty \dots \int_0^\infty e^{(A-\gamma^2 BP)^\top t_1} G^\top e^{(A-\gamma^2 BP)^\top t_2} G^\top \dots e^{(A-\gamma^2 BP)^\top t_i} L e^{(A-\gamma^2 BP) t_i} \dots G e^{(A-\gamma^2 BP) t_1} dt_1 \dots dt_i$$

with P the positive definite solution of Eq. (9.7).

Proof. Applying Lemma (9.1) and the fact that $\operatorname{tr}([A, B]) = 0$, we get:

$$\dot{P} = -\gamma^2 \int_0^\infty e^{(A-\gamma^2 BP)t_1} K\dot{B}Pe^{(A-\gamma^2 BP)^\top t_1} dt_1$$
$$- \gamma^2 \int_0^\infty \int_0^\infty e^{(A-\gamma^2 BP)t_2} Ge^{(A-\gamma^2 BP)t_1} P\dot{B}$$
$$Pe^{(A-\gamma^2 BP)^\top t_1} G^\top e^{(A-\gamma^2 BP)^\top t_2} dt_1 dt_2 - \dots$$

Using the above, we obtain

$$\begin{aligned} D_\Omega(J) &= \operatorname{tr}(L\dot{P}) \\ &= -\gamma^2 \operatorname{tr}(\int_0^\infty Pe^{(A-\gamma^2 BP)^\top t_1} Le^{(A-\gamma^2 BP)t_1} Pdt_1 \dot{B}) \\ &- \gamma^2 \operatorname{tr}(\int_0^\infty \int_0^\infty Pe^{(A-\gamma^2 BL)^\top t_1} G^\top e^{(A-\gamma^2 BP)^\top t_2} L \\ &e^{(A-\gamma^2 BP)t_2} Ge^{(A-\gamma^2 BP)t_1} dt_1 dt_2 \dot{B}) - \dots \\ &= -\gamma^2 \operatorname{tr}([\sum_i M_i, B]\Omega), \end{aligned}$$

where $M_i := PR_iP$ and we set R_i as in the statement of the Lemma. Note that $\sum_i M_i$ converges since it is a linear transformation of a convergent series. Hence $D_\Omega(J)$ is well-defined. ■

Next, we compute the gradient ∇J_γ of the function $J_\gamma(B)$.

Theorem 9.2. *The gradient ∇J_γ of the function J_γ with respect to the metric $\langle \cdot, \cdot \rangle$ defined above is*

$$\nabla(J_\gamma(B)) = \gamma^2 [B, [B, M]],$$

where $M := PRP$, and P, R satisfy

$$A^\top P + PA + Q + G^\top PG - \gamma^2 PBP = 0,$$
$$(A - \gamma^2 BP)R + R(A - \gamma^2 BP)^\top + GRG^\top + L = 0.$$

Proof. The gradient ∇J_γ of J_γ satisfies

$$\langle \nabla J_\gamma, \Delta \rangle = D(J_\gamma)$$

for all vector fields $\Delta \in T\Gamma, \Delta = [B, \Omega], \Omega \in \operatorname{skew}(n)$. Using the definition

of the inner product given in Eq. (9.12), we obtain

$$\operatorname{tr}(\bar{\operatorname{ad}}_B{}^{-1}(\nabla J_\gamma(B))\Omega) = \gamma^2 \operatorname{tr}([M,B](\Omega+\Theta)),$$

where $\Theta \in \ker \bar{\operatorname{ad}}_B$ is arbitrary and M is as defined in the statement of the Theorem. Using the easily verified relation

$$\operatorname{tr}([A,B]C]) = \operatorname{tr}(A[B,C]),$$

we get $\operatorname{tr}([M,B]\Theta) = \operatorname{tr}(M[B,\Theta])$. Since $\Theta \in \ker \bar{\operatorname{ad}}_B$,

$$\operatorname{tr}([M,B](\Omega+\Theta)) = \operatorname{tr}([M,B]\Omega) + \operatorname{tr}(M[B,\Theta]) = 0,$$

and therefore

$$\operatorname{tr}(\bar{\operatorname{ad}}_B{}^{-1}(\nabla J_\gamma(B))\Omega) = -\gamma^2 \operatorname{tr}([M,B]\Omega)$$

for all $\Omega \in \operatorname{skew}(n)$. Since $-\operatorname{tr}(\Omega_1\Omega_2)$ is a non-degenerate inner-product on $\operatorname{skew}(n)$, this implies $\bar{\operatorname{ad}}_B{}^{-1}(\nabla J_\gamma(B)) = \gamma^2 \operatorname{ad}_B(M)$ and

$$\nabla(J_\gamma(B)) = \gamma^2 \operatorname{ad}_B \operatorname{ad}_B M = \gamma^2 [B,[B,M]].$$

as announced. ■

We record the immediate Corollary

Corollary 9.1. *The critical points of the function $J_\gamma(B)$ satisfy the equality*

$$[B,M] = 0,$$

where M is as defined in Theorem 9.2 .

Proof. The critical points of a function are exactly the points where its gradient vanishes. Since B is symmetric and $[B,M]$ is skew symmetric, $\gamma^2[B,[B,M]] = 0$ implies

$$[B,M] = 0.$$

as announced. ■

9.3 Convergence of gradient descent

We aim to find an optimal actuator via a gradient descent

$$\dot{B} = -\gamma^2[B,[B,M]] \tag{9.13}$$

with M defined in Theorem 9.2. It is not too difficult to see that the function $J_\gamma(B)$ is not convex, and hence we need to argue for the convergence of the method. We do so by showing that J_γ generically for the parameters A, G, Q has a *unique* minimum, and hence gradient descent will converge to that minimum from almost all initial value $B(0)$.

To this end, define the function

$$J_\gamma^* := \frac{1}{\gamma^2}(J_\gamma - J_0), \text{ with } J_0 := \operatorname{tr}(LP_0),$$

where P_0 is the positive definite solution of the equation

$$A^\top P_0 + P_0 A + Q + G^\top P_0 G = 0.$$

Furthermore, set

$$J_0^* := \lim_{\gamma \to 0} J_\gamma^*.$$

We know from Lemma 9.2 that $J_\gamma^*(B)$ is analytic in both γ and B and it clearly has the same critical points as $J_\gamma(B)$ for fixed $\gamma \neq 0$, since the two functions differ by a constant. Therefore, if we show that the critical points of the function J_0^* are *non-degenerate*, then it will follow that J_γ has the same number of critical points and the same corresponding signatures as J_0^* for small $\gamma \neq 0$.

In order to do this, first we first establish the following result

Proposition 9.1. *Let A and G be so that the assumption* (9.6) *is satisfied. Suppose also that there exists $x \in \mathbb{R}^n$, such that the pair (A, x) is controllable. Then, generically for all positive definite Q, and positive semi-definite L of rank* 1*, the function J_0^* has $\binom{n}{m}$ critical points.*

The derivation uses Theorem 3.6.1 in [47] and follows strictly the proof of Proposition 1 in [21], and we thus omit it here. We now evaluate the Hessian of J_0^*, that is the derivative of the gradient of J_0^*, to check that it is indeed

non-degenerate. Recall that the Hessian is a symmetric bilinear form taking its argument in $T_B\Gamma$. We have the following result:

Proposition 9.2. *The Hessian $H_{J_0^*}$ of the function J_0^* satisfies the equality*

$$H_{J_0^*}(\Delta_1, \Delta_2) = \mathrm{tr}([M_0, \Omega_1][B, \Omega_2])$$

at critical points B of J_γ^*, where $\Delta_1 = [B, \Omega_1]$ and $\Delta_2 = [B, \Omega_2]$ for $\Omega_i \in \mathrm{skew}(n)$ and the matrix $M_0 := P_0 R_0 P_0$ where P_0 positive definite solution of

$$A^\top P_0 + P_0 A + Q + G^\top P_0 G = 0,$$

and R_0 the positive definite solution of

$$A R_0 + R_0 A^\top + G^\top R_0 G + L = 0.$$

Proof. Let $F : \Gamma \to \mathbb{R}$ be a twice differentiable function. We have the general formula for the Hessian [44] H_F of F evaluated in the directions Δ_1, Δ_2:

$$H_F(\Delta_1, \Delta_2) = \Delta_1 \cdot \Delta_2 \cdot F + D_{\Delta_1}\Delta_2 \cdot F,$$

where Δ_1 and Δ_2 are arbitrary vector fields on $T\Gamma$ and $D_{\Delta_1}\Delta_2$ is the covariant derivative of Δ_2 along Δ_1. It is easy to see that second term on the right side of the formula above vanishes at the critical points of F, since $D_{\Delta_1}\Delta_2 \cdot F = \langle \nabla F, D_{\Delta_1}\Delta_2 \rangle = 0$ when $\nabla F = 0$. Hence we just need to evaluate

$$H^*_{J_0^*}(\Delta_1, \Delta_2) = \Delta_1 \cdot \Delta_2 \cdot J_0^*.$$

To proceed, we note that from Theorem 9.2 and the definition of J_0^* (recall that J_0 is constant) we get $\nabla J_0^* = [B, [B, M_0]$ where $M_0 := P_0 R_0 P_0$ and P_0, R_0 are as in the statement of the Proposition.

From the definition of the gradient and the inner product used, we have

$$\Delta_2 \cdot J_0^* = \langle \nabla J_0^*, \Delta_2 \rangle = \mathrm{tr}(\Omega_2 [B, M_0]).$$

Next we evaluate $D_1 \cdot D_2 \cdot J_0$, which is the derivative of $\mathrm{tr}(\Omega_2[B, M_0])$ in the

direction D_1. This is easily seen to be :

$$\begin{aligned} D_1 \cdot D_2 \cdot J_0 &= \mathrm{tr}([M_0, [B, \Omega_1]]\Omega_2) \\ &= \mathrm{tr}([M_0, \Omega_1][B, \Omega_2]). \end{aligned}$$

This concludes the proof of the proposition. ■

Recall that the signature of a bilinear form is a triplet of integers (n_+, n_-, n_0) with entries the number of positive, negative and zero eigenvalues of the bilinear form. The bilinear form is non-degenerate if $n_0 = 0$. The next step is to compute the signature of the bilinear form $H^* : (\Omega_1, \Omega_2) \to tr([M, \Omega_1][B, \Omega_2])$, which gives us the sign of the eigenvalues of the Hessian of J_0^0 at the critical points. We need to introduce the number of distinct partitions of an integer bounded by an integer: to this end, let n, k and l be positive integers. We call a partition of l into k parts a set of k (strictly) positive integers whose sum is l. We call the partition *distinct* if no integer in the sum is repeated. Finally, we say that the partition is bounded by n if not number in the sum is larger than n. We denote by $Q_n(k, l)$ the number of distinct partitions of k into l parts, bounded by n. For example $Q_4(9, 3) = 3$, since we have $9 = 3 + 3 + 3 = 4 + 4 + 1 = 4 + 3 + 2$. We have the following result:

Proposition 9.3. *The function J_0^* has $Q_n(m, n_+ + \frac{m(m+1)}{2})$ critical points with signatures $(n_+, n_-, 0)$, where (n_+, n_-) are all pairs for which $n_+ + n_- = d$, and $Q_n(k, l)$ is the number of ways to partition l into k parts no larger than n. Furthermore, exactly one of these critical points is a minimum. Furthermore, no critical points is degenerate generically for the parameters A, G, Q.*

The proof of Proposition 9.3 follows the lines of the proof of Theorem 3 in [21]. This proposition proves Theorem 9.1

REFERENCES

[1] J. Baillieul and P. Antsaklis, "Control and communication challenges in networked Real-Time systems," *Proceedings of the IEEE*, vol. 95, no. 1, pp. 9–28, 2007.

[2] Ali Jadbabaie, Jie Lin, and A. Stephen Morse, "Coordination of groups of mobile autonomous agents using nearest neighbor rules," *IEEE Transactions on Automatic Control*, vol. 48, no. 6, pp. 988–1001, 2003.

[3] S. Sundaram and C. N. Hadjicostis, "Distributed function calculation via linear iterative strategies in the presence of malicious agents," *IEEE Transactions on Automatic Control*, vol. 56, no. 7, pp. 1495–1508, 2011.

[4] A. Nedic and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Transactions on Automatic Control*, vol. 54, pp. 48–61, 2009.

[5] J. P. Desai, J. P. Ostrowski, and V. Kumar, "Modeling and control of formations of nonholonomic mobile robots," *Robotics and Automation, IEEE Transactions on*, vol. 17, no. 6, pp. 905–908, 2001.

[6] M. Mesbahi and M. Egerstedt, *Graph Theoretic Methods in Multiagent Networks*. Princeton University Press, 2010.

[7] J. Shamma, *Cooperative Control of Distributed Multi-Agent Systems*. Wiley, 2008.

[8] R. May, "Will a large complex system be stable?" *Nature*, vol. 238, no. 413-414, 1972.

[9] M.-A. Belabbas, "Sparse stable systems," *Systems & Control Letters*, vol. 62, pp. 981–987, 2013.

[10] A. Kirkoryan, "Sparse matrix spaces," M.S. thesis, University of Illinois at Urbana–Champaign, 2017.

[11] C. Johnson, J. Maybee, D. Olesky, and P. van den Driessche, "Nested sequences of principal minors potential stability," *Linear Algebra and its Applications*, vol. 262, pp. 243–257, 1997.

[12] A. Hurwitz, "Ueber die Bedingungen, unter welchen eine Gleichung nur Wurzeln mit negativen reellen Theilen besitzt," *Mathematische Annalen*, vol. 46, no. 2, pp. 273–284, 1895.

[13] P. Hall, "On representatives of subsets," *J. London Math. Soc.*, vol. 10, no. 1, pp. 26–30, 1935.

[14] D. Dummit and R. M. Foote, *Abstract Algebra*. Wiley, 2003.

[15] F. Gantmacher, *Theory of Matrices*. New York: Chelsea, 1960.

[16] B. Bollobás, *Random Graphs*, 2nd ed., ser. Cambridge Studies in Advanced Mathematics. Cambridge, UK: Cambridge University Press, 2001, no. 73.

[17] P. Erdős and A. Rényi, "On random matrices," *Publications of the Mathematical Institute of the Hungarian Academy of Sciences, Series A*, vol. VIII, no. 3, pp. 455–460, 1963.

[18] A. Schrijver, *Polyhedral Combinatorics and Combinatorial Optimization, vol. A*. Springer, 2003.

[19] R. Zippel, *Effective Polynomial Computation*. Springer, 1993.

[20] L. Valiant, "The complexity of computing the permanent," *Theoretical Computer Science*, vol. 8, pp. 189–201, 1979.

[21] M.-A. Belabbas, "Geometric methods for optimal sensor design," *Proc. R. Soc. A*, vol. 472, no. 2185, p. 20150312, 2016.

[22] F. Pasqualetti, S. Zampieri, and F. Bullo, "Controllability metrics, limitations and algorithms for complex networks," *IEEE Transactions on Control of Network Systems*, vol. 1, no. 1, pp. 40–52, 2014.

[23] X. Chen and M.-A. Belabbas, "Optimal actuator design for minimizing the worst-case control energy," *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 9991–9996, 2017, 20th IFAC World Congress.

[24] N. Darivandi, K. Morris, and A. Khajepour, "An algorithm for LQ optimal actuator location," *Smart Materials and Structures*, vol. 22, no. 3, p. 035001, 2013. [Online]. Available: http://stacks.iop.org/0964-1726/22/i=3/a=035001

[25] F. Fahroo and M. A. Demetriou, "Optimal actuator/sensor location for active noise regulator and tracking control problems," *Journal of Computational and Applied Mathematics*, vol. 114, no. 1, pp. 137–158, 2000.

[26] K. Morris, "Linear-quadratic optimal actuator location," *IEEE Transactions on Automatic Control*, vol. 56, no. 1, pp. 113–124, January 2011.

[27] H. Zhang, R. Ayoub, and S. Sundaram, "Sensor selection for kalman filtering of linear dynamical systems: Complexity, limitations and greedy algorithms," *Automatica*, vol. 78, pp. 202–210, 2017.

[28] G. Sagnol and R. Harman, *Optimal Designs for Steady-State Kalman Filters*. Cham: Springer International Publishing, 2015, pp. 149–157.

[29] N. K. Dhingra, M. R. Jovanović, and Z.-Q. Luo, "An admm algorithm for optimal sensor and actuator selection," in *Decision and Control (CDC), 2014 IEEE 53rd Annual Conference on*. IEEE, 2014, pp. 4039–4044.

[30] H. Singhal and G. Michailidis, "Optimal experiment design in a filtering context with application to sampled network data," *The Annals of Applied Statistics*, vol. 4, no. 1, pp. 78–93, 2010. [Online]. Available: http://www.jstor.org/stable/27801580

[31] M. Shamaiah, S. Banerjee, and H. Vikalo, "Greedy sensor selection: Leveraging submodularity," in *Decision and Control (CDC), 2010 49th IEEE Conference on*. IEEE, 2010, pp. 2572–2577.

[32] H. Kwakernaak and R. Sivan, *Linear Optimal Control Systems*. New York: Wiley–Interscience, 1972.

[33] S. Ditlevsen and A. Samson, "Introduction to stochastic models in biology," in *Stochastic biomathematical models*. Springer, 2013, pp. 3–35.

[34] D. Lemons and P. Langevin, *An Introduction to Stochastic Processes in Physics*, ser. Johns Hopkins Paperback. Johns Hopkins University Press, 2002.

[35] B. Øksendal, *Stochastic Differential Equations: An Introduction with Applications*, ser. Hochschultext / Universitext. Springer, 2003.

[36] L. Schwartz, *Analyse II. Calcul différentiel et équations différentielles*, ser. Enseignement des sciences. Paris: Hermann, 1992, no. 43.

[37] S. Helgason, *Differential Geometry, Lie Groups, and Symmetric Spaces*. American Mathematical Society, 2001.

[38] R. Brockett, "Differential geometry and the design of gradient algorithms," in *Proceedings of Symposia in Pure Mathematics*. American Mathematical Society, 1993, pp. 69–93.

[39] U. Helmke and J. B. Moore, *Optimization and dynamical systems*. London: Springer, 1994.

[40] D. F. Delchamps, "Analytic stabilization and the algebraic riccati equation," in *Proceedings of the 22nd IEEE Conference on Decision and Control, December 1983*, 1983, pp. 1396–1401.

[41] J. Dieudonné, *Treatise on Analysis*. New York: Academic Press, 1972, vol. III, translated from French.

[42] G. P. Bessa, L. Jorge, L. Mari, and J. F. Montenegro, "Spectrum estimates and applications to geometry," in *Topics in Modern Differential Geometry*, S. Haesen and L. Verstraelen, Eds. Paris: Atlantis Press, 2017.

[43] E. de Souza and S. P. Bhattacharyya, "Controllability, observability and the solution of $AX - XB = C$," *Linear Algebra and its Applications*, vol. 39, pp. 167–188, 1981.

[44] J. Jost, *Riemannian Geometry and Geometric Analysis*. Springer, 2011.

[45] R. Brockett, *Finite dimensional linear systems*. John Wiley & Sons, 1970.

[46] J. Geromel, "Convex analysis and global optimization of joint actuator location and control problems," *IEEE Transactions on Automatic Control*, vol. 34, no. 7, pp. 711–720, July 1989.

[47] T. Damm, *Rational Matrix Equations in Stochastic Control*, ser. Lecture Notes in Control and Information Sciences. Berlin: Springer–Verlag, 2004, no. 297.

[48] P. Benner and T. Damm, "Lyapunov equations, energy functionals, and model order reduction of bilinear and stochastic systems." *SIAM journal on control and optimization*, pp. 686–711, 2011.

[49] D. Delchamps, "Analytic feedback control and the algebraic Riccati equation," *IEEE Transactions on Automatic Control*, vol. 29, no. 11, pp. 1031–1033, 1984.

[50] S. Bittanti, A. J. Laub, and J. C. Willems, *The Riccati Equation*. Springer Science & Business Media, 2012.

[51] J. M. Lee, *Introduction to smooth manifolds*, 2nd ed., ser. Graduate texts in mathematics. New York, NY [u.a.]: Springer, 2012.

[52] M.-A. Belabbas, "Geometric methods for optimal sensor design," *Proceedings of the Royal Society, Series A Math Phys Eng Sci*, vol. 472, no. 2185, p. 20150312, Jan 2016.

[53] R. Brockett, "Dynamical systems that sort lists, solve linear programming problems and diagonalize symmetric matrices," *Linear Algebra Appl*, vol. 146, pp. 79–91, 1991.