University of California
Santa Barbara

Network systems: Social networks, Epidemics, Optimization and Contraction Theory

A dissertation submitted in partial satisfaction
of the requirements for the degree

Doctor of Philosophy
in
Electrical and Computer Engineering

by

Pedro A. Cisneros

Committee in charge:

Professor Francesco Bullo, Chair
Professor Ambuj Singh
Professor Jason Marden
Professor Joao Hespanha

June 2021

The dissertation of Pedro A. Cisneros is approved.

Professor Ambuj Singh

Professor Jason Marden

Professor Joao Hespanha

Professor Francesco Bullo, Committee Chair

April 2021

Network systems: Social networks, Epidemics, Optimization and Contraction Theory

by

Pedro A. Cisneros

Para mi familia: Pedro, Lourdes, Sergio y abuelos.

Acknowledgements

I am deeply grateful to Prof. Francesco Bullo, my advisor, for all the help and good advising during my PhD studies. Without him, this accomplishment could not be possible. I am also grateful to my family for their unconditional support, especially my parents and brother. I am also grateful to my friend Mary. I am also grateful to all the professors, students and collaborators whose interactions and help have been very helpful during my studies, with a special mention to Dr. Saber Jafarpour, Prof. Ambuj Singh, and Prof. Alex Petersen. For the funding, I am also grateful for the support from the ARO MURI project, grant number W911NF-15-1-0577

I am grateful to God for all the blessings throughout my studies. Gloria in excelsis Deo, in saecula saeculorum.

Curriculum Vitæ

Pedro A. Cisneros

Education

2021	Ph.D. in Electrical and Computer Engineering, University of California, Santa Barbara.
2020	M.A. in Mathematical Statistics, University of California, Santa Barbara.
2018	M.S. in Electrical and Computer Engineering, University of California, Santa Barbara.

Published and submitted work:

P. Cisneros-Velarde, K Chan, F Bullo. "Polarization and fluctuations in signed social networks." **IEEE Transactions on Automatic Control**. 2020.
P. Cisneros-Velarde, N.E. Friedkin, A.V. Proskurnikov, F. Bullo. "Structural Balance via Gradient Flows over Signed Graphs." **IEEE Transactions on Automatic Control**. 2020.
P. Cisneros-Velarde, A. Petersen, S.Y. Oh. "Distributionally Robust Formulation and Model Selection for the Graphical Lasso." **AISTATS**, 2020.
P. Cisneros-Velarde, F. Bullo. "Signed Network Formation Games and Clustering Balance." **Dynamic Games and Applications**. 2020.
P. Cisneros-Velarde, F. Bullo. "Distributed Wasserstein Barycenters via Displacement Interpolation." Preprint. 2020.
P. Cisneros-Velarde, S. Jafarpour, F Bullo. "Contraction Theory for Dynamical Systems on Hilbert Spaces." Preprint. 2020.
P. Cisneros-Velarde, F Bullo. "Multi-group SIS Epidemics with Simplicial and Higher-Order Interactions." Preprint. 2020.
P. Cisneros-Velarde, S Jafarpour, F Bullo. "Distributed and time-varying primal-dual dynamics via contraction analysis." Preprint. 2020.
P. Cisneros-Velarde, F Bullo. "A Network Formation Game for the Emergence of Hierarchies." Preprint. 2020.
S. Jafarpour, **P. Cisneros-Velarde**, F Bullo. "Weak and Semi-Contraction Theory with Application to Network Systems." **IEEE Transactions on Automatic Control**. 2021
W. Mei, **P. Cisneros-Velarde**, G. Chen, N.E. Friedkin, F. Bullo. "Dynamic social balance and convergent appraisals via homophily and influence mechanisms." **Automatica**. 2019.

Abstract

Network systems: Social networks, Epidemics, Optimization and Contraction Theory

by

Pedro A. Cisneros

In this thesis, I will first present mathematical models that explain the evolution of interpersonal relationships in a social network, represented by a signed graph, converging to structures that have a long history in sociology - namely, structural and clustering balance. Then, I will present a simple model for the evolution of opinions over signed graphs, including the aforementioned special structures. Then, I will present an important phenomenon that occurs on the susceptible-infected-susceptible (SIS) model of epidemics: the emergence of a new epidemic domain of bistability when higher-order interaction among individuals are considered on the contact network. Then, I will present an algorithm for the computation of Wasserstein barycenters, and show a connection with the theory of opinion dynamics. Finally, the last part of this thesis is devoted to the study and application of contraction theory, an important tool that certifies incremental stability. We study its expansion to dynamical systems on Hilbert spaces, as well as its application to various optimization problems and settings.

Contents

Chapter 1

Structural Balance via Gradient Flows over Signed Graphs

1.1 Introduction

Problem description and motivation

Signed graphs represent networked systems with interactions classified as positive or negative, e.g., cooperation or antagonism, promotion or inhibition, attraction or repulsion. Such graphs naturally arise in diverse fields, e.g., political science [88], communication studies [103] and biology [106]. In sociology [69, 62], they are used to represent friendly or antagonistic relationships, whereby signed edges may be interpreted as interpersonal sentiment appraisals. In the work by Heider [76], each individual appraises all other individuals either positively (friends, allies) or negatively (enemies, rivals). Heider postulated four famous axioms: (i) "the friend of a friend is a friend," (ii) "the enemy of a friend is an enemy," (iii) "the friend of an enemy is an enemy," and (iv) "the enemy of an enemy is a friend." Violations of these axioms lead to cognitive tensions and disso-

nances that the individuals strive to resolve; in this sense, Heider's axioms are consistent with the general theory of cognitive dissonance [67]. A signed network satisfying Heider's axioms is called *structurally balanced* and can have only two possible configurations: either all of its members have positive relationships with each other and become a unique faction, or there exist two factions in which members of the same faction are friends but enemies with every other member in the other faction. We refer to [69, 62] for textbook treatment and to [177] for a recent comprehensive survey.

Whereas Heider's theory describes the qualitative emergence of structural balance as the result of tension-resolving cognitive mechanisms, it does not provide a quantitative description of these mechanisms and dynamic models explaining the emergence of balance. The aim to fill this gap has given rise to the important research area of *dynamic structural balance*. The Kułakowski et al. [97] model postulates an influence process, whereby any individual i updates her appraisal of individual j based on what others positively or negatively think about j. The Traag et al. [164] model postulates a homophily process, whereby any individual i updates her appraisal of j according to how much she agrees with j on the appraisals of their common acquaintances. Both models explain convergence to structural balance under certain assumptions on the initial state (see below for more information). Remarkably, both models assume the existence of so-called *self-appraisals* (loops in the signed graph) that strongly influence the system dynamics. Self-appraisals can be interpreted as individuals' positive or negative opinions of themselves.

A second line of research, consistent with dissonance theory, has focused on formulating social balance via appropriate energy functions. The work [120] proposes an energy function for binary appraisal matrices with global minima that represent structurally stable configurations; it is argued that a dynamic structural balance model should aim to navigate through this energy landscape and look for its minima. Some models

(e.g., [14, 15]) were designed precisely to achieve this task. The work [63] computes a distance to balance via a combinatorial optimization problem, inspired by Ising models.

The purpose of this paper is threefold. First, we aim to propose a more parsimonious model of the influence process establishing structural balance, that is, a model without self-appraisal weights. Our argument for dropping these variables is that balance theory axioms do not include self-appraisals, and the inclusion of such appraisals amounts to an additional assumption and introduces unnecessary complexities. Second, we aim to connect the literature on dynamic structural balance with the literature treating social balance as an optimization problem. Finally, we aim to emphasize through numerical simulations that our parsimonious model does not suffer from a key limitation present in the Kułakowski et al. model, namely that the Kułakowski et al. model cannot predict the emergence of structural balance from asymmetric initial configurations.

Further comments on the state of the art

We now present a summary of the current literature on dynamic structural balance. Historically, the first models appeared in the physics community [14, 15, 147]. These models borrowed some concepts from statistical physics and had the particularity of assuming that the appraisals between individuals are binary valued (either $+1$ or -1). At the same time, they rely on hard-wired random mechanisms for the asynchronous updates of the appraisals that lack a sociological insightful interpretation.

Another type of proposed models is based on discrete- and continuous-time dynamical systems with real-valued appraisals. The seminal models of this kind are due to Kułakowski et al. [97] (later analyzed more formally by [119]) and Traag et al. [164]. Models with real-valued appraisals capture not only signs, but also magnitudes of positive or negative sentiments. All these models adopt synchronous updating and stipulate sociological meaningful rules for the updating of appraisals, based on either influ-

ence or homophily processes. The following facts are known about the Kułakowski et al. influence-based and the Traag et al. homophily-based models: the set of well-behaved initial conditions that lead the social network towards social balance for the first model is a subset of the set of normal matrices, while the second model can work under generic initial conditions. Similar results are obtained by [122] for two discrete-time models based on influence and homophily respectively: influence-based processes do not perform well under generic initial conditions (in contrast to the homophily-based processes). Finally, only the models proposed in [122] and a variation of the model by Kułakowski et al. proposed in the early work [97], have a bounded evolution of appraisals, whereas the others have finite escape time.

Recent work has also started to focus on dynamic models for other relevant configuration of signed graphs, e.g., configurations that satisfy only a subset of the four Heider's axioms. The work [70] provides a parsimonious model explaining the emergence of a generalized version of structural balance from any initial configuration; this model is based on an influence process of positive contagion whereby influence is accorded only to positively-appraised individuals. A second model in this area is proposed by [92]. Finally, there has been a third type of models that propose the emergence of structural balance or other generalized balance structures for undirected graphs from a game theoretical perspective [167, 115, 43].

Contributions

First of all, we contribute by proposing two new dynamic models that do not adopt the long-standing assumption of self-appraisals and describe the evolution of signed networks without self-loops. We argue that the introduction of self-weights is poorly justified and that a model without them is a more faithful representation of Heider's theory. The first model, called the *pure-influence model*, is a modification of the classic model by

Kułakowski et al. which is obtained by eliminating self-appraisals (and thus reducing the system's dimension). Analysis of its convergence properties reduces to the analysis of our second model, called the *projected pure-influence model*, which arises as a projection of the first model onto the unit sphere. This second model has a self-standing interest, since it enjoys bounded evolution of the appraisals, while the first model shares the finite escape time property of the classic model by Kułakowski et al.

Our second contribution is to build a bridge between dynamic structural balance and structural balance as an optimization problem. We propose an energy function inspired by [120], namely the *dissonance function*, which measures the degree at which Heider's axioms are violated among the individuals of a social network. We show that this energy function has global minima that correspond to signed graphs satisfying structural balance in the case of real-valued appraisals (restricted on the unit sphere). Moreover, we show that our (projected) pure-influence model is the gradient system of the dissonance function in the case of undirected signed graphs, and hence the critical points of the dissonance function are the equilibria of our dynamical system. Thus, we establish a novel connection between dynamic structural balance and the characterization of structural balance as the minima of an energy function. Remarkably, our derivations show that this property of our models is enabled by the elimination of self-appraisals. Thus, the models contributed in this paper may be considered as both an interpersonal influence process and an extremum seeking dynamics for the dissonance function.

Our third and more detailed contribution is the mathematical analysis of the projected pure-influence model in the cases where the initial appraisal matrix is symmetric. In particular, we provide a complete characterization of the critical points of the dissonance function (i.e., the equilibrium points of the projected pure-influence model). This characterization relies upon a special submanifold of the Stiefel manifold and its properties. Along with the characterization of the critical points, we analyze their local stability

properties and provide some results on convergence towards structural balance.

Our final contribution is a Monte Carlo numerical study of the convergence of our models to structural balance under generic initial conditions in both the symmetric and the asymmetric case. For the symmetric case, our numerical result is comparable to, but stronger than, what has already been proved for the Kułakowski et al. model: our models converge to structural balance under generic symmetric initial conditions. One key advantage of our models, as compared with those by Kułakowski et al., is that convergence to structural balance emerges under generic asymmetric initial conditions. Based on these numerical results, we formulate relevant conjectures.

Paper organization

Section 6.2 presents preliminary concepts. Section 1.3 presents our models and shows they are gradient flows. Section 1.4 and Section 1.5 contain an analysis of equilibria and important convergence results, respectively. Section 1.6 contains numerical results and conjectures. Finally, Section 1.7 contains some concluding remarks.

1.2 Preliminaries

1.2.1 Signed weighted digraphs

Given an $n \times n$ matrix $X = (x_{ij})$ with entries taking values in $[-\infty, \infty]$, let $G(X)$ denote the signed directed graph where the directed edge $i \to j$ exists if and only if $x_{ij} \neq 0$, and x_{ij} represents its signed weight. The directed graph $G(X)$ is complete if X has no zero entries, except for the main diagonal. $G(X)$ has no self-loops if and only if X has zero diagonal entries. Let x_{i*} denote the ith row of the matrix X and x_{*i} the ith column of the matrix X. Let $\text{sign}(X) = (\text{sign}(x_{ij}))$, where $\text{sign} : [-\infty, \infty] \to \{-1, 0, +1\}$

is as usual

$$\operatorname{sign}(x) = \begin{cases} -1, & \text{if } x < 0, \\ 0, & \text{if } x = 0, \\ +1, & \text{if } x > 0. \end{cases}$$

Given a sequence $a_1, \ldots, a_n$, let $B = \operatorname{diag}(a_1, \ldots, a_n)$ denote the diagonal $n \times n$ matrix (b_{ij}), where $b_{ii} = a_i$ and $b_{ij} = 0$ for $i \neq j$. For an $n \times n$ matrix X, define $\operatorname{diag}(X) = \operatorname{diag}(x_{11}, \ldots, x_{nn})$. For a vector $v \in \mathbb{R}^n$, define $\operatorname{diag}(v) = \operatorname{diag}(v_1, \ldots, v_n)$. Let $\mathbb{0}_n$ denote the $n \times 1$ vector of zeros, and $\mathbb{0}_{n \times n}$ the $n \times n$ matrix with zero entries.

Let $\succ$ and $\prec$ denote "entry-wise greater than" and "entry-wise less than," respectively.

A *triad* (if it exists) is a cycle between three nodes in $G(X)$. The *sign* of a triad is defined by the sign of the product of the weights composing a triad. For example, the triad $i \to j \to k \to i$ has sign $\operatorname{sign}(x_{ij}x_{jk}x_{ki})$.

A real-valued matrix Z is *irreducible* if its graph $G(Z)$ is strongly connected (a directed path between every two nodes exists) and *reducible* otherwise. If Z is reducible, a permutation matrix P exists such that the matrix

$$PZP^\top = \begin{bmatrix} Z_1 & * & \ldots & * \\ 0 & Z_2 & \ldots & * \\ \vdots & & & \\ 0 & & & Z_k \end{bmatrix}$$

is upper-triangular with irreducible blocks Z_i (some of them can be 1×1 matrices). If $Z = Z^\top$, the latter matrix is block-diagonal matrix $PZP^\top = \operatorname{diag}(Z_1, \ldots, Z_k)$ and the graphs $G(Z_i)$ are the *connected components* of the graph $G(Z)$.

1.2.2 Sets of matrices and the Frobenius inner product

Given two matrices $A, B \in \mathbb{R}^{n\times n}$, their Frobenius inner product is defined by $\langle\langle A, B\rangle\rangle_F = \text{trace}(B^\top A)$; the inducednorm is $\|A\|_F = \sqrt{\langle\langle A, A\rangle\rangle_F}$. Some important properties for the trace operator are: $\text{trace}(A) = \text{trace}(A^\top)$, $\text{trace}(AB) = \text{trace}(BA)$, and, for all $d \in \mathbb{N}$, $\text{trace}(A^d) = \sum_{i=1}^n \lambda_i^d$ where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A.

Let $\mathbb{R}^{n\times n}_{\text{zero-diag}}$ be the set of $n \times n$ real matrices with zero diagonal entries, and $\mathbb{R}^{n\times n}_{\text{zero-diag,symm}}$ be the set of symmetric matrices belonging to $\mathbb{R}^{n\times n}_{\text{zero-diag}}$. Let $\mathbb{S}^{n\times n}$ be the unit sphere in $\mathbb{R}^{n\times n}$, that is $A \in \mathbb{S}^{n\times n}$ if and only if $A \in \mathbb{R}^{n\times n}$ with $\|A\|_F = 1$. Similarly, we define the sets $\mathbb{S}^{n\times n}_{\text{zero-diag}} = \mathbb{R}^{n\times n}_{\text{zero-diag}} \cap \mathbb{S}^{n\times n}$ and $\mathbb{S}^{n\times n}_{\text{zero-diag,symm}} = \mathbb{R}^{n\times n}_{\text{zero-diag,symm}} \cap \mathbb{S}^{n\times n}$.

Let $\mathbb{R}^{n\times n}_{\text{diag}}$ be the set of all real diagonal matrices and $\mathbb{R}^{n\times n}_{\text{sk-symm}}$ be the set of all skew-symmetric matrices. Then, we have the following orthogonal decomposition of $\mathbb{R}^{n\times n}$ equipped with the Frobenius inner product:

$$\mathbb{R}^{n\times n} = \mathbb{R}^{n\times n}_{\text{sk-symm}} \oplus \mathbb{R}^{n\times n}_{\text{zero-diag,symm}} \oplus \mathbb{R}^{n\times n}_{\text{diag}}. \tag{1.1}$$

1.2.3 A review on structural balance

Throughout the paper we deal with social networks composed of $n \geq 3$ individuals, although the definition of structural balance (Definition 1.2.3) is formally applicable to the case of degenerate networks with $n = 1$ or $n = 2$ nodes.

Definition 1.2.1 (Appraisal matrix and network) *We let the entry x_{ij} of the matrix $X \in \mathbb{R}^{n\times n}$ denote the appraisal (or qualitative evaluation) held by individual i of individual j. The sign of x_{ij} indicates if the relationship is positive ($+1$), negative (-1) or of indifference (0). The magnitude of x_{ij} indicates the strength of the relationship. x_{ii} can be interpreted as i's self-appraisal. We call X the* appraisal matrix, *and $G(X)$ the* appraisal network.

Definition 1.2.2 (Heider's axioms and social balance notions) *The* Heider's axioms *are*

H1) A friend of a friend is a friend,

H2) An enemy of a friend is an enemy,

H3) A friend of an enemy is an enemy,

H4) An enemy of an enemy is a friend.

An appraisal network $G(X)$ is structurally balanced in Heider's sense*, if it is complete and satisfies axioms H1)-H4).*

Consider a complete appraisal network $G(X)$. We call a *faction* any group of agents whose members positively appraise each other. We say two factions are *antagonistic* if every representative from one faction negatively appraise every representative of the other faction. It can be shown [76, 74, 38] that Heider's structural balance condition for $G(X)$ with $n \geq 3$ nodes holds if and only if either the individuals constitute a single faction or can be partitioned into two antagonistic factions. The possession of the latter property may thus be considered as an alternative definition of structural balance (and is formally applicable to graphs without triads).

Definition 1.2.3 (Structural balance) *A complete appraisal network $G(X)$ is said to satisfy* structural balance*, if $G(X)$ is composed by one faction or two antagonistic factions; or, whenever $n \geq 3$, equivalently, that all triads are positive, i.e., $x_{ij}x_{jk}x_{ki} > 0$ for any different $i, j, k \in \{1, \ldots, n\}$.*

Notice that a structurally balanced graph is always sign-symmetric: $\text{sign}(x_{ij}) = \text{sign}(x_{ji})$ for any $i \neq j$. For simplicity we will say that a matrix X *corresponds to* structural balance whenever $G(X)$ satisfies structural balance.

1.3 Proposed models and representation as gradient flows

In this section we propose our models defining them over the set of symmetric matrices. We postponed the general asymmetric setting to Section 1.6.

1.3.1 Pure-influence model

We propose our new dynamic model solely based on interpersonal appraisals.

Definition 1.3.1 (Pure-influence model) *The* pure-influence model *is a system of differential equations on the set of zero-diagonal matrices* $\mathbb{R}^{n\times n}_{\text{zero-diag}}$ *defined by*

$$\dot{x}_{ij} = \sum_{\substack{k=1\\k\neq i,j}}^{n} x_{ik}x_{kj}, \tag{1.2}$$

for any $i, j \in \{1, \dots, n\}$ *and* $i \neq j$*. Here* x_{ij}*,* $i \neq j$*, are the off-diagonal entries of a zero-diagonal matrix* $X \in \mathbb{R}^{n\times n}_{\text{zero-diag}}$*. In equivalent matrix form, the previous equations read:*

$$\dot{X} = X^2 - \operatorname{diag}(X^2), \qquad X(0) \in \mathbb{R}^{n\times n}_{\text{zero-diag}}. \tag{1.3}$$

We interpret X as the interpersonal appraisal matrix. While system (1.2) does not define the evolution of self-appraisals, the matrix reformulation (1.3) ensures $\operatorname{diag}(\dot{X}) = \mathbb{0}_{n\times n}$ and, since $X(0) \in \mathbb{R}^{n\times n}_{\text{zero-diag}}$ means $\operatorname{diag}(X(0)) = \mathbb{0}_{n\times n}$, we have $\operatorname{diag}(X(t)) = \mathbb{0}_{n\times n}$ for all positive times t.

Our model is a modification of the classical model proposed by Kułakowski et al. [97], where self-appraisals play a crucial role in the dynamics of the interpersonal appraisals.

Definition 1.3.2 (Kułakowski et al. model) *The* Kułakowski et al. model *is a system of differential equations on the state space* $\mathbb{R}^{n\times n}$ *defined by*

$$\dot{x}_{ij} = \sum_{k=1}^{n} x_{ik}x_{kj} = x_{ij}(x_{ii} + x_{jj}) + \sum_{\substack{k=1\\k\neq i,j}}^{n} x_{ik}x_{kj}, \tag{1.4a}$$

$$\dot{x}_{ii} = x_{ii}^2 + \sum_{\substack{k=1\\k\neq i}}^{n} x_{ik}x_{ki}, \tag{1.4b}$$

for any $i \neq j \in \{1,\dots,n\}$. *In equivalent matrix form, the previous equations read:* $\dot{X} = X^2$.

Remark 1.3.1 (The problem with self-appraisals) *The introduction of self-appraisals in model* (1.4) *is objectionable on several grounds. The first conceptual problem is that self-appraisals are not considered in any definition of structural balance in the social sciences. Heider's axioms in Definition 1.2.2 do not take into account self-appraisals: social balance is a function of only interpersonal appraisals. Moreover, once self-appraisals are introduced, one needs to postulate why and how self-appraisals affect interpersonal appraisals, i.e., justify the choice of the first addendum for the right hand side of* (1.4a). *Finally, one needs to postulate how they evolve, i.e., justify the choice for the right hand side of* (1.4b). *In summary, the pure influence model* (1.2) *avoids these difficulties and stays closer to the foundations of structural balance, in which individuals are attending only to interpersonal appraisals. Even though* $\dot{X} = X^2$ *may appear mathematically simpler or more elegant than* $\dot{X} = X^2 - \operatorname{diag}(X^2)$, *we believe the latter model is actually more parsimonious, lower dimensional, and more faithful to Heiders' axioms.*

One easily notices the following important property of the pure-influence model (1.3): the right-hand side is an analytic function of X so that the equation enjoys (local) existence and uniqueness of the solutions. A second property is that, if $X(0) = X(0)^\top$,

then $X(t) = X(t)^\top$ for all subsequent times. This implies that the pure-influence model is well defined over the set of symmetric (zero diagonal) matrices $\mathbb{R}^{n\times n}_{\text{zero-diag,symm}}$.

1.3.2 Dissonance function

We introduce and study the properties of a useful *dissonance function* that summarize the total amount of cognitive dissonances [67] among the members of a social network due to the lack of satisfaction of Heider's axioms. Recall that, according to Definition 1.2.3, a triad $i \to j \to k \to i$ satisfies the axioms if and only if $x_{ij}x_{jk}x_{ki} > 0$.

Definition 1.3.3 (Dissonance function) *The* dissonance function $\mathcal{D} : \mathbb{R}^{n\times n}_{\text{zero-diag}} \to \mathbb{R}$ *is*

$$\mathcal{D}(X) = -\sum_{\substack{i,j,k=1\\ i\neq j, j\neq k, k\neq i}}^{n} x_{ij}x_{jk}x_{ki} = -\operatorname{trace}(X^3) = -\sum_{i=1}^{n}\lambda_i^3, \tag{1.5}$$

where $\{\lambda_i\}_{i=1}^n$ *is the set of eigenvalues of* X.

We plot $\mathcal{D}$ in a low-dimensional setting in Figure 1.1.

Energy landscapes in social balance theory are studied in [120, 63]. Our proposed dissonance function is the extension to $\mathbb{R}^{n\times n}_{\text{zero-diag}}$ of the energy function proposed by [120] for the setting of binary-valued symmetric appraisal matrices. For binary-valued appraisals, the global minima of $\mathcal{D}$ correspond to networks that satisfy structural balance, since all triads are positive (Definition 1.2.3). Thus, $\mathcal{D}$ naturally measures to which extent Heider's axioms are violated in a complete graph.

Lemma 1.3.2 (Properties of the dissonance function) *Consider the dissonance function* $\mathcal{D}$ *and pick* $X \in \mathbb{R}^{n\times n}_{\text{zero-diag}}$. *Then*

(i) $\mathcal{D}$ *is analytic and attains its maximum and minimum values on any compact matrix subset of* $\mathbb{R}^{n\times n}_{\text{zero-diag}}$,

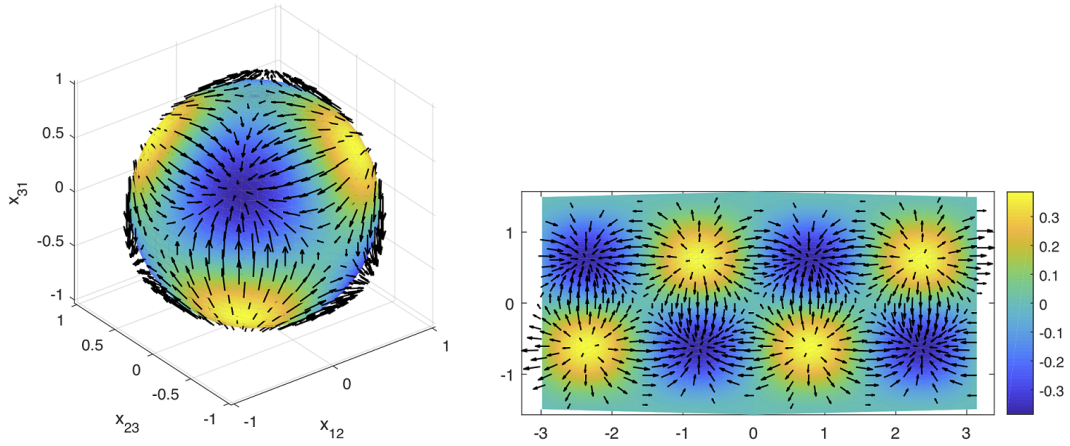


Figure 1.1: For $n = 3$, an arbitrary symmetric unit-norm zero-diagonal matrix $X \in \mathbb{S}^{n\times n}_{\text{zero-diag,symm}}$ is described by (x_{12}, x_{23}, x_{31}) with these coordinates living in the sphere $x_{12}^2 + x_{23}^2 + x_{31}^2 = 1$. In the upper figure, we plot this sphere with a heatmap, with dark blue being the lowest value and light yellow the largest value, according to the evaluation of the dissonance function $\mathcal{D}(X)$. The function has four global minima corresponding to the four possible configurations of $G(X)$ satisfying structural balance, and we can qualitatively appreciate the convergence of solution trajectories to these minima in the superimposed vector field on the sphere. The lower figure is a stereographic projection of the upper figure.

(ii) *if $G(X)$ satisfies structural balance, then $\mathcal{D}(X) < 0$,*

(iii) $\mathcal{D}(X) = \mathcal{D}(X^\top)$,

(iv) $\mathcal{D}(X) = -\langle\langle X^2, X^\top \rangle\rangle_F$.

Additionally, if $\|X\|_F = 1$, that is, $X \in \mathbb{S}^{n\times n}_{\text{zero-diag}}$, then

(v) $-1 \leq \mathcal{D}(X) \leq 1$.

Proof: Here we show only property (v), since the other properties follow easily from

the definition of $\mathcal{D}$. We note:

$$\begin{aligned}\left\|X^2\right\|_F^2 &= \sum_{i,j=1}^{n}(X^2)_{ij}^2 = \sum_{i,j=1}^{n}(X_{i*}X_{*j})^2 \\ &\leq \sum_{i,j=1}^{n}\|X_{i*}\|_2^2\|X_{*j}\|_2^2 = \Big(\sum_{i=1}^{n}\|X_{i*}\|_2^2\Big)\Big(\sum_{j=1}^{n}\|X_{*j}\|_2^2\Big) \\ &= \Big(\sum\nolimits_{i,k=1}^{n} x_{ik}^2\Big)^2 = \|X\|_F^2 = 1.\end{aligned}$$

Now, note that the Frobenius norm on the set of matrices coincides with the Euclidean norm of a single vector obtained by stacking the column vectors of the matrix. Then, by the Cauchy-Schwarz inequality applied to the inner-product $\langle\!\langle \cdot, \cdot \rangle\!\rangle_F$, it follows that: $|D(X)| = |\langle\!\langle X^2, X \rangle\!\rangle_F| \leq \|X^2\|_F \|X\|_F \leq (\|X\|_F)^3 \leq 1$ when $\|X\|_F \leq 1$. ■

1.3.3 Transcription on the unit sphere and the projected pure-influence model

We start by noting a simple fact. Given a trajectory $X : \mathbb{R}_{\geq 0} \to \mathbb{R}^{n\times n}_{\text{zero-diag}} \setminus \{\mathbb{0}_{n\times n}\}$, there exist unique trajectories $\eta : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ and $Z : \mathbb{R}_{\geq 0} \to \mathbb{S}^{n\times n}_{\text{zero-diag}}$ such that $X(t) = \eta(t)\mathcal{Z}(t)$, where $\eta(t) = \|X(t)\|_F$ and $\mathcal{Z}(t) = X(t)/\|X(t)\|_F$.

Theorem 1.3.3 (Transcription of the pure-influence model) *The pure-influence model* (1.2) *with initial conditions in* $\mathbb{R}^{n\times n}_{\text{zero-diag,symm}}$ *can be expressed as the following system of differential equations:*

$$\begin{aligned}\dot{\mathcal{Z}} &= \eta\mathcal{P}_{\mathcal{Z}^\perp}(\mathcal{Z}^2 - \operatorname{diag}(\mathcal{Z}^2)) \\ &= \eta(\mathcal{Z}^2 - \operatorname{diag}(\mathcal{Z}^2) + \mathcal{D}(\mathcal{Z})\mathcal{Z}), \qquad (1.6a)\\ \dot{\eta} &= -\mathcal{D}(\mathcal{Z})\eta^2, \qquad (1.6b)\end{aligned}$$

where $\eta : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ *and* $\mathcal{Z} : \mathbb{R}_{\geq 0} \to \mathbb{S}^{n\times n}_{\text{zero-diag,symm}}$. *Here* $\mathcal{P}_{\mathcal{Z}^\perp}$ *is the orthogonal projection onto* $\text{span}\{\mathcal{Z}\}^\perp$ *in the vector space of square matrices with the Frobenius inner product.*

Proof: Since $\dot{X} = \dot{\eta}\mathcal{Z} + \eta\dot{\mathcal{Z}}$ and $X^2 - \text{diag}(X^2) = \eta^2\left(\mathcal{Z}^2 - \text{diag}(\mathcal{Z}^2)\right)$, equation (1.3) can be written as

$$\dot{\eta}\mathcal{Z} + \eta\dot{\mathcal{Z}} = \eta^2\left(\mathcal{Z}^2 - \text{diag}(\mathcal{Z}^2)\right). \tag{1.7}$$

Differentiating the equality $\|\mathcal{Z}(t)\|_F^2 = \langle\!\langle \mathcal{Z}(t), \mathcal{Z}(t)\rangle\!\rangle_F = 1$, one shows that $\langle\!\langle \mathcal{Z}(t), \dot{\mathcal{Z}}(t)\rangle\!\rangle_F = 0$, that is, $\mathcal{Z}(t) \perp \dot{\mathcal{Z}}(t)$. Computing the Frobenius inner product with $\mathcal{Z}(t)$ on both sides of (1.7), equation (1.6b) is immediate:

$$\dot{\eta} = \eta^2\langle\!\langle \mathcal{Z}(t), \mathcal{Z}^2(t) - \text{diag}(\mathcal{Z}^2(t))\rangle\!\rangle_F = \eta^2\langle\!\langle \mathcal{Z}(t), \mathcal{Z}^2(t)\rangle\!\rangle_F = -\mathcal{D}(\mathcal{Z}(t))\eta^2. \tag{1.8}$$

where we have used the fact that $\mathcal{Z}(t)$ is symmetric, and that $\text{diag}(\mathcal{Z}(t)) = \mathbb{0}_{n\times n}$ and hence $\langle\!\langle \mathcal{Z}(t), \text{diag}(\mathcal{Z}^2(t))\rangle\!\rangle_F = \text{trace}(\mathcal{Z}(t)^\top \text{diag}(\mathcal{Z}^2(t))) = 0$. Substituting (1.8) into equation (1.7), one arrives at $\dot{\mathcal{Z}} = \eta\left(\mathcal{Z}^2 - \text{diag}(\mathcal{Z}^2) + \mathcal{D}(\mathcal{Z})\right)$.

Given $Y \in \mathbb{R}^{n\times n}$, let $\mathcal{P}_{\mathcal{Z}}(Y) = \langle\!\langle Y, \mathcal{Z}\rangle\!\rangle_F \mathcal{Z}$, i.e., $\mathcal{P}_{\mathcal{Z}}$ is the orthogonal projection operator onto the linear space spanned by $\mathcal{Z}$; and let $\mathcal{P}_{\mathcal{Z}^\perp}(Y) = Y - \mathcal{P}_{\mathcal{Z}}(Y) = Y - \langle\!\langle Y, \mathcal{Z}\rangle\!\rangle_F \mathcal{Z}$ be the orthogonal projection onto the space perpendicular to the linear space spanned by $\mathcal{Z}$. Then, we observe that $\mathcal{P}_{\mathcal{Z}^\perp}(\mathcal{Z}) = 0$ and $\mathcal{P}_{\dot{\mathcal{Z}}^\perp}(\mathcal{Z}) = \dot{\mathcal{Z}}$. Using these results, we apply $\mathcal{P}_{\mathcal{Z}^\perp}$ to both sides of (1.7) and obtain $\dot{\mathcal{Z}} = \eta\mathcal{P}_{\mathcal{Z}^\perp}(\mathcal{Z}^2 - \text{diag}(\mathcal{Z}^2))$. This concludes the proof of equations (1.6). ■

In what follows, we are primarily interested in the dynamics (1.6a), describing the behavior of the bounded component $\mathcal{Z}(t)$. From Lemma 1.8.1 we observe that η is a time-scale change for (1.6a) and so, for our convenience, we get rid of it and obtain the following dynamical system on the unit sphere.

Definition 1.3.4 (Projected pure-influence model) *The* projected pure-influence model

is a system of differential equations on the manifold $\mathbb{S}^{n\times n}_{\text{zero-diag,symm}}$ *defined by*

$$\dot{Z} = Z^2 - \operatorname{diag}(Z^2) + \mathcal{D}(Z)Z. \tag{1.9}$$

Given a solution $Z(t)$ to (1.9) with initial condition $Z(0)$, Lemma 1.8.1 in the Appendix shows that $Z(t)$ is a time-scaled version of a solution $\mathcal{Z}(t)$ to (1.6a) with initial condition $\mathcal{Z}(0) = Z(0)$, where η in (1.6b) can have any positive initial condition. Therefore, there is a solution $X(t)$ to (1.3) that is both a scaled and time-scaled version of $Z(t)$.

Similarly, projecting onto the unit sphere leads to a new model based on the Kułakowski et al. model.

Definition 1.3.5 (Projected Kułakowski et al. model) *The* projected Kułakowski et al. model *is a system of differential equations on the manifold* $\mathbb{S}^{n\times n}_{\text{zero-diag,symm}}$ *defined by*

$$\dot{Z}(t) = Z^2 + \mathcal{D}(Z)Z. \tag{1.10}$$

1.3.4 Pure-influence is the gradient flow of the dissonance function

We now let $\operatorname{grad}\mathcal{D}$ denote the gradient vector field of the dissonance function $\mathcal{D}$ on the manifold $\mathbb{R}^{n\times n}_{\text{zero-diag}}$ equipped with the Riemannian metric tensor $\langle\!\langle \cdot, \cdot \rangle\!\rangle_F$. We also let $\mathcal{D}\big|_{\mathbb{S}^{n\times n}_{\text{zero-diag,symm}}}$ denote the restriction of $\mathcal{D}$ onto the manifold $\mathbb{S}^{n\times n}_{\text{zero-diag,symm}}$. We now present the first of our main results.

Theorem 1.3.4 (The pure-influence models are gradient flows) *Consider the pure-influence model* (1.2) *with* $X(0) \in \mathbb{R}^{n\times n}_{\text{zero-diag,symm}}$ *and the projected pure-influence model* (1.9) *with* $Z(0) \in \mathbb{S}^{n\times n}_{\text{zero-diag,symm}}$. *Then*

(i) $t \mapsto X(t)$ remains in the set $\mathbb{R}^{n\times n}_{\text{zero-diag,symm}}$ and

$$\dot{X} = -\tfrac{1}{3}\operatorname{grad}\mathcal{D}(X), \tag{1.11}$$

(ii) $t \mapsto Z(t)$ remains in the set $\mathbb{S}^{n\times n}_{\text{zero-diag,symm}}$ and

$$\dot{Z} = -\tfrac{1}{3}\mathcal{P}_{Z^\perp}\big(\operatorname{grad}\mathcal{D}(Z)\big) = -\tfrac{1}{3}\operatorname{grad}\mathcal{D}\Big|_{\mathbb{S}^{n\times n}_{\text{zero-diag,symm}}}(Z). \tag{1.12}$$

In other words, the projected pure-influence model (1.9) is, modulo a constant factor, the gradient flow of the dissonance function $\mathcal{D}$ restricted to the manifold of zero-diagonal unit-norm symmetric matrices $\mathbb{S}^{n\times n}_{\text{zero-diag,symm}}$.

Proof: [Proof of Theorem 1.3.4] The forward invariance of the set of symmetric matrices in both statements is immediate from the solution uniqueness. To prove equation (1.12), we adopt the slight abuse of notation $\operatorname{grad}\mathcal{D}(Z) = \operatorname{grad}\mathcal{D}\Big|_{\mathbb{S}^{n\times n}_{\text{zero-diag,symm}}}(Z)$. With this notation, $Z \mapsto \operatorname{grad}\mathcal{D}(Z)$ is [77, pages 15-17] the unique vector field on $\mathbb{S}^{n\times n}_{\text{zero-diag,symm}}$ such that

$$\frac{d}{dt}\mathcal{D}(Z(t)) = \langle\!\langle \operatorname{grad}\mathcal{D}(Z(t)), \dot{Z}(t)\rangle\!\rangle_F \tag{1.13}$$

for any differentiable $Z : [0,\infty) \to \mathbb{S}^{n\times n}_{\text{zero-diag,symm}}$. Here, both $\operatorname{grad}\mathcal{D}(Z(t))$ and $\dot{Z}(t)$ belong to the tangent space to the manifold $\mathbb{S}^{n\times n}_{\text{zero-diag,symm}}$. Now, using the various properties of the trace inner product (e.g., $\dot{Z}(t) \perp Z(t)$), we compute

$$\begin{aligned}
\dot{\mathcal{D}}(Z(t)) &= -(\operatorname{trace}(\dot{Z}(t)Z(t)Z(t)) + \operatorname{trace}(Z(t)\dot{Z}(t)Z(t))) \\
&\quad + \operatorname{trace}(Z(t)Z(t)\dot{Z}(t)) \\
&= -3\operatorname{trace}(\dot{Z}(t)Z^2(t)) = -3\langle\!\langle \dot{Z}(t), Z^2(t)\rangle\!\rangle_F \\
&= -3\langle\!\langle \dot{Z}(t), Z^2(t) - \operatorname{diag}(Z^2(t)) + \mathcal{D}(Z(t))Z(t)\rangle\!\rangle_F.
\end{aligned}$$

Recalling that $Z^2 - \operatorname{diag}(Z^2) + \mathcal{D}(Z)Z \overset{(1.6a)}{=} P_{Z^\perp}(Z^2 - \operatorname{diag}(Z^2))$ belongs to the tangent space to the manifold $\mathbb{S}^{n\times n}_{\text{zero-diag,symm}}$ at the point $Z(t)$, one arrives at the equality

$$\operatorname{grad}\mathcal{D}(Z) = -3\big(Z^2 - \operatorname{diag}(Z^2) + \mathcal{D}(Z)Z\big).$$

This concludes the proof of statement (ii). Finally, equation (1.11) can be proved in a similar way. ■

1.4 Classification of symmetric equilibria

We here give the complete classification of the symmetric equilibria in the projected pure-influence model (1.9); the classification of general asymmetric equilibria remains an open problem. Thanks to Theorem 1.3.4, all symmetric equilibria of the projected pure-influence model are critical points of the dissonance function $\mathcal{D}$. We start with the equilibrium equation:

$$Z^2 + \mathcal{D}(Z)Z - \operatorname{diag}(Z^2) = \mathbb{0}_{n\times n}, \quad Z \in \mathbb{S}^{n\times n}_{\text{zero-diag,symm}}. \tag{1.14}$$

Note that the equilibria Z^* with $\mathcal{D}(Z^*) = 0$ correspond to equilibria of the original system (1.3) $X(t) \equiv X^* = \eta(0)Z^*$, whereas the others with $\mathcal{D}(Z^*) \neq 0$ lead to

$$X(t) = \eta(t)Z^*, \quad \eta(t) = \frac{\eta(0)}{1 + t\eta(0)\mathcal{D}(Z^*)}$$

defined for $t \in [0, \frac{1}{\eta(0)\mathcal{D}(Z^*)})$ if $\mathcal{D}(Z^*) < 0$ (for which the solution is unbounded) or for $t \geq 0$ if $\mathcal{D}(Z^*) > 0$.

1.4.1 Normalized Stiefel matrices

To start with, we introduce a special important manifold of non-square matrices that we will use throughout the paper.

Definition 1.4.1 (Normalized Stiefel matrices) *A matrix $V \in \mathbb{R}^{n\times k}$, for $k \leq n$, is* normalized Stiefel *(*nSt*), if*

(i) the columns of V are pairwise orthogonal unit vectors, i.e., $V^\top V = I_k$;

(ii) the norm of each row is the same (obviously, it must be $\sqrt{k/n} \leq 1$): $\operatorname{diag}(VV^\top) = n^{-1}kI_n$.

Let $\mathrm{nSt}(n,k) \subseteq \mathbb{R}^{n\times k}$ denote the set of normalized Stiefel matrices.

In general, the rows of an nSt matrix *need not* be orthogonal. We recall from [90] the notion of *compact Stiefel manifold*, denoted by $\mathrm{St}(k,n) = \{X \in \mathbb{R}^{n\times k} \mid X^\top X = I_k\}$.

Lemma 1.4.1 (Characterization of nSt matrices) *The set $\mathrm{nSt}(n,k)$, $k \leq n$, is a compact and analytic submanifold of $\mathbb{R}^{n\times k}$ of dimension $(k-1)n + 1 - k(k+1)/2$, and it is also a submanifold of the compact Stiefel manifold (and thus, $\mathrm{nSt}(n,k) \subseteq \mathrm{St}(k,n)$). Moreover,*

(i) $\mathrm{nSt}(n,n)$ is the set of orthogonal matrices,

(ii) for $k = 1$, the matrix V is nSt *if and only if*

$$V = \frac{1}{\sqrt{n}} \begin{bmatrix} s_1 \\ \vdots \\ s_n \end{bmatrix}, \tag{1.15}$$

for any numbers $s_i \in \{-1, +1\}$, $i \in \{1, \ldots, n\}$,

(iii) for $k = 2$, the matrix V is nSt *if and only if*

$$V = \sqrt{\frac{2}{n}} \begin{bmatrix} \cos\alpha_1 & \sin\alpha_1 \\ \vdots & \vdots \\ \cos\alpha_n & \sin\alpha_n \end{bmatrix}, \tag{1.16}$$

for any set of angles $\alpha_1, \ldots, \alpha_n$ satisfying

$$\sum_{m=1}^{n} e^{2\alpha_m \sqrt{-1}} = 0. \tag{1.17}$$

We postpone the proof of Lemma 1.4.1 to Appendix 1.8.1. We remark that in the case of $n = k = 2$, the constraint (1.17) implies that $2\alpha_2 = \pi + 2\pi s + 2\alpha_1$, where $s \in \mathbb{Z}$, that is, $\alpha_2 = \pi/2 + \pi s + \alpha_1$ and $\cos\alpha_2 = (-1)^{s+1}\sin\alpha_1$, $\sin\alpha_2 = (-1)^s \cos\alpha_1$. Thus, the matrices in $\mathrm{nSt}(2,2)$ are orthogonal 2×2 matrices (representing proper or improper rotations):

$$V = \begin{bmatrix} \cos\alpha_1 & \sin\alpha_1 \\ -\varepsilon\sin\alpha_1 & \varepsilon\cos\alpha_1 \end{bmatrix}, \quad \varepsilon \in \{-1, +1\}.$$

For a general k, it is difficult to give a closed-form description of all matrices from $\mathrm{nSt}(n,k)$. However, there are simple examples of matrices from $\mathrm{nSt}(n,k)$ in the case where $n = 2k$, including every matrix of the form

$$V = \frac{1}{\sqrt{2}} \begin{bmatrix} U_1 \\ U_2 \end{bmatrix},$$

where U_i are orthogonal $k \times k$ matrices.

1.4.2 Technical results

The classification of equilibria relies on the following technical results that will be proved in Appendix 1.8.1.

Lemma 1.4.2 *Suppose that $Z^2 - 2\alpha Z = \beta I_n$ for some symmetric $n \times n$ matrix Z with $\mathrm{diag}(Z) = \mathbb{0}_{n\times n}$ and scalars α, β. Then Z can be decomposed as*

$$Z = pVV^\top - qI_n = Z^\top \tag{1.18}$$

for some $V \in \mathrm{nSt}(n,k)$ $(1 \le k < n)$ and constants $p, q \ge 0$ such that $pk = qn$, $2\alpha = p - 2q$ and $\beta = q(p-q)$. Namely, $p = 2\sqrt{\alpha^2+\beta}$, $\quad q = \sqrt{\alpha^2+\beta} - \alpha$.

Corollary 1.4.3 *Given a matrix $Z = Z^\top$ with $\mathrm{diag}(Z) = \mathbb{0}_{n\times n}$, the matrix $Z^2 - 2\alpha Z$ is diagonal with s different eigenvalues $\beta_1 < \ldots < \beta_s$ of multiplicities $n_1, \ldots, n_s$ respectively $(n_1 + n_2 + \ldots + n_s = n)$ if and only if there exists such a permutation matrix S that*

$$SZS^{-1} = \mathrm{diag}(Z_1, \ldots, Z_s),$$

where each Z_i is decomposed as (1.18) with parameters p_i, q_i, V_i, where $V_i \in \mathrm{nSt}(n_i, k_i)$ for some $k_i < n_i$ and

$$p_i = 2\sqrt{\alpha^2+\beta_i}, \quad q_i = \sqrt{\alpha^2+\beta_i} - \alpha. \tag{1.19}$$

Thus, for irreducible *$Z = Z^\top$ the matrix $Z^2 - 2\alpha Z$ is diagonal if and only if Z is decomposed as (1.18) with $p, q \ge 0$.*

1.4.3 Classification of irreducible symmetric equilibria

Theorem 1.4.4 (Irreducible equilibria for the projected pure-influence model) *For the projected pure-influence model* (1.9)*,*

(i) all irreducible symmetric equilibria are of the form

$$Z^* = pVV^\top - qI_n, \tag{1.20}$$

with $V \in \mathrm{nSt}(n,k)$*,* $1 \le k < n$*, and*

$$p = \sqrt{\frac{n}{k(n-k)}}, \quad q = \sqrt{\frac{k}{n(n-k)}}; \tag{1.21}$$

(ii) Z^* *has* k *positive eigenvalues with value* $p-q$ *and* $n-k$ *negative eigenvalues with value* $-q$*;*

(iii) the dissonance function satisfies

$$\mathcal{D}(Z^*) = -\frac{n-2k}{\sqrt{kn(n-k)}}, \tag{1.22}$$

and the right-hand side is monotonically increasing in $k \in \{1,\dots,n-1\}$ *(see Figure 1.2).*

Proof: We start by proving a technical statement. Pick $V \in \mathrm{nSt}(n,k)$, p,q real numbers and set $\theta = p - 2q$. Then, the matrix $Z = pVV^\top - qI_n = Z^\top$ satisfies the following properties:

(a) $Z^2 - \theta Z = q(p-q)I_n$, and thus $\mathrm{diag}(Z^2) = \theta\,\mathrm{diag}(Z) + q(p-q)I_n$;

(b) for any $p \neq 0$, the matrix Z has two eigenvalues $p-q$ and $(-q)$ whose multiplicities are k and $(n-k)$ respectively;

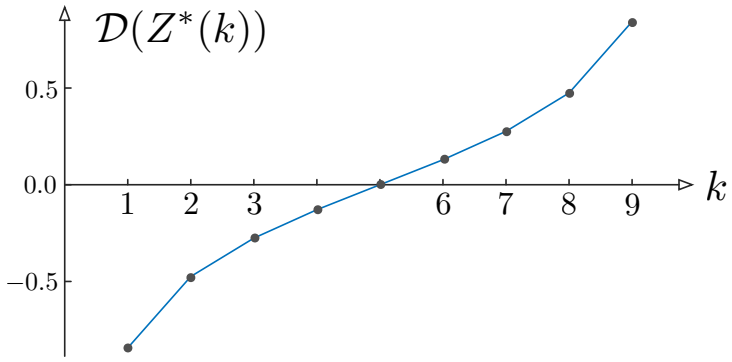


Figure 1.2: For a network with size $n = 10$, the dissonance function $\mathcal{D}$ evaluated on all irreducible symmetric equilibria with $k \in \{1, \dots, 9\}$ positive eigenvalues, according to equation (1.22).

(c) the eigenspaces corresponding to $p - q$ and $-q$ are the image of V and the kernel of $V^\top$ respectively;

(d) $\operatorname{diag}(Z) = \mathbb{0}_{n\times n}$ if and only if $pk = qn$; in this situation, $\operatorname{trace}(Z^2) = q(p-q)n$ and $\mathcal{D}(Z) = -\operatorname{trace}(Z^2 Z^\top) = -\theta n q(p-q)$.

To prove (a), recall that $V^\top V = I_k$ and therefore

$$Z^2 = p^2 VV^\top VV^\top + q^2 I_n - 2pqVV^\top = p\theta VV^\top + q^2 I_n = \theta Z + (pq - q^2) I_n.$$

To prove (b) and (c), notice that for any vector $z = Vy$ one has $VV^\top z = V(V^\top V)y = Vy = z$, and thus $Zz = (p-q)z$. The space of such vectors is nothing else than the image of V and has dimension k (recall that the columns of V are orthogonal, and hence are linearly independent). If $V^\top z = 0$, then $Zz = -qz$, and the dimension of $\ker(V^\top)$ is $(n-k)$. Since $Z = Z^\top$ and $p - q \neq -q$ (except for the case where $p = q = 0$ and $Z = 0$, which is trivial), the two eigenspaces are orthogonal and their sum coincides with $\mathbb{R}^n$. Hence, there are no other eigenvalues. To prove (d), note first $p \operatorname{diag}(VV^\top) = (pk/n) I_n$, and thus $\operatorname{diag}(Z) = \mathbb{0}_{n\times n}$ if and only if $pk/n = q$. Using statement (a), one shows that in this situation $\operatorname{diag}(Z^2) = q(p-q)I_n$ and hence $\operatorname{trace}(Z^2) = q(p-q)n$. Thanks to (a),

$Z^3 = \theta Z^2 + q(p-q)Z \Longrightarrow \operatorname{trace}(Z^3) = \theta \operatorname{trace}(Z^2) = \theta nq(p-q)$, which finishes the proof of (d).

Now, to prove the statement (i) of the theorem, let Z^* be an irreducible symmetric solution to equation (1.14). For $\alpha = -\mathcal{D}(Z^*)/2$, the matrix $(Z^*)^2 - 2\alpha Z^* = \operatorname{diag}(Z^{*2})$ is diagonal. Since Z^* is irreducible, it follows from Corollary 1.4.3 that Z^* can be decomposed as (1.20) with some $p, q \geq 0$. Then, from (a) and (d), it also follows that Z^* satisfies equation (1.14) if and only if $pk = qn$ (which comes from $\operatorname{diag}(Z^*) = \mathbb{0}_{n\times n}$) and $pq - q^2 = 1/n$ (which comes from $\operatorname{trace}(Z^{*2}) = 1$). This implies that $q = \sqrt{\frac{k}{n(n-k)}}$ and $p = \sqrt{\frac{n}{k(n-k)}}$.

Finally, statement (ii) follows from (b); and (iii) is obtained by substituting the values of p and q into the definition of the dissonance function (1.5) and noting that the smooth function $\kappa \mapsto -\frac{n-2\kappa}{\sqrt{n\kappa(n-\kappa)}}$ has positive derivative on $(0, n)$. ■

1.4.4 Classification of reducible symmetric equilibria

The next theorem generalizes Theorem 1.4.4 and characterizes all symmetric equilibria for the projected pure-influence model and its proof can be found in Appendix 1.8.1.

Theorem 1.4.5 (All equilibria for the projected pure-influence model) *The matrix Z^* is an equilibrium* (1.14) *of the projected pure-influence model if and only if a permutation matrix S exists such that:*

(i) $SZ^*S^{-1} = \operatorname{diag}(Z_1^*, \ldots, Z_s^*)$, $s \geq 1$, $Z_i^* = Z_i^{*\top} \in \mathbb{R}^{n_i \times n_i}$;

(ii) the blocks Z_i^* admit representation (1.18)*: $Z_i^* = p_i V_i V_i^\top - q_i I_{n_i}$, where $p_i, q_i \geq 0$ and $V_i \in \operatorname{nSt}(n_i, k_i)$, $1 \leq k_i < n_i$;*

(iii) the sign $\varepsilon = \operatorname{sign}(n_i - 2k_i) \in \{-1, 0, 1\}$ is the same for all $i = 1, \ldots, s$ such that $Z_i^* \neq \mathbb{0}_{n_i \times n_i}$ and

(iv) each block $Z_i^* \neq \mathbb{0}_{n_i \times n_i}$ *is irreducible and the corresponding coefficients* p_i, q_i *have the form*

$$p_i = 2\sqrt{\alpha^2 + \beta_i}, \quad q_i = \sqrt{\alpha^2 + \beta_i} - \alpha, \tag{1.23}$$

where

(a) for $\varepsilon \neq 0$, α *and* β_i *are determined from*

$$\begin{aligned} \alpha &= \varepsilon \left(\sum_{i:Z_i \neq 0} \frac{4k_i n_i (n_i - k_i)}{(n_i - 2k_i)^2} \right)^{-1/2}, \\ \beta_i &= \alpha^2 \frac{4n_i k_i - 4k_i^2}{(n_i - 2k_i)^2}; \end{aligned} \tag{1.24}$$

(b) for $\varepsilon = 0$, $\alpha = 0$, *for all* i, *and* β_i *are chosen in such a way that* $\sum_{i:Z_i \neq 0} \beta_i n_i = 1$.

Remark 1.4.6 *Let* Z^* *be a reducible equilibrium for the projected pure-influence model such that* $G(Z^*)$ *is composed of* m *(disconnected) subgraphs that satisfy structural balance. According to Definition 1.2.3,* $G(Z^*)$ does not *satisfy structural balance since this definition requires* $G(Z^*)$ *to be complete.*

1.4.5 Structural balance and equilibria

We now characterize the equilibria corresponding to structural balance and how they minimize the dissonance function.

Corollary 1.4.7 (Balanced equilibria of the projected pure-influence model) *For the projected pure-influence model* (1.9), *let* $Z^* \in \mathbb{S}^{n \times n}_{\text{zero-diag}}$ *be an equilibrium point with a single positive eigenvalue. Then,*

(i) after a relabelling of the agents, Z^* has the form

$$Z^* = \left[\begin{array}{c|c} Z' & \mathbb{0}_{n_1\times(n-n_1)} \\ \hline \mathbb{0}_{(n-n_1)\times n_1} & \mathbb{0}_{(n-n_1)\times(n-n_1)} \end{array}\right] \tag{1.25}$$

with $n_1 \leq n$ and

$$Z' = \frac{1}{\sqrt{n_1(n_1-1)}}(ss^\top - I_{n_1}), \tag{1.26}$$

for some $s \in \{-1,+1\}^{n_1}$; and thus, for any fixed n_1, there are only 2^{n_1-1} different equilibria (with a single positive eigenvalue),

(ii) $G(Z')$ satisfies structural balance, with the binary vector s characterizing the distribution of the individuals in the single faction or in the two factions, and

(iii) if $G(Z^*)$ is a connected graph, then $G(Z^*)$ satisfies structural balance (being thus complete) and Z^* is a global minimizer to the optimization problem:

$$\begin{aligned} \underset{Z\in\mathbb{R}^{n\times n}}{\text{minimize}} \quad & \mathcal{D}(Z) \\ \text{subject to} \quad & Z \in \mathbb{S}^{n\times n}_{\text{zero-diag,symm}} \end{aligned}$$

and satisfies $\mathcal{D}(Z^*) = -\frac{n-2}{\sqrt{n(n-1)}}$.

Proof:

Consider a permutation of indices from Theorem 1.4.5. Since Z^* has only one positive eigenvalue, it can have only one non-zero diagonal block $Z^*_i = Z'$. Statement (i) now follows from (1.20),(1.21) (with $k = 1, n = n_1$) and (1.15).

Regarding statement (ii), observe that for any different i, j and k,

$$z'_{ij}z'_{jk}z'_{ki} = \frac{(s_is_j)(s_js_k)(s_ks_i)}{(n_1(n_1-1))^{3/2}} = \frac{1}{(n_1(n_1-1))^{3/2}} > 0.$$

This inequality implies $\operatorname{sign}(z'_{ij}) = \operatorname{sign}(z'_{jk}z'_{ki})$ and thus we know that Z' satisfies structural balance. It is immediate to see that any i and j such that $s_i = s_j$ correspond to the same faction in the network $G(Z')$. This completes the proof for (ii).

Regarding statement (iii), we notice that the smooth function $\eta \mapsto -\frac{\eta-2}{\sqrt{\eta(\eta-1)}}$ has negative derivative for $\eta > 3/2$. Hence, the value of $\mathcal{D}(Z^*) = \mathcal{D}(Z') = -\frac{n_1-2}{\sqrt{n_1(n_1-1)}}$ at equilibrium (1.25) with one positive eigenvalue is minimal when $Z' = Z^*$ and $n_1 = n$, that is, the matrix is irreducible. Now, let us focus on the points that vanish the gradient of $\mathcal{D}$, i.e., the equilibria of the projected pure-influence model. Permuting the agents, we may confine ourselves to equilibria described in Theorem 1.4.5 that have s blocks of size n_i with $k_i < n_i$ positive eigenvalues, $i \in \{1, \dots, s\}$. To see why this is true, in the proof of Theorem 1.4.4 it was shown that $\mathcal{D}(Z_i^*) = -2\alpha n q_i(p_i - q_i) = -2\alpha\beta_i$. Next, if $\varepsilon = -1$, then $\alpha < 0$ and $D(Z^*) > 0$. If $\varepsilon = \operatorname{sign}(n_i - 2k_i) = 0$ for all $Z_i^* \neq 0$, then $\mathcal{D}(Z^*) = \sum_i \mathcal{D}(Z_i^*) = 0$. As we know, the minimal value should be negative, so such equilibria cannot be global minimizers. Therefore, we may assume that $\varepsilon = 1$, that is, $k_i < n_i/2$ for all such i that $Z_i^* \neq 0$. Assume, without loss of generality, that $Z_1^*, \dots, Z_m^* \neq 0$ and $Z_{m+1}^*, \dots, Z_s^* = 0$. Denote $k_1 + \cdots + k_m = k'$ and $n_1 + \cdots + n_m = n' \leq n$. Note that the function $f(\xi) = \xi(1-\xi)/(1-2\xi)^2$ is convex on $(0, 1/2)$. Therefore, Jensen's inequality implies

$$\frac{1}{n}\sum_{i=1}^{m} \frac{k_i n_i (n_i - k_i)}{(n_i - 2k_i)^2} \;=\; \sum_{i=1}^{m} \frac{n_i}{n} f\Big(\frac{k_i}{n_i}\Big) \;\geq\; f\Big(\sum_{i=1}^{m} \frac{k_i}{n'}\Big) \;=\; f\Big(\frac{k'}{n'}\Big) \;=\; \frac{k'(n'-k')}{(n'-2k')^2},$$

and, in turn,

$$\mathcal{D}(Z^*) = -\left(\sum_{i=1}^{m} \frac{k_i n_i (n_i - k_i)}{(n_i - 2k_i)^2}\right)^{-1/2} \geq -\frac{n' - 2k'}{\sqrt{k'n'(n'-k')}}.$$

We know, however from Theorem 1.4.4 that the right-hand side is minimal when $k' = 1$,

in which case the minimal value, as we have seen in the beginning in the proof, is achieved at $n' = n$. Hence, the irreducible equilibrium with one positive eigenvalue is the global minimizer of $\mathcal{D}^*$. ■

Remark 1.4.8 *Let Z^* denote an equilibrium point with one positive eigenvalue. Then, $-Z^*$ has one negative eigenvalue and does not correspond to structural balance. All such $-Z^*$ correspond to isolated critical points of $\mathcal{D}$.*

1.4.6 Examples of equilibria with two positive eigenvalues

Let Z^* be any equilibrium of the projected pure-influence model parameterized by $\mathrm{nSt}(n, 2)$ matrices, so that it has two positive eigenvalues. Let us assume first that it is irreducible. Then, another class of equilibria is found using the parametrization (1.16). It can be easily shown that

$$Z^* = \sqrt{\frac{2}{n(n-2)}}(\theta_{ij})_{i,j=1}^n, \quad \theta_{ij} = \begin{cases} 0, \, i = j \\ \cos(\alpha_i - \alpha_j), \, i \neq j. \end{cases}$$

Here the angles α_i should satisfy the relation (1.17). Interestingly, many of such matrices do not correspond to structural balance. Consider, for example, the case where the unit vectors in (1.17) constitute a regular n-gon: $\alpha_i = \frac{\pi(i-1)}{n}$, $i = 1, \ldots, n$. For any pair $i, j > i$ the entry z_{ij} is negative if $(j - i) > n/2$, positive if $j - i < n/2$ and zero if $j - i = n/2$ (possible only for even n). If n is odd, the graph is complete, otherwise, the pairs of nodes $(i, i + n/2)$ for $i = 1, \ldots, n/2$ are not connected. For example, in the smallest

dimension $n = 3$, by setting $\alpha_1 = 0$, $\alpha_2 = \pi/3$ and $2\pi/3$, we obtain the equilibrium

$$Z^* = \frac{1}{\sqrt{6}} \begin{bmatrix} 0 & +1 & -1 \\ +1 & 0 & +1 \\ -1 & +1 & 0 \end{bmatrix}$$

which does not correspond to structural balance. Actually, in the case where $n = 3$ or $n \geq 5$, the graph always contains imbalanced triads. For instance, for $n \geq 3$ being odd the nodes $i = 1$, $j = (n-1)/2$ and $\ell = (n+3)/2$ always constitute such a triad: $z_{i\ell} < 0$, whereas $z_{ij}, z_{j\ell} \geq 0$. For an even number $n \geq 6$, one may take $i = 1$, $j = n/2$, $\ell = n/2+2$. In the case $n = 4$, the equilibrium Z^* corresponds to an incomplete cyclic graph such that $\mathcal{D}(Z^*) = 0$:

$$Z^* = \frac{1}{2\sqrt{2}} \begin{bmatrix} 0 & \frac{1}{\sqrt{2}} & 0 & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} & 0 \\ 0 & \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} & 0 \end{bmatrix}.$$

For the reducible matrix case, since Z^* has two positive eigenvalues, $G(Z^*)$ contains two disconnected subgraphs that satisfy structural balance with possibly other isolated nodes.

1.5 Convergence to balanced equilibria and stability analysis

We now provide convergence results for our models towards equilibria that correspond to structural balance. We present a supporting lemma and then our main theorem.

Lemma 1.5.1 *Assume that the solution of* (1.2) *satisfies* $x_{i*}(t_0) = \mathbb{0}_{1\times n}$ *at some* $t_0 \geq 0$,

that is, in the graph $G(X(t_0))$ node i does not communicate to any other node. Then, $x_{i*}(t) \equiv \mathbb{0}_{1\times n}$ for any $t \geq 0$. The same holds for the solutions of (1.9).

Proof: Since the right-hand sides of (1.2) and (1.9) are analytic, any solution is a real-analytic function of time. Assuming that $x_{ij}(t_0) = 0$ for all j, one finds that $\dot{x}_{ij}(t_0) = 0$. Differentiating (1.2), it is easy to show that $\ddot{x}_{ij}(t_0) = 0$, and so on, $x_{ij}^{(m)}(t_0) = 0$ for any $m \geq 1$. In view of analyticity, one has $x_{ij}(t) \equiv 0$ for any t. Similarly, $z_{ij}(t_0) = 0\,\forall j$ entails that $z_{ij}(t) \equiv 0$ for any solution of (1.9). ■

Theorem 1.5.2 (Convergence results and dynamical properties) *Consider the pure-influence model* (1.2) *with an initial condition $X(0) \in \mathbb{R}^{n\times n}_{\text{zero-diag,symm}}$ and the projected pure-influence model* (1.9) *with initial condition $Z(0) = \frac{X(0)}{\|X(0)\|_F}$. Then,*

(i) the solution $Z(t)$ converges to a single critical point of the dissonance function $\mathcal{D}$;

(ii) the number of negative eigenvalues of $Z(t)$ is non-decreasing.

Moreover, if $X(0)$ has one positive eigenvalue, then

(iii) $\lim_{t\to+\infty} Z(t) = Z^*$, where Z^* is as in (1.26)*, so that $G(Z(t))$ or one of its connected components (while the rest of nodes are isolated) reaches structural balance in finite time;*

(iv) $X(t)$ achieves the same sign structure as Z^* in finite time;

(v) nonzero entries of $X(t)$ diverge to infinity in finite time.

Proof: For convenience, throughout this proof, let us denote $W(t) = \frac{X(t)}{\|X(t)\|_F}$, i.e., $X(t) = \eta(t)W(t)$ with $\eta(t)$ evolving according to (1.6a) and $W(t)$ evolving according to (1.6b). From the construction of the transcription of the pure-influence model in Theorem 1.3.3, we have that $\eta(t) = \|X(t)\|_F$ and so $\eta(t) > 0$ for all well-defined $t \geq 0$.

Moreover, Lemma 1.8.1 let us conclude that $W(t) = Z(\int_0^t \eta(s)ds)$ for all $t \geq 0$, and thus the solution $X(t)$ is well defined.

To prove (i), recall that (1.9) is a gradient flow dynamics of the analytic function $\mathcal{D}$, and the trajectory $Z(t)$ stays on a compact manifold and, in particular, is bounded. The classical result of Łojasiewicz [3] implies convergence of the trajectory to a single fixed point.

To prove (ii), we enumerate the eigenvalues of $Z(t)$ in the descending order $\lambda_1(t) \geq \lambda_2(t) \ldots \geq \lambda_n(t)$ and consider the corresponding orthonormal bases of eigenvectors $v_i(t)$. Since $Z_i(t)v_i(t) = \lambda_i(t)v_i(t)$ and $v_i(t)^\top v_i(t) = 1$, we obtain $\dot{Z}v_i + Z\dot{v}_i = \dot{\lambda}_i v_i + \lambda_i \dot{v}_i$ and $\dot{v}_i(t)^\top v_i(t) = 0$. Therefore,

$$\dot{\lambda}_i = v_i^\top \dot{Z} v_i + v_i^\top Z \dot{v}_i = v_i^\top \dot{Z} v_i + \lambda_i v_i^\top \dot{v}_i = v_i^\top \dot{Z} v_i,$$

entailing the following differential equation

$$\dot{\lambda}_i = \lambda_i^2 + \mathcal{D}(Z)\lambda_i - v_i^\top \operatorname{diag}(Z^2) v_i. \tag{1.27}$$

Notice that all diagonal entries of $\operatorname{diag}(Z^2)$ are nonnegative. Now, due to Lemma 1.5.1, if the ith row of X was initially the zero vector, then it will continue being the same for all times and also for Z; and, moreover, $\operatorname{diag}(Z^2)_{ii} = 0$ and there exists a zero eigenvalue with its associated eigenvector having zero entries in all the positions of the entries where $\operatorname{diag}(Z^2)$ are positive. Then, it immediately follows from (1.27) that if $\lambda_i(0) = 0$ due to $Z(0)$ having a row being the zero vector $\mathbb{0}_{1\times n}$, then $\dot{\lambda}_i = 0$.

Now, let $\mathcal{N}$ be the set of indices i such that $\operatorname{diag}(Z^2)_{ii} > 0$. Thus, for any $i \in \mathcal{N}$, if

λ_i crosses the real axis at time t^*, i.e., $\lambda(t^*) = 0$, then

$$\dot{\lambda}_i(t^*) = -(v_i(t^*))^\top \operatorname{diag}(Z^2(t^*))v_i(t^*) < 0. \tag{1.28}$$

Therefore, if $\lambda_i(t_0) \leq 0$ for some $t_0 \geq 0$, then $\lambda_i(t) \leq 0$ for all $t \geq t_0$. This finishes the proof for (ii).

Notice that since $\operatorname{trace}(Z(t)) = 0$ and $Z(t) = Z(t)^\top \neq \mathbb{0}_{n\times n}$, then $Z(t)$ has at least one positive eigenvalue. Then, equation (1.28) implies that

$$\Lambda := \left\{ Z \in \mathbb{S}^{n\times n}_{\text{zero-diag,symm}} \mid Z \text{ has only one positive eigenvalue} \right\}$$

is forward invariant and, in particular, the limit $Z^* = \lim_{t\to\infty} Z(t)$ (existing in view of statement (i)) belongs to Λ. Since Z^* is a critical point of $\mathcal{D}$ (or, in view of Theorem 1.3.4, the equilibrium of (1.9)), it has the structure described by Corollary 1.4.7.

By continuity of the flow $Z(t)$, there is a finite time τ such that $G(Z(t))$ has the same sign structure as $G(Z^*)$ for all $t \geq \tau$. This finishes the proof for (iii).

Now we prove the last two statements of the theorem. Knowing the convergence result from (iii), Lemma 1.8.1 tells us that introducing the term η as in the transcribed system (1.6a) to the projected pure-influence model has the simple effect of altering the convergence rate properties for $Z(t)$. Therefore, there always exist a finite time $\tau^* \geq 0$ such that, for any $t \geq \tau^*$, $W(t)$ satisfies the sign properties of statement (iii) regarding structural balance. Moreover, the fact that $X(t) = \eta(t)W(t)$ and $\eta(t) \geq 0$ by construction, immediately implies (iv). Now, let $g(t) := -\mathcal{D}(W(t))$, and notice that $g(t)$ is a strictly positive continuous function for all (well-defined) $t \geq \tau^*$. Now, from equation (1.6b), we have the system $\dot{\eta}(t) = g(t)\eta^2(t)$, with solution $\eta(t) = \frac{\eta(\tau)}{1-\eta(\tau)\int_\tau^t g(s)ds}$ for $t \geq \tau$. Then, since $\int_\tau^t g(s)ds$ is a monotonic strictly increasing function on $t \geq \tau$,

we have that $\eta(t) \to +\infty$ as $t \to t^*$, where $t^* > \tau^*$ is some finite time such that $\int_\tau^{t^*} g(s)ds = \frac{1}{\eta(\tau)}$ (note that $t^* > \tau^*$ holds from the relationship $W(t) = Z(\int_0^t \eta(s)ds)$). Then, we conclude that the solution $\eta(t)$ and the entries of $X(t)$ diverge in some finite time t^*, which proves (v). ■

Corollary 1.5.3 *Consider the same conditions as in Theorem 1.5.2, i.e., the projected pure-influence model with initial condition $Z(0) \in \mathbb{S}^{n\times n}_{\text{zero-diag,symm}}$ having one positive eigenvalue. If $\mathcal{D}(Z(0)) < -\frac{n-3}{\sqrt{(n-1)(n-2)}}$, then $G(Z(t))$ eventually reaches structural balance.*

The previous theorem immediately implies that the set of irreducible equilibria with a single positive eigenvalue is (locally) asymptotically stable. We present further results on the stability of equilibria.

Lemma 1.5.4 (Further results on stability of the equilibria) *Consider a symmetric equilibrium point Z^* for the projected pure-influence model* (1.9)*. Without loss of generality, assume that Z^* has no row equal to the zero vector*[1]*. If $\mathcal{D}(Z^*) \geq 0$, then Z^* is an unstable equilibrium point and does not correspond to structural balance.*

Proof: Write the analytic projected influence system (1.9) as $\dot{Z} = f(Z) := Z^2 - \text{diag}(Z^2) + \mathcal{D}(Z)Z$, thereby defining $f : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$, and compute

$$\frac{\partial f_{ij}(Z)}{\partial z_{ij}} = \mathcal{D}(Z) + \frac{\partial \mathcal{D}(Z)}{\partial z_{ij}} z_{ij},$$
$$\frac{\partial \mathcal{D}(Z^*)}{\partial z_{ij}} = -3 \sum\nolimits_{\substack{k=1 \\ k\neq i,j}}^{n} z^*_{ik} z^*_{kj}.$$

Now, the Jacobian of f, denoted by Df, is a $(n^2 - n) \times (n^2 - n)$ matrix (since we do not consider self-appraisals). Let $Df(Z^*)$ be the Jacobian evaluated at Z^* and let $\{\lambda_i\}_{i=1}^{n^2-n}$

[1]If Z^* had a row equal to the zero vector, then, in the lemma statement, we would replace n by $n_1 < n$, where n_1 is the number of rows of Z^* that are not equal to the zero vector.

be the set of its eigenvalues. Then, we compute

$$\sum_{i=1}^{n^2-n} \lambda_i = \operatorname{trace}(Df(Z^*)) = \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j\neq i}}^{n} \frac{\partial f_{ij}(Z^*)}{\partial z_{ij}}$$
$$= (n^2-n)\mathcal{D}(Z^*) + 3\mathcal{D}(Z^*) = (n^2-n+3)\mathcal{D}(Z^*).$$

Since $n^2 - n + 3 > 0$ for $n \geq 3$, we draw the following conclusions for $\mathcal{D}(Z^*) \geq 0$: (i) $Df(Z^*)$ contains at least one positive eigenvalue and so the equilibrium point Z^* is unstable; (ii) at least one triad in $G(Z^*)$ is unbalanced and so Z^* does not correspond to structural balance. ■

1.6 Simulation results and conjectures

The generic convergence of trajectories to the minima of $\mathcal{D}$ (or, equivalently, the convergence from almost all initial conditions) is an open problem. However, we present strong numerical evidence that support such claim. We first remark that, from the proof of Theorem 1.3.3, the projected pure-influence model (1.9) can be generalized over any asymmetric matrix in $\mathbb{S}^{n\times n}_{\text{zero-diag}}$ by replacing $\mathcal{D}(Z)$ by $-\operatorname{trace}(Z^\top Z^2)$ and this is the model we will refer throughout this section.

A *generic asymmetric initial condition* $X(0)$ for the pure-influence model (1.2) is a matrix that is generated with each entry independently sampled from a uniform distribution with support $[-100, 100]$, and its diagonal entries set to zero. A *generic symmetric initial condition* is similarly constructed by only sampling the upper triangular entries of the matrix. For the projected pure-influence model, we say $Z(0) = \frac{X(0)}{\|X(0)\|_F}$ is a (non-)symmetric generic initial condition depending on how $X(0)$ was generated. We immediately see from the proof of Theorem 1.5.2, that $Z(t)$ converges to social balance if and only if $X(t)$ converges to social balance. Indeed, given that $X(t)$ diverges at some

finite time $\bar{t}$, we have $Z(\infty) = \frac{X(\bar{t}^-)}{\|X(\bar{t}^-)\|_F}$.

For a fixed network size n, we use a Monte Carlo method [162] to estimate the probability p of the event "under a generic asymmetric initial condition $Z(0)$, $Z(t)$ converges to structural balance in finite time". We estimate p by performing N independent simulations (i.e., each simulation generates a new independent initial condition) and obtaining the proportion $\hat{p}_N$, also known as the empirical probability, of times that the simulation indeed had $Z(t)$ converging to structural balance in finite time. For any accuracy $1-\epsilon \in (0,1)$ and confidence level $1-\eta \in (0,1)$ we have that $|\hat{p}_N - p| < \epsilon$ with probability greater than $1-\eta$ if the Chernoff bound $N \geq \frac{1}{2\epsilon^2} \log \frac{2}{\eta}$ is satisfied. For $\epsilon = \eta = 0.01$, the bound is satisfied by $N = 27000$. We performed the $N = 27000$ independent simulations with $n \in \{5, 6\}$, and found that $\hat{p}_N = 1$. Our observations let us conclude that *for generic asymmetric initial condition $Z(0)$ and $n \in \{5,6\}$, with 99% confidence level, there is at least 0.99 probability that $Z(t)$ converges to structural balance in finite time.*

Similarly, we performed the same Monte Carlo analysis for generic symmetric initial conditions with $n \in \{3, 5, 6, 15\}$, and found for that $\hat{p}_N = 1$ for all n. Therefore, we conclude that *for any symmetric generic initial condition $Z(0)$ and $n \in \{3,5,6,15\}$, with 99% confidence level, there is at least 0.99 probability that $Z(t)$ converges to structural balance in finite time.*

We report three more observations and then state a resulting conjecture. First, remarkably, we found that all of our simulations (for any type of random initial condition) that converged to structural balance in finite time, did it by converging to an equilibrium point having only one positive eigenvalue inside the set of scale-symmetric matrices, which is a superset of the set of symmetric matrices (see Appendix 1.8.2). Second, we did not perform experiments for larger sizes of n due to computational constraints. Third, unfortunately, for $n = 3$, we did find randomly-generated asymmetric initial conditions whose numerically-computed solutions do not converge to structural balance.

Conjecture 1 (Convergence from generic initial conditions) *Consider the pure-influence model* (1.2) *with some initial condition* $X(0)$*, and the projected pure-influence model* (1.9) *with initial condition* $Z(0) = \frac{X(0)}{\|X(0)\|_F}$*. Then,*

(i) under generic asymmetric initial conditions, $\lim_{t\to+\infty} Z(t) = Z^*$ *for a sufficiently large* n*,*

(ii) under generic symmetric initial conditions, $\lim_{t\to+\infty} Z(t) = Z^*$ *for any* n*,*

where Z^* *is scale-symmetric (and particularly symmetric for (ii)) corresponding to structural balance. Then,* $Z(t)$ *reaches structural balance in finite time. Moreover,* $X(t)$ *reaches structural balance in finite time with same sign structure as* Z^**, and also diverges in finite time.*

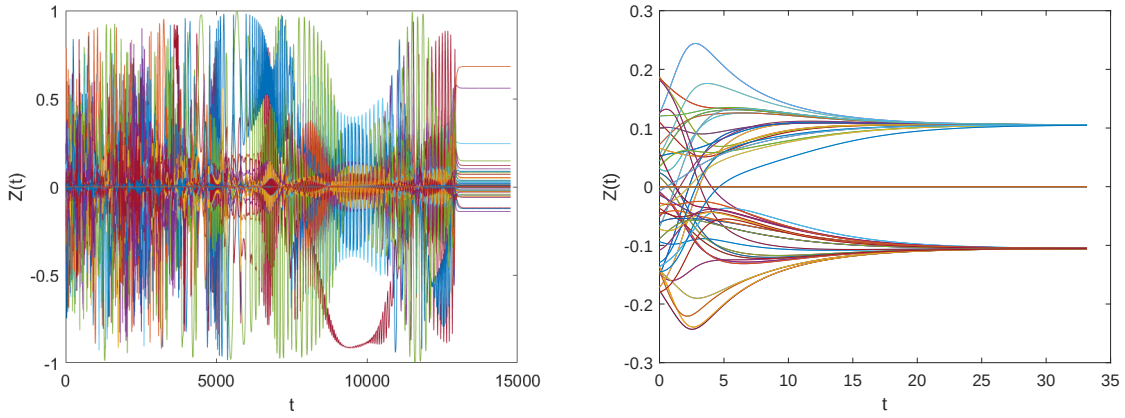
Similarly, we performed the same simulation analysis for the Kułakowski et al. model (1.4), which converges to structural balance if and only if the projected Kułakowski model (1.10) does. To generate a generic initial condition for this system, we generated an $n \times n$ matrix with each entry independently sampled from a uniform distribution with support $[-100, 100]$, and then divide it by its Frobenius norm. We performed $N = 27000$ independent simulations with $n \in \{5, 6\}$, and found that *for generic initial condition* $Z(0)$ *and* $n = 5$*, only* 16.94% *converged to structural balance, and for* $n = 6$*, only* 11.50% *converged to structural balance.*

Also, for $n = 3$, not all simulations converged to structural balance. We remark that not all of the networks for which the system converged and did not satisfy structural balance were complete, some of them were networks with only self-loops, e.g., Figure 2.3(a). Similarly, we performed the same Monte Carlo analysis for symmetric initial conditions with $n \in \{3, 5, 6, 15\}$. Our results show that *for symmetric generic initial condition,* $Z(0)$ *did not always converge to structural balance for* $n = 3$*, but, for* $n \in \{5, 6, 15\}$*, with*

99% *confidence level, there is at least* 0.99 *probability that* $Z(t)$ *converges to structural balance in finite time.*

These Monte Carlo results are expected, since it has been formally proved that the Kułakowski et al. model converges to structural balance only under generic symmetric initial conditions as $n \to \infty$ [119] and negative results for asymmetric conditions are given by [164].

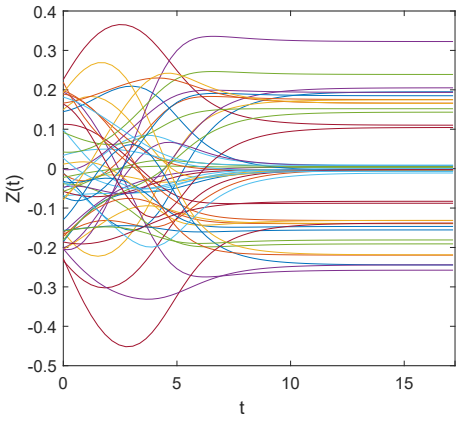
See Figure 3.4 for a comparison of trajectories of the pure-influence model in both generic and symmetric generic initial conditions. Figure 1.4 shows a comparison between our projected pure-influence model, which does not consider self-appraisals, and the projected influence model, which considers self-appraisals. Note how not considering self-appraisals drastically changes the convergence time as well as the dynamic behavior of the interpersonal appraisals.



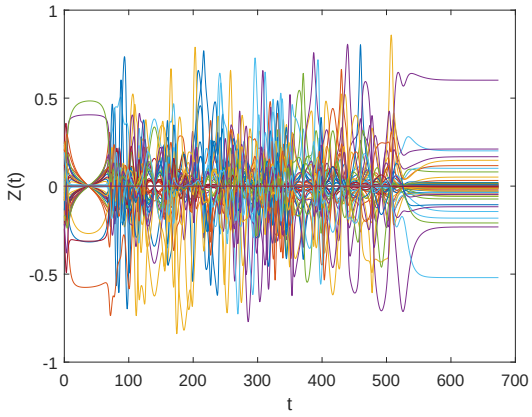
(a) Projected pure-influence model (1.9) with generic asymmetric initial condition

(b) Projected pure-influence model (1.9) with generic symmetric initial condition

Figure 1.3: Convergence to structural balance for a network of size $n = 10$. We plot the evolution of all the entries of $Z(t)$.



(a) Projected influence model (1.10) with generic asymmetric initial condition



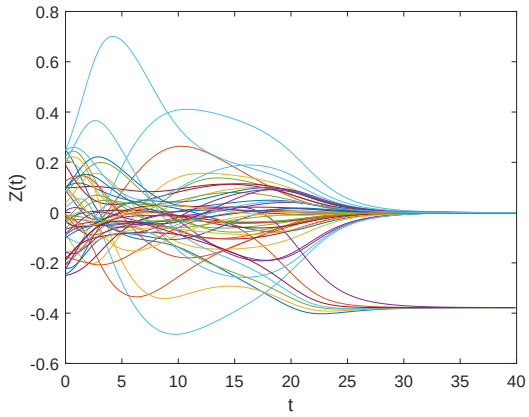
(b) Projected pure-influence model (1.9) with generic asymmetric initial condition

Figure 1.4: Convergence comparison for a network of size $n = 7$ (a) with and (b) without the consideration of self-appraisals. We first generated an $n \times n$ random matrix W with each entry independently sampled from a uniform distribution with support $[-100, 100]$. Then, for (a), we normalize this matrix to have unit Frobenius norm and used it as the initial condition. For (b), we set the diagonal entries of W to zero and then normalize it to have unit Frobenius norm and use it as the initial condition. In this example, (a) did not converged to structural balance, whereas (b) did. We plot the evolution of all the entries of the appraisal matrix.

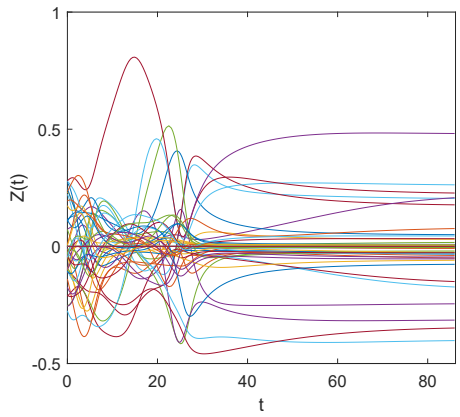
1.7 Conclusion

We propose two new dynamic structural balance models that incorporates more psychologically plausible assumptions than previous models in the literature, based on a modification by a model proposed by Kułakowski et al. We have established important convergence properties for these models and also that, most importantly, they correspond to gradient systems over an energy function that characterizes the violations of Heider's axioms for the symmetric case. We also expanded our results to a set of asymmetric matrices called scale-symmetric. Numerical results illustrates that, under generic initial conditions, our models converges to structural balance (for sufficiently large n) and thus have better convergence properties than the previous model by Kułakowski et al.

As future work, we propose to further study the general case of asymmetric (and non-



(a) Projected influence model (1.10) with generic asymmetric initial condition



(b) Projected pure-influence model (1.9) with generic asymmetric initial condition

Figure 1.5: Convergence comparison for a network of size $n = 7$ (a) with and (b) without the consideration of self-appraisals. The setting is the same one as in Figure 1.4, but with a different random initial condition. (a) converged to a network with only diagonal negative entries (all interpersonal appraisals go to zero), whereas (b) converged to structural balance.

scale-symmetric) equilibria and the convergence properties of our models under arbitrary initial conditions. For example, numerical simulations of the projected pure-influence model from generic (asymmetric) initial conditions illustrate how this system features transient chaos before converging towards an equilibrium. Future work will focus on models with a more sociologically justified transient behavior. Finally, one could study the removal of the self-appraisals in other dynamical structural balance models, like the homophily-based model by Traag el al. [164].

1.8 Appendix

1.8.1 Supporting results and proofs

Lemma 1.8.1 *Let $x(t)$ be the solution to $\dot{x} = f(x)$ from initial condition $x(0)$, with f being a continuously differentiable vector field. Let η be a positive continuous scalar*

function. Then, $z(t)$ is the solution to $\dot{z} = \eta(t)f(z)$ with initial condition $z(0) = x(0)$ if and only if $z(t) = x(\int_0^t \eta(s)ds)$.

Proof: Consider the time transformation $\bar{t}(t) = \int_0^t \eta(s)ds$, which is well-defined since it is continuous and monotonically increasing on t (recall that $\eta(s) > 0$ for $s \in [0, t]$), with $\bar{t} = 0$ if and only if $t = 0$. Now, from the chain rule, it follows that

$$\frac{dz}{dt} = \frac{dx(\bar{t})}{d\bar{t}}\frac{d\bar{t}}{dt} = f(z)\eta(t), \quad z(0) = x(0).$$

This finishes proof of the "if" part. The "only if" part follows from the uniqueness theorem. ■

Proof: [**Proof of Lemma 1.4.1**] First, to prove that the set $\mathrm{nSt}(n, k)$, $k \leq n$ is a submanifold of the compact Stiefel manifold, define the smooth map $\Phi : \mathrm{St}(n, k) \to \mathbb{R}^n$ by $X \mapsto (\|X_{i*}\|_2^2, \ldots, \|X_{n*}\|_2^2)^\top$, where X_{i*} is the ith row of X. Then, we have that $\mathrm{nSt}(n, k) = \Phi^{-1}((k/n, \ldots, k/n)^\top)$ and it is easy to prove the mapping Φ has constant rank n. Thus, we use the Constant-Rank Level Set Theorem [102] to conclude our claim. The properties of compactness and analyticity are immediate from the definition of the set $\mathrm{nSt}(n, k)$, $k \leq n$.

Now, notice that conditions ((i)) and ((ii)) from Definition 1.4.1 impose, in total, $\frac{k(k+1)}{2} + n$ constraints on kn independent variables, however, these constraints are linearly dependent: one of them can be removed (for instance, if one requires condition (i) from Definition 1.4.1, then suffices to constrain only sums of $n-1$ rows, whereas the remaining sum automatically equals k/n)). Whenever $k \leq n$ and $n \geq 3$, one has $\frac{k(k+1)}{2} + n - 1 < kn$, which implies that the set $\mathrm{nSt}(n, k)$ has the dimension $(k-1)n + 1 - k(k+1)/2$.

Statements (i) and (ii) are immediate. Now regarding (iii), it is obvious that each row has norm $\sqrt{k/n}$ if and only if V can be written as (1.16). Notice now the columns are unit vectors if and only if $\sum_{m=1}^n \cos^2 \alpha_i = n/2 = \sum_{m=1}^n \sin^2 \alpha_i$, which in turn holds if

and only if $\sum_m \cos 2\alpha_m = 2\sum_m \cos^2 \alpha_m - n = 0$. Similarly, the columns are orthogonal if and only if $\sum_{m=1}^{n} \cos\alpha_i \sin\alpha_i = 0 = \frac{1}{2}\sum_m \sin 2\alpha_m$. These two constraints are equivalent to (1.17). ■

Proof: [**Proof of Lemma 1.4.2**] The case where $\alpha = \beta = 0$ is trivial: $Z = 0$ and it obviously can be decomposed as in (1.18) with $p = q = 0$. Notice that every eigenvalue of $Z = Z^\top$ corresponds to the eigenvalue $\lambda^2 - 2\alpha\lambda$ of $Z^2 - 2\alpha Z$, and hence $\lambda^2 - 2\alpha\lambda - \beta = 0$. Therefore, $\alpha^2 + \beta \geq 0$ (otherwise, eigenvalues of Z would be complex). Furthermore, $\alpha^2 + \beta \neq 0$ (otherwise, $\lambda = \alpha$ would be the only eigenvalue of Z of multiplicity n, and one would have $\text{trace}(Z) = \alpha n$, entailing that $\alpha = \beta = 0$). Denoting $\Delta = \sqrt{\alpha^2 + \beta}$, the matrix Z has two different eigenvalues $\alpha + \Delta$ and $\alpha - \Delta$, denote their multiplicities by k and $n - k$. Then $(\alpha + \Delta)k + (\alpha - \Delta)(n - k) = 0$. Denoting $q = \Delta - \alpha$ and $p = 2\Delta > 0$, one has $(p - q)k - q(n - k) = 0$ or, equivalently, $pk = qn$ thus, $q > 0$.

Consider the orthonormal eigenvectors $v_1, \ldots, v_k$, corresponding to the eigenvalue $p - q = \alpha + \Delta$ and orthonormal eigenvectors $w_1, \ldots, w_{n-k}$, corresponding to $-q = \alpha - \Delta$. The sequence $v_1, \ldots, v_k, w_1, \ldots, w_{n-k}$ constitutes an orthonormal basis of eigenvectors for the operator Z. Stacking the columns v_i and w_i, one obtains $n \times k$ and $n \times (n - k)$ matrices $V = (v_1, \ldots, v_k)$, $W = (w_1, \ldots, w_{n-k})$. The matrix $[V, W]$ is orthonormal and diagonalizes Z, that is, $Z[V, W] = [V, W] \begin{bmatrix} (p-q)I_k & 0 \\ 0 & -qI_{n-k} \end{bmatrix}$ and thus $Z = (p - q)VV^\top - qWW^\top$. Since $VV^\top + WW^\top = I_n$, Z is decomposed as (1.18). It remains to notice that $V^\top V = I_k$ by definition of the orthonormal basis and $\text{diag}(VV^\top) = (q/p)I_n = (k/n)I_n$ since, by (1.18), $\text{diag}(Z) = \mathbb{0}_{n\times n}$. To finish the proof, notice that $p - 2q = 2\alpha$ and $\beta = \Delta^2 - \alpha^2 = (\Delta - \alpha)(\Delta + \alpha) = q(p - q)$. ■

Proof: [**Proof of Corollary 1.4.3**] Let $f(z) = z^2 - 2\alpha z$, $z \in \mathbb{C}$. It suffices to show that, if $f(Z) = \text{diag}(\beta_1 I_{n_1}, \ldots, \beta_s I_{n_s})$, then $Z = \text{diag}(Z_1, \ldots, Z_s)$, where $f(Z_i) = \beta_i I_{n_i}$. This statement will be proved for any analytic function $f(z)$. It is well known

that the spectrum of $f(Z)$ consists of all points $f(\lambda)$, where λ is an eigenvalue of Z. Consider the set of eigenvalues of Z that belong to $f^{-1}(\beta_i)$ and let $\mathcal{X}_i$ be the sum of corresponding eigenspaces. Then $\mathcal{X}_i$ is invariant under the operator Z, and $\mathbb{R}^n = \oplus_{i=1}^s \mathcal{X}_i$ (the sum is orthogonal). Also, $f(Z)x = \beta_i x$ for any $x \in \mathcal{X}_i$. For any basis vector $e_r = (0, \ldots, 1, \ldots, 0)^\top$ consider the decomposition $e_r = \oplus_{i=1}^s e_r^i$, $e_r^i \in \mathcal{X}_i$. Then $Ze_r = \oplus_{i=1}^s Ze_r^i$, $Ze_r^i \in \mathcal{X}_i$ and $f(Z)e_r = \oplus_{i=1}^s f(Z)e_r^i = \oplus_{i=1}^s \beta_i e_r^i$. Suppose that $1 \leq r \leq n_1$. Then $f(Z)e_r = \beta_1 e_r$. Since $\beta_1, \ldots, \beta_s$ are pairwise different, we have $e_r = e_r^1$ and $e_r^2 = \ldots = e_r^s = 0$. Similarly, for $n_1 + n_2 + \ldots + n_{j-1} + 1 \leq r \leq n_1 + n_2 + \ldots + n_{j-1} + n_j$ one has $e_r = e_r^j$ $(j = 2, \ldots, s)$.

In other words, each $\mathcal{X}_i$ contains n_i basis vectors e_r, where $n_1 + n_2 + \ldots + n_{i-1} + 1 \leq r \leq n_1 + n_2 + \ldots + n_{i-1} + n_i$ and thus $\dim \mathcal{X}_i \geq n_i$. Recalling that $n_1 + \ldots + n_s = n$, one shows that $\dim \mathcal{X}_i = n_i \, \forall i$ and thus $\mathcal{X}_i$ is spanned by the corresponding basis vectors. Since $\mathcal{X}_i$ is invariant under Z, $Z = \mathrm{diag}(Z_1, \ldots, Z_s)$, where the block Z_i has dimension $n_i \times n_i$. Obviously, $f(Z_i) = \beta_i I_{n_i}$. The statement of Corollary is now immediate from Lemma 1.4.2. ■

Proof: [**Proof of Theorem 1.4.5**] We prove the necessity first. Denoting $2\alpha = -\mathcal{D}(Z)$. By assumption, $Z^2 - 2\alpha Z$ is diagonal. Statements (i) and (ii) follow from Corollary 1.4.3, entailing also that p_i, q_i can be represented as (1.23) with some β_i. Since $Z_i^2 = 2\alpha Z_i + \beta_i I_{n_i}$ and $\mathrm{diag}(Z_i) = \mathbb{0}_{n_i \times n_i}$, one has $\mathrm{trace}\, Z_i^2 = \beta_i n_i$, therefore

$$\sum\nolimits_{i=1}^{s} \beta_i n_i = \mathrm{trace}(Z^2) = 1. \tag{1.29}$$

Recall also that for each i one has $p_i k_i = q_i n_i$ or, equivalently,

$$\frac{2k_i}{n_i} = \frac{\sqrt{\alpha^2 + \beta_i} - \alpha}{\sqrt{\alpha^2 + \beta_i}} = 1 - \frac{\alpha}{\sqrt{\alpha^2 + \beta_i}} \quad \forall i : p_i, q_i \neq 0.$$

(if $\alpha = 0$, one always has $p_i, q_i \neq 0$, otherwise it is possible that $\beta_i = 0$ and then $Z_i = 0$). This implies condition 3 ($\varepsilon = \operatorname{sign}\alpha$) and allows to determine α, β_i. In the case where $\varepsilon \neq 0$ notice that $n_i - 2k_i \neq 0$ for any i such that $Z_i \neq 0$. Thus

$$\frac{\beta_i + \alpha^2}{\alpha^2} = \frac{n_i^2}{(n_i - 2k_i)^2} \Longleftrightarrow \beta_i = \alpha^2 \frac{4n_i k_i - 4k_i^2}{(n_i - 2k_i)^2}.$$

In view of (1.29), one obtains that

$$\alpha = \varepsilon \left(\sum_{i: Z_i \neq \mathbb{0}_{n_i \times n_i}} \frac{4k_i n_i (n_i - k_i)}{(n_i - 2k_i)^2} \right)^{-1/2},$$

which entails (1.24). In the case of $\alpha = 0$, one has $p_i = 2\sqrt{\beta_i}, q_i = \sqrt{\beta_i}$ for any i, and (1.29) implies that $\sum_i q_i^2 n_i = 1$. This finishes the proof of statement (iv).

The proof of sufficiency is similar. For any i such that $Z_i \neq 0$, the coefficients p_i, q_i have the form (1.23) (if $\varepsilon \neq 0$, this is implied by (iv)a, otherwise we choose $\alpha = 0$ and $\beta_i = q_i^2 = p_i^2/4$). Therefore, we have $Z_i^2 - 2\alpha Z_i = \beta_i Z_i$ and, in particular, $Z^2 - 2\alpha Z$ is diagonal. A straightforward computation shows that $p_i k_i = q_i n_i$ and thus $\operatorname{diag}(Z_i) = \mathbb{0}_{n_i \times n_i} \, \forall i$, in particular, $\operatorname{diag}(Z) = \mathbb{0}_{n \times n}$. Also, $\operatorname{diag}(Z_i^2) = \beta_i I_{n_i}$, and statement (iv) now implies that $\operatorname{trace} Z^2 = 1$. It remains to notice that $Z_i^3 = 2\alpha Z_i^2 + \beta_i Z_i$, and hence $\operatorname{trace}(Z_i^3) = 2\alpha\beta_i n_i$. Hence, $\mathcal{D}(Z) = -\operatorname{trace}(Z^3) = -2\alpha$, $Z^2 + \mathcal{D}(Z)Z$ is a diagonal matrix, and Z is an equilibrium (1.14). ■

1.8.2 Scale-symmetric matrices

We now generalize our results for symmetric appraisal networks to a class of asymmetric matrices. We define the sets of *scale-symmetric* matrices

$$\begin{aligned}\mathbb{R}^{n\times n}_{\text{zero-diag,dss}} &= \{A \in \mathbb{R}^{n\times n}_{\text{zero-diag}} | \text{there exists } \gamma \succ \mathbb{0}_n \text{ such that} \\ &\qquad A\,\mathrm{diag}(\gamma) = (A\,\mathrm{diag}(\gamma))^\top\}, \\ \mathbb{S}^{n\times n}_{\text{zero-diag,dss}} &= \mathbb{S}^{n\times n}_{\text{zero-diag}} \cap \mathbb{R}^{n\times n}_{\text{zero-diag,dss}}.\end{aligned}$$

Note that $\mathbb{S}^{n\times n}_{\text{zero-diag,dss}} \supset \mathbb{S}^{n\times n}_{\text{zero-diag,symm}}$ and

$$\begin{aligned}\mathbb{S}^{n\times n}_{\text{zero-diag,dss}} &= \bigcup_{\gamma\succ\mathbb{0}_n} \mathbb{S}^{n\times n}_{\text{zero-diag,dss}}(\gamma), \\ \mathbb{S}^{n\times n}_{\text{zero-diag,dss}}(\gamma) &= \left\{A \in \mathbb{S}^{n\times n}_{\text{zero-diag}} \mid A\,\mathrm{diag}(\gamma) = (A\,\mathrm{diag}(\gamma))^\top\right\}.\end{aligned}$$

.

Lemma 1.8.2 *Consider any $\gamma \succ \mathbb{0}_n$ and some matrix $A \in \mathbb{R}^{n\times n}$ such that $A\,\mathrm{diag}(\gamma) = \mathrm{diag}(\gamma)A^\top$. Then,*

(i) A has real eigenvalues and it is diagonalizable,

(ii) $\mathrm{trace}(A^2) = 0$ if and only if $A = 0$.

Proof: Since $A\,\mathrm{diag}(\gamma)$ is symmetric, then $A' = \mathrm{diag}(\gamma)^{-1/2}A\,\mathrm{diag}(\gamma)^{1/2}$ is also symmetric and thus has real eigenvalues and its eigenvectors form an orthogonal basis. Now, let (λ, v) be an eigenpair for A'. Then, by defining $u = \mathrm{diag}(\gamma)^{1/2}v$, we observe that $Au = \lambda u$, and so $(\lambda, \mathrm{diag}(\gamma)v)$ is an eigenpair for A. Hence the eigenvectors of A form a basis, and thus A is diagonizable. This proves (i).

Observe that

$$A = \mathrm{diag}(\gamma)A^\top\,\mathrm{diag}(\gamma)^{-1}.$$

Then, $\text{trace}(A^2) = \text{trace}(A\,\text{diag}(\gamma)A^\top\,\text{diag}(\gamma)^{-1})$. From simple algebraic operations, it can be found that $\text{trace}(A^2) = \sum_{i=1}^n \sum_{j=1}^n \frac{\gamma_j}{\gamma_i} a_{ij}^2$. Since $\frac{\gamma_i}{\gamma_j} > 0$, $\text{trace}(A^2) = 0$ if and only if $A = 0$. This proves (ii). ■

In view of Lemma 1.8.2, a matrix A is scale-symmetric if and only if $A = D^{-1}A_sD$, where $D > 0$ is a positive diagonal matrix (in Lemma 1.8.2, $D = \text{diag}(\gamma^{-1/2})$ for some $\gamma \succ \mathbb{0}_n$) and A_s a symmetric matrix.

Recall the invariance property of the pure-influence model (1.2): if $X(0) = X(0)^\top$, then $X(t) = X(t)^\top$ for all $t > 0$. We are now ready to provide a more general version of this property: If $D > 0$ is a diagonal matrix and $X(t)$ is a solution, then $DX(t)D^{-1}$ is also a solution. For this reason, if $X(0) = DX_s(0)D^{-1}$ is a scale-symmetric matrix with some $X_s(0) = X_s(0)^\top$, then the solution $X(t) = DX_s(t)D^{-1}$ is scale-symmetric. A similar result holds for the projected pure-influence model (1.9). Indeed, all of the theoretical results obtained in this paper for symmetric appraisal matrices, can be generalized to scale-symmetric appraisal matrices. For example, if $X(0) \in \mathbb{R}^{n\times n}_{\text{zero-diag,dss}}$ ($Z(0) \in \mathbb{S}^{n\times n}_{\text{zero-diag,dss}}$) then $t \mapsto \mathcal{D}(X(t))$ ($t \mapsto \mathcal{D}(Z(t))$) is monotonically nondecreasing in $\mathbb{R}^{n\times n}_{\text{zero-diag,dss}}$ ($\mathbb{S}^{n\times n}_{\text{zero-diag,dss}}$).

Chapter 2

Polarization and Fluctuations in Signed Social Networks

2.1 Introduction

There have been various opinion dynamics models in the literature [5, 144]. Opinions can be modeled as real numbers taking values in the closed interval $[0, 1]$, where 0 means an agent completely disagrees with a particular issue, and 1 that she completely agrees. One important question to answer is how the evolution and final distribution of opinions in a social network depend on the underlying network's topology and of the (positive) influence structure among the individuals. More recently, signed graphs were introduced into the opinion dynamics literature. Signed graphs represent a natural way to model positive and negative relationships among individuals. For example, a sociological relevant concept is *structural balance*, in which the members of a social network can either have only positive relationships or be divided in two factions in which members of the same faction have positive relationships but negative ones with members of the other faction. The seminal work by Altafini [7] proposed a continuous time model over a signed graph

where the opinions can take any real value. It is shown that when the underlying graph satisfies structural balance (and assuming that it is strongly connected), the opinions converge to bipartite consensus and polarize, i.e., all opinions have the same absolute value with their signs indicating which agents belong to the same faction (if there is one faction, all opinions have the same sign). A discrete-time signed opinion model which is a counterpart of the Altafini model has also been proposed [78, 124], in which bipartite consensus is also attained under structural balance. These two models have initiated a lot of research in the field of signed opinion dynamics, and are, arguably, the most popular models in the literature. Extensions of these models and further analysis have been done in the literature, as can be noted in the recent work [108] and the references therein. Note, however, that both Altafini models and their extensions present an unrealistic opinion vanishing behavior (i.e., the opinions converge to zero) whenever the property of structural balance is lost in the underlying social network, with the underlying graph still being strongly connected.

Another class of models in opinions dynamics was proposed by Li et al. [104] and is based on an extension of the voter model to signed graphs. In this model, individuals initially take binary opinion values (e.g., 0 and 1). Then, at each subsequent time step, an individual is selected according to some process and updates her opinions by copying the same or the opposite opinion of one of her neighbors according to the sign of their relationship. By design, opinions cannot vanish under generic signed networks; however, the opinion values are simply discrete. Whenever the graph satisfies structural balance, they showed that the opinions polarize: one faction takes one value, while the other faction takes the remaining one. Recently, Lin et al. [107] proposed a model which can be regarded as an extension to the one from Li et al. In this model, opinions can take m different discrete values from a set S. Then, an individual will copy the same opinion from a positive neighbor, but when facing a negative one, will randomly select an opinion

different from that neighbor from the set S.

In this paper, we propose a novel opinion model over signed graphs. We assume that the opinions are real numbers taking value in a closed interval and each edge of the graph indicates the friendly or antagonistic relationship between two individuals. Our model is inspired by the *boomerang effect* studied in social psychology [46, 35, 1], which aims to explain why in some situations where two individuals engage in communication, they may not end up being in a better agreement but rather their attitudes become more dissentive, i.e., their opinions do not go in the *intended direction* (e.g., consensus or agreement) but in the *opposite direction* (e.g., polarization). The early work [84] suggested that this phenomenon can be explained by *"the relative distance between subjects' attitudes and position of communication"*. Our model is motivated by the empirical observations in the social sciences (e.g., from the study of interpersonal attraction [17]) that two friendly agents will be closer in their attitudes and perspectives than two unfriendly agents. Specifically, we make the following assumption: whenever two agents who have a positive relationship interact, they are more agreeable and their opinions will become closer or even be in consensus, i.e., the opinion *changes in the intended direction*. On the other hand, whenever two agents with a negative relationship interact, the differences in their opinions will be more polarized after the interaction because of their increasing disagreement, i.e., the opinion *changes in the opposite direction*. Our opinion model captures such behavior mathematically, and we call it the *affine boomerang model*. Mathematically, our proposed model is an affine model, which makes it remarkably simple, and its dynamics are self-explanatory. Besides a linear model like the discrete Altafini model, this is, arguably, the next simplest model structurally.

Our second contribution is a formal analysis of our proposed model: under certain conditions on the sign structures of the network that corresponds to structural balance, our model expresses opinion polarization, i.e., the opinions of two groups converge to

opposite extreme values of the closed interval.

Finally, it is important to compare our model and the aforementioned models in the literature. Our model has the property that opinions do not necessarily vanish whenever the graph is not balanced, but, for example, can continue fluctuating inside the closed interval. The vanishing behavior, which we mentioned happens in both types of Altafini models and their extensions, has been interpreted as if the agents in the network become neutral or indifferent towards a specific topic. In the case of three antagonistic groups in a connected network, this would mean that all groups will end up having a zero valued opinion, i.e., they will have consensus on *not having an opinion*. This might be difficult to interpret. Instead, our proposed model predicts that two groups will polarize their opinions and the third one will continue fluctuating its opinions since its members observe people they dislike having opposite opinions. Thus, this third group does not settle down to a definite opinion and its members are persistently disagreeing with each other. This is, arguably, more intuitive since individuals of a social network can always hold an opinion, independently of whether their network is balanced or not. Moreover if we have an unbalanced network that differs from a balanced one in just the sign of one edge, it is not clear why that would drive the whole social network towards an indifferent opinion. Instead, our model suggests that opinions may fluctuate around extreme values of opinion, which is more intuitive since the underlying social network is *approximately* balanced.

2.2 The model

A *signed graph* G is an undirected graph with signed edges, i.e., with edge weights equal to either $+1$ or -1. Let $\mathcal{E} = \mathcal{E}_+ \cup \mathcal{E}_-$ be the edge set of G, where $\mathcal{E}_+$ is the set of positive edges and $\mathcal{E}_-$ the set of negative edges. G is complete when there exists an edge

between any pair of vertices. A path from vertex i to j in G is a sequence of edges that connect a sequence of distinct vertices starting from i and finishing at j. A connected component is any subgraph such that all of its vertices are connected to each other by paths, but they are not connected to any other vertex of G. G is connected whenever it has only one connected component. The abbreviation *i.o.* stands for *infinitely often*.

We model the structure of a social network composed by agents as a graph. Then, throughout the paper, we use the words *graph* and *network* interchangeably, as well as the terms *vertex* and *agent*. Each agent in the network holds an opinion about a particular statement of a discussion topic, and her opinion describes how much she agrees with it. An agent i has an opinion $x_i \in [o_{\min}, o_{\max}]$: $x_i = o_{\max}$ whenever i completely agrees with the statement being discussed, and $x_i = o_{\min}$ whenever she completely disagrees with it. The opinion vector $x \in [o_{\min}, o_{\max}]^n$ has in its ith entry the opinion x_i of agent i.

Definition 2.2.1 (Sign arrangement property) *Given a connected signed graph* $G = (\{1, \dots, n\}, \mathcal{E}_+ \cup \mathcal{E}_-)$ *with* $n \geq 3$, *let* $G_+ = (\{1, \dots, n\}, \mathcal{E}_+)$. *For* $k \in \mathbb{N}$, *we say that* G *satisfies the* k-sign arrangement property *if*

(i) G_+ *has* $k \geq 1$ *connected components, and*

(ii) each negative edge connects vertices belonging to different connected components of G_+.

If this property holds, then each connected component of G_+ *is a* faction.

Based on the works [38, 54] in the sociological literature, we definite the notion of *structural* and *clustering balance* for connected graphs.

Definition 2.2.2 (Structural and clustering balance) *Consider a connected signed graph* G *with* $n \geq 3$. *Assume the vertices of* G *can be partitioned in* m *groups such that*

each positive edge joins two vertices from the same group and each negative edge joins vertices from different groups. We say that G satisfies

(i) structural balance *if $m \leq 2$, and*

(ii) clustering balance *if $m \geq 3$.*

The following result follows immediately from the previous definitions.

Lemma 2.2.1 *Let G be a complete signed graph. G satisfies the k-sign arrangement if and only if it satisfies structural balance when $k \leq 2$ or clustering balance when $k \geq 3$.*

Note that a signed graph satisfying the k-sign arrangement property does not need to be complete.

Definition 2.2.3 (Affine boomerang model) *Let $G = (\{1, \ldots, n\}, \mathcal{E}_+ \cup \mathcal{E}_-)$ be a signed graph. Assume that each agent has an initial opinion $x_i(0) \in [o_{\min}, o_{\max}]$, $o_{\min} < o_{\max}$, and a self-weight $a_i \in (0, 1)$. At each time step $t \in \mathbb{Z}_{\geq 0}$, select randomly an edge of G; assume each edge $\{i, j\}$ has a time-invariant positive selection probability p_{ij}. Update the opinions of the two agents i and j according to:*

$$x_i(t+1) = \begin{cases} a_i x_i(t) + (1 - a_i) x_j(t), & \text{if } \{i,j\} \in \mathcal{E}_+, \\ a_i x_i(t) + (1 - a_i) o_{\min}, & \\ \qquad \text{if } \{i,j\} \in \mathcal{E}_- \text{ and } x_i(t) < x_j(t), & \\ a_i x_i(t) + (1 - a_i) o_{\max}, & \\ \qquad \text{if } \{i,j\} \in \mathcal{E}_- \text{ and } x_i(t) \geq x_j(t), & \end{cases} \tag{2.1}$$

and similarly for agent j.

The following remarks interpret and elaborate on the various features of the model.

Remark 2.2.2 (Bounded evolution) *In our model, opinions take values on an arbitrary closed interval* $[o_{\min}, o_{\max}]$*. From a sociological (and intuitive) point of view, it is plausible to have bounded opinions since there is no clear interpretation of diverging opinions. Indeed, bounded opinions are present throughout the literature on opinion dynamics. The case* $o_{\min} = -\theta$ *and* $o_{\max} = \theta$*, for* $\theta > 0$*, is characteristic in the literature on bipartite consensus (e.g., [7, 156]), and the case* $o_{\min} = 0$ *and* $o_{\max} = 1$ *is characteric in the literature of opinion dynamics over graphs with positive weights (e.g., [5]) or bounded-confidence models (e.g., [39]).*

Remark 2.2.3 (Asynchronous updating) *Our model features* asynchronous updating *of the opinions since only two opinions are updated simultaneously and independently per time step, instead of all opinions at once (which would be* synchronous updating*). This type of updating has been studied in other previous opinion models, e.g., in the Deffuant-Weisbuch model [39] and in the gossip model [31]. A classic strategy is to assign homogeneous selection probability to each edge in the graph.*

Remark 2.2.4 (Magnitude of the boomerang effect) *Assume* i *and* j *are two agents with an antagonistic relationship and, without loss of generality, assume* $x_i(t) > x_j(t)$*. As stated in the last case of equation* (4.3)*, when* i *and* j *interact, the opinion* $x_i(t+1)$ *jumps towards the extreme opinion* $o_{\max}$ *and does so independently of the opinion difference* $d_{ij}(t) = |x_i(t) - x_j(t)|$*. Note that this is a simplifying assumption in the sense that the jump magnitude (i.e., the magnitude of the boomerang effect) could be assumed to be directly proportionally to* $d_{ij}(t)$*, inversely proportional to* $d_{ij}(t)$*, or , more generally, a positive function of* $d_{ij}(t)$ *ensuring boundedness of the evolutions. This simplifying assumption is justified because, to the best of our knowledge, no empirical evidence or psychological theory is available about the jump magnitude or about whether the magnitude should even depend upon* d_{ij} *at all. In this sense, our independence assumption is*

arguably the simplest (and therefore preferrable) model.

2.3 Model analysis

In this section, we first introduce the two main theoretical results from our paper, namely, that the affine boomerang model can explain, given certain conditions on the underlying social network, both the polarization of opinions and their persistent fluctuations. Additionally, we present some numerical results on how opinions attempt to polarize when these conditions on the underlying social network are relaxed.

2.3.1 Theoretical results

Theorem 2.3.1 (Consensus and polarization in signed graphs) *Consider a network satisfying the k-sign arrangement property. Consider the evolution of the affine boomerang model* (4.3) *with initial opinion vector $x(0) \in [o_{\min}, o_{\max}]^n$. Then*

(i) Consensus: if $k = 1$, then, with probability one, $\lim_{t\to\infty} x(t) = c\mathbb{1}_n$, where c is a random convex combination of the entries of $x(0)$.

(ii) Polarization: if $k = 2$, then, with probability one, $\lim_{t\to\infty} x_i(t) = o_{\min}$ for each agent i of one of the two factions and $\lim_{t\to\infty} x_j(t) = o_{\max}$ for each j of the other faction.

Proof: Formally, at any time step, the selected edge is a discrete random variable over some probability space $(\Omega', \mathcal{F}', \mathbb{P}')$ with Ω' being the set of all edges on the graph, $\mathcal{F}'$ the power set, and $\mathbb{P}'[\{i, j\}] = p_{ij}$. Let $\omega(t)$ be the random edge selected at time t, then, the collection of random variables $\{\omega(t) \mid t \in \mathbb{Z}_{\geq 0}\}$ forms a stochastic process of an independent sequence of random variables. Then, an adequate probability space $(\Omega, \mathcal{F}, \mathbb{P})$

can be constructed with $\Omega = \prod_{t\in\mathbb{Z}_{\geq 0}} \Omega'$, $\mathcal{F}$ being the product of σ-algebras $\mathcal{F}'$ over $t \in \mathbb{Z}_{\geq 0}$, and $\mathbb{P}$ being the product probability measure $\prod_{t\in\mathbb{Z}_{\geq 0}} \mathbb{P}'$. Therefore, given the sequence of edges $\{s(t)\}_{t\in S}$ with some finite set $S \subset \mathbb{Z}_{\geq 0}$, $\mathbb{P}[\{\omega \in \Omega \mid \omega(t) = s(t), t \in S\}] = \prod_{t\in S} \mathbb{P}'[s(t)]$.

We start by considering the case $k = 1$. In this case, the model is a linear system of the form $x(t+1) = W(t)x(t)$, where $W(t)$ is a random matrix that takes, at each time step, the value $W_{ij} = I_{n\times n} - (1-a_i)e_i(e_i - e_j)^\top - (1-a_j)e_j(e_j - e_i)^\top$ whenever the edge $\{i, j\}$ is selected to be updated with probability p_{ij} (here, e_i is the ith column of the identity matrix $I_{n\times n}$). With probability one, $W(t)$ is a row stochastic matrix with a strictly positive diagonal for any t; moreover $W(t)$ is independent and identically distributed for any t. Thus, $\mathbb{E}[W(t)]$ (with respect to $\mathbb{P}'$) is a row stochastic matrix that, when interpreted as an adjacency matrix, corresponds to a connected undirected network. Under these assumptions, [32, Theorem 13.1] implies the first statement of the theorem.

Now we prove the case $k = 2$. For any opinion vector $x \in [o_{\min}, o_{\max}]^n$, we define the variable $Z : [o_{\min}, o_{\max}]^n \to \{1, 2\}$ as

(C1) $Z(x) = 1$ when there is no value $\tau > 0$ such that one faction has all of its opinions above τ and the other faction has them equal or below it;

(C2) $Z(x) = 2$ when there exists a value $\tau > 0$ such that one faction has all of its opinions above τ and the other faction has them equal or below it.

Clearly, Z exhausts all possible situations for the values of the opinion vector x, and, moreover, induces a partition over the set $[o_{\min}, o_{\max}]^n$: $[o_{\min}, o_{\max}]^n = \cup_{m=1}^{2} Z^{-1}(m)$ and $Z^{-1}(1) \cap Z^{-1}(2) = \emptyset$.

Now, let us remark that, from the random selection process of the edges, it immediately follows that $\{x(t)\}_{t>0}$ is a random process over the probability space $(\Omega, \mathcal{F}, \mathbb{P})$; and,

moreover, it is a Markov process, i.e., $\mathbb{P}[x(t) \in Z^{-1}(m) \,|\, x(t-1) = c_{t-1}, \ldots, x(0) = c_o] = \mathbb{P}[x(t) \in Z^{-1}(m) \,|\, x(t-1) = c_{t-1}]$ for any $m \in \{1, 2\}$. Observe that, with probability one, $x(t) \in [o_{\min}, o_{\max}]^n$ for any t since $x(0) \in [o_{\min}, o_{\max}]^n$.

Now, assume that $x(t) \in Z^{-1}(2)$ for some $t < \infty$. Let F_1 be the faction such that $x_i(t) \leq \tau$ for any $i \in F_1$; and F_2 the one such that $x_i(t) > \tau$ for any $i \in F_2$. Let $\theta_{F_1}(t) = \max_{i \in F_1} x_i(t)$ and $\theta_{F_2}(t) = \min_{i \in F_2} x_i(t)$. If at $t+1$ some $i \in F_1$ and $j \in F_2$ are selected, we have that $x_i(t+1) < x_i(t)$ and $x_j(t+1) > x_j(t)$. On the other hand, if at $t+1$ both i and j belong to the same faction with $x_i(t) \leq x_j(t)$, we have that $x_i(t) \leq x_i(t+1), x_j(t+1) \leq x_j(t)$, with equality if and only if $x_i(t) = x_j(t)$. From these two observations it is easy to show that $\theta_{F_1}(t+1) \leq \theta_{F_1}(t)$ with probability one; i.e., $\{\theta_{F_1}(s)\}_{s \geq t}$ is a non-decreasing sequence which is lower bounded by $o_{\min}$. This implies convergence of $\{\theta_{F_1}(s)\}_{s \geq t}$ to some lower bound $c_{\min}$ with probability one. Now, for any $\epsilon > 0$ and $t^* \geq t$, there exists some finite $T > 0$ such that if the sequence of edges $\{(\theta_{F_1}(s), k(s))\}_{s=t^*}^{t^*+T}$ with $k(s) \in F_2$ for $t^* \leq s \leq t^* + T$ is selected, then $|\theta_{F_1}(t^* + T) - o_{\min}| < \epsilon$. Such sequence has a positive probability of being selected sequentially by the affine boomerang model for any t^*, from which it follows that $c_{\min} = o_{\min}$. Therefore, there is polarization for any $i \in F_1$ towards $o_{\min}$. A similar reasoning leads to the proof that $\{\theta_{F_2}(s)\}_{s \geq t}$ has an analogous increasing monotonic behavior and thus that there is polarization for $i \in F_2$ towards $o_{\max}$ with probability one. In conclusion, if $x(t) \in Z^{-1}(2)$ for $t \geq 0$, then polarization occurs with probability one and we say that $Z^{-1}(2)$ is an *absorbing set* since the opinion vector cannot escape from it once it enters this set.

Therefore, to finish the proof of the theorem, we only need to prove that, given $x(t) \in Z^{-1}(1)$ at any time t, there always exists (with probability one) a finite sequence of edges such that eventually $x(t^*) \in Z^{-1}(2)$ for some $t < t^* < \infty$. Then, since any finite sequence of edges has positive probability of being selected sequentially by the affine boomerang model and $Z^{-1}(2)$ is an absorbing set, it follows that $\mathbb{P}[x(t) \in$

$Z^{-1}(1)$ *i.o.* $|\, x(0) \in Z^{-1}(1)] = 0$; and this finishes the proof for item (ii) of the theorem. Therefore, it suffices to prove that $\mathbb{P}[x(t+T) \in Z^{-1}(2)$ for some $T > 0 \,|\, x(t) = x_o] = 1$ for any $x_o \in Z^{-1}(1)$. So, let us fix any $x_o \in Z^{-1}(1)$. Let $\mathcal{T}_{1\to 2}(t) = \inf\{t^* > t : x(t^*) \in Z^{-1}(2) | x(t) = x_o\}$ be the first time, after starting in $x_o \in Z^{-1}(1)$ at time t, at which the opinion vector enters the set $Z^{-1}(2)$. If we show that $\mathbb{P}[\mathcal{T}_{1\to 2}(t) < \infty] = 1$ for any t, then we have finished the proof.

By the Markov property, we only need to show that $\mathbb{P}[\mathcal{T}_{1\to 2}(0) < \infty] = 1$. We start by noticing that, by the finite-time proximity property (Lemma 2.5.1), there exists a sequence of edges $s(0), \ldots, s(\tau - 1)$ for some $\tau > 0$ such that $x(\tau) \in Z^{-1}(2)$. Let $\gamma_o := \min_{\{i,j\}\in\mathcal{E}} p_{ij}$. Then,

$$\begin{aligned}
\mathbb{P}[x(\tau) \in Z^{-1}(2) | x(0) = x_o] &\geq \mathbb{P}[\omega(0) = s(0) | x(0) = x_o] \\
&\quad \times \mathbb{P}[\omega(1) = s(1) \,|\, x(0) = x_o, \omega(0) = s(0)] \ldots \\
&\quad \times \mathbb{P}[\omega(\tau - 1) = s(\tau - 1)\} \,| \\
&\qquad x(0) = x_o, \omega(\ell) = s(\ell) \text{ for } \ell \in [0, \tau - 2]] \\
&= \mathbb{P}'[s(0)]\mathbb{P}'[s(1)] \ldots \mathbb{P}'[s(\tau - 1)] \\
&\geq (\gamma_o)^\tau,
\end{aligned} \tag{2.2}$$

where the first inequality comes from a repetitive application of the conditional probability and the following equality comes from the independence of the underlying stochastic process. Let $\Gamma > 0$ be any integer and $A_\ell = \{x(t) \notin Z^{-1}(2), t \in [\ell, \ell + \tau]\}$, then $\mathbb{P}[A_0 | x(0) = x_o] \leq 1 - \gamma_o^\tau$. Likewise, in a way similar to how we obtained expression (2.2),

we compute

$$\begin{aligned}\mathbb{P}[\mathcal{T}_{1\to 2}(0) \geq (\tau+1)\Gamma] &= \mathbb{P}[x(t) \notin Z^{-1}(2), t \in [0,(\tau+1)\Gamma-1] | x(0)=x_o] \\ &= \mathbb{P}[\cap_{\ell=0}^{\Gamma-1} A_{\ell(\tau+1)} | x(0)=x_o] \\ &= \mathbb{P}[A_0 | x(0)=x_o] \\ &\quad \times \prod_{\ell=1}^{\Gamma-1} \mathbb{P}[A_{\ell(\tau+1)} | x(0)=x_o, \cap_{0\leq \ell' \leq \ell} A_{\ell'(\tau+1)}] \\ &\leq (1-\gamma_o^\tau)^\Gamma =: \gamma^\Gamma.\end{aligned}$$

Now, we observe that $\sum_{t=1}^{\infty} \mathbb{P}[\mathcal{T}_{1\to 2}(0) \geq (\tau+1)t] \leq \sum_{t=1}^{\infty} \gamma^t = \frac{\gamma}{1-\gamma} < \infty$ because of geometric series since $0 < \gamma < 1$. Then, by the first Borel-Cantelli lemma, we conclude that $\mathbb{P}[\mathcal{T}_{1\to 2}(0) < \infty] = 1$. This concludes the proof. ■

A consequence of Theorem 2.3.1 is that a complete social network that satisfies structural balance with two factions ends up having its agents with totally opposite opinions. This agrees with the intuitive result that antagonistic groups are expected to develop polarized opinions, as shown by other models in the literature [104, 108]. Also, as expected, if there are no negative relationships between the agents (i.e., there is only one faction), all agents reach consensus. Finally, we remark that, in our model, since the opinions converge to $o_{\min}$ and $o_{\max}$ in the case of polarization, the agents' final opinions can be more extreme than the most extreme initial opinions. This phenomenon does not arise in models proposed in the literature on bipartite consensus and based on weighted averaging of opinions (e.g., in the Altafini model).

Lemma 2.3.2 (Fluctuations) *Consider a network satisfying the k-sign arrangement property with $k \geq 3$ factions $\{F_1, \ldots, F_k\}$ and such that there exists at least one negative edge between any pair of factions. Consider the boomerang opinion dynamics model* (4.3) *with $x_i(0) = o_{\min}$ for any $i \in F_1$, $x_i(0) = o_{\max}$ for any $i \in F_2$, and $x_i(0) \in (o_{\min}, o_{\max})$*

for any $i \in F_k$, $k \geq 3$. Then, for any $0 < \epsilon < (o_{\max} - o_{\min})/2$ and any $i \in F_k$, $k \geq 3$,

$$\mathbb{P}[x_i(t) \in (o_{\min}, o_{\min} + \epsilon) \cup (o_{\max} - \epsilon, o_{\max}) \text{ i.o.}] = 1.$$

Proof: Note that $x_i(t) \in (o_{\min}, o_{\max})$ for any $t \geq 0$ and any $i \in F_k$, $k \geq 3$, with probability one. Pick a positive $\epsilon < (o_{\max} - o_{\min})/2$ and define the intervals $A_\epsilon^\ell = (o_{\min}, o_{\min} + \epsilon)$, $A_\epsilon^u = (o_{\max} - \epsilon, o_{\max})$ and $A_\epsilon^c = [o_{\min} + \epsilon, o_{\max} - \epsilon]$. Note that these three intervals are non-empty and form a partition of $(o_{\min}, o_{\max})$.

Now, take any $i \in F_k$, $k \geq 3$. First, define the random stopping times $\tau_{c\to\ell}(t) = \inf\{t^* > t \,|\, x_i(t^*) \in A_\epsilon^\ell | x_i(t) \in A_\epsilon^c\}$, $\tau_{\ell\to u}(t) = \inf\{t^* > t \,|\, x_i(t^*) \in A_\epsilon^u | x_i(t) \in A_\epsilon^\ell\}$ and $\tau_{u\to\ell}(t) = \inf\{t^* > t \,|\, x(t^*) \in A_\epsilon^\ell | x_i(t) \in A_\epsilon^u\}$. Note that, if the pair $\{i, j\}$ is chosen, then the opinion of i is always pushed towards $o_{\max}$ if $j \in F_1$, and always pushed towards $o_{\min}$ if $j \in F_2$ (this follows from the fact that for any $k \in F_1 \cup F_2$, $x_k(t) = x_k(0)$ for all $t \geq 0$ with probability one). Then, following a reasoning similar to the one adopted in the proof of Theorem 2.3.1, we conclude that $\mathbb{P}[\tau_{c\to\ell}(t) < \infty] = \mathbb{P}[\tau_{\ell\to u}(t) < \infty] = \mathbb{P}[\tau_{u\to\ell}(0) < \infty] = 1$ for any $t \geq 0$, from which the result follows. ■

Note that the conditions for the underlying signed network in this lemma are immediately satisfied if the network is complete and satisfies clustering balance. This lemma is interpreted as follows. Assume there are multiple antagonistic groups of people such that for any two groups there exist two members that can communicate with each other. Additionally, assume that only two groups are already polarized in the opinion spectrum with the rest having opinions at intermediate values (i.e., mathematically, in the interval $(o_{\min}, o_{\max})$). Then, these non-polarized groups will have their opinions always fluctuating at intermediate values, i.e., their opinions do not polarize or reach consensus at some specific value. Intuitively, since the boomerang effect is persistent on the agents with intermediate values, these agents cannot settle on a definite opinion since they con-

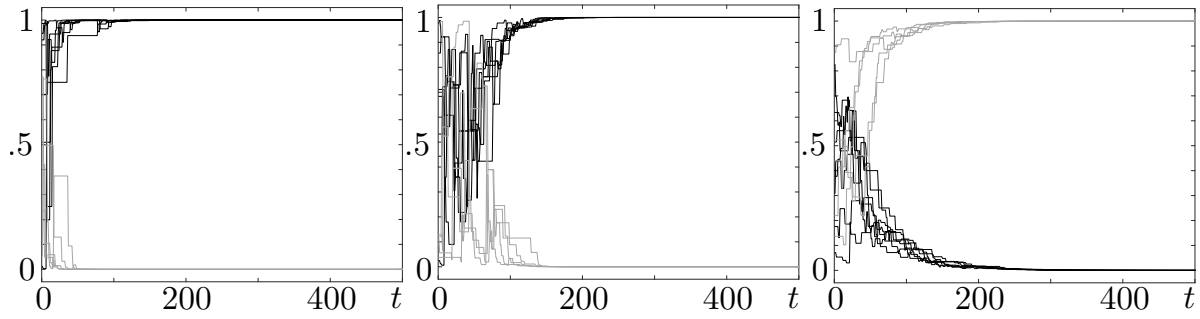
tinue to interact with antagonistic agents on both ends of the spectrum. This behavior of opinion fluctuation has been observed in other models in the presence of stubborn agents who forbid the consensus of opinions among the agents [4]. Our work is the first one to propose a persistent fluctuating behavior based on the structure of friendly and antagonistic relationships in a social network.

2.3.2 Numerical results

For a complete graph satisfying structural balance, which is a particular case satisfying the conditions of Theorem 2.3.1, Figure 3.4 shows some example evolutions for self-weights $a_i = a \in (0, 1)$ for any agent i. We observed that, in general, the larger the self-weights, the more time the polarization process takes.

Figure 2.2 shows examples where the underlying signed network has three factions. Remarkably, under generic initial conditions (which are weaker initial conditions than the ones in Corollary 2.3.2), two factions tend to polarize and the opinions of the third one show persistent fluctuations.

Finally, we provide numerical evidence of the behavior under networks that are the result of perturbations on balanced networks. Consider the situation where a complete and balanced social network with two antagonistic factions is randomly perturbed by flipping the sign of some of its edges. Intuitively, for small perturbations, we would expect that opinions, though not being able to perfectly polarize, would still "attempt" to be in such a state and fluctuate near extreme values. Figure 2.3 shows some examples confirming this phenomenon.



(a) $a = 0.25$ (b) $a = 0.5$ (c) $a = 0.75$

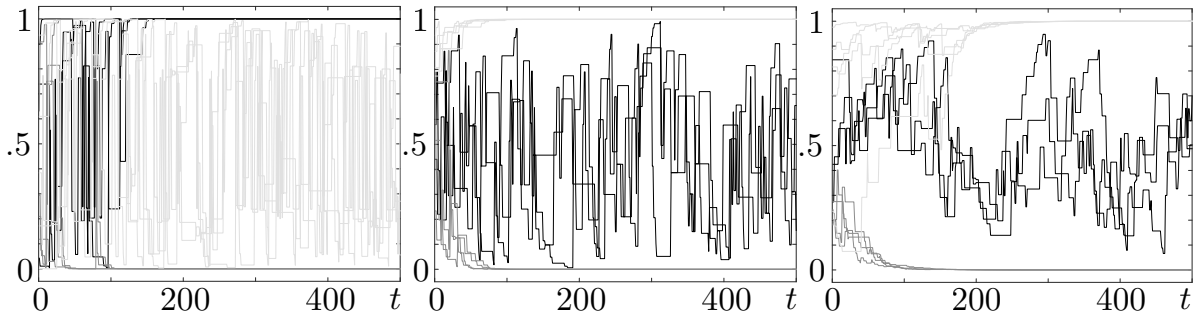
Figure 2.1: Opinion evolution with $o_{\min} = 0$ and $o_{\max} = 1$ for a complete graph satisfying structural balance with two factions of 5 (light gray) and 7 (black) agents respectively. All agents are assumed to have the same self-weight a, and the edges to be updated are chosen uniformly. All simulations have randomly sampled initial conditions.

2.4 Conclusion

We have proposed a novel simple model for opinion dynamics over signed graphs. This model provides intuitive behavior and results on the opinion evolution under sociologically relevant sign structures of the underlying social network. Future work may be the inclusion of directional updating (i.e., updating one opinion at a time) in the model, as well as its analysis under relevant directed network structures. Another open direction for research is an analytical understanding of the transient time and convergence analysis for the polarization of opinions of the factions in a balanced network.

2.5 Appendix

Lemma 2.5.1 (Finite-time proximity property) *Consider the same assumptions as in Theorem 2.3.1 with a network satisfying the* 2*-sign arrangement property. There exists a finite sequence of edges such that, if they are selected sequentially by our affine boomerang model, then, inside the interval* $[o_{\min}, o_{\max}]$*, the opinions of any two vertices become arbitrarily close if they belong to the same faction, or arbitrarily apart if they*



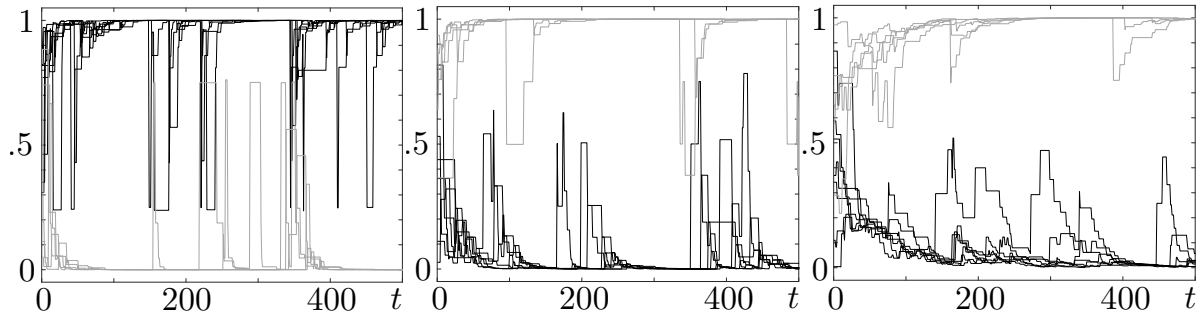
(a) $a = 0.25$ (b) $a = 0.5$ (c) $a = 0.75$

Figure 2.2: Opinion evolution with $o_{\min} = 0$ and $o_{\max} = 1$ for a complete graph satisfying clustering balance with three clusters of three, four and five agents (i.e., twelve curves are plotted). The black curves correspond to the opinions of the cluster of three agents, the medium gray curves to the cluster of four, and the light gray curves to the cluster of five. Two of the clusters polarized their opinions (to 0 and 1), while the third one shows permanent fluctuations in its opinions. The shown plots were chosen so that the cluster with four agents always end up oscillating. All agents are assumed to have the same self-weight a, and the edges to be updated are chosen uniformly. All simulations have randomly sampled initial conditions.

belong to different ones.

Proof: Since the network satisfies the 2-sign arrangement, for any i and j that belong to the same faction, there exists a nonempty collection of paths $\mathcal{P}^{+}_{i\leftrightarrow j}$ between i and j in which each path contains only positive edges. Let $p \in \mathcal{P}^{+}_{i\leftrightarrow j}$, then, from statement (i) from Theorem 2.3.1, we observe that if we only update pair of vertices present along the path p, then they can become arbitrarily close. Then, we can construct a finite sequence of edges such that it includes only edges from one or more different paths in $\mathcal{P}^{+}_{i\leftrightarrow j}$ in a sufficient number so that i and j become arbitrarily close. This proves the first part of the lemma.

Now, we consider the case where i and j belong to different factions. Notice that equation (4.3) clearly shows that we can always make the opinions of two vertices joined by a negative edge arbitrarily apart by continuously sampling such edge. Let $\mathcal{P}^{-}_{i\leftrightarrow j}$ be the nonempty collection of paths between i and j. Due to the structure of the network,



(a) $a = 0.25$ (b) $a = 0.5$ (c) $a = 0.75$

Figure 2.3: Opinion evolution with $o_{\min} = 0$ and $o_{\max} = 1$ for a complete graph that originally satisfied structural balance with two factions of 5 (light gray) and 7 (black) agents and is now under a perturbation of 3 of its edges having the opposite sign. All agents are assumed to have the same self-weight a, and the edges to be updated are chosen uniformly. All simulations have randomly sampled initial conditions.

any $p \in \mathcal{P}^{-}_{i \leftrightarrow j}$ must have an odd number of negative edges. Then, p can be constructed by appropriately concatenating sequences of positive edges with sequences of negative edges. From our discussion above, we can make the opinions of the agents participating in any of these positive sequences (if any) arbitrarily close, and the opinions of the agents in any of the negative edges arbitrarily apart. Then, it is possible to come up with a finite sequence of edges such that i and j become arbitrarily apart. This finishes the proof of the lemma. ■

Chapter 3

Multi-group SIS Epidemics with Simplicial and Higher-Order Interactions

3.1 Introduction

The study and modeling of the spread of infectious diseases in contact networks has a long history of development and is of major relevance today. A first class of models are called *scalar models*, where a single population is studied. The epidemiological evolution in this single population is represented by the dynamics of one or more scalar values that represent a specific proportion of the population (e.g., a scalar value can represent the proportion of currently infected people). We refer to the work [81] for a survey on these type of models. The basic assumption on these models is that the whole population is homogeneous, i.e., every individual in the population has the same probability of interaction. However, in view of this shortcoming, *network* or *multi-group models* were introduced, in which several homogeneous populations, also called groups,

interact with each other according to an underlying contact network. Thus, these models can capture different kinds of heterogeneity, e.g., age structures, spatial diversity and social behavior. The epidemics is then modulated by the different model parameters (e.g., the recovery rate from a disease) that each population may have, and the connectivity of the underlying network and the strength of its connections. Thus, the propagation of the epidemic is now a network process.

Multi-group epidemic models have a longstanding history that can be traced back to the seminal works [80, 100]. A recent interpretation as an approximation of Markov-chain models is given by [150]. Degree-based versions of the model have been analyzed through statistical mechanics in the physics community [140, 60]. Stability analyses by the controls community include [64, 95]. Much recent work by the control community has focused on (i) control of epidemic dynamics in multi-group models, e.g. [175, 133], (ii) extensions of epidemics on time-varying graphs across populations, e.g. [135, 138], (iii) extensions to multi-competitive viruses on multi-group models, e.g. [139], and (iv) game-theoretical analysis on multi-group models, e.g. [83, 136]. Finally, we mention the recent surveys [123, 133].

In this work, we focus on the *Susceptible-Infected-Susceptible* (SIS) model for the propagation of infectious diseases in the context of social contagion. SIS models are applicable to diseases that have the possibility of a repeated reinfection, i.e., those in which a person does not develop permanent immunity after recovery [118]. Some examples of these diseases are ghonorrea, chlamydya, the common cold, etc. In the scalar SIS model, the population can be divided in two fractions: those who are infected and those who are susceptible to become infected [81]. In the multi-group SIS model, each node of the graph can be interpreted as either (i) an individual and its associated scalar variable as the infection probability, or (ii) as a homogeneous group of individuals and the associated scalar variable is the fraction of infected individuals. The type of interaction among the

individuals or populations defines the social contagion mechanism.

In SIS models, it is important to investigate conditions under which the system converges or not to a *disease-free* equilibrium, i.e., a state in which all populations become healthy/uninfected (or equivalently, the probability of any person of being infected becomes zero) or to an *endemic* equilibrium, i.e., a state in which all populations maintain a (nonzero) fraction of its members always infected (or equivalently, the probability of any person of being infected remains nonzero).

Nonlinear incidence and simplicial contagion models The vast majority of the literature on multi-group SIS models (and other epidemic models in general) considers only that the interaction between populations (or individuals) is pairwise, i.e., the social contagion occurs only through the edges that connect them. Equivalently, in the context of scalar models, this prevalent assumption is understood as the *incidence rate*, i.e., the rate of new infections, being bilinear in the proportions of infected and susceptible people (because the rate is simply the product of both proportions). The idea of considering nonlinear incidence rates in epidemic scalar models can be traced back to the late eighties [109].

From a network-science viewpoint, the recent work by Iacopini et al. [87] elaborates on the idea of nonlinear incidence models and considers higher-orders of interaction in the social contagion of a disease. Since its publication, the work [87] has received considerable interest and much attention is now focused on higher-order interactions and simplicial models. We now elaborate on these ideas. Consider three populations or individuals i, j, k. If the pairwise interactions $\{i, j\}$ or $\{i, k\}$ occur, then there is a certain susceptibility of i to be infected. However, if the whole group $\{i, j, k\}$ interact together, then the likelihood of infection for i may increase since now the simultaneous interaction effect by j and k are aggregated to the single pairwise interactions we previously

described. We can consider $\{i, j, k\}$ as a hyperedge. An important class of hypergraphs is a *simplicial complex*, which is a hypergraph that contains all nonempty subsets of hyperedges as hyperedges. In a simplicial complex, a hyperedge with d vertices forms a $(d-1)$-simplex, and the simplicial complex is said to be of dimension $d-1$ if d is the largest number of vertices in any of its simplices (i.e, in its largest simplex). As an example, if $\{i, j, k\}$ is a 2-simplex, then $\{i, j\}$, $\{j, k\}$, $\{i, k\}$, $\{i\}$, $\{j\}$, $\{k\}$ are simplices. Thus, a simplex $\{i, j, k\}$ can be understood as a set of nodes that form a triad. Note that if $\{i, j\}$, $\{j, k\}$ and $\{i, k\}$ belong to a simplicial complex, then $\{i, j, k\}$ is not necessarily a simplex. We refer to [75] for a general and extensive treatment of simplicial complexes. Starting from these ideas, the work [87] proposes a new SIS model that considers the evolution of the epidemic with an underlying simplicial complex of dimension 2, as opposed to the classical SIS model that has up to 1-simplices. However, [87] performs the analysis of a mean-field approximation which becomes a scalar model. A different derivation of the SIS model over simplicial complexes was recently introduced in [121] from a Markov-chain and mean-field approximation perspective up to 2-simplices. Also recently, Jhun et al. [91] consider the multi-group SIS model and restrict their analysis to a mean-field approximation of the model for a special class of simplicial complexes, namely, an infinite hypergraph composed of hyperedges of the same size corresponding to simplicial complexes of the same dimension.

As discussed by [87], the adoption of simplicial interactions in modeling contagion bears some similarities with the modeling ideas behind linear threshold models by Granovetter [73] in sociology, where individuals adopt innovations only when a certain fraction of their contacts have earlier adopted that innovation. Moreover, simplicial and higher-order graphical models may be more accurate than simpler pairwise contagion models to describe transmission events during large gatherings or other social aggregation phenomena [93, 55]. Overall, the study of simplicial and higher-order interactions

is well motivated by the observation that these structures are ubiquitous and play an important role in real-world social networks [28, 86, 22, 152]. We refer to the excellent recent survey [19] for an overview of the emerging field of networks with higher-order interactions.

Problem statement We now state what is, to the best of our knowledge, an outstanding open problem. Namely, no work in the current literature establishes a formal analysis of the dynamical behavior of a general multi-group SIS model with higher-order interaction terms over general classes of (hyper)graphs. An example of such model could be an SIS model with interactions described by a finite simplicial complex. Our paper responds to this need. The analysis of such a model may help better understand the effect of higher-order interaction terms on the dynamics of social contagion in societies with large gatherings or other social aggregation phenomena.

Contributions As main contribution of this paper, we consider the simplicial SIS model and analyze its dynamical behavior. In particular, we identify conditions on the parameters of the model that allow us to conclude the existence and asymptotic behavior of a disease-free and/or endemic equilibrium. In particular, we prove that the model, according to different regimes in its parameter space, can have its dynamic behavior classified in three domains: (i) *disease-free domain:* where convergence to a disease-free equilibrium is guaranteed as well as the nonexistence of endemic equilibria; (ii) *bistable domain:* where, depending on the initial amount of infection across populations, convergence to a disease-free or endemic equilibria may occur; and (iii) *endemic domain:* where the disease-free equilibrium is unstable and a unique endemic equilibrium is asymptotically stable. While the conditions given in our main theorem (Theorem 3.5.2) does not exhaust all possible values of the system parameters, we include numerical results that il-

lustrate the tightness of our derived conditions. Despite this gap, our sufficient conditions rigorously establish the crucial qualitative behavior of transition between the disease-free domain and the bistable domain. To the best of our knowledge, this transition was formally proved only for the scalar version of the simplicial SIS model in [87].

As second contribution, we propose an iterative algorithm which computes an endemic state through monotone convergence when the system is in either the bistable or the endemic domain according to the presented sufficient conditions. We remark that obtaining a closed form expression for an endemic equilibrium appears to be intractable and, indeed, for the classical multi-group SIS model the best-known result is a monotonic convergent iteration, see [123, Theorem 4.3].

As third contribution, we present a general multi-group SIS model with higher-order interactions, generalizing the two dimensional simplicial SIS model. Analyzing this generalized model, we prove that the existence of the bistable domain is a general phenomenon resulting from higher-order interactions. While the treatment becomes more cumbersome, we show that our analysis techniques are still applicable.

As minor contributions, we provide numerical examples that illustrate the behavioral domains of the simplicial SIS model and present two interesting conjectures about the features of the epidemic diagram. Moreover, we present a self-contained formal review of previously known results for the scalar version of the simplicial SIS model; this review facilitates the comparison between the scalar and the multi-group models.

We conclude by mentioning that, to prove our results, we make use of the theory of Metzler matrices and positive systems, fixed-point analysis of continuous mappings, and exponential convergence with matrix measures and Lyapunov theory. We review a little known result for exponential convergence combining the theory of matrix measures for positive systems with the theory of solution estimates (Coppel's inequality) for systems with continuously differentiable vector fields. We remark that previous works that ana-

lyze the classical multi-group SIS model have used specialized cases of this exponential convergence result, e.g., see [64, Theorem 2.7].

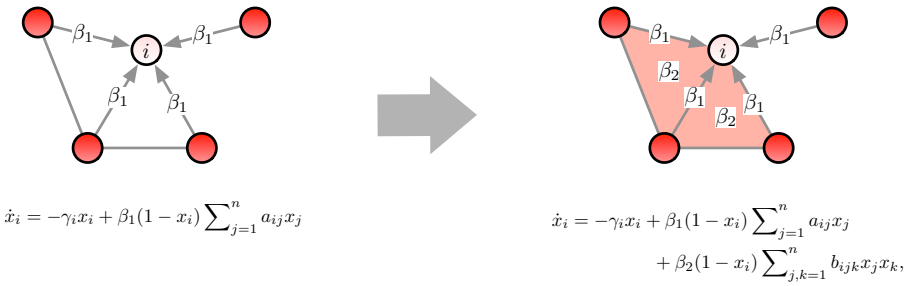


Figure 3.1: From pairwise to simplicial interactions in the multi-group SIS epidemic model: the left figure corresponds to the classical version and the right one to the simplicial SIS model.

Paper organization Section 6.2 is the preliminaries and notation. Section 3.4 introduces the simplicial SIS model. Section 3.5 presents the dynamical analysis. Section 3.7 presents numerical examples, and Section 6.7 is the conclusion.

3.2 Preliminaries and notation

3.2.1 General notation

Given $A \in \mathbb{R}^{n\times n}$, let $\rho(A)$ denote its spectral radius and let $A \geq 0$ mean that all its elements are non-negative. A nonnegative matrix A is irreducible if for any $i, j \in \{1, \ldots, n\}$, there exists a $k = k(i, j) \leq n - 1$ such that the ij entry of A^k is positive. Alternatively, if $A \geq 0$ is regarded as a weighted adjacency matrix of some directed graph $\mathcal{G}$, A is irreducible if and only if the graph $\mathcal{G}$ is strongly connected. If $A \geq 0$ is irreducible, then, by the Perron-Frobenius theorem [82, Theorem 8.4.4.], its eigenvalue with largest magnitude $\lambda_{\max}(A)$ is real, simple, and equal to $\rho(A) > 0$. This eigenvalue is called the Perron-Frobenius or dominant eigenvalue and has associated left and right

Perron-Frobenius or dominant eigenvectors with positive entries (normalized to have unit sum, by convention).

Let $\|\cdot\|$ denote an arbitrary norm, $\|\cdot\|_p$ denote the ℓ_p-norm, and $\|\cdot\|_{p,Q} := \|Q\cdot\|_p$ with Q being a positive definite matrix denote a weighted ℓ_p-norm. When the argument of a norm is a matrix, we refer to its respective induced matrix norm. Given two vectors $x, y \in \mathbb{R}^n$, we denote $x \ll y$ when $x_i < y_i$ for every i; $x \leq y$ when $x_i \leq y_i$ for every i; and $x < y$ when $x \leq y$ and $x \neq y$.

Let I_n be the $n \times n$ identity matrix, $\mathbb{1}_n, \mathbb{0}_n \in \mathbb{R}^n$ be the all-ones and all-zeros vector with n entries respectively. Let $\mathbb{0}_{n\times n}$ be the $n \times n$ zero matrix. Let $\mathrm{diag}(X_1, \ldots, X_N) \in \mathbb{R}^{\sum_{i=1}^N n_i \times \sum_{i=1}^N n_i}$ represent a block-diagonal matrix whose elements are the matrices $X_1 \in \mathbb{R}^{n_1\times n_1}, \ldots, X_N \in \mathbb{R}^{n_N\times n_N}$. Given a vector $x \in \mathbb{R}^n$, $\mathrm{diag}(x) = \mathrm{diag}(x_1, \cdots, x_n)$. Let $\mathbb{R}_{\geq 0}$ be the set of non-negative real numbers. Given $x_i \in \mathbb{R}^{k_i}$, for $i \in \{1, \ldots, N\}$, we let $(x_1, \ldots, x_N) = \begin{bmatrix} x_1^\top & \ldots & x_N^\top \end{bmatrix}$.

Finally, we recall a classic monotonicity property. If A and A' are square matrices of the same dimension,

$$0 \leq A \leq A' \quad \Longrightarrow \quad \rho(A) \leq \rho(A'). \tag{3.1}$$

3.2.2 Matrix measures

Given $A \in \mathbb{R}^{n\times n}$ and norm $\|\cdot\|$ on $\mathbb{R}^n$, its associated matrix measure is $\mu(A) = \lim_{h\to 0^+} \frac{\|I_n+hA\|-1}{h}$ [168, 47]. Given $x \in \mathbb{R}^n$ and $\xi \gg \mathbb{0}_n$, the weighted ℓ_∞-norm is $\|x\|_{\infty,\mathrm{diag}(\xi)} = \|\mathrm{diag}(\xi)x\|_\infty$ its associated matrix measure is

$$\mu_{\infty,\mathrm{diag}(\xi)}(A) = \max_{i\in\{1,\ldots,n\}} \Big(a_{ii} + \xi_i \sum\nolimits_{j=1, j\neq i}^{n} |a_{ij}|/\xi_j \Big).$$

Given a Metzler matrix $M \in \mathbb{R}^{n\times n}$ and a scalar b,

$$M\xi \leq b\xi \quad \Longleftrightarrow \quad \mu_{\infty,\mathrm{diag}(\xi)^{-1}}(M) \leq b. \tag{3.2}$$

3.3 Exponential convergence and matrix measures

The following result combines the matrix measure results shown above with the Coppel's inequality as stated in [168, Theorem 22, (Chapter 2, page 52)]. To the best of our knowledge, this connection and the result in [168] have not been explicitly exploited before. This result will be useful for the paper's main theorem.

Theorem 3.3.1 (Exponential convergence from Coppel's inequality) *Consider a smooth dynamical system $\dot{x} = f(x)$ with a convex compact invariant set $\mathcal{X}$ and an equilibrium point $x^* \in \mathcal{X}$. Write the system as*

$$\dot{x} = D(x, x^*)(x - x^*). \tag{3.3}$$

where $D(x, x^*) \in \mathbb{R}^{n\times n}$ is a function of x and x^*. Let $\|\cdot\|$ be a norm and μ be its associated matrix measure μ. If $\mu(D(x, x^*)) \leq -c$ for any $x \in \mathcal{X}$, then x^* is the unique exponentially stable equilibrium point in $\mathcal{X}$ and exponential convergence is attained at rate c. Moreover, $V(x) = \|x - x^*\|$ is a global Lyapunov function for x^* in $\mathcal{X}$.

Proof: First, it is always possible [168, Lemma 17, Chapter 2, page 52] to write f in the form (3.3) using the fundamental theorem of calculus and the convexity of $\mathcal{X}$. Second, as argued in [48, Chapter 1, page 3], since the right-hand derivative of $x(t) - x^*$

is $\dot{x}(t)$ at any $t \geq 0$, the right-hand derivative $\frac{d^+}{dt} \|x - x^*\|$ exists and moreover

$$\begin{aligned}\frac{d^+}{dt} \|x - x^*\| &= \lim_{h \to 0^+} \frac{\|x - x^* + h\dot{x}\| - \|x - x^*\|}{h} \\ &\leq \lim_{h \to 0^+} \frac{\|I_n + hD(x, x^*)\| - 1}{h} \|x - x^*\| \\ &\leq \mu(D(x, x^*)) \|x - x^*\| \leq -c \|x - x^*\|,\end{aligned}$$

where the second inequality follows from Coppel's inequality as in [48, Theorem 3, Chapter 3] and in [168, Theorem 22, Chapter 2, page 52], and the third inequality follows from the negative matrix measure assumption. Therefore, applying Grönwall's inequality, any trajectory $x(t)$ starting in $\mathcal{X}$ satisfies $\|x(t) - x^*\| \leq e^{-ct} \|x(0) - x^*\|$. Moreover, x^* is the unique globally exponentially stable equilibrium in $\mathcal{X}$.

Finally, observe that $V(x) = \|x - x^*\|$, $x \in \mathcal{X}$, is a Lyapunov function with respect to x^* since (i) it is globally proper, i.e., for each $\ell > 0$, the set $\{x \in \mathcal{X} \mid V(x) \leq \ell\}$ is compact (since $\mathcal{X}$ is compact), (ii) it is positive definite on $\mathcal{X}$, (iii) strictly decreasing for any $x \neq x^*$ on $\mathcal{X}$. This finishes the proof. ■

3.4 The Simplicial SIS model

We study the following multi-group deterministic model, which can be regarded as a mean-field approximation of a more realistic stochastic model.

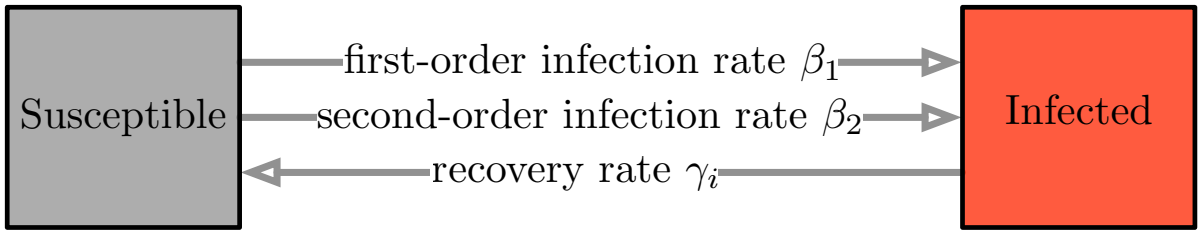


Figure 3.2: Simplicial SIS as a compartmental model

Definition 3.4.1 (The simplicial SIS model) *Assume $x \in [0,1]^n$, and let $\beta_1, \beta_2 > 0$ and $\gamma_i > 0$, $i \in \{1, \dots, n\}$. Then, the* simplicial SIS model *is, for any $i \in \{1, \dots, n\}$,*

$$\dot{x}_i = -\gamma_i x_i + \beta_1(1 - x_i) \sum\nolimits_{j=1}^{n} a_{ij} x_j + \beta_2(1 - x_i) \sum\nolimits_{j,k=1}^{n} b_{ijk} x_j x_k, \tag{3.4}$$

or, in its matrix form, with $\Gamma = \mathrm{diag}(\gamma_1, \cdots, \gamma_n)$, the model is

$$\dot{x} = -\Gamma x + \beta_1(I_n - \mathrm{diag}(x))Ax \quad + \beta_2(I_n - \mathrm{diag}(x))(x^\top B_1 x, \cdots, x^\top B_n x)^\top \tag{3.5}$$

where $B_i = \begin{bmatrix} b_{i11} & \cdots & b_{i1n} \\ \vdots & & \vdots \\ b_{in1} & \cdots & b_{inn} \end{bmatrix}$, $i \in \{1, \dots, n\}$, and $A = (a_{ij})$ are nonnegative matrices.

We now provide some remarks about this definition.

Remark 3.4.1 (Interpretation of Definition 3.4.1) *(i) Matrix $A \geq 0$ represents the pairwise contact rate between the agents: $a_{ij} > 0$ if agent i (i.e., population or individual) is in contact with j; and the magnitude of a_{ij} indicates the contact frequency: the larger, the more positive effect on the infection spread. Now, for matrix $B_i \geq 0$, $b_{ijk} > 0$ if agent i can have a simultaneous interaction with j and k, and the magnitude of b_{ijk} indicates the strength of the interaction. Thus, the elements of B_i indicate higher-order interaction effects that two agents jointly have over i. This is a key structural difference with the classical multi-group SIS model, see Figure 3.1. Finally, $a_{ii} > 0$ and $b_{iii} > 0$ indicate different orders on the effect of actions taken by i that increase the effect of the infection, and $b_{ijj} > 0$ indicates the higher-order effects of j's actions over i.*

(ii) If our model is strictly defined over a simplicial complex, then A and B_i should be symmetric and have joint restrictions on their elements. However, in our work, we do not restrict A or B_i to be symmetric and consider a more general mathematical model.

We keep the term simplicial *in the title of the model since the special case of simplicial complexes inspired the more general model.*

(iii) The parameter γ_i is the recovery rate of agent i from the infection. Parameters β_1 and β_2 are the infection rates at which an agent may get infected due to pairwise or higher-order interactions respectively. Figure 3.2 shows how these parameters modulate the proportion of infected and susceptible people inside a population, or equivalently, the changes in the probability for an individual to be infected or susceptible.

We revisit the qualitative behavioral domains that a multi-group SIS model with higher-order terms must display.

Definition 3.4.2 (Epidemic domains) *Consider the simplicial SIS model with fixed parameters Γ, A and B_i for all $i \in \{1, \ldots, n\}$. According to the values of parameters (β_1, β_2), the system is in the:*

(i) Disease-free domain: *the disease-free equilibrium $\mathbb{0}_n$ is the unique equilibrium and globally stable.*

(ii) Bistable domain: *the disease-free equilibrium is locally asymptotically stable and there exists an endemic equilibrium $x^* \gg \mathbb{0}_n$ which is also locally asymptotically stable.*

(iii) Endemic domain: *the disease-free equilibrium is unstable and there exists a unique endemic equilibrium that is asymptotically stable in $[0, 1]^n \setminus \{\mathbb{0}_n\}$.*

The following theorem describes the behavior of the scalar version of the simplicial SIS model in [87]; although [87] does not state its results as a theorem, we present them as such for comparison purposes.

Theorem 3.4.2 (Dynamics of the scalar model in [87]) *Consider the* scalar simplicial SIS model

$$\dot{y} = -\gamma y + \beta_1(1-y)y + \beta_2(1-y)y^2 \tag{3.6}$$

with $y \in [0,1]^n$ *and* $\gamma, \beta_1, \beta_2 > 0$. *Then, the set* $[0,1]$ *is invariant and* 0 *is an equilibrium point. Define* $v_c(\beta_2/\gamma) = 2\sqrt{\frac{\beta_2}{\gamma}} - \frac{\beta_2}{\gamma}$ *and the two variables* $\nu_\pm = \frac{1}{2}(1 - \frac{\beta_1}{\beta_2}) \pm \frac{1}{2\beta_2}\sqrt{(\beta_1-\beta_2)^2 - 4\beta_2(\gamma-\beta_1)}$. *Moreover,*

Disease-free domain: *If either* $\frac{\beta_2}{\gamma} \leq 1$ *and* $\frac{\beta_1}{\gamma} \leq 1$, *or* $\frac{\beta_2}{\gamma} > 1$ *and* $\frac{\beta_1}{\gamma} < v_c(\beta_2/\gamma)$, *then*

(i) 0 *is the unique equilibrium point in* $[0,1]$,

(ii) 0 *is globally asymptotically stable in* $[0,1]^n$.

Bistable domain: *If* $\frac{\beta_2}{\gamma} > 1$ *and* $v_c(\beta_2/\gamma) < \frac{\beta_1}{\gamma} < 1$, *then* $\nu_-, \nu_+ \in (0,1]$ *and*

(iii) 0 *is locally asymptotically stable in* $[0, \nu_-)$,

(iv) ν_+ *is a locally asymptotically stable equilibrium in* $(\nu_-, 1]$, *and*

(v) ν_- *is an unstable equilibrium.*

Endemic domain: *If* $\frac{\beta_1}{\gamma} > 1$, *then*

(vi) 0 *is unstable,*

(vii) ν_+ *is the unique equilibrium in* $(0,1]$ *and is globally asymptotically stable in* $(0,1]$.

Notice the polynomial resemblance of the scalar model in (3.6) and our multi-group simplicial model in (3.5).

3.5 Analysis of the model

First, we establish properties of the model independently from their parameter values.

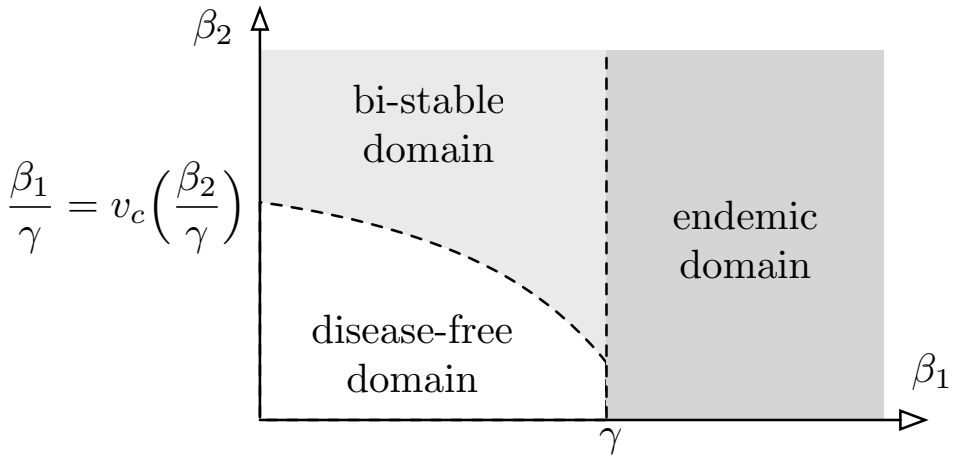


Figure 3.3: Epidemic diagram for the scalar simplicial SIS model (see Theorem 3.4.2).

Lemma 3.5.1 (General properties of the simplicial SIS model) *Consider the simplicial SIS model with an irreducible $A \geq 0$ and arbitrary $B_i \geq 0$. Then,*

(i) The set $[0,1]^n$ is an invariant set.

(ii) If $x(0) > \mathbb{0}_n$, then $x(t) \gg \mathbb{0}_n$ for any $t > 0$.

(iii) The origin $\mathbb{0}_n$ is an equilibrium of the system and there are no other equilibria on the boundary of the set $[0,1]^n$.

Proof: Let $f(x)$ be the right-hand side of equation (3.5). We first prove statement (i). Following Nagumo's theorem [26, Theorem 4.7] we analyze the vector field at the boundary of $[0,1]^n$. From equation (3.4), we see that 1) $f_i(x) \geq 0$ for all $x \in [0,1]^n$ such that $x_i = 0$ for some $i \in \{1,\dots,n\}$; 2) $f_i(x) < 0$ for all $x \in [0,1]^n$ such that $x_i = 1$ for some $i \in \{1,\dots,n\}$; from which it follows that $[0,1]^n$ is an invariant set. This proves statement (i).

Set the change of variables $y = e^{\Gamma t}x$. Then, from equation (3.5),

$$\dot{y} = \operatorname{diag}(e^{\gamma_1 t},\cdots,e^{\gamma_n t})(I_n - \operatorname{diag}(x))(\beta_1 Ax + \beta_2(x^\top B_1 x,\cdots,x^\top B_n x)^\top). \tag{3.7}$$

Since $x(0) \in [0,1]^n$, notice that $\dot{y}(t) \geq \mathbb{0}_n$ for any $t \geq 0$, and so there is the monotonicity property $y(t_1) \geq y(t_0)$ for any $t_1, t_0 \geq 0$. Now, we prove statement (ii) by contradiction. Let us assume that $x(0) > \mathbb{0}_n$, which implies $y(0) > \mathbb{0}_n$, and that there exists some $i \in \{1, \ldots, n\}$ and $T > 0$ such that $y_i(T) = 0$. Then, because of the monotonicity property, $y_i(t) = 0$ for all $t \in [0, T]$, which implies that $x_i(t) = 0$. Then, from the equilibrium equation of (3.7), we have that $0 = \beta_1 e^{\gamma_i t} \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij} e^{-\gamma_j t} y_j(t) + \beta_2 e^{\gamma_i t} \sum_{\substack{j=1 \\ j \neq i}}^{n} \sum_{\substack{k=1 \\ k \neq i}}^{n} b_{ijk} e^{-\gamma_j t} e^{-\gamma_k t} y_j(t) y_k(t)$ for any $t \in [0, T]$, and since all terms are non-negative, it follows that $y_j(t) = 0$ for $t \in [0, T]$ and all j such that $a_{ij} > 0$. Then, for any such j, we repeat the same analysis we just did and find that $y_k(T) = 0$ for all $t \in [0, T]$ and all k such that $a_{jk} > 0$. Then, since A is irreducible, we could continue repeating this procedure and finally obtain $y(t) = \mathbb{0}_n$ for all $t \in [0, T]$. This gives a contradiction, since we had that $y(0) > \mathbb{0}_n$ because of $x(0) > \mathbb{0}_n$. Then, $y(t) \gg \mathbb{0}_n$ implies $x(t) \gg \mathbb{0}_n$ for $t > 0$ and finish the proof of statement (ii).

Finally, we prove statement (iii). First, let us introduce the functions $h_+(z) = \frac{z}{1+z}$ for any $z \in \mathbb{R}_{\geq 0}$ and $h_-(z) = \frac{z}{1-z}$ for $z \in [0, 1)^n$. We also introduce $H_+(y) = (h_+(y_1), \ldots, h_+(y_n))^\top$ for $y \geq \mathbb{0}_n$, and $H_-(y) = (h_-(y_1), \ldots, h_-(y_n))^\top$ for $y \in [0, 1)^n$.

Now, it is immediate from equation (3.5) that $\mathbb{0}_n$ is an equilibrium point, and observe that there is no equilibrium point x^* such that $x_i^* = 1$ for some $i \in \{1, \ldots, n\}$, since that would imply that $(f(x^*))_i < 0$. Now, assume x^* is an equilibrium point such that $x_i^* = 0$ for some $i \in \{1, \ldots, n\}$. Let $B_{x^*} := ({x^*}^\top B_1 x^*, \cdots, {x^*}^\top B_n x^*)^\top$. First, from the

equilibrium equation of the system (3.5), since $x^* \ll \mathbb{1}_n$, we obtain

$$\begin{aligned}
&\mathbb{0}_n = -\Gamma x^* + (I_n - \operatorname{diag}(x^*))(\beta_1 A x^* + \beta_2 B_{x^*}) \\
&\iff (I_n - \operatorname{diag}(x^*))^{-1} x^* = \Gamma^{-1}(\beta_1 A x^* + \beta_2 B_{x^*}) \\
&\iff H_-(x^*) = \Gamma^{-1}(\beta_1 A x^* + \beta_2 B_{x^*}) \\
&\iff H_+(\Gamma^{-1}(\beta_1 A x^* + \beta_2 B_{x^*})) = x^*,
\end{aligned}$$

and so $x_i^* = h_+(\frac{\beta_1}{\gamma_i}\sum_{j=1}^n a_{ij}x_j^* + \frac{\beta_2}{\gamma_i}{x^*}^\top B_i x^*)$. Then, since $x_i^* = 0$, this implies that $x_j^* = 0$ for all j such that $a_{ij} > 0$. Then, since A is irreducible, we could iterate this procedure and conclude that $x^* = \mathbb{0}_n$. Therefore, any equilibrium point at the boundary of $[0,1]^n$ must be the origin. This proves statement (iii). ■

From an epidemiological perspective, Lemma 3.5.1 shows two important things for the well-posedness of the simplicial SIS model: 1) each entry of the state vector of the model can represent a proportion or probability; 2) there cannot exist another type of equilibria than disease-free or endemic ones. Now we present our main result.

Theorem 3.5.2 (The simplicial SIS model and its different epidemiological domains) *Consider the simplicial SIS model with an irreducible $A \geq 0$ and arbitrary $B_i \geq 0$. Define $\mathbf{1}_B \in \{0,1\}^n$ by $(\mathbf{1}_B)_i = 1$ if $B_i \neq \mathbb{0}_{n\times n}$ and $(\mathbf{1}_B)_i = 0$ otherwise.*

Disease-free domain: *If*

$$\rho(\beta_1\Gamma^{-1}A + \beta_2\Gamma^{-1}(\mathbb{1}_n^\top B_1, \cdots, \mathbb{1}_n^\top B_n)^\top) < 1,$$

then

(i) $\mathbb{0}_n$ is the unique equilibrium point in $[0,1]^n$,

(ii) $\mathbb{0}_n$ *is globally exponentially stable in* $[0,1]^n$ *with Lyapunov function* $V(x) = \|x\|_{1,\mathrm{diag}(v)\Gamma^{-1}} = v^\top \Gamma^{-1} x$, *where* v *is the dominant left eigenvector of* $\beta_1\Gamma^{-1}A + \beta_2\Gamma^{-1}(\mathbb{1}_n^\top B_1, \cdots, \mathbb{1}_n^\top B_n)^\top$.

Bistable domain: *If* $\beta_1\rho(\Gamma^{-1}A) < 1$ *and*

$$\min_{i \text{ s.t. } B_i \neq \mathbb{0}_{n\times n}} \Big(\frac{2\beta_1}{\gamma_i}(A\mathbf{1}_B)_i + \frac{\beta_2}{\gamma_i}\mathbf{1}_B^\top B_i \mathbf{1}_B\Big) \geq 4,$$

then

(iii) $\mathbb{0}_n$ *is a locally exponentially stable equilibrium,*

(iv) there exists an equilibrium point $x^* \gg \mathbb{0}_n$ *such that* $x_i^* \geq \frac{1}{2}$ *for any* i *such that* $B_i \neq \mathbb{0}_{n\times n}$, *and*

(v) any such equilibrium point x^* *is locally exponentially stable.*

Endemic domain: *If* $\beta_1\rho(\Gamma^{-1}A) > 1$, *then*

(vi) $\mathbb{0}_n$ *is an unstable equilibrium,*

(vii) there exists an equilibrium point $x^* \gg \mathbb{0}_n$ *in* $[0,1]^n$, *and*

(viii) if β_2 *is sufficiently small, then* x^* *is unique in* $(0,1]^n$ *and it is globally exponentially stable in* $[0,1]^n \backslash \{\mathbb{0}_n\}$, *with Lyapunov function* $V(x) = \|x - x^*\|_{\infty,\mathrm{diag}(x^*)^{-1}}$, $x \in \mathcal{X}$.

Moreover, if $\beta_1\rho(\Gamma^{-1}A) < 1$, *then the system is either in the disease-free domain or in the bi-stable domain.*

Remark 3.5.3 (About Theorem 3.5.2) *(i) Pick* β_1 *satisfying* $\beta_1\rho(\Gamma^{-1}A) < 1$. *Assume either that each* B_i *is non-zero, or that each non-zero* B_i *has a positive* i*th diagonal entry. Then there exists some* $\hat{\beta}_2 > 0$ *such that the second condition for the bistable*

domain is satisfied for $\beta_2 = \hat{\beta}_2$ and the simplicial SIS model is in the bistable domain for any $\beta_2 \geq \hat{\beta}_2$.

(ii) Compared to the scalar model in Theorem 3.4.2, the sufficient conditions in Theorem 3.5.2 defining the different domains for the simplicial SIS model do not exhaust all the possible values for (β_1, β_2). Despite this gap, our theorem rigorously establishes the following crucial qualitative behavior: assume there exist parameters (β_1, β_2) that satisfy the sufficient condition for the bistable region in Theorem 3.5.2, then we can show the system can transition from the disease-free domain to the bistable domain (and vice versa) by modifying β_2. This transition, presented as a novelty for the scalar model, is also a novelty of the simplicial SIS model.

(iii) In the literature on the classical multi-group SIS model, where only the disease-free and endemic domains exist, the number $\beta_1\rho(\Gamma^{-1}A)$ is known as the reproduction number *and its value has been used to determine whether the system is in the endemic domain or not. This number has a similar role for the simplicial SIS model. Indeed, if all higher-order interaction matrices B_i are equal to zero, then our theorem reduces to and restates some properties of the classical multi-group SIS model, e.g., see [123, Theorems 4.2 and 4.3].*

(iv) In the classical SIS multi-group model, the work [64] uses the Lyapunov function $V(x) = \|x - x^*\|_{1,\mathrm{diag}(x^*)}$ to show asymptotic convergence to the a unique endemic state $x^* \in [0,1]^n \setminus \{0\}_n$. Note that Theorem 3.3.1 generalizes [64, Theorem 2.7].

Proof: [Proof of Theorem 3.5.2] Let us consider the functions H_+ and h_+ introduced in the proof of Lemma 3.5.1. Let $\bar{A} := \beta_1\Gamma^{-1}A$, $\bar{B}_i := \frac{\beta_2}{\gamma_i}B_i$ for $i \in \{1, \ldots, n\}$, and let $\dot{x} := f(x)$. We introduce the following result: if $\mathbb{0}_n \leq y \ll z$ and $C \geq 0$ an $n \times n$ irreducible matrix, then $H_+(Cy) \ll H_+(Cz)$. This follows from the fact that, since C is irreducible, there exists at least one positive entry in some off-diagonal entry in any row

of C, and so $C(z-y) \gg \mathbb{0}_n$. Then, $Cz \gg Cy$, and since h_+ is monotonically increasing, then $H_+(Cy) \ll H_+(Cz)$. Similarly, if $\mathbb{0}_n \leq y \leq z$ and $C \geq 0$ an $n \times n$ matrix (not necessarily irreducible), then $H_+(Cy) \leq H_+(Cz)$. We use these results throughout the rest of this proof.

We first prove fact (i). Let x^* be an equilibrium different than the origin. From the proof of Lemma 3.5.1, x^* is an equilibrium point if and only if

$$H_+(\bar{A}x^* + ({x^*}^\top \bar{B}_1 x^*, \cdots, {x^*}^\top \bar{B}_n x^*)^\top) = x^*,$$

i.e., if and only if x^* is the fixed point of the map $H(x) := H_+(\bar{A}x + (x^\top \bar{B}_1 x, \cdots, x^\top \bar{B}_n x)^\top)$. Now, observe that

$$\begin{aligned} H(x^*) &\leq \bar{A}x^* + ({x^*}^\top \bar{B}_1 x^*, \cdots, {x^*}^\top \bar{B}_n x^*)^\top \\ &\leq \bar{A}x^* + (\mathbb{1}_n^\top \bar{B}_1 x^*, \cdots, \mathbb{1}_n^\top \bar{B}_n x^*)^\top \end{aligned}$$

where the first inequality follows from $h_+(z) \leq z$ for $z \in (0,1]$ and the second one from $x^* \in [0,1]^n$. Now, observe that if $\mathbb{0}_n \leq x \leq y$ then $\mathbb{0}_n \leq H(x) \leq \bar{A}x + (\mathbb{1}_n^\top \bar{B}_1 x, \cdots, \mathbb{1}_n^\top \bar{B}_n x)^\top \leq \bar{A}y + (\mathbb{1}_n^\top \bar{B}_1 y, \cdots, \mathbb{1}_n^\top \bar{B}_n y)^\top$; and so, the kth iteration of the map H satisfies: $\mathbb{0}_n \leq H^k(x^*) \leq (\bar{A} + (\mathbb{1}_n^\top \bar{B}_1, \cdots, \mathbb{1}_n^\top \bar{B}_n)^\top)^k x^*$. Now, assume by contradiction that $x^* \neq \mathbb{0}_n$. Then, from our previous calculations, $0 \leq \left\| H^k(x^*) - H^k(0) \right\| \leq \left\| (\bar{A} + (\mathbb{1}_n^\top \bar{B}_1, \cdots, \mathbb{1}_n^\top \bar{B}_n)^\top)^k \right\| \|x^*\|$ since $H^k(\mathbb{0}_n) = \mathbb{0}_n$ and where the last inequality follows from the definition of induced norms. Now, by hypothesis, we have that $\rho(\bar{A} + (\mathbb{1}_n^\top \bar{B}_1, \cdots, \mathbb{1}_n^\top \bar{B}_n)^\top) < 1$, and so, it follows that $\lim_{k\to\infty}(\bar{A} + (\mathbb{1}_n^\top \bar{B}_1, \cdots, \mathbb{1}_n^\top \bar{B}_n)^\top)^k = \mathbb{0}_{n\times n}$. Then, by the Sandwich theorem, $\lim_{k\to\infty} \left\| H^k(x^*) - H^k(0) \right\| = 0$ but recalling that $H^k(x^*) = x^*$ since x^* is a fixed point of H, we obtain $\|x^*\| = \mathbb{0}_n$, which is a contradiction. Then, $\mathbb{0}_n$ is the unique fixed point in $[0,1]^n$ for the map H, and thus, the unique

equilibrium point for the system.

Now we prove fact (ii). Since $\bar{A} \geq 0$ is irreducible, let $v \gg \mathbb{0}_n$ be the left Perron-Frobenius eigenvector of $\bar{A} + (\mathbb{1}_n^\top \bar{B}_1, \cdots, \mathbb{1}_n^\top \bar{B}_n)^\top \geq 0$ [82, Theorem 8.4.4.], and let $\lambda := \rho(\bar{A} + (\mathbb{1}_n^\top \bar{B}_1, \cdots, \mathbb{1}_n^\top \bar{B}_n)^\top)$ be its eigenvalue. Set $y = v^\top \Gamma^{-1} x$, then $\dot{y} = v^\top \Gamma^{-1} \dot{x}$ and

$$\begin{aligned}
\dot{y} &\leq -v^\top x + v^\top (\bar{A}x + (x^\top \bar{B}_1 x, \cdots, x^\top \bar{B}_n x)^\top) \\
&\leq (-1+\lambda) v^\top x \\
&= (-q+\lambda) v^\top \Gamma \Gamma^{-1} x \leq (-1+\lambda)(\min_i \gamma_i) y,
\end{aligned}$$

where the first inequality follows from $v^\top \Gamma^{-1}(I_n - \mathrm{diag}(x)) \leq v^\top \Gamma^{-1}$ for any $x \in [0,1]^n$. Set $q := (-1+\lambda)(\min_i \gamma_i) < 0$. Then, the Comparison Lemma [94] implies $y(t) \leq y(0) e^{qt}$ for $t \geq 0$. From this it follows that $x_i(t) \leq \frac{v^\top x(0)}{v_i} e^{qt}$ and so $\|x(t)\|_1 \leq C_o e^{qt}$ for some constant $C_o > 0$, which finally implies that $\mathbb{0}_n$ is globally exponentially stable in $[0,1]^n$.

Next we prove fact (iii). First, observe that the Jacobian evaluated at the equilibrium point $\mathbb{0}_n$ is $\dot{x} = (-\Gamma + \beta_1 A)x$. Since A is irreducible, let $v \gg \mathbb{0}_n$ be the right Perron-Frobenius eigenvector of $\beta_1 \Gamma^{-1} A$; and let ρ denote its associated eigenvalue. Note that $-\Gamma + \beta_1 A$ is Metzler and $(-\Gamma + \beta_1 A)v = (-1+\rho)\Gamma v \ll \mathbb{0}_n$ since $-1+\rho < 0$ by assumption. Using [33, Theorem 15.17], we conclude that the matrix $-\Gamma + \beta_1 A$ is Hurwitz and so the origin is locally exponentially stable.

Now we prove fact (iv). First, we introduce the following result: for any $\alpha > 1$, $h_+(\alpha z) \geq z$ with $z \geq 0$ if and only if $z \leq 1 - \frac{1}{\alpha}$. Now, consider the vector $\mathbf{1}_B$ as in the theorem statement and define $Y = \left\{ y \in [0,1]^n \mid \frac{1}{2}\mathbf{1}_B \leq y \leq \mathbb{1}_n \right\}$ and $\theta := \min_{i \text{ s.t. } B_i \neq \mathbb{0}_{n \times n}} \left(\frac{2\beta_1}{\gamma_i} (A\mathbf{1}_B)_i + \frac{\beta_2}{\gamma_i} \mathbf{1}_B^\top B_i \mathbf{1}_B \right)$. Note that $\theta \geq 4$ by hypothesis. Let $y \in Y$,

then

$$\begin{aligned} H(y) &= H_+(\bar{A}y + (y^\top \bar{B}_1 y, \cdots, y^\top \bar{B}_n y)^\top) \\ &\geq H_+\Big(\frac{1}{2}\bar{A}\mathbf{1}_B + \frac{1}{4}(\mathbf{1}_B^\top \bar{B}_1, \cdots, \mathbf{1}_B^\top \bar{B}_n)^\top \mathbf{1}_B\Big), \end{aligned} \tag{3.8}$$

where the monotonicity of the function h_+ implies the inequality. Now, the ith entry of the argument of H_+ in right-hand side of (3.8) is $\frac{1}{4}(\frac{2\beta_1}{\gamma_i}\sum_{j=1}^n a_{ij}(\mathbf{1}_B)_j + \frac{\beta_2}{\gamma_i}\mathbf{1}_B^\top B_i \mathbf{1}_B)$. When $B_i \neq \mathbb{0}_{n\times n}$, we can lower bound the ith entry by $\frac{1}{4}\theta$; and when $B_i = \mathbb{0}_{n\times n}$, by 0. Therefore, from (3.8),

$$H(y) \geq H_+\Big(\frac{1}{4}\theta\mathbf{1}_B\Big) \geq \frac{1}{2}\mathbf{1}_B$$

where the last inequality follows from our statement at the beginning of the paragraph. Now, from the fact that $h_+(z) \leq 1$ for any $z \geq 0$, we have that $H(y) = H_+\bar{A}y + (y^\top \bar{B}_1 y, \cdots, y^\top \bar{B}_n y)^\top) \leq \mathbb{1}_n$. Then, we conclude that $H : Y \to Y$, and so H is a continuous map that maps Y into itself. The Brouwer Fixed-Point Theorem (e.g., see [155, Theorem 4.5]) implies that there exists $y^* \in Y$ such that $H(y^*) = y^*$, i.e., an equilibrium point y^* for the system which belongs to Y. This equilibrium point y^* is not guaranteed to be unique. Moreover, from statement (iii) of Lemma 3.5.1, we conclude that no entry of y^* can be zero, and so $y^* \gg \mathbb{0}_n$.

Now, we prove fact (v). Let x^* be an equilibrium of the system such that $x^* \geq \frac{1}{2}\mathbf{1}_B$ with $x^* \gg \mathbb{0}_n$. Evaluating the Jacobian of the system at x^*, namely $Df(x^*)$, we obtain

$$\begin{aligned} Df(x^*) = &-\Gamma + \beta_1(I_n - \operatorname{diag}(x^*))A - \beta_1 \operatorname{diag}(Ax^*) \\ &+ \beta_2(I_n - \operatorname{diag}(x^*))O_1(x^*) - \beta_2 O_2(x^*), \end{aligned}$$

with

$$O_1(x^*) := ({x^*}^\top(B_1 + B_1^\top), \cdots, {x^*}^\top(B_n + B_n^\top))^\top$$

and

$$O_2(x^*) := \operatorname{diag}(x^{*\top} B_1 x^*, \cdots, x^{*\top} B_n x^*)^\top.$$

Clearly, $Df(x^*)$ is a Metzler matrix. Now, observe that

$$\begin{aligned} Df(x^*)x^* = &-\beta_1 \operatorname{diag}(Ax^*)x^* + \beta_2(I_n - \operatorname{diag}(x^*))(x^{*\top} B_1 x^*, \cdots, x^{*\top} B_n x^*)^\top \\ &- \beta_2 \operatorname{diag}(x^{*\top} B_1 x^*, \cdots, x^{*\top} B_n x^*)x^*, \end{aligned} \tag{3.9}$$

where we simplified terms by using the equilibrium equation for the system (3.5). Let $(Df(x^*)x^*)_i$ be the ith entry of the left-hand side of equation (3.9). Then

$$\begin{aligned} (Df(x^*)x^*)_i = &-\beta_1 \left(\sum_{i=1}^{n} a_{ij} x_j^* \right) x_i^* \\ &+ \beta_2(1 - 2x_i^*)(x^{*\top} B_i x^*). \end{aligned} \tag{3.10}$$

First, consider $B_i \neq \mathbb{0}_{n\times n}$. Then, it follows that $x_i^* \geq \frac{1}{2}$ and that $(1 - 2x_i^*) \leq 0$. In turn we obtain, in (3.10),

$$(Df(x^*)x^*)_i \leq -\left(\beta_1 \min_j \left(\sum_{i=1}^{n} a_{ij} x_j^* \right) \right) x_i^*.$$

On the other hand, if $B_i = \mathbb{0}_{n\times n}$, then $(Df(x^*)x^*)_i = -\beta_1 \left(\sum_{i=1}^{n} a_{ij} x_j^* \right) x_i^*$ in (3.10). Therefore, from these two cases, we conclude $Df(x^*)x^* \leq -dx^*$ for some $d > 0$ since A is irreducible. Then, since $x^* \gg \mathbb{0}_n$ [33, Theorem 15.17] implies that $Df(x^*)$ is Hurwitz, and so x^* is locally exponentially stable.

Now we prove fact (vi). First we prove that $\mathbb{0}_n$ is an unstable equilibrium. The linearization respect to the equilibrium point $\mathbb{0}_n$ is $\dot{x} = (-\Gamma + \beta_1 A)x$. Let $v \gg \mathbb{0}_n$ be the right Perron-Frobenius vector of the matrix $\beta_1 \Gamma^{-1} A$, and let ρ be its associated eigenvalue. Now, since $-\Gamma + \beta_1 A$ is Metzler, $\rho > 1$, and A is irreducible; we invoke [33,

E10.15] to conclude that the leading eigenvalue of $-\Gamma + \beta_1 A$ is strictly positive.

Next we prove fact (vii). Define $Y = \{y \in [0,1]^n \mid c \leq y \leq \mathbb{1}_n\}$ for a fixed $c = \alpha v$ and $0 < \alpha < 1$ small enough so that $c \leq \left(1 - \frac{1}{\rho}\right)\mathbb{1}_n$, which is well-posed since $\rho > 1$ by assumption. Let $y \in Y$, then

$$H(y) = H_+(\bar{A}y + (y^\top \bar{B}_1 y, \cdots, y^\top \bar{B}_n y)^\top) \geq H_+(\bar{A}y) \geq H_+(\alpha\rho v) = H_+(\rho c) \geq c$$

where the inequalities are similar to the ones used in the the proof of fact (iv). Since we know also that $H(y) \leq \mathbb{1}_n$, the Brouwer Fixed-Point Theorem implies that there exists some $y^* \in Y$ such that $H(y^*) = y^*$, i.e., there exists an equilibrium point $y^* \in Y$ for the system and, by construction, $y^* \gg \mathbb{0}_n$.

Now, we prove fact (viii). First, we prove that Y can be made a forward-invariant set for the system (3.5). If $x \in Y$, then $x_i \in [c_i, 1]$. Then, we can use Nagumo's theorem [26, Theorem 4.7] and analyze the vector field at the boundary of Y, which is an n-dimensional rectangle. As in the proof for statement (i) of Lemma 3.5.1, we have that $f_i(x) < 0$ for all $x \in Y$ such that $x_i = 1$ for some $i \in \{1, \ldots, n\}$. Then, we need to analyze only the case where $x \in Y$ with $x_i = c_i = \alpha v_i$ for some $i \in \{1, \ldots, n\}$. Consider such x. Then,

$$\begin{aligned}
f_i(x) &= -\gamma_i c_i + \beta_1(1 - c_i)\sum_{j=1}^{n} a_{ij}x_j + \beta(1 - c_i)x^\top B_1 x \\
&\geq -\gamma_i c_i + \beta_1(1 - c_i)\sum_{j=1}^{n} a_{ij}c_j \\
&= -\alpha\gamma_i v_i + \alpha\gamma_i(1 - \alpha v_i)\frac{\beta_1}{\gamma_i}\sum_{j=1}^{n} a_{ij}v_j \\
&= \alpha\gamma_i(-1 + (1 - \alpha v_i)\rho(\bar{A}))v_i,
\end{aligned}$$

and so $f_i(x) \geq 0$ if $\rho(\bar{A}) \geq \frac{1}{1-\alpha v_i}$. Then, if $\rho(\bar{A}) \geq \frac{1}{1-\alpha \max_i v_i}$, we conclude that Y

is forward invariant. Now, by construction of Y, we can make the parameter $\alpha > 0$ arbitrarily small, and since $\rho(\bar{A}) > 1$ by assumption, then we conclude that Y is forward invariant. Indeed, since $c \to \mathbb{0}_n$ as $\alpha \to 0$, we can define the positively invariant set Y to include any initial condition in $(0, 1]^n$. Moreover, from statement (ii) of Lemma 3.5.1, we conclude that any trajectory starting in $[0, 1]^n \setminus \{\mathbb{0}_n\}$ eventually enters the positive invariant set Y.

Now, let x^* be an equilibrium point of the system belonging to Y, so that $x^* \gg \mathbb{0}_n$ and let us consider the system (3.5) starting in the set Y. By subtracting the right-hand side of the equilibrium equation $\mathbb{0}_n = f(x^*)$, we can express the same equation (3.5) as

$$\dot{x} = \Lambda(x, x^*)(x - x^*) + \beta_2 \Omega(x, x^*)$$

with

$$\Lambda(x, x^*) := -\Gamma + \beta_1(I_n - \operatorname{diag}(x^*))A - \beta_1 \operatorname{diag}(Ax)$$

and

$$\begin{aligned}\Omega(x, x^*) := (I_n - \operatorname{diag}(x))(x^\top B_1 x, \cdots, x^\top B_n x)^\top \\ - (I_n - \operatorname{diag}(x^*))(x^{*\top} B_1 x^*, \cdots, x^{*\top} B_n x^*)^\top,\end{aligned}$$

and after some calculations,

$$\begin{aligned}\Omega(x, x^*) = \Bigg((I_n - \operatorname{diag}(x^*)) \begin{bmatrix} x^\top B_1^\top + x^{*\top} B_1 \\ \vdots \\ x^\top B_n^\top + x^{*\top} B_n \end{bmatrix} \\ - \operatorname{diag}(x^\top B_1 x, \cdots, x^\top B_n x)\Bigg)(x - x^*).\end{aligned}$$

Then, we can have the alternative expression for (3.5) as

$$\dot{x} = \mathcal{D}(x, x^*)(x - x^*)$$

with $\mathcal{D}(x, x^*) := (\mathcal{D}_1(x, x^*) + \mathcal{D}_2(x, x^*))$ and

$$\begin{aligned}
\mathcal{D}_1(x, x^*) &:= -\Gamma + \beta_1(I_n - \operatorname{diag}(x^*))A + \beta_2(I_n - \operatorname{diag}(x^*))({x^*}^\top B_1, \cdots, {x^*}^\top B_n)^\top, \\
\mathcal{D}_2(x, x^*) &:= -\beta_1 \operatorname{diag}(Ax) + \beta_2(I_n - \operatorname{diag}(x^*))(x^\top B_1^\top, \cdots, x^\top B_n^\top)^\top \\
&\quad - \beta_2 \operatorname{diag}(x^\top B_1 x, \cdots, x^\top B_n x).
\end{aligned}$$

Now, from the equilibrium equation $\mathbb{0}_n = f(x^*)$, we notice that $\mathcal{D}_1(x, x^*)x^* = \mathbb{0}_n$. Since $x \in Y$, notice that $-\operatorname{diag}(Ax)x^* \leq -\operatorname{diag}(Ac)x^*$ and $-\operatorname{diag}(x^\top B_1 x, \cdots, x^\top B_n x)x^* \leq -\operatorname{diag}(c^\top B_1 c, \cdots, c^\top B_n c)x^* \leq \mathbb{0}_n$. Using these results, we obtain

$$\begin{aligned}
\mathcal{D}_2(x, x^*)x^* &\leq -\beta_1 \operatorname{diag}(Ac)x^* \\
&\quad + \beta_2(I - \operatorname{diag}(x^*))(x^\top B_1^\top x^*, \cdots, x^\top B_n^\top x^*)^\top \\
&\leq (-\beta_1 \operatorname{diag}(Ac) \\
&\quad + \beta_2(I - \operatorname{diag}(x^*))(\mathbb{1}_n^\top B_1^\top, \cdots, \mathbb{1}_n^\top B_n^\top)^\top)x^*.
\end{aligned}$$

Now, since A is irreducible and $c \gg \mathbb{0}_n$, for a fixed value of $\beta_1 > 0$, there exists $\beta_2 > 0$ sufficiently small so that $\mathcal{D}_2(x, x^*)x^* \leq -dx^*$ for some constant $d > 0$. Therefore, we have shown that $\mathcal{D}(x, x^*)x^* \leq -dx^*$ for any $x \in Y$. Since $\mathcal{D}(x, x^*)$ is Metzler (because both $\mathcal{D}_1(x, x^*)$ and $\mathcal{D}_2(x, x^*)$ are Metzler) and Y is a convex compact forward-invariant set, we can use expression (3.2) along with Theorem 3.3.1. Then, we conclude that x^* is the unique globally exponentially stable equilibrium point in Y, and, as a consequence of statement (iii) from Lemma 3.5.1, it has the same property over the set $[0, 1]^n \setminus \{\mathbb{0}_n\}$.

This finishes the proof of fact (viii).

The last claim of the theorem follows from the proof of fact (ii) which states that $\beta_1\rho(\Gamma^{-1}A) < 1$ implies $\mathbb{0}_n$ is locally exponentially stable, and thus we are in either the disease-free or bistable domain. ■

Theorem 3.5.4 (Algorithm for computing an endemic equilibrium) *Consider the simplicial SIS model and assume that the system parameters satisfy the sufficient conditions in Theorem 3.5.2 for the system to be in either the bistable or endemic domain. Define the map* $H_+ : \mathbb{R}^n_{\geq 0} \to \mathbb{R}^n_{\geq 0}$ *by* $H_+(z) = (\frac{z_1}{1+z_1}, \cdots, \frac{z_n}{1+z_n})^\top$ *and* $y_0 \in (0,1)^n$ *by*

$$y_0 = \begin{cases} \frac{1}{2}\mathbf{1}_B, & \text{for the bistable domain,} \\ (1-\frac{1}{\rho})u, & \text{for the endemic domain,} \end{cases}$$

with (ρ, u) *being the dominant right eigenpair of* $\beta_1\Gamma^{-1}A$ *and* $\|u\|_\infty = 1$. *Then the sequence* $(y_k)_{k\in\mathbb{N}} \subset (0,1)^n$ *defined by*

$$y_{k+1} = H_+\big(\beta_1\Gamma^{-1}Ay_k + \beta_2\Gamma^{-1}(y_k^\top B_1 y_k, \cdots, y_k^\top B_n y_k)^\top\big)$$

is monotonic nondecreasing and $\lim_{k\to\infty} y_k = x^*$ *is an endemic equilibrium (satisfying* $y_0 \ll x^* \ll \mathbb{1}_n$*).*

Proof: Let $f(x) := \beta_1\Gamma^{-1}Ax + \beta_2\Gamma^{-1}(x^\top B_1 x, \cdots, x^\top B_n x)^\top$ for $x \in [0,1]^n$. From the proof of Theorem 3.5.2, there exists an endemic state x^* which satisfies $H_+(f(x^*)) = x^*$. Now, we also know that $H_+(f(y_0)) \geq y_0$, and so $y_1 \geq y_0$. Similarly, we note that $y_2 = H_+(f(y_1)) \geq H_+(f(y_0)) = y_1$, which follows from the entry-wise monotonicity of H_+ and $y_1 \geq y_0$. Then, by induction, we obtain that $y_{k+1} = H_+(f(y_k)) \geq y_k$ for $k \geq 0$. Now, notice that $y_k \leq \mathbb{1}_n$ for $k \geq 0$, which let us conclude that $(y_i(k))_k$ is a monotonically

non-decreasing bounded sequence with upper bound 1. Then, $\lim_{k\to\infty} y_k = x^*$, with x^* an equilibrium point of the system in Y and away from $\mathbb{1}_n$ due to Lemma 3.5.1. ■

3.6 Analysis of higher-order models

We extend the simplicial SIS model to the setting of multiple arbitrary high-order interactions.

Definition 3.6.1 (The general higher-order SIS model) *Assume $x \in [0,1]^n$, and let $\beta_1, \cdots, \beta_{n-1} > 0$ and $\gamma_i > 0$, $i \in \{1, \ldots, n\}$. Then, the* general higher-order SIS model *is, for any $i \in \{1, \ldots, n\}$,*

$$\dot{x}_i = -\gamma_i x_i + \beta_1(1 - x_i)\sum_{j=1}^{n} a_{ij}x_j + (1 - x_i)\sum_{k=2}^{n-1}\beta_k \sum_{i_1,\ldots,i_k=1}^{n} b_{ii_1\cdots i_k}x_{i_1}\cdots x_{i_k},$$

where $b_{ii_1\cdots i_k} \geq 0$ for any $i \in \{1, \ldots, n\}$ and $k \in \{2, \cdots, n-1\}$, and $A = (a_{ij})$ is a nonnegative matrix.

We believe it is straightforward to extend the analysis of the simplicial SIS model in Lemma 3.5.1 to the general higher-order SIS model in this definition. The reason is that the Lemma 3.5.1's proof essentially depends on matrix A and so is independent of any higher-order interaction; therefore, we omit it here in the interest of brevity. Similarly, under appropriate changes on the sufficient conditions that define each behavioral domain, parallel results to Theorem 3.5.2 can be obtained. In the interest of brevity, we only focus on establishing that a bistable domain also exists for arbitrary higher-order interactions. For convenience, define the shorthand:

$$b_i^* := \sum_{k=2}^{n-1}\beta_k\Big(\sum_{i_1,\ldots,i_k=1}^{n} b_{ii_1\cdots i_k}\Big).$$

Proposition 3.6.1 (Bistable domain in higher-order interactions) *Consider the general higher-order SIS model with an irreducible $A \geq 0$ and arbitrary $b_{ii_1\cdots i_k} \geq 0$ for any $i \in \{1,\dots,n\}$ and $k \in \{2\cdots, n-1\}$. Define $\mathbf{1}_{b^*} \in \{0,1\}^n$ by $(\mathbf{1}_{b^*})_i = 1$ if $b_i^* > 0$ and $(\mathbf{1}_{b^*})_i = 0$ otherwise. If $\beta_1\rho(\Gamma^{-1}A) < 1$ and*

$$\min_{i \text{ s.t. } b_i^*\neq 0} \left(\frac{\beta_1}{\gamma_i}(A\mathbf{1}_{b^*})_i \; + \; \sum_{k=2}^{n-1} \frac{\beta_k}{\gamma_i}\Big(\frac{n-2}{n-1}\Big)^{k-1} \sum_{i_1,\dots,i_k=1}^{n} b_{ii_1\cdots i_k} \prod_{\ell=1}^{k} (\mathbf{1}_{b^*})_{i_\ell} \right) \quad \geq \quad n \; - \; 1,$$

then

(i) $\mathbb{0}_n$ is a locally exponentially stable equilibrium,

(ii) there exists an equilibrium point $x^* \gg \mathbb{0}_n$ such that $x_i^* \geq \frac{n-2}{n-1}$ for any i such that $b_i^* \neq 0$, and

(iii) any such equilibrium point x^* is locally exponentially stable.

Proof: Consider the functions H_+ and h_+ introduced in the proof of Lemma 3.5.1. Let $\bar{A} := \beta_1\Gamma^{-1}A$. The proof for fact (i) is the same as in Theorem 3.5.2. Now, we prove fact (ii). Define $Y = \big\{y \in [0,1]^n \mid \frac{n-2}{n-1}\mathbf{1}_{b^*} \leq y \leq \mathbb{1}_n\big\}$. Rewrite the second inequality assumption in the proposition statement as $\theta \geq n-1$, where θ is a shorthand for the minimum term. For a point $y \in Y$, we compute

$$\begin{aligned}(H_+(y))_i &= h_+\Big((\bar{A}y)_i + \sum_{k=2}^{n-1} \frac{\beta_k}{\gamma_i} \sum_{i_1,\dots,i_k=1}^{n} b_{ii_1\cdots i_k} \prod_{\ell=1}^{k} y_{i_\ell}\Big) \\ &\geq h_+\left(\frac{n-2}{n-1}((\bar{A}\mathbf{1}_B)_i + \sum_{k=2}^{n-1} \frac{\beta_k}{\gamma_i}\Big(\frac{n-2}{n-1}\Big)^{k-1} \times \sum_{i_1,\dots,i_k=1}^{n} b_{ii_1\cdots i_k} \prod_{\ell=1}^{k} (\mathbf{1}_{b^*})_{i_\ell})\right),\end{aligned}$$

where the inequality follows from the monotonicity of the function h_+. Whenever $b_i^* \neq 0$, we can lower bound the expression in (**??**) by $h_+(\frac{n-2}{n-1}\theta)$; and whenever $b_i^* = 0$, we can

lower bound it by $h_+(0) = 0$. Therefore, as in the proof of Theorem 3.5.2, we obtain

$$H(y) \geq H_+\Big(\frac{n-2}{n-1}\theta \mathbf{1}_{b^*}\Big) \geq \frac{n-2}{n-1}\mathbf{1}_{b^*}.$$

Then, following the proof for the bistable domain of Theorem 3.5.2, we obtain that there exists an equilibrium point $y^* \in Y$ such that $y \gg \mathbb{0}_n$.

Now we prove fact (iii). Let $x^* \gg \mathbb{0}_n$ be an equilibrium satisfying $x^* \geq \frac{n-2}{n-1}\mathbf{1}_{b^*}$. Evaluating the Jacobian of the system at x^*, namely $Df(x^*)$, and after some algebraic work (similar to the one done in the proof of Theorem 3.5.2), we observe that $Df(x^*)$ is a Metzler matrix and, moreover, that

$$(Df(x^*)x^*)_i = -\beta_1(Ax^*)_i x_i^* + \sum_{k=2}^{n-1}\big((k-1) - kx_i^*\big)\beta_k\big(\sum_{i_1,\dots,i_k=1}^{n} b_{ii_1\cdots i_k}\prod_{\ell=1}^{k} x_{i_\ell}\big).$$

First, if $b_i^* \neq 0$, then $x_i^* \geq \frac{n-2}{n-1}$ and $(k-1-kx_i^*) \leq 0$, since $\frac{k-1}{k} \leq \frac{n-2}{n-1} \leq x^*$ for $k \in \{2, \cdots, n-1\}$. In turn,

$$(Df(x^*)x^*)_i \leq -\left(\beta_1 \min_j\left(\sum_{i=1}^{n} a_{ij}x_j^*\right)\right)x_i^*.$$

On the other hand, if $b_i^* = 0$, then $(Df(x^*)x^*)_i = -\beta_1\left(\sum_{i=1}^{n} a_{ij}x_j^*\right)x_i^*$. Therefore, from these two cases and recalling that A is irreducible, we have $Df(x^*)x^* \leq -dx^*$ for some $d > 0$. Finally, since $x^* \gg \mathbb{0}_n$, [33, Theorem 15.17] implies that $Df(x^*)$ is Hurwitz and, therefore, x^* is locally exponentially stable. ■

3.7 Numerical example

In Figure 3.4, we present two numerical examples of the behavior of the simplicial SIS model. First, we verify the existence of a parameter region under which the sufficient

conditions of Theorem 3.5.2 cannot be applied. We can readily observe the transition from the disease-free domain to the bistable domain as we increase β_2 for a fixed β_1, as mentioned in Remark 3.5.3. Also, notice that the sufficient condition for determining the endemic domain in Theorem 3.5.2 is tight. We also remark that the sufficient condition for determining the bistable region captures most of the true parameter region in these simulations.

From our numerical simulations we propose the following conjectures, which are consistent with the behavior observed in the scalar model.

Conjectures 3.7.1 (Behaviors in the bistable and endemic domains) *For the simplicial SIS model,*

(i) in the bistable domain, at fixed β_2, the domain of attraction of the disease-free equilibrium $x^* = \mathbb{0}_n$ decreases as β_1 increases. Once $\beta_1 = \frac{1}{\rho(\Gamma^{-1}A)}$, a bifurcation occurs and the origin becomes an unstable equilibrium point in the endemic domain;

(ii) in the endemic domain, the endemic equilibrium is unique and globally stable for any value of β_2.

3.8 Conclusion

In this paper, we formally analyze the simplicial SIS model and establish its different behavioral domains. As seen in a previous scalar model, we show the existence of the bistable domain and its possible transition from the disease-free domain by changing the model parameters. This feature makes our model qualitatively different from the classical multi-group SIS model. We also show that the bistable domain exists for any multi-group SIS model with higher-order interactions.

As future work, we plan to study control strategies for the mitigation of the epidemic in the simplicial SIS model; e.g., how to drive the system to the origin whenever it is in the bistable domain. More generally, we also plan to study the aggregation of higher order interaction terms in other epidemiological models, where we believe our approach based on Coppel's inequalities can also be useful. Finally, it is relevant to provide a more comprehensive characterization of the model parameters β_1 and β_2, and thus prove the tight transition between the disease-free and the bistable domains illustrated by our simulations.

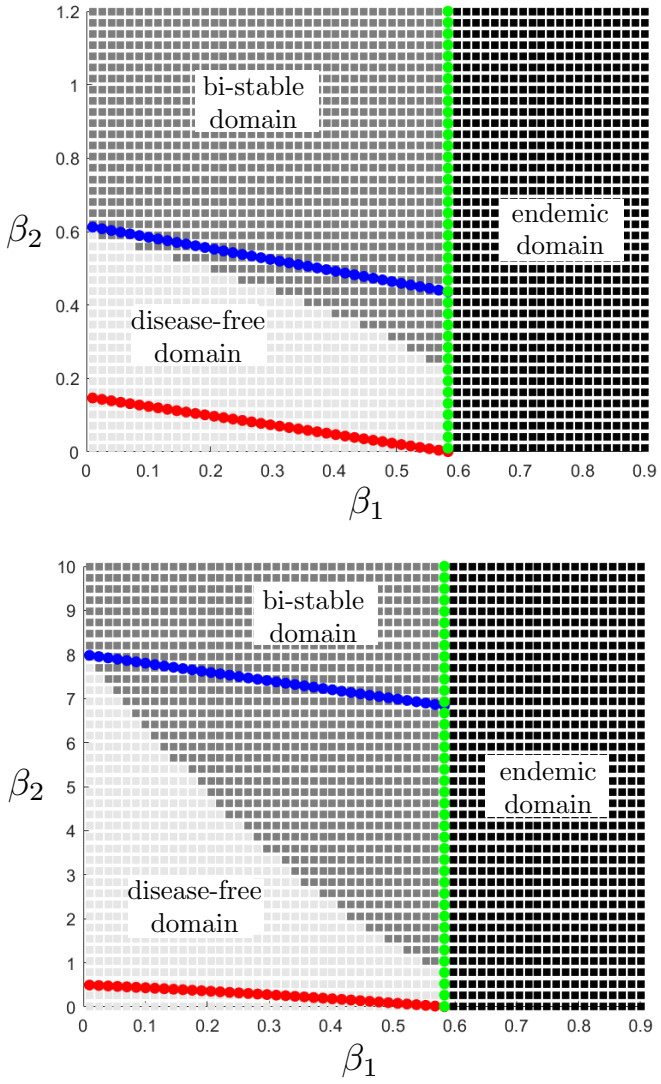


Figure 3.4: Consider the simplicial SIS model with its parameters (Definition 3.4.1). For the upper figure, we randomly generated and fixed six irreducible matrices $A \in \{0,1\}^{5\times 5}$ and $B_i \in \{0,1\}^{5\times 5}$, $B_i \neq \mathbb{0}_n$, $i \in \{1,\dots,5\}$; and fixed $\Gamma = 2I_5$. The light-gray/gray/black region corresponds to the disease-free/bistable/endemic domain from the simulation. Regarding the sufficient conditions established by Theorem 3.5.2, all the region to the right of the green line correspond to the endemic domain, all the region above the blue line and left to the green line to the bistable domain, and all the region below the red curve to the disease-free domain. For the lower figure, we considered the same settings as in the upper figure, but with the difference that this time we set $B_i = \mathbb{0}_{n\times n}$ for $i \in \{2,\cdots,5\}$.

Chapter 4

Distributed Wasserstein Barycenters via Displacement Interpolation

4.1 Introduction

Problem statement and motivation

There has been strong interest in the theoretical study and practical application of Wasserstein barycenters over the last decade. In this paper, we characterize the evolution of a distributed system where all the computing units or *agents* hold a probability measure, interact through pairwise communication, and perform *displacement interpolations* in the Wasserstein space. These pairwise interactions are asynchronous[1] and stochastic. We study the conditions under which the agents' measures will asymptotically achieve consensus and, additionally, consensus on a Wasserstein barycenter of the agents' initial measures. Moreover, we are interested in computing both the standard Wasserstein

[1] A distributed system is *synchronous* when all computing units perform their computations altogether in every time step (assuming the system computes in a discrete sequence of time steps); otherwise, it is *asynchronous*.

barycenter and randomized weighted versions of it – as a result of the stochastic interactions. Finally, we consider both undirected and directed communication graphs. To the best of our knowledge, these problems have not been studied in the literature on the distributed computation of Wasserstein barycenters.

Asynchronous pairwise algorithms are inherently robust to communication failures and do not require synchronization of the whole multi-agent system. Pairwise interactions are also important because they have the potential of reducing the computation complexity of each agent, which otherwise may need to perform more complex local computations. Indeed, displacement interpolations have the practical advantage that they may have a closed form expression, e.g., in the Gaussian case.

Literature review

Wasserstein barycenters and their applications The Wasserstein barycenter of a set of measures can be interpreted as an interpolation or weighted Fréchet mean of multiple measures in the Wasserstein space; it is intimately related to the theory of *optimal transport* [169, 137]. In this interpolation, each measure has an associated positive *weight* that indicates its importance in the computation of the barycenter, the collection of all weights form convex coefficients. When all weights are equal, we obtain the *standard* Wasserstein barycenter; otherwise, we obtain a *weighted* one.

There has been a strong interest in the theoretical study of Wasserstein barycenters over the last decade; e.g., uniqueness results and the connection to multi-marginal optimal transport problems in [6]; the study of interpolated discrete measures with finite support and its relationship with linear programming in [37, 13]; the characterization of the barycenter as a fixed point of an operator and the proposal of iterative computation procedures in [8]; the study of consistency and other statistical properties in [101]. For

further information, we refer to the recent introductory book [137].

Along with the theoretical progress, many applications of Wasserstein barycenters have emerged, as well as computational or numerical approaches for computing them. For example, Wasserstein barycenters have found applications in economics [36], image processing [146, 127], computer graphics [29], physics [34], statistics [159, 154], machine learning [50, 153], signal processing [24, 20], and biology [72]. On the other hand, examples of computational approaches include: exact algorithms [37, 45], algorithms that use entropic regularization [51, 52], and algorithms based on approximations of Wasserstein distances [30]. Finally, the particular case of interpolating two measures, i.e., the *displacement interpolation* (which defines the pairwise interactions in our distributed algorithm), is interesting in its own right because of its applications in partial differential equations and geometry [170, 151], and fluid mechanics [21].

Moreover, the Wasserstein barycenter has been interpreted as a denoised version of an original signal whose sensor measurements are each of the noisy probability distributions that are being interpolated; thus, the barycenter has found multiple applications as an *information fusion* algorithm [72, 24, 29, 45]. In a related setting, randomized barycenters could also be of practical interest. For example, consider we want to estimate the interpolation resulting from the measurements of various sensors of unknown accuracy or noise level. Then, a randomized barycenter will randomly weight each sensor and may provide different estimates of the true measurement.

Finally, we mention that the problem studied in this paper contributes to the fields of randomized consensus algorithms (e.g., [33, Chapter 13]) and of consensus in spaces other than the classic Euclidean space. Moreover, our study of stochastic asynchronous pairwise interactions also contributes to the field of opinion dynamics, since this type of interactions is also used in classic opinion models, e.g., see [56, 5]. Indeed, in our paper, we argue that the displacement interpolation is a more suitable modeling approach for

the non-Bayesian updating for the beliefs of individuals in a social network, compared to the classic averaging approaches in the literature.

Distributed algorithms for Wasserstein barycenters To the best of our knowledge, there is only a recent and growing literature on distributed algorithms for Wasserstein barycenters. The idea of computing Wasserstein barycenters in a distributed way was first studied by Bishop and Doucet [25]. Their work formally shows consensus towards the Wasserstein barycenter of the agents' initial measures. However, in order to compute such consensus, each agent needs to fully compute the Wasserstein barycenter resulting from its own measure and the measures from all its neighbors at each iteration. The work [25] formally studies the case of probability measures on the real line. Finally, the results in [25] assume that the communication between agents is deterministic, but flexible enough to consider both synchronous and asynchronous deterministic updating. The work assumes agents are connected by an undirected graph.

The recent work [166] focuses on the design and distributed implementation of a numerical solver that approximates the standard Wasserstein barycenter when all the measures are discrete, through the use of entropic regularization. Moreover, the recent work [61] from the same authors proposes another distributed solver for an approximate Wasserstein barycenter with the difference that the agents' measures may correspond to continuous distributions. Indeed, its framework is *semi-discrete*, in that the measures to be interpolated can be continuous, but the sought measure that serves as a proxy for the barycenter, is restricted to be a discrete measure with finite support. Therefore, we observe that the distributed algorithms from both works [166, 61] compute an approximate or a proxy of the true barycenter. Remarkably, both works exploit the dual formulation of the Wasserstein barycenter optimization problem to propose their numerical algorithms. Finally, both works require synchronous updating, i.e., all the agents

need to communicate at the same time with all their neighbors at every time step, and all the computations are performed over an undirected graph.

Our paper is more in line with the spirit of [25], in the sense that we propose a theoretical formulation and analysis that prove how to generate Wasserstein barycenters from distributed computations. We do not propose specific designs of numerical solvers for the local computations of the agents, as it is instead performed in [166, 61]. Indeed, since the local computations in our algorithm are displacement interpolations at every time step, any numerical method that can solve optimal transport problems can be used, including for example any of the numerical algorithms mentioned above.

Contributions

In this paper we propose the algorithm *PaWBar* (*Pa*irwise distributed algorithm for *W*asserstein *Bar*ycenters), where the agents update their measures via pairwise stochastic and asynchronous interactions implementing displacement interpolations (in contrast to the deterministic communication requirements in [25]). The algorithm has two versions: a *directed* and a *symmetric* version. As main contribution of this paper, we establish conditions under which both algorithms compute randomized and standard Wasserstein barycenters respectively. In the directed case, we prove that every time the algorithm is run, a barycenter with random convex weights is asymptotically generated as a result of the stochastic selection of the pairwise interactions. It is easy to characterize the first two moments of these random weights. During any pairwise directed interaction, only one agent updates its probability measure. On the other hand, in the symmetric case, both agents update their measures to equal a consensus value during their pairwise interaction. Although the interactions are stochastic, we prove that the asymptotically computed Wasserstein barycenter is the standard one (with probability

one). In contrast to the works [166, 61], our algorithm does not require all the agents to synchronously update their measures at every time step. Also in contrast to [166, 61], our framework provides convergence guarantees towards the computation of the barycenter independently from the numerical implementation of the local computations. We also remark that work [25] is different from ours because it dictates that each agent at every time step must locally compute the full Wasserstein barycenter of its neighboring agents' measures, which could be as complex as the centralized or direct computation of the barycenter of all the agents' measures.

We now elaborate on the convergence results. We first prove convergence to a randomized or standard Wasserstein barycenter for a class of discrete measures on $\mathbb{R}^d$, $d \geq 1$. In particular, we show that the obtained barycenter interpolates the agents' measures attained at some random finite time. However, if the initial measures are sufficiently close in the Wasserstein space, then such time is zero with probability one, i.e., there is an interpolation of the initial measures. For the particular case where these discrete measures are on $\mathbb{R}$, the interpolation of the initial measures occurs with probability one no matter how arbitrarily distant these measures are from each other.

We then prove convergence to a randomized or standard Wasserstein barycenter for a class of measures that are absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$. As corollaries, we prove convergence of continuous probability distributions on the real line, and of a class of multivariate Gaussian distributions. In the case of these Gaussian distributions, we also provide simpler closed form expressions for the computations of the PaWBar algorithm, and a simplified expression of the converged barycenter. We also conjecture that the convergence to Wasserstein barycenters holds for general absolutely continuous measures, and we present supporting numerical evidence for the general multivariate Gaussian case.

Moreover, in all the cases mentioned above, the convergence results are proved with

the following underlying communication graphs: a strongly connected digraph for the directed PaWBar algorithm and a connected undirected graph for the symmetric case. For randomized barycenters, we characterize their random convex coefficients by the limit product of random stochastic matrices.

Finally, we prove a general consensus result for the case where the initial measures on $\mathbb{R}^d$ are of arbitrary nature; and our proofs make strong use of metric and geodesic properties of the Wasserstein space. The results are proved over a cycle graph for the directed PaWBar algorithm and over a line graph for the symmetric case. We also prove the consensus measure satisfies a known necessary condition for certain Wasserstein barycenters.

Paper organization

Section 6.2 has notation and preliminary concepts. Section 4.3 has the proposed *PaW-Bar algorithm* and its theoretical analysis. Section 4.4 presents the proofs for Section 4.3. Section 4.5 presents the connection between our algorithm and opinion dynamics. Section 6.7 is the conclusion.

4.2 Notation and preliminary concepts

Let $z = (z_i, \dots, z_n)^\top$ denote a vector $z \in \mathbb{R}^n$ with ith entry z_i, $i \in \{1, \dots, n\}$. Let $\|\cdot\|_2$ denote the Euclidean distance. The vector $e_i \in \mathbb{R}^n$ has all of its entries zero but the ith entry is one. Let $\mathbb{1}_n, \mathbb{0}_n \in \mathbb{R}^n$ be the all-ones and all-zeros vectors respectively, and I_n be the $n \times n$ identity matrix. A nonnegative matrix $A \in \mathbb{R}^{n \times n}$ is *row-stochastic* if $A\mathbb{1}_n = \mathbb{1}_n$, and *doubly-stochastic* if additionally $A^\top \mathbb{1}_n = \mathbb{1}_n$. The operator $\circ$ the composition of functions, and $\otimes$ the Kronecker product.

The numbers $\lambda_1, \dots, \lambda_n$ are called *convex coefficients* if $\lambda_i \geq 0$, $i \in \{1, \dots, n\}$, and

$\sum_{i=1}^{n} \lambda_i = 1$. The vector $\lambda := (\lambda_1, \ldots, \lambda_n)^\top$ is called a *convex vector*.

Let $V = \{1, \ldots, n\}$, $n \geq 2$, be the finite set of agents. We assume the agents are connected according to the graph $G = (V, E)$, so that V becomes the set of nodes and E the set of edges. When the elements of E are ordered pairs, i.e., $(i, j) \in E$ for some $i, j \in V$, G is a directed graph or *digraph*. In particular, $(i, j) \in E$ means that i and j are connected with a directed edge starting from i and pointing to j. When the elements of E are unordered pairs, i.e., $\{i, j\} \in E$ for some $i, j \in V$, G is said to be an undirected graph. The edges of an undirected graph have no sense of direction. When a scalar value is assigned to every edge of G, then G is *weighted*. An undirected graph G is a line graph whenever, after an appropriate labeling of the nodes, $E = \{\{1, 2\}, \{2, 3\}, \ldots, \{n-1, n\}\}$. A digraph G is a cycle whenever, after an appropriate labeling of the nodes, $E = \{(1, 2), (2, 3), \ldots, (n-1, n), (n, 1)\}$. A digraph is strongly connected when, for any $i, j \in V$, it is possible to go from i to j by traversing the edges according to their direction; e.g., a cycle graph is strongly connected. An undirected graph is connected whenever it is possible to go from one node to another by traversing the edges in any direction; e.g., a line graph is connected.

We denote the set of all probability measures on $\Omega \subseteq \mathbb{R}^d$ by $\mathcal{P}(\Omega)$, and we define the subset of measures $\mathcal{P}^2(\Omega) = \left\{\mu \in \mathcal{P}(\Omega) \mid \int_\Omega \|x\|_2^2 \, d\mu(x) < \infty\right\}$. Consider $\mu \in \mathcal{P}(\mathbb{R})$. For $\Omega = \mathbb{R}$, we denote by F_μ the cumulative distribution function, i.e., $F_\mu(x) = \mu((-\infty, x])$. We denote by $\#$ the *push-forward operator*, which, for any Borel measurable map $\mathcal{M} : \Omega \to \Omega$, defines the linear operator $\mathcal{M}_\# : \mathcal{P}(\Omega) \to \mathcal{P}(\Omega)$ characterized by $(\mathcal{M}_\#\mu)(B) = \mu(\mathcal{M}^{-1}(B))$ for any Borel set $B \subseteq \Omega$. We denote the support of μ by $\mathrm{supp}(\mu)$.

We briefly review relevant concepts on optimal transport and Wasserstein barycenters.

Given $\mu, \nu \in \mathcal{P}^2(\Omega)$, the 2-Wasserstein distance[2] between μ and ν is

$$W_2(\mu,\nu) = \left(\inf_{\gamma \in \Pi(\mu,\nu)} \int_{\Omega\times\Omega} \|x-y\|_2^2 \, d\gamma(x,y) \right)^{1/2} \tag{4.1}$$

where $\Pi(\mu,\nu)$ is the set of probability measures on $\Omega \times \Omega$ with marginals μ and ν, i.e., if $\gamma \in \Pi(\mu,\nu)$, then $(\pi_1)_{\#}\gamma = \mu$ and $(\pi_2)_{\#}\gamma = \nu$ with $\pi_1(x,y) = x$ and $\pi_2(x,y) = y$. The optimization problem that defines the Wasserstein distance (i.e., the right-hand side of (4.1)) is an *(Monge-Kantorovich) optimal transport problem.* Any solution to an optimal transport problem is an *optimal transport plan*, and we let $\gamma^{\text{opt}}(\mu,\nu)$ denote an optimal transport plan between measures μ and ν. Given $\gamma^{\text{opt}}(\mu,\nu)$ such that $\nu = T_{\#}\mu$, we say that γ^{opt} solves the *Monge optimal transport problem* and the map T is called the *optimal transport map* from μ to ν. The Wasserstein space of order 2 is the space $\mathcal{P}^2(\Omega)$ endowed with the distance W_2. In this paper, we will consider $\Omega = \mathbb{R}^d$, $d \geq 1$.

Given convex coefficients $\lambda_1, \ldots, \lambda_n$ and probability measures $\mu_1, \ldots \mu_n \in \mathcal{P}^2(\Omega)$, Ω convex, $n \geq 2$, the *Wasserstein barycenter problem* is defined by the following convex problem[3]

$$\min_{\nu \in \mathcal{P}^2(\Omega)} \sum_{i=1}^{n} \lambda_i W_2^2(\nu, \mu_i). \tag{4.2}$$

A Wasserstein barycenter of the measures $\{\mu_i\}_{i=1}^n$ with weights $\{\lambda_i\}_{i=1}^n$ is any measure that solves equation (4.2), i.e., a *minimizer* of (4.2).

The *displacement interpolation* between the measures $\mu, \nu \in \mathcal{P}^2(\Omega)$ is the curve $\mu_\lambda = (\pi_\lambda)_{\#}\gamma^{\text{opt}}(\mu,\nu)$, $\lambda \in [0,1]$, where $\pi_\lambda : \Omega \times \Omega \to \Omega$ is defined by $\pi_\lambda(x,y) = (1-\lambda)x + \lambda y$. The curve π_λ is known to be a *constant-speed geodesic curve* in the Wasserstein space connecting $\mu_0 = \mu$ to $\mu_1 = \nu$ [151]. Moreover, for a fixed $\lambda \in [0,1]$, it is known to be the solution to the Wasserstein barycenter problem $\min_{\rho \in \mathcal{P}_2(\Omega)}((1-\lambda)W_2^2(\rho,\mu_1) + \lambda W_2^2(\rho,\mu_2))$.

[2]For simplicity, we refer to it as the *Wasserstein distance*.

[3]Some works in the literature multiply the functional to be minimized in (4.2) by a factor $\frac{1}{2}$, but the set of minimizers is the same in either problem.

When there exists an optimal transport map $\mathcal{M}$, the displacement interpolation can be written as $(\pi_\lambda)_\# \mu = ((1-\lambda)Id + \lambda\mathcal{M})_\# \mu$, $\lambda \in [0,1]$, where Id is the identity operator.

4.3 Proposed algorithm and analysis

4.3.1 The PaWBar algorithm

Let $\mu_i(t) \in \mathcal{P}^2(\mathbb{R}^d)$, $i \in V$, represent the measure of agent i at time $t \in \{0, 1, \dots\}$. Our proposed *PaWBar* (*P*airwise distributed *a*lgorithm for *W*asserstein *Bar*ycenters) algorithm has two versions.

Definition 4.3.1 (Directed PaWBar algorithm) *Let G be a weighted directed graph with weight $a_{ij} \in (0,1)$ for $(i,j) \in E$. Assume $\mu_i(0) := \mu_{i,0} \in \mathcal{P}^2(\mathbb{R}^d)$ for every $i \in V$. At each time t, execute:*

(i) select a random edge $(i,j) \in E$ of G, independently according to some time-invariant probability distribution, with all edges having a positive selection probability;

(ii) update the measure of agent i by

$$\mu_i(t+1) := (\pi_{a_{ij}})_\# \gamma^{\text{opt}}(\mu_i(t), \mu_j(t)) \tag{4.3}$$

where $\pi_{a_{ij}} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ is defined by $\pi_{a_{ij}}(x,y) = (1-a_{ij})x + a_{ij}y$.

Definition 4.3.2 (Symmetric PaWBar algorithm) *Let G be an undirected graph. Assume $\mu_i(0) := \mu_{i,0} \in \mathcal{P}^2(\mathbb{R}^d)$ for every $i \in V$. At each time t, execute:*

(i) select a random edge $\{i,j\} \in E$ of G, independently according to some time-invariant probability distribution, with all edges having a positive selection probability;

(ii) update the measures of agents i and j by

$$\mu_i(t+1) = \mu_j(t+1) := (\pi_{1/2})_{\#}\gamma^{\text{opt}}(\mu_i(t), \mu_j(t)) \tag{4.4}$$

where $\pi_{1/2} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ is defined by $\pi_{1/2}(x, y) = \frac{1}{2}(x + y)$.

Remark 4.3.1 (Well-posedness) *The PaWBar algorithm is well-posed since the displacement interpolation between any two measures in $\mathcal{P}^2(\mathbb{R}^d)$ provides measures in $\mathcal{P}^2(\mathbb{R}^d)$ [151, Theorem 5.27].*

Remark 4.3.2 (Symmetry in the interpolated measure) *Since $\pi_{1/2}(x, y) = \pi_{1/2}(y, x)$ for any $x, y \in \mathbb{R}^d$, the update rule (4.4) of the symmetric PaWBar algorithm is equivalent to $\mu_i(t+1) := (\pi_{1/2})_{\#}\gamma^{\text{opt}}(\mu_i(t), \mu_j(t))$ and $\mu_j(t+1) := (\pi_{1/2})_{\#}\gamma^{\text{opt}}(\mu_j(t), \mu_i(t))$.*

For simplicity, we call *edge selection process* the underlying stochastic process of edge selection by the PaWBar algorithm, whose realizations are the infinite sequence of selected edges chosen every time the PaWBar algorithm is run. When a result is stated *with probability one*, it is to be understood with respect to the induced measure by the edge selection process. The following concept and proposition are useful for our results.

Definition 4.3.3 (Evolution random matrix) *Consider the edge selection process from the PaWBar algorithm. Define the* evolution random matrix $A(t)$ *by:*

$$A(t) = \begin{cases} I_n - a_{ij}e_ie_i^\top - (1 - a_{ij})e_ie_j^\top, & \text{if } (i,j) \in E \text{ is chosen,} \\ I_n - \frac{1}{2}(e_ie_i^\top + e_je_j^\top + e_ie_j^\top + e_je_i^\top), & \text{if } \{i,j\} \in E \text{ is chosen.} \end{cases}$$

Proposition 4.3.3 (Convergence of products of evolution random matrices) *Consider the PaWBar algorithm. For the directed case with a strongly connected digraph:*

$$\lim_{t\to\infty}\prod_{\tau=0}^{t}A(\tau)=\mathbb{1}_n\lambda^\top$$

for some random convex vector λ with probability one. For the symmetric case with a connected undirected graph,

$$\lim_{t\to\infty}\prod_{\tau=0}^{t}A(\tau)=\frac{1}{n}\mathbb{1}_n\mathbb{1}_n^\top$$

with probability one.

Proof: We first prove the directed case. Observe that: (i) the sequence of random matrices $(A(t))_t$ is i.i.d due to the edge selection process, (ii) $A(t)$ has strictly positive diagonal entries and is row-stochastic for any time t with probability one, (iii) $\mathbb{E}[A(t)]$, the expected value of the evolution random matrix, corresponds to the adjacency matrix of a strongly connected weighted digraph. Then, from these conditions (i)-(iii), we can apply [33, Theorem 13.1] and obtain the sought convergence. The proof for the symmetric case follows from applying [33, Corollary 13.2] instead. ■

4.3.2 Analysis of discrete measures

Theorem 4.3.4 (Wasserstein barycenters for discrete measures) *Consider initial measures $\{\mu_{i,0}\}_{i\in V}$, such that $\mu_{i,0}=\frac{1}{N}\sum_{j=1}^{N}\delta_{x_j^i}$, with $x_1^i,\dots,x_N^i\in\mathbb{R}^d$ being distinct points; i.e., $\mu_{i,0}$ is a discrete uniform measure.*

(i) Consider the directed PaWBar algorithm with an underlying strongly connected digraph G; then, with probability one, for any $i\in V$,

$$W_2(\mu_i(t),\mu_\infty)\to 0 \text{ as } t\to\infty, \tag{4.5}$$

where the discrete uniform measure μ_∞ solves the barycenter problem

$$\mu_\infty \in \arg\min_{\nu\in\mathcal{P}^2(\mathbb{R}^d)} \sum_{i=1}^{n} \lambda_i W_2(\nu, \mu_i(T))^2$$

with $\lambda = (\lambda_1, \ldots, \lambda_n)^\top$ being a random convex vector satisfying $\prod_{\tau=1}^{\infty} A(\tau) = \mathbb{1}_n \lambda^\top$ with probability one, and $T \geq 0$ being some finite random time. If $\max_{i,j\in V} W_2(\mu_{i,0}, \mu_{j,0})$ is sufficiently small, then $T = 0$ with probability one.

(ii) Consider the symmetric PaWBar algorithm with an underlying connected undirected graph G; then, with probability one, for any $i \in V$,

$$W_2(\mu_i(t), \mu_\infty) \to 0 \text{ as } t \to \infty, \tag{4.6}$$

where the discrete uniform measure μ_∞ solves the barycenter problem

$$\mu_\infty \in \arg\min_{\nu\in\mathcal{P}^2(\mathbb{R}^d)} \sum_{i=1}^{n} W_2(\nu, \mu_i(T))^2 \tag{4.7}$$

with $T \geq 0$ being some finite random time. If $\max_{i,j\in V} W_2(\mu_{i,0}, \mu_{j,0})$ is sufficiently small, then $T = 0$ with probability one.

Corollary 4.3.5 (Wasserstein barycenters for discrete measures on $\mathbb{R}$) *Consider initial measures $\{\mu_{i,0}\}_{i\in V}$, such that $\mu_{i,0} = \frac{1}{N}\sum_{j=1}^{N} \delta_{x_j^i}$, with $x_1^i, \ldots, x_N^i \in \mathbb{R}$ such that $x_1^i < \cdots < x_N^i$. Then, the directed, respectively symmetric, PaWBar algorithm computes a randomized, respectively standard, Wasserstein barycenter of the initial measures.*

Remark 4.3.6 (Discussion of our results) *(i) The setting of Theorems 4.3.4 and Corollary 4.3.5 has found applications in computational geometry, computer graphics and digital image processing; e.g., see [146, 51, 30].*

(ii) In Theorem 4.3.4, if all initial measures are sufficiently close in the Wasserstein space, then the PaWBar algorithm computes one of their Wasserstein barycenters. This sufficient condition is not a problem in practical applications where the barycenter is used as an interpolation among measures that are known to be similar (e.g., measurements of the same object under noise). Moreover, Corollary 4.3.5 tells us that the initial measures in $\mathcal{P}(\mathbb{R})$ could be arbitrarily distant from each other and still the PaWBar algorithm will compute one of their barycenters.

4.3.3 Analysis of absolutely continuous measures

We consider measures that are absolutely continuous with respect to the Lebesgue measure. For any such measures $\mu, \nu \in \mathcal{P}^2(\mathbb{R}^d)$, there exists a unique optimal transport map from μ to ν, which we denote by T_μ^ν; and we also denote $T_\mu^\mu = Id$. It is also known that there exists a unique Wasserstein barycenter when all the interpolated measures are absolutely continuous with respect to the Lebesgue measure [6].

We focus on the class of measures that form a *compatible* collection. According to [137, Definition 2.3.1], a collection of absolutely continuous measures $\mathbb{C} \subset \mathcal{P}^2(\mathbb{R}^d)$ is *compatible* if for all $\nu, \mu, \gamma \in \mathcal{C}$, we have $(T_\gamma^\mu \circ T_\nu^\gamma)_\# \nu = (T_\nu^\mu)_\# \nu$.

Theorem 4.3.7 (Wasserstein barycenters for continuous measures in $\mathcal{P}^2(\mathbb{R}^d)$) *Consider initial measures $\{\mu_{i,0}\}_{i\in V}$ that are absolutely continuous with respect to the Lebesgue measure and that form a compatible collection. Let $\gamma \in \{\mu_{i,0}\}_{i\in V}$.*

(i) Consider the directed PaWBar algorithm with an underlying strongly connected digraph G; then, with probability one, for any $i \in V$,

$$W_2(\mu_i(t), \mu_\infty) \to 0 \text{ as } t \to \infty,$$

where the absolutely continuous measure $\mu_\infty = \left(\sum_{j=1}^n \lambda_j T_\gamma^{\mu_{j,0}}\right)_\# \gamma$ *is the barycenter*

$$\mu_\infty = \arg\min_{\nu \in \mathcal{P}^2(\mathbb{R}^d)} \sum_{i=1}^n \lambda_i W_2(\nu, \mu_{i,0})^2 \tag{4.8}$$

with $\lambda = (\lambda_1, \ldots, \lambda_n)^\top$ *being a random convex vector satisfying* $\prod_{\tau=1}^\infty A(\tau) = \mathbb{1}_n \lambda^\top$ *with probability one.*

(ii) Consider the symmetric PaWBar algorithm with an underlying connected undirected graph G; then, with probability one, for any $i \in V$*,*

$$W_2(\mu_i(t), \mu_\infty) \to 0 \text{ as } t \to \infty,$$

where the absolutely continuous measure $\mu_\infty = \left(\frac{1}{n}\sum_{j=1}^n T_\gamma^{\mu_{j,0}}\right)_\# \gamma$ *is the barycenter*

$$\mu_\infty = \arg\min_{\nu \in \mathcal{P}^2(\mathbb{R}^d)} \sum_{i=1}^n W_2(\nu, \mu_{i,0})^2.$$

The following corollary considers examples of measures relevant to our previous theorem. We use the term distribution and measure interchangeably for well-known probability measures with continuous distributions.

Corollary 4.3.8 (Examples of Wasserstein barycenters) *Consider that initially either*

(i) all agents have a probability measure in $\mathcal{P}^2(\mathbb{R})$ *with continuous distribution; or*

(ii) one agent has the standard Gaussian distribution on $\mathcal{P}^2(\mathbb{R}^d)$ *and any other agent* $i \in V$ *has a Gaussian distribution* $\mu_{i,0} = \mathcal{N}(m_{i,0}, \Sigma_{i,0})$ *with* $m_{i,0} \in \mathbb{R}^d$ *and* $\Sigma_{i,0} \in \mathbb{R}^{d\times d}$ *being a positive definite matrix with the joint commutative property* $\Sigma_{i,0}\Sigma_{j,0} = \Sigma_{j,0}\Sigma_{i,0}$ *for any* $j \in V$*.*

Consider the directed PaWBar algorithm with an underlying strongly connected digraph G. Then, with probability one, $W_2(\mu_i(t), \mu_\infty) \to 0$ as $t \to \infty$ for any $i \in V$, where μ_∞ is the Wasserstein barycenter of the initial measures. In particular,

- *for case (i), $\mu_\infty = \left(\sum_{j=1}^n \lambda_j F^{-1}_{\mu_{j,0}} \circ F_{\mu_{i,0}}\right)_\# \mu_{i,0} = \left(\sum_{j=1}^n \lambda_j F^{-1}_{\mu_{j,0}}\right)_\# \mathcal{L}$, with $\mathcal{L}$ being the Lebesgue measure on $[0,1]$; and*
- *for case (ii), $\mu_\infty = \mathcal{N}(m_\infty, \Sigma_\infty)$ with $m_\infty = \sum_{j=1}^n \lambda_j m_{j,0}$ and $\Sigma_\infty \in \mathbb{R}^{d\times d}$ being a positive definite matrix that satisfies $\Sigma_\infty = \sum_{j=1}^n \lambda_j (\Sigma_\infty^{1/2} \Sigma_{j,0} \Sigma_\infty^{1/2})^{1/2}$;*

with $\lambda = (\lambda_1, \ldots, \lambda_n)^\top$ being a random convex vector such that $\prod_{\tau=1}^\infty A(\tau) = \mathbb{1}_n \lambda^\top$ with probability one.

Moreover, all the previous results also hold for the symmetric PaWBar algorithm when G is a connected undirected graph, with the difference that the barycenters are now the standard one, i.e., with $\lambda = (\frac{1}{n}, \ldots, \frac{1}{n})^\top$ in the previous bullet points.

The work [8] proposes a non-distributed iterative algorithm tailored to compute the Wasserstein barycenter of Gaussian distributions. However, to the best of our knowledge, the PaWBar algorithm is the first algorithm that proposes a distributed computation of randomized and standard Gaussian barycenters with closed-form iteration steps, as indicated in the following proposition.

We remark that the distance interpolation between two Gaussian distributions is a curve of Gaussian distributions.

Proposition 4.3.9 (PaWBar algorithm for Gaussian distributions) *Assume any agent $i \in V$ has an initial distribution $\mu_{i,0} = \mathcal{N}(m_{i,0}, \Sigma_{i,0})$ with $m_{i,0} \in \mathbb{R}^d$ and $\Sigma_{i,0} \in \mathbb{R}^{d\times d}$ being a positive definite matrix. At any time t, let $m_i(t)$ and $\Sigma_i(t)$ be the mean and covariance matrix associated to agent $i \in V$.*

(i) For the directed PaWBar algorithm, if $(i,j) \in E$ is selected at time t, update the Gaussian distribution of agent i according to:

$$\begin{aligned} m_i(t+1) &:= (1-a_{ij})m_i(t) + a_{ij}m_j(t), \\ \Sigma_i(t+1) &:= (1-a_{ij})^2\Sigma_i(t) + a_{ij}^2\Sigma_j(t) \\ &\quad + a_{ij}(1-a_{ij})\Big((\Sigma_i(t)\Sigma_j(t))^{\frac{1}{2}} + (\Sigma_j(t)\Sigma_i(t))^{\frac{1}{2}}\Big). \end{aligned} \tag{4.9}$$

(ii) For the symmetric PaWBar algorithm, if $\{i,j\} \in E$ is selected at time t, update the Gaussian distributions of agents i and j according to:

$$\begin{aligned} m_i(t+1) &:= m_j(t+1) = \frac{1}{2}(m_i(t) + m_j(t)), \\ \Sigma_i(t+1) &:= \Sigma_j(t+1) = \frac{1}{4}\Big(\Sigma_i(t) + \Sigma_j(t) + (\Sigma_i(t)\Sigma_j(t))^{\frac{1}{2}} + (\Sigma_j(t)\Sigma_i(t))^{\frac{1}{2}}\Big). \end{aligned} \tag{4.10}$$

Then, the results stated in Corollary 4.3.8, under the conditions indicated therein, hold.

Proof: For statement (i), equation (4.9) results from the displacement interpolation between Gaussian distributions, as seen in [41], and thus implements equation (4.3). Case (ii) follows similarly. ■

Remark 4.3.10 (Examples of measures for Theorem 4.3.7) *We refer to [137, Section 2.3] and [27] for more examples of measures that form compatible collections and their statistical applications.*

Remark 4.3.11 (Further propertiesof the randomized Wasserstein barycenter) *The results in [161] can be applied to characterize the mean and covariance matrix associated to the random convex vector present in the randomized Wasserstein barycenter in Theorem 4.3.4, Corollary 4.3.5, Theorem 4.3.7, and Corollary 4.3.8. This characterization numerically depends on the values of the time-invariant probabilities associated with the edge selection process.*

We propose the following conjecture.

Conjecture 2 (Computation under more general continuous measures) *The convergence results of the PaWBar algorithm in Theorem 4.3.7 also hold for general absolutely continuous measures, i.e., ones which do not necessarily form a compatible collection.*

We provide some numerical evidence that Conjecture 2 is true at least for the case where all agents initially have multivariate Gaussian distributions that do not form a compatible collection. In the numerical evidence presented in Figure 4.1 and Figure 4.2, we use the updates presented in Proposition 4.3.9, and we only focus on the evolution of the agents' covariance matrices (the mean values evolve linearly and are easy to verify they converge to the mean of the Wasserstein barycenter). In this Gaussian setting, we also performed similar simulations to the ones in both figures but using the symmetric PaWBar algorithm with connected graphs; and we obtained convergence to the standard Wasserstein barycenter.

4.3.4 Analysis of general measures

So far, we presented convergence results to a Wasserstein barycenter for classes of discrete (Theorem 4.3.4 and Corollary 4.3.5) and absolutely continuous (Theorem 4.3.7 and Corollary 4.3.8) measures. In these cases an optimal transport map exists between any two agents' measures at every time. Now we analyze the PaWBar algorithm on general measures in $\mathcal{P}^2(\mathbb{R}^d)$. Examples of this setting include cases where there may not exist an optimal transport map between two or more initial measures, or cases where there could exist a mix of discrete and absolutely continuous initial measures. We prove that agents converge to consensus with probability one in more restricted graph topologies,

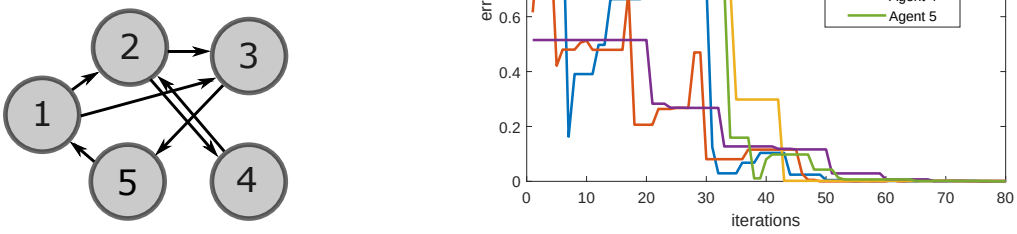


Figure 4.1: Consider five agents that initially have multivariate Gaussian distributions on $\mathbb{R}^5$, with their covariance matrices being randomly generated. On the left, we present the underlying digraph over which the PaWBar algorithm is run. The weight associated to all edges is 0.75. We first fix a realization of the edge selection process by fixing the seed of the random number generator in our scientific software. Then, we compute the covariance matrix Σ_∞ of the Wasserstein barycenter that would be obtained with the weights corresponding to the entries of any row of the product of evolution random matrices after a long period of time. The numerical computation of Σ_∞ follows the scheme proposed in [8]. On the right, each of the five plotted curves corresponds to the evolution of the error quantity $\|\Sigma_i(t) - \Sigma_\infty\|_F$ for each agent $i \in \{1, \dots, 5\}$, where $\Sigma_i(t)$ is the value of agent i's covariance matrix at iteration t, and $\|\cdot\|_F$ is the Frobenius norm. All agents asymptotically reach consensus and their covariance matrices become the covariance matrix of the randomized Wasserstein barycenter, thus giving evidence for the veracity of Conjecture 2 at least for the Gaussian case.

and that the consensus measure satisfies a necessary condition known to characterize barycenters for certain classes of measures.

Theorem 4.3.12 (Consensus result for general measures) *Consider the PaWBar algorithm with an underlying graph G which is either*

(i) a cycle graph for the directed case, or

(ii) a line graph for the symmetric case;

and with the agents having initial measures $\mu_{i,0} \in \mathcal{P}^2(\mathbb{R}^d)$, $i \in V$. Then, with probability

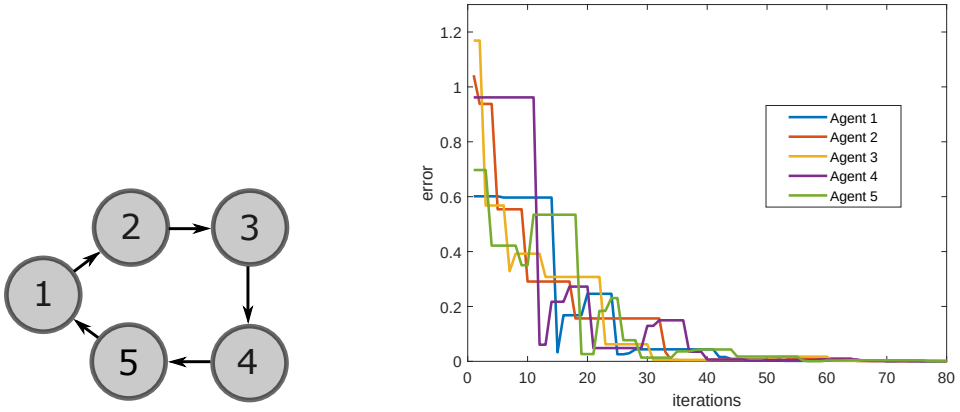


Figure 4.2: Consider five agents that initially have multivariate Gaussian distributions on $\mathbb{R}^5$, with their covariance matrices being randomly generated. The setting and methodology for computing the plot on the right is similar to the one described in Figure 4.1, with the difference that now the underlying digraph is a cycle (as seen on the left). All agents asymptotically reach consensus and their covariance matrices become the covariance matrix of the randomized Wasserstein barycenter.

one, for any $i \in V$,

$$W_2(\mu_i(t), \mu_\infty) \to 0 \text{ as } t \to \infty, \tag{4.11}$$

where $\mu_\infty \in \mathcal{P}^2(\mathbb{R}^d)$ *is a random measure whose possible values depend on the realization of the edge selection process. If* $\mu_{i,0} = \mu_{j,0}$ *for any* $i, j \in V$, *then* $\mu_\infty = \mu_{i,0}$ *with probability one.*

Moreover, for either the directed or symmetric case, and with probability one,

$$\operatorname{supp}(\mu_\infty) \subseteq \operatorname{cl}\left\{ \sum_{i=1}^n \lambda_i x_i \;\middle|\; x_i \in \operatorname{supp}(\mu_{i,0}), \lambda_i \geq 0, i \in V, \text{ and } \sum_{i=1}^n \lambda_i = 1 \right\}, \tag{4.12}$$

with cl *indicating the closure operation for sets.*

Remark 4.3.13 (Theorem 4.3.12 and Wasserstein barycenters) *Equation* (4.12) *is known to be satisfied when* μ_∞ *is a Wasserstein barycenter of measures that are absolutely continuous with respect to the Lebesgue measure, or discrete with finite support [6, 13]. However, the characterization of the consensus value in our theorem does*

not state any sufficient condition under which the converged consensus random measure is a Wasserstein barycenter or not: this is an open problem for further research.

4.4 Proofs of results in Section 4.3

4.4.1 Proofs of results in Subsection 4.3.2

Proof: [Proof of Theorem 4.3.4] Since measures $\mu_{i,0}$ and $\mu_{j,0}$, $i,j \in V$, are discrete uniform, we have $W_2^2(\mu_{i,0},\mu_{j,0}) = \min_{\sigma\in\Sigma_N} \frac{1}{N}\sum_{k=0}^{N} \left\| x_k^i - x_{\sigma(k)}^j \right\|_2^2$ [169, 146], with Σ_N being the set of all possible permutations of the elements in $\{1,\ldots,N\}$; i.e., any permutation map $\sigma \in \Sigma_N$ is a bijective function $\sigma : \{1,\ldots,N\} \to \{1,\ldots,N\}$. Then, the displacement interpolation between any of the initial measures provides discrete measures. Now, consider $\sigma \in \Sigma_N$ from solving the optimal transport problem from $\mu_{i,0}$ to $\mu_{j,0}$, i.e., $\gamma^{\text{opt}}(\mu_{i,0},\mu_{j,0}) = \frac{1}{N}\sum_{k=0}^{N} \delta_{(x_k^i, x_{\sigma(k)}^j)}$. Consider two arbitrary points $(x_{k_1}^i, x_{\sigma(k_1)}^j), (x_{k_2}^i, x_{\sigma(k_2)}^j) \in \text{supp}(\gamma^{\text{opt}}(\mu_{i,0},\mu_{j,0}))$. Then, for any $a_{ij} \in (0,1)$, the displacement interpolation implies the existence of some $z_{k_1}(a_{ij}), z_{k_2}(a_{ij}) \in \text{supp}((\pi_{a_{ij}})_{\#}\mu_{i,0})$, such that $z_{k_1}(a_{ij}) = (1-a_{ij})x_{k_1}^i + a_{ij}x_{\sigma(k_1)}^j$ and $z_{k_2}(a_{ij}) = (1-a_{ij})x_{k_2}^i + a_{ij}x_{\sigma(k_2)}^j$. Now, since the optimal transport plan $\gamma^{\text{opt}}(\mu_{i,0},\mu_{j,0})$ has cyclically monotone support [169, Section 2.3], we can follow the treatment in [170, Chapter 8] and conclude that there exists no $a_{ij} \in (0,1)$ such that $z^{k_1}(a_{ij}) = z^{k_2}(a_{ij})$. As a consequence, $\text{supp}((\pi_{a_{ij}})_{\#}\mu_{i,0})$ has N (different) elements for any possible edge weight a_{ij}, i.e., $(\pi_{a_{ij}})_{\#}\mu_{i,0}$ is a discrete uniform measure. It is easy to show by induction that in either the directed or symmetric PaWBar algorithm, $\mu_i(t)$ is a discrete uniform distribution for any $i \in V$ and time t with probability one.

We now introduce some notation. Given $A \in \mathbb{R}^{m\times m}$, let $\text{diag}^{i,k}(A) \in \mathbb{R}^{km\times km}$ be the $k \times k$ block-diagonal matrix such that its ith block has the matrix A and the rest of blocks are I_m. Given $A, B \in \mathbb{R}^{m\times m}$, let $\text{diag}^{ij,k}(A,B) \in \mathbb{R}^{km\times km}$ be the $k \times k$ block-

diagonal matrix such that its ith and jth blocks are the matrices A and B respectively, and the rest of blocks are I_m. Let $\mathbf{x}^i(t) \in \mathbb{R}^{Nd}$ be a vector stacking the elements of $\text{supp}(\mu_i(t))$, which we call the *support vector*. Note that since the measures are discrete uniform at every time t (with probability one), the order of the elements $x_k^i(t) \in \mathbb{R}^d$, $k \in \{1, \dots, N\}$, in the vector $\mathbf{x}^i(t)$ can be arbitrary; but for convenience we denote it as $\mathbf{x}^i(t) = (x_1^i(t), \cdots, x_N^i(t))^\top$. For any $i, j \in V$ and time t, let $\sigma_{ij,t} \in \Sigma_N$ be an optimal transport map from $\mu_i(t)$ to $\mu_j(t)$; and let $\sigma_{ii,t}(k) = k$ and $\sigma_{ji,t} = \sigma_{ij,t}^{-1}$.

We now focus on proving statement (i). Assume $(i, j) \in E$ is selected at time t. Then, $x_k^i(t+1) = (1 - a_{ij})x_k^i(t) + a_{ij}x^j_{\sigma_{ij,t}(k)}(t)$, $k \in \{1, \dots, N\}$, i.e.,

$$\mathbf{x}^i(t+1) = (1 - a_{ij})\mathbf{x}^i(t) + a_{ij}(P(t) \otimes I_d)\mathbf{x}^j(t) \tag{4.13}$$

with the permutation matrix $P(t) \otimes I_d$ defined by the permutation matrix $P(t) \in \{0, 1\}^{N \times N}$ whose kth row is $e^\top_{\sigma_{ij,t}(k)}$. Indeed, with $Q_{ij,t} = P(t) \otimes I_d$,

$$W_2^2(\mu_i(t), \mu_j(t)) = \frac{1}{N} \left\| \mathbf{x}^i(t) - Q_{ij,t}\mathbf{x}^j(t) \right\|_2^2.$$

We make the following claim:

(i.a) for any $i^*, j^* \in V$, $i^* \neq j^*$, $\epsilon > 0$ and time t, the event "$W_2(\mu_{i^*}(t+T), \mu_{j^*}(t+T)) < \epsilon$ for some finite $T > 0$" has positive probability.

We prove the claim. Define $d(i, j, \sigma^i, \sigma^j, t_i, t_j) := \left(\sum_{k=1}^N \frac{1}{N} \left\| x^i_{\sigma^i(k)}(t_i) - x^j_{\sigma^j(k)}(t_j) \right\|_2^2 \right)^{\frac{1}{2}}$, for $i, j \in V$, $\sigma^i, \sigma^j \in \Sigma_N$. Consider any $i^*, j^* \in V$ and $\epsilon > 0$. Since G is strongly connected, there exists a shortest directed path $\mathcal{P}_{i^* \to j^*}$ from i^* to j^* of some length L. Let $\mathcal{P}_{i^* \to j^*} = ((i^*, \ell_1), \dots, (\ell_{L-1}, j^*))$. Now, pick positive numbers $\epsilon_1, \dots, \epsilon_L$ such that $\sum_{i=1}^L \epsilon_i < \epsilon$. Consider any time t. Then, we can first select T_1 times the edge (ℓ_{L-1}, j^*)

so that

$$d(\ell_{L-1}, j^*, \sigma^{-1}_{\ell_{L-1}j^*,t+T_1}, Id, t+T_1, t) = (1-a_{\ell_{L-1}j^*})^{T_1} d(\ell_{L-1}, j^*, \sigma^{-1}_{\ell_{L-1}j^*,t}, Id, t, t) < \epsilon_L.$$

Then, we can select T_2 times the edge (ℓ_{L-2}, ℓ_{L-1}) so that

$$d(\ell_{L-2}, \ell_{L-1}, \sigma^{-1}_{\ell_{L-2}\ell_{L-1},t+T_1+T_2} \circ \sigma^{-1}_{\ell_{L-1}j^*,t+T_1}, \sigma^{-1}_{\ell_{L-1}j^*,t+T_1}, t+T_1+T_2, t+T_1) < \epsilon_{L-1},$$

and we can continue like this until finally selecting T_L times the edge (i^*, ℓ_1) such that $d(i^*, \ell_1, \sigma, \sigma^{-1}_{\ell_1\ell_2,t+\sum_{i=0}^{L-1} T_i} \circ \cdots \circ \sigma^{-1}_{\ell_{L-1}j^*,t}, t+T, t+\sum_{i=1}^{L-1} T_i) < \epsilon_1$ with $\sigma = \sigma^{-1}_{i^*\ell_1,t+T} \circ \cdots \circ \sigma^{-1}_{\ell_{L-1}j^*,t}$ and $T = \sum_{i=1}^{L} T_i$. Then,

$$W_2(\mu_{i^*}(t+T), \mu_{j^*}(t+T)) \leq (\sum_{k=1}^{N} \frac{1}{N} \left\| x^{i^*}_{\sigma(k)}(t+T) - x^{j^*}_k(t+T) \right\|_2^2)^{\frac{1}{2}} < \sum_{i=1}^{L} \epsilon_i < \epsilon.$$

where the first inequality follows by definition of the Wasserstein distance, and the second inequality from both the triangle inequality and the fact that $x^{j^*}_k(t+T) = x^{j^*}_k(t)$, $x^{\ell_{L-1}}_k(t+T) = x^{\ell_{L-1}}_k(t+T_1)$, $\ldots$, $x^{\ell_1}_k(t+T) = x^{\ell_1}_k(t+\sum_{i=1}^{L-1} T_i)$. Moreover, our construction implies

$$W_2(\mu_p(t+T), \mu_{j^*}(t+T)) < \epsilon \text{ for any } p \in \mathcal{P}_{i^* \to j^*}. \tag{4.14}$$

Now, consider any $m \in V$ and construct a directed acyclic subgraph $G' = (V, E')$, $E' \subset E$, of G as follows: m is the unique node with zero out-degree (i.e, $(m, i) \notin E'$ for any $i \in V$) and there exists a unique directed path from any node $i \in V \setminus \{m\}$ to m. Such subgraph G' exists because G is strongly connected. Consider any $\epsilon > 0$ and time t. Then, the selection process just described above can make all nodes $\bar{m}$ with zero in-degree in G' (i.e., any $\bar{m} \in V$ such that $(i, \bar{m}) \notin E'$ for any $i \in V \setminus \{\bar{m}\}$) satisfy $W_2(\mu_{\bar{m}}(t+T), \mu_m(t+T)) = W_2(\mu_{\bar{m}}(t+T), \mu_m(t)) < \frac{\epsilon}{2}$ for some T. Then, as a consequence

of (4.14), $W_2(\mu_i(t+T), \mu_j(t+T)) < \frac{\epsilon}{2}$ for any $i \in V$, and the triangle inequality then implies $W_2(\mu_i(t+T), \mu_j(t+T)) < \epsilon$ for any $j \in V$. Finally, for any $i, j \in V$, the event "$W_2(\mu_i(t+T), \mu_j(t+T)) < \epsilon$ for some $T > 0$" has a positive probability to occur at any time t because any selection of a finite sequence of edges has positive probability to occur at any time t. This finishes the proof of claim (i.a).

Now, note that the event in result (i.a), due to its persistence, will eventually happen with probability one. Assume it happens at time t. Then, we claim that ϵ in this event could have been chosen so that, for any time $t' \geq t$ and any $i, j, p \in V$,

(i.b) $\sigma_{ij,t'} = \sigma_{ij,t}$, and

(i.c) $\sigma_{ij,t'} = \sigma_{ip,t'} \circ \sigma_{pj,t'}$.

Now we prove the claim. Firstly, note that from (i.a) and the fact that the measures are discrete uniform at every time with probability one, we can consider a small enough ϵ such that for any $i, j \in V$ and any permutation map $\sigma \neq \sigma_{ij,t}$,

$$\begin{aligned} &W_2(\mu_i(t), \mu_j(t)) = \Big(\frac{1}{N}\sum_{k=1}^{N} \Big\|x_k^i(t) - x^j_{\sigma_{ij,t}(k)}(t)\Big\|_2^2\Big)^{\frac{1}{2}} < \epsilon \quad \text{and} \\ &2\epsilon < \Big(\frac{1}{N}\sum_{k=1}^{N} \Big\|x_k^i(t) - x^j_{\sigma(k)}(t)\Big\|_2^2\Big)^{\frac{1}{2}}. \end{aligned} \tag{4.15}$$

Such choice of ϵ implies that $\sigma_{ip,t} \circ \sigma_{pj,t} = \sigma_{ij,t}$ for any $i, j, p \in V$; otherwise, if $\sigma_{ip,t} \circ \sigma_{pj,t} \neq \sigma_{ij,t}$, then we obtain a contradiction:

$$\begin{aligned} 2\epsilon &< \Big(\frac{1}{N}\sum_{k=1}^{N} \Big\|x_k^i(t) - x^j_{\sigma_{ip,t}\circ\sigma_{pj,t}(k)}(t)\Big\|_2^2\Big)^{\frac{1}{2}} = \Big(\frac{1}{N}\sum_{k=1}^{N} \Big\|x^i_{\sigma_{pi,t}(k)}(t) - x^j_{\sigma_{pj,t}(k)}(t)\Big\|_2^2\Big)^{\frac{1}{2}} \\ &\leq \Big(\frac{1}{N}\sum_{k=1}^{N} \Big\|x^i_{\sigma_{pi,t}(k)}(t) - x^p_k(t)\Big\|_2^2\Big)^{\frac{1}{2}} + \Big(\frac{1}{N}\sum_{k=1}^{N} \Big\|x^p_k(t) - x^j_{\sigma_{pj,t}(k)}(t)\Big\|_2^2\Big)^{\frac{1}{2}} < 2\epsilon. \end{aligned}$$

We just proved that (i.c) holds for $t' = t$. Note that (i.b) for $t' = t$ is trivial. Now,

assume any $(i^*, j^*) \in E$ is selected at time t. Then, for any $j \in V \setminus \{i^*, j^*\}$, using the identity $Q_{j^*j,t} = Q_{j^*i^*,t}Q_{i^*j,t}$ from (i.c) for $t' = t$ implies

$$\begin{aligned}
&\left\|\mathbf{x}^{i^*}(t+1) - Q_{i^*j,t}\mathbf{x}^j(t+1)\right\|_2 \qquad (4.16)\\
&\quad\leq (1-a_{i^*j^*})\left\|\mathbf{x}^{i^*}(t) - Q_{i^*j,t}\mathbf{x}^j(t)\right\|_2 + a_{i^*j^*}\left\|Q_{i^*j^*,t}\mathbf{x}^{j^*}(t) - Q_{i^*j,t}\mathbf{x}^j(t)\right\|_2\\
&\quad= (1-a_{i^*j^*})\left\|\mathbf{x}^{i^*}(t) - Q_{i^*j,t}\mathbf{x}^j(t)\right\|_2 + a_{i^*j^*}\left\|\mathbf{x}^{j^*}(t) - Q_{j^*j,t}\mathbf{x}^j(t)\right\|_2\\
&\quad< (1-a_{i^*j^*})\epsilon\sqrt{N} + a_{i^*j^*}\epsilon\sqrt{N} = \epsilon\sqrt{N};
\end{aligned}$$

likewise, we immediately obtain $\frac{1}{\sqrt{N}}\left\|\mathbf{x}^i(t+1) - Q_{ij,t}\mathbf{x}^j(t+1)\right\|_2 < \epsilon$ for any $i \in V \setminus \{i^*, j\}$, and $\frac{1}{\sqrt{N}}\left\|\mathbf{x}^{i^*}(t+1) - Q_{i^*j^*,t}\mathbf{x}^{j^*}(t+1)\right\|_2 < (1-a_{i^*j^*})\epsilon < \epsilon$. In summary, $\frac{1}{\sqrt{N}}\left\|\mathbf{x}^i(t+1) - Q_{ij,t}\mathbf{x}^j(t+1)\right\|_2 < \epsilon$ for any $i, j \in V$, which implies $\sigma_{ij,t+1} = \sigma_{ij,t}$ for any $i, j \in V$; i.e., (i.b) holds for $t' = t+1$. Now, to prove claim (i.c) holds for $t' = t+1$, we must first prove that (4.15) holds for time $t+1$.

Set $\mathbf{y}^1(t) := \mathbf{x}^1(t), \mathbf{y}^2(t) = Q_{12,t}\mathbf{x}^2(t), \ldots, \mathbf{y}^n(t) = Q_{1n,t}\mathbf{x}^n(t)$ (this labeling is arbitrary and any other $i \in V \setminus \{1\}$ could have been chosen to define $Q_{i1}, \ldots, Q_{in}$) and $\mathbf{y}^i(t) = (y_1^i(t), \ldots, y_N^i(t))^\top$, $y_1^i(t) \in \mathbb{R}^d$, $i \in \{1, \ldots, n\}$. For any $k \in \{1, \ldots, N\}$, let $\mathcal{L}_k(t)$ be the convex hull of the set $\{y_k^1(t), \ldots, y_k^n(t)\}$. For any $p, q \in \{1, \ldots, N\}$, define the distance between $\mathcal{L}_p(t)$ and $\mathcal{L}_q(t)$ as $d_{pq}(t) = \inf_{w_1 \in \mathcal{L}_p(t), w_2 \in \mathcal{L}_q(t)} \|w_1 - w_2\|_2$. Assuming that $(i^*, j^*) \in E$ is selected at time t, our result (i.b) for $t' = t+1$ and (4.13) imply that $y_k^{i^*}(t+1) = (1-a_{i^*j^*})y_k^{i^*}(t) + a_{i^*j^*}y_k^{j^*} \in \mathcal{L}_k(t)$, $k \in \{1, \ldots, N\}$. Obviously, for any $j \in V \setminus \{i^*\}$, $y_k^j(t+1) = y_k^j(t) \in \mathcal{L}_k(t)$, $k \in \{1, \ldots, N\}$. Then, $\mathcal{L}_i(t+1) \subseteq \mathcal{L}_i(t)$ for any $i \in \{1, \ldots, N\}$, and thus $d_{pq}(t) \leq d_{pq}(t+1)$ for any $p, q \in \{1, \ldots, N\}$. Now, for any $i, j \in V$, time $\tau \geq t$, and permutation map $\sigma \neq Id$, we have $d_{k\sigma(k)}(\tau) \leq \left\|y_k^i(\tau) - y_{\sigma(k)}^j(\tau)\right\|_2 \Longrightarrow (\sum_{k=1}^N d_{k\sigma(k)}^2(\tau))^{\frac{1}{2}} \leq (\sum_{k=1}^N \left\|y_k^i(\tau) - y_{\sigma(k)}^j(\tau)\right\|_2^2)^{\frac{1}{2}}$. Now, we need to consider two cases. In the first case we consider

$2\epsilon < \min_{\bar{\sigma}\in\Sigma_N,\bar{\sigma}\neq Id}(\frac{1}{N}\sum_{k=1}^{N} d^2_{k\bar{\sigma}(k)}(t))^{\frac{1}{2}}$. Then,

$$2\epsilon < (\frac{1}{N}\sum_{k=1}^{N} d^2_{k\sigma(k)}(t+1))^{\frac{1}{2}} \leq (\frac{1}{N}\sum_{k=1}^{N} \left\| y^i_k(t+1) - y^j_{\sigma(k)}(t+1)\right\|^2_2)^{\frac{1}{2}}$$

for any permutation map $\sigma \neq Id$; and (i.b) for $t' = t+1$ and (i.c) for $t' = t$ imply $2\epsilon < (\frac{1}{N}\sum_{k=1}^{N} \left\| x^i_k(t+1) - x^j_{\sigma'(k)}(t+1)\right\|^2_2)^{\frac{1}{2}}$ with $\sigma' = \sigma_{i1,t+1}\circ\sigma\circ\sigma_{1j,t+1} \neq \sigma_{ij,t+1}$. Thus, (4.15) holds for time $t+1$ in this first case. Now, we consider the second case $2\epsilon \geq \min_{\bar{\sigma}\in\Sigma_N,\bar{\sigma}\neq Id}(\frac{1}{N}\sum_{k=1}^{N} d^2_{k\bar{\sigma}(k)}(t))^{\frac{1}{2}}$. Then, due to G being strongly connected and $\{d_{pq}(\tau)\}_{\tau\geq t}$ being a nondecreasing sequence for any $p,q \in \{1,\ldots,N\}$, we can follow the proof of result (i.a) and arbitrarily reduce the diameter of the set $\mathcal{L}_k$ for any $k \in \{1,\ldots,N\}$ at some future time $\bar{t}$, i.e., $\mathcal{L}(\bar{t}) \subset \mathcal{L}(t)$. This diameter reduction can be chosen such that $d_{ij}(\bar{t}) > d_{ij}(t)$ for any $i,j \in \{1,\ldots,N\}$, and this increase on the distances between sets can be done so that $2\epsilon' < \min_{\bar{\sigma}\in\Sigma_N,\bar{\sigma}\neq Id}(\frac{1}{N}\sum_{k=1}^{N} d^2_{k\bar{\sigma}(k)}(\bar{t}))^{\frac{1}{2}}$ for some $0 < \epsilon' < \epsilon$. In other words, we are in the first case at time $\bar{t}$. After this change, we will never be in the second case again for any time after $\bar{t}$ with probability one. In summary, we just proved the conditions in equation (4.15) can be made to hold for time $t+1$, and so (i.c) holds for $t' = t+1$.

Now, assume results (i.b) and (i.c) hold for time $t' = \tau \geq t$, and (4.15) holds for time τ. Following the proof just presented above, we easily establish that (i.b) and (i.c) hold for $t' = \tau+1$ and that (4.15) holds for time $\tau+1$. Then, by induction, we proved our initial claim about (i.b) and (i.c).

Now, set $\mathbf{x}(t) = (\mathbf{x}^1(t),\ldots,\mathbf{x}^n(t))^\top \in \mathbb{R}^{nNd}$. Assume $(i,j) \in E$ is selected at any time $t \geq 0$. Then, (4.13) becomes

$$\mathbf{x}(t+1) = B(t)\mathbf{x}(t) \tag{4.17}$$

with the row-stochastic matrix $B(t) = \mathrm{diag}^{j,n}(P(t)^\top \otimes I_d)(A(t)\otimes I_{Nd})\,\mathrm{diag}^{j,n}(P(t)\otimes I_d)$

(recall the permutation matrix $P(t)$ was defined after (4.13)).

Consider an initial vector $\mathbf{x}(0)$ and a fixed realization of the edge selection process. Now consider $\mathbf{x}'(0) = \text{diag}(P_1 \otimes I_d, \ldots, P_n \otimes I_d)\mathbf{x}(0)$ with arbitrary permutation matrices $P_1, \ldots, P_n \in \{0,1\}^{N\times N}$. Notice that both $\mathbf{x}(0)$ and $\mathbf{x}'(0)$ represent the supports of the same group of measures $\{\mu_{i,0}\}_{i\in V}$ but may be the case that $\mathbf{x}(0) \neq \mathbf{x}'(0)$. We claim that

$$\mathbf{x}'(t) = \text{diag}(P_1 \otimes I_d, \ldots, P_n \otimes I_d)\mathbf{x}(t) \text{ for any time } t. \tag{4.18}$$

To prove this claim, first recall that we have a fixed realization of the edge selection process. Assume $(i,j) \in E$ is selected at time $t = 0$ and obtain $\mathbf{x}(1) = B(0)\mathbf{x}(0)$. Likewise, $\mathbf{x}'(1) = B'(0)\mathbf{x}'(0)$, with $B'(0) = \text{diag}^{j,n}(P'(0)^\top \otimes I_d)(A(0)\otimes I_{Nd})\,\text{diag}^{j,n}(P'(0)\otimes I_d)$, is the update that results if the algorithm starts with initial vector $\mathbf{x}'(0)$. Then,

$$P'(0) \otimes I_d = (P_i \otimes I_d)(P(0) \otimes I_d)(P_j^\top \otimes I_d) = (P_i P(0) P_j^\top) \otimes I_d.$$

After some algebraic work, we obtain $B'(0) = \text{diag}^{ij,n}(P_i, P_j)B(0)\,\text{diag}^{ij,n}(P_i^\top, P_j^\top)$. Then,

$$\begin{aligned}
\mathbf{x}'(1) &= \text{diag}^{ij,n}(P_i, P_j)B(0)\,\text{diag}^{ij,n}(P_i^\top, P_j^\top)\mathbf{x}'(0) \\
&= \text{diag}^{ij,n}(P_i, P_j)B(0)\,\text{diag}^{ij,n}(P_i^\top, P_j^\top)\,\text{diag}(P_1, \ldots, P_n)\mathbf{x}(0) \\
&= \text{diag}(P_1, \ldots, P_n)B(0)\mathbf{x}(0) = \text{diag}(P_1, \ldots, P_n)\mathbf{x}(1).
\end{aligned}$$

Finally (4.18) is easily proved by induction, and the claim is proved.

Now, from results (i.a), (i.b) and (i.c), there exists some random time $T > 0$ such that, with probability one: for any time $t \geq T$ and any $i, j, p \in V$, $\sigma_{ij,t} = \sigma_{ij,T}$ and $\sigma_{ij,t} = \sigma_{ip,t} \circ \sigma_{pj,t}$. Consider such time T, which is now a deterministic function of $\mathbf{x}(0)$ since we have a fixed realization of the edge selection process. Without loss of generality,

as a consequence of (4.18), we can assume we started the algorithm with the initial support vectors $\{Q_{1i,T}\mathbf{x}^i(0)\}_{i\in V}$ at time $t = 0$. Then, it is easy to prove that $B(t) = A(t) \otimes I_{Nd}$ for any $t \geq T$, i.e., $B(t)$ has an associated permutation matrix $P(t) = I_N$. Then, Proposition 4.3.3 let us conclude that $\lim_{t\to\infty} \prod_{\tau=T}^{t} B(\tau) = (\mathbb{1}_n \lambda^\top) \otimes I_{Nd}$ for some convex vector λ. Thus, $\mathbf{x}^i(\infty) = \sum_{j=1}^{n} \lambda_j \mathbf{x}^j(T)$, $i \in V$.

It remains to prove that μ_∞ corresponds to a Wasserstein barycenter. Let us formulate the Wasserstein barycenter problem $\min_{\nu\in\mathcal{P}^2(\mathbb{R}^d)} \sum_{i=1}^{n} \lambda_i W_2(\nu, \mu_i(T))^2$. Since the measures $\{\mu_i(T)\}_{i\in V}$ have finite support, any barycenter is a discrete measure with finite support [13]. Moreover, since all the measures are uniform, we can consider a minimizer with a discrete uniform distribution. We now prove that μ_∞ is such a minimizer. Firstly, by construction and the fact that (4.15) holds for $t \geq T$, we have

$$\begin{aligned} W_2(\mu_\infty, \mu_i(T))^2 &= \frac{1}{N} \sum_{j=1}^{N} \left\| \sum_{k=1}^{n} \lambda_k x_j^k(T) - x_j^i(T) \right\|_2^2 \\ &< 2\epsilon < \frac{1}{N} \sum_{j=1}^{N} \left\| \sum_{k=1}^{n} \lambda_k x_{\sigma^k(j)}^k(T) - x_{\sigma^i(j)}^i(T) \right\|_2^2 \end{aligned}$$

for any $\sigma^i \in \Sigma_N$, $\sigma_i \neq Id$, $i \in V$. Then,

$$\begin{aligned} \sum_{i=1}^{n} \lambda_i W_2(\mu_\infty, \mu_i(T))^2 &< \frac{1}{N} \sum_{i=1}^{n} \lambda_i \sum_{j=1}^{N} \left\| \sum_{k=1}^{n} \lambda_k x_{\sigma^k(j)}^k(T) - x_{\sigma^i(j)}^i(T) \right\|_2^2 \\ &\leq \frac{1}{N} \sum_{i=1}^{n} \lambda_i \sum_{j=1}^{N} \left\| y_i - x_{\sigma^i(j)}^i(T) \right\|_2^2 \end{aligned}$$

for any $y = (y_1, \ldots, y_n)^\top \in \mathbb{R}^{Nd}$. The last inequality of the previous expression is proved by using first optimality conditions to minimize the differentiable and strictly convex functional of y (i.e., set the gradient with respect to y equal to the zero vector). Now, take $y_i \neq y_j \in \mathbb{R}^d$ for any $i, j \in \{1, \ldots, N\}$, define the discrete uniform

measure $\nu = \frac{1}{N}\sum_{j=1}^{N}\delta_{y_j}$, $y_j \in \mathbb{R}^d$; and let $\bar{\sigma}^i$, $i \in V$, be such that $W_2(\nu, \mu_i(T))^2 = \frac{1}{N}\sum_{i=1}^{n}\lambda_i \sum_{j=1}^{N}\left\|y_i - x^i_{\bar{\sigma}^i(j)}(T)\right\|_2^2$. Then, our recent analysis implies that: (1) if there exists $i \in V$ such that $\bar{\sigma}_i \neq Id$, then $\sum_{i=1}^{n}\lambda_i W_2(\mu_\infty, \mu_i(T))^2 < \sum_{i=1}^{n}\lambda_i W_2(\nu, \mu_i(T))^2$; (2) if $\bar{\sigma}_i = Id$ for all $i \in V$, then $\sum_{i=1}^{n}\lambda_i W_2(\mu_\infty, \mu_i(T))^2 \leq \sum_{i=1}^{n}\lambda_i W_2(\nu, \mu_i(T))^2$. Given the generality of ν, cases (1) and (2) together imply that μ_∞ is a Wasserstein barycenter.

Finally, all of our previous results hold with probability one because we considered an arbitrary realization of the edge selection process for our analysis (note that λ now becomes a random convex vector). This concludes the proof of statement (i).

We now focus on proving statement (ii). Assume $\{i, j\} \in E$ is selected at time t. Without loss of generality, the update of the PaWBar algorithm can be set as $x^i_k(t+1) = \frac{1}{2}x^i_k(t) + \frac{1}{2}x^j_{\sigma_{ij,t}(k)}(t)$ and $x^j_k(t+1) = \frac{1}{2}x^j_k(t) + \frac{1}{2}x^i_{\sigma_{ji,t}(k)}(t)$, $k \in \{1, \ldots, N\}$; i.e.,

$$\begin{aligned}\mathbf{x}^i(t+1) &= \frac{1}{2}\mathbf{x}^i(t) + \frac{1}{2}(P(t)\otimes I_d)\mathbf{x}^j(t),\\ \mathbf{x}^j(t+1) &= \frac{1}{2}\mathbf{x}^j(t) + \frac{1}{2}(P(t)^\top \otimes I_d)\mathbf{x}^i(t)\end{aligned} \tag{4.19}$$

recalling that the permutation matrix $P(t) \in \{0,1\}^{N\times N}$ has $e^\top_{\sigma_{ij,t}(k)}$ as its kth row.

We make the following claim:

(ii.a) for any $i^*, j^* \in V$, $i^* \neq j^*$, $\epsilon > 0$ and time t, the event "$W_2(\mu_{i^*}(t+T), \mu_{j^*}(t+T)) < \epsilon$ for some finite $T > 0$" has positive probability.

Now, we prove the claim. Let us fix a spanning tree G' of G. For any $i, j \in V$, let $\mathcal{P}_{i-j}$ denote the unique path between i and j in G'. Let

$$U(t) = \max_{i,j\in V}\sum_{\{p,q\}\in\mathcal{P}_{i-j}} W_2(\mu_p(t), \mu_q(t)).$$

Let $\{k, \ell\} \in \arg U(t)$ and $\mathcal{P}_{k-l} = (\{k, p_1\}, \ldots, \{p_{L-1}, \ell\})$, i.e., edge $\{k, p_1\}$ is followed by

$\{p_1, p_2\}$ and so on until $\{p_{L-1}, \ell\}$. *Case 1)* $W_2(\mu_k(t), \mu_{p_1}(t)) \neq 0$. For simplicity we also assume $W_2(\mu_i(t), \mu_j(t)) \neq 0$ for any $\{i, j\} \in \mathcal{P}_{k-\ell}$; otherwise, if there exists $\{i^*, j^*\} \in \mathcal{P}_{k-\ell}$ such that $W_2(\mu_{i^*}(t), \mu_{j^*}(t)) = 0$, we would need to use a similar analysis to Case 2) which will be treated later. Select $\{k, p_1\}$ at time t. If $\mathcal{P}_{k-\ell}$ contains only one element, then $p_1 = \ell$ and $W_2(\mu_k(t+1), \mu_\ell(t+1)) = 0 < U(t) = W_2(\mu_k(t), \mu_\ell(t))$. Now, consider $\mathcal{P}_{k-\ell}$ contains two or more elements. Set $\mathcal{U}(t) = \sum_{\{p,q\} \in \mathcal{P}_{k-\ell} \setminus \{\{k,p_1\},\{p_1,p_2\}\}} W_2(\mu_p(t), \mu_q(t))$ (with $p_2 = \ell$ and $\mathcal{U}(t) = 0$ if $\mathcal{P}_{k-\ell}$ only has two elements). Then

$$
\begin{aligned}
\sum_{\{p,q\} \in \mathcal{P}_{k-\ell}} W_2(\mu_p(t+1), \mu_q(t+1)) &= \mathcal{U}(t) + \frac{1}{\sqrt{N}} \left\| \mathbf{x}^{p_1}(t+1) - Q_{p_1 p_2, t+1} \mathbf{x}^{p_2}(t) \right\|_2 \\
&\leq \mathcal{U}(t) + \frac{1}{\sqrt{N}} \left\| \mathbf{x}^{p_1}(t+1) - Q_{p_1 p_2, t} \mathbf{x}^{p_2}(t) \right\|_2 \\
&\leq \mathcal{U}(t) + \frac{1}{2} W_2(\mu_{p_1}(t), \mu_{p_2}(t)) + \frac{1}{2\sqrt{N}} \left\| Q_{p_1 k, t} \mathbf{x}^k(t) - Q_{p_1 p_2, t} \mathbf{x}^{p_2}(t) \right\|_2 \\
&\leq \mathcal{U}(t) + \frac{1}{2} W_2(\mu_{p_1}(t), \mu_{p_2}(t)) + \frac{1}{2\sqrt{N}} \left\| Q_{p_1 k, t} \mathbf{x}^k(t) - \mathbf{x}^{p_1}(t) \right\|_2 \\
&\quad + \frac{1}{2\sqrt{N}} \left\| \mathbf{x}^{p_1}(t) - Q_{p_1 p_2, t} \mathbf{x}^{p_2}(t) \right\|_2 \\
&= \mathcal{U}(t) + W_2(\mu_{p_1}(t), \mu_{p_2}(t)) + \frac{1}{2} W_2(\mu_k(t), \mu_{p_1}(t)),
\end{aligned}
$$

and so $\sum_{\{p,q\} \in \mathcal{P}_{k-\ell}} W_2(\mu_p(t+1), \mu_q(t+1)) < U(t)$. Therefore, for any length of $\mathcal{P}_{k-\ell}$, if $U(t+1) \leq \sum_{\{p,q\} \in \mathcal{P}_{k-\ell}} W_2(\mu_p(t+1), \mu_q(t+1))$, then $U(t+1) < U(t)$. If $U(t+1) > \sum_{\{p,q\} \in \mathcal{P}_{k-\ell}} W_2(\mu_p(t+1), \mu_q(t+1))$, then we can choose $\{\bar{k}, \bar{\ell}\} \in \arg U(t+1)$ and, using the analysis just presented, obtain $\sum_{\{p,q\} \in \mathcal{P}_{\bar{k}-\bar{\ell}}} W_2(\mu_p(t+2), \mu_q(t+2)) < U(t+1)$. If this does not imply $U(t+2) < U(t)$, we can keep iterating this procedure until, eventually, obtain $U(t+T) < U(t)$ for some $T > 0$. *Case 2)* $W_2(\mu_k(t), \mu_{p_1}(t)) = 0$. In this case, we do not select the edge $\{k, p_1\}$, but we consecutively check the edges along $\mathcal{P}_{k-\ell}$ starting from $\{k, p_1\}$ and look for the first $\{i^*, j^*\} \in \mathcal{P}_{k-\ell}$ such that $W_2(\mu_{i^*}(t), \mu_{j^*}(t)) \neq 0$. We select this edge and a similar analysis to Case 1) implies that $\sum_{\{p,q\} \in \mathcal{P}_{k \to \ell}} W_2(\mu_p(t+1), \mu_q(t+$

1)) $\leq U(t)$. Then, we select the edge previous to $\{i^*, j^*\}$ and continue to successively select the preceding edges until reaching the first edge $\{k, p_1\}$. Once this edge is selected, say at time $\bar{t}$, the proof of case Case 1) let us conclude that $\sum_{\{p,q\}\in\mathcal{P}_{k-\ell}} W_2(\mu_p(\bar{t}+1), \mu_q(\bar{t}+1)) < U(\bar{t})$, and we can continue the analysis of Case 1) until we have that $U(t+T) < U(\bar{t}) \leq U(t)$ for some $T > 0$. In conclusion, we proved the existence of some finite sequence of selected edges such that $U(t+T) < U(t)$ for some $T > 0$. Moreover, we can iterate selections of such sequences to arbitrarily reduce the value of $U(t)$ after some finite time. Finally, claim (ii.a) follows from the fact that $\max_{i,j\in V} W_2(\mu_i(t), \mu_j(t)) \leq U(t)$ and that any finite sequence of edges has a positive probability of being consecutively selected at any time t.

We can now follow the same analysis as in the proof of statement (i) of the theorem (using result (ii.a) and its proof instead of (i.a)) to conclude that results (i.b) and (i.c) also hold for the symmetric PaWBar algorithm.

Now, assume $\{i, j\} \in E$ is selected at time $t \geq 0$. Then, (4.19) becomes

$$\mathbf{x}(t+1) = C(t)\mathbf{x}(t), \tag{4.20}$$

with the matrix $C(t) = \operatorname{diag}^{i,n}(P(t) \otimes I_d)(A(t) \otimes I_{Nd}) \operatorname{diag}^{i,n}(P^\top(t) \otimes I_d)$.

We fix a realization of the edge selection process. Consider any initial vector $\mathbf{x}(0)$ and $\mathbf{x}(0)' = \operatorname{diag}(P_1 \otimes I_d, \ldots, P_n \otimes I_d)\mathbf{x}(0)$ with arbitrary permutation matrices $P_1, \ldots, P_n \in \{0,1\}^{N\times N}$. As in the proof of statement (i), it is possible to prove that $\mathbf{x}'(t) = \operatorname{diag}(P_1 \otimes I_d, \ldots, P_n \otimes I_d)\mathbf{x}(t)$ for any t after some algebraic work. From here, we can closely follow the proof of statement (i) to conclude the proof for statement (ii). ■

Proof: [Proof of Corollary 4.3.5] We follow the notation and proof of Theorem 4.3.4. Note that the entries of $\mathbf{x}^i(0) = (x_1^i, \ldots, x_N^i)^\top$, $i \in V$, are sorted in ascending order. Then, $W_2^2(\mu_{i,0}, \mu_{j,0}) = \frac{1}{N}\sum_{k=0}^{N}(x_k^i - x_k^j)^2$ for $i, j \in V$. Now, consider the directed PaWBar

algorithm and that $(i,j) \in E$ is selected at time $t = 0$. Then $\mathbf{x}^i(1) = (1 - a_{ij})\mathbf{x}^i(0) + a_{ij}\mathbf{x}^j(0)$ and $\mathbf{x}^i(1)$ has its entries sorted in ascending order. Then, it is easy to prove by induction that, at every time t, $\mathbf{x}^i(t)$ for any $i \in V$ is sorted in ascending order with probability one. Considering $\mathbf{x}(t) = (\mathbf{x}^i(t), \ldots, \mathbf{x}^n(t))^\top \in \mathbb{R}^{nN}$, we have that $\mathbf{x}(t+1) = (A(t) \otimes I_N)\mathbf{x}(t)$ and so $\mathbf{x}(t) = (\prod_{i=0}^{t} A(i) \otimes \mathbb{1}_N)\mathbf{x}(0)$. Then, we conclude the proof for the directed PaWBar algorithm by using Proposition 4.3.3 and the fact that $\sum_{i=1}^n \lambda_i \mathbf{x}^i(0) \in \arg\min_{\substack{y \in \mathbb{R}^d \\ y_i < \cdots < y_n}} \frac{1}{N} \sum_{i=1}^n \lambda_i \sum_{k=0}^N (y_k - x_k^i)^2$ for any convex vector $\lambda \in \mathbb{R}^n$. The symmetric case is proved similarly. ■

4.4.2 Proofs of results in Subsection 4.3.3

Proof: [Proof of Theorem 4.3.7] Consider the following notation: for any $i, j, k \in V$, denote $(T_{\alpha_2}^{\alpha_3} \circ T_{\alpha_1}^{\alpha_2})_{\#}\alpha_1 = (T_{\alpha_1}^{\alpha_3})_{\#}\alpha_1$ by $T_{\alpha_2}^{\alpha_3} \circ T_{\alpha_1}^{\alpha_2} = T_{\alpha_1}^{\alpha_3}$ for absolutely continuous measures $\alpha_1, \alpha_2, \alpha_3 \in \mathcal{P}^2(\mathbb{R}^d)$.

First of all, it is known that a displacement interpolation between any of the absolutely continuous initial measures results in a curve of absolutely continuous measures [151]. Then, with probability one, $\mu_i(t)$ is absolutely continuous for every $i \in V$ and time t. This also implies all measures up to any time form a compatible collection with probability one. To see this, fix any $\gamma \in \{\mu_{i,0}\}_{i \in V}$ and assume any $(i,j) \in E$ is selected at time zero. Clearly, $\mu_i(1) = (T_\gamma^{\mu_i(1)})_{\#}\gamma = (T_\gamma^{\mu_i(1)} \circ T_{\mu_{i,0}}^{\gamma})_{\#}\mu_{i,0}$ and $\mu_i(1) = (T_{\mu_{i,0}}^{\mu_i(1)})_{\#}\mu_{i,0}$, thus *(a)* $T_{\mu_{i,0}}^{\mu_i(1)} = T_\gamma^{\mu_i(1)} \circ T_{\mu_{i,0}}^{\gamma}$ and $T_{\mu_i(1)}^{\mu_{i,0}} = T_\gamma^{\mu_{i,0}} \circ T_{\mu_i(1)}^{\gamma}$. To obtain the second equality in (a), we used the identity $(T_{\alpha_2}^{\alpha_1})^{-1} = T_{\alpha_1}^{\alpha_2}$ that holds for two absolutely continuous measures $\alpha_1, \alpha_2 \in \mathcal{P}^2(\mathbb{R}^d)$. Now, $(T_\gamma^{\mu_i(1)})_{\#}\gamma = (T_{\mu_{i,0}}^{\mu_i(1)})_{\#}\mu_{i,0} = (T_{\mu_{i,0}}^{\mu_i(1)})_{\#}(T_\gamma^{\mu_{i,0}})_{\#}\gamma = (T_{\mu_{i,0}}^{\mu_i(1)} \circ T_\gamma^{\mu_{i,0}})_{\#}\gamma$, and so *(b)* $T_\gamma^{\mu_i(1)} = T_{\mu_{i,0}}^{\mu_i(1)} \circ T_\gamma^{\mu_{i,0}}$ and $T_{\mu_i(1)}^{\gamma} = T_{\mu_{i,0}}^{\gamma} \circ T_{\mu_i(1)}^{\mu_{i,0}}$. Now, we do a push-forward operation under the map $T_{\mu_i(1)}^{\mu_{i,0}}$ on both sides of the first equality in (b) to obtain $(T_{\mu_i(1)}^{\mu_{i,0}} \circ T_\gamma^{\mu_i(1)})_{\#}\gamma = (T_{\mu_i(1)}^{\mu_{i,0}} \circ T_{\mu_{i,0}}^{\mu_i(1)} \circ T_\gamma^{\mu_{i,0}})_{\#}\gamma \Longrightarrow (T_{\mu_i(1)}^{\mu_{i,0}} \circ T_\gamma^{\mu_i(1)})_{\#}\gamma = (T_\gamma^{\mu_{i,0}})_{\#}\gamma$,

and so *(c)* $T_\gamma^{\mu_{i,0}} = T_{\mu_i(1)}^{\mu_{i,0}} \circ T_\gamma^{\mu_i(1)}$ and $T_{\mu_{i,0}}^{\gamma} = T_{\mu_i(1)}^{\gamma} \circ T_{\mu_{i,0}}^{\mu_i(1)}$. Then, from results (a)-(c) and the fact that the selected edge at time zero was arbitrary, we conclude that $\{\mu_1(1), \dots, \mu_n(1)\} \cup \{\mu_1(0), \dots, \mu_n(0)\}$ forms a compatible collection with probability one. Indeed, we can easily prove by induction that, with probability one, $\cup_{\tau=0}^{t}\{\mu_i(\tau)\}_{i\in V}$ is a compatible collection.

Now, assume any $(i, j) \in E$ is selected at time zero. Then,

$$\begin{aligned}
\mu_i(1) &= ((1-a_{ij})Id + a_{ij}T_{\mu_{i,0}}^{\mu_{j,0}})_\# \mu_{i,0} = ((1-a_{ij})Id + a_{ij}T_\gamma^{\mu_{j,0}} \circ T_{\mu_{i,0}}^{\gamma})_\# \mu_{i,0} \\
&= ((1-a_{ij})T_\gamma^{\mu_{i,0}} \circ T_{\mu_{i,0}}^{\gamma} + a_{ij}T_\gamma^{\mu_{j,0}} \circ T_{\mu_{i,0}}^{\gamma})_\# \mu_{i,0} \\
&= \big(((1-a_{ij})T_\gamma^{\mu_{i,0}} + a_{ij}T_\gamma^{\mu_{j,0}}) \circ T_{\mu_{i,0}}^{\gamma}\big)_\# \mu_{i,0} \\
&= ((1-a_{ij})T_\gamma^{\mu_{i,0}} + a_{ij}T_\gamma^{\mu_{j,0}})_\# (T_{\mu_{i,0}}^{\gamma})_\# \mu_{i,0} = ((1-a_{ij})T_\gamma^{\mu_{i,0}} + a_{ij}T_\gamma^{\mu_{j,0}})_\# \gamma,
\end{aligned} \tag{4.21}$$

where the second equality follows from the property of measures in a compatible collection. Moreover, using the recent result that $\cup_{\tau=0}^{t}\{\mu_i(\tau)\}_{i\in V}$ is a compatible collection with probability one, we can follow a similar derivation to (4.21) and prove by induction that $(T_\gamma^{\mu_i(t+1)})_\#\gamma = ((1-a_{ij})T_\gamma^{\mu_i(t)} + a_{ij}T_\gamma^{\mu_j(t)})_\#\gamma$ for every t. This is equivalent to

$$T_\gamma^{\mu_i(t+1)}(x) = (1-a_{ij})T_\gamma^{\mu_i(t)}(x) + a_{ij}T_\gamma^{\mu_j(t)}(x) \tag{4.22}$$

for any $x \in \text{supp}(\gamma)$. Define $T_x(t) := (T_\gamma^{\mu_1(t)}(x), \dots, T_\gamma^{\mu_n(t)}(x))^\top$ for any $x \in \text{supp}(\gamma)$. Then, (4.22) can be expressed as $T_x(t+1) = (A(t) \otimes I_d)T_x(t)$. Then, Proposition 4.3.3 implies that $\lim_{t\to\infty} T_x(t) = (\mathbb{1}_n\lambda^\top \otimes I_d)T_x(0)$ for some random convex vector $\lambda = (\lambda_1, \dots, \lambda_n)^\top$ with probability one. Thus, we conclude the following consensus result: for any $i \in V$ and any $x \in \text{supp}(\gamma)$,

$$\lim_{t\to\infty} T_\gamma^{\mu_i(t)}(x) = \sum_{j=1}^{n} \lambda_j T_\gamma^{\mu_{j,0}}(x) \Longrightarrow \lim_{t\to\infty} \mu_i(t) = \big(\sum_{j=1}^{n} \lambda_j T_\gamma^{\mu_{j,0}}\big)_\# \gamma.$$

Then, defining $\mu_\infty := \left(\sum_{j=1}^n \lambda_j T_\gamma^{\mu_j(t)}\right)_\# \gamma$ we conclude from [137, Theorem 3.1.9] that the measure μ_∞ is the unique solution to the barycenter problem with convex vector λ, i.e., equation (4.8) is proved. This concludes the proof of statement (i). Statement (ii) is proved with a similar analysis. ■

Proof: [Proof of Corollary 4.3.8] We focus on proving the results for the directed PaWBar algorithm. First, consider the initial measures as in case (i). As mentioned in [137, Section 2.3], the set of absolute continuous measures in $\mathcal{P}^2(\mathbb{R})$ forms a compatible collection. Then, for any $i \in V$, we use the well-known property that $\mu_i(t) = (F^{-1}_{\mu_i(t)})_\# \mathcal{L}$, with $\mathcal{L}$ being the Lebesgue measure on $[0, 1]$. Moreover, the solution to the Monge optimal transport problem from $\mu_{i,0}$ to $\mu_{j,0}$ for any $i, j \in V$ provides the so-called Brenier's map $F^{-1}_{\mu_{j,0}} \circ F_{\mu_{i,0}}$ [151, Theorem 2.5]. We use these two results in the expression for the Wasserstein barycenter μ_∞ in statement (i) of Theorem 4.3.7 and conclude the proof.

Now we consider case (ii). We refer to [137, Section 2.3] for the proof that shows that these particular Gaussian distributions form a compatible collection. Finally, Theorem 4.3.7 implies the convergence to the Wasserstein barycenter and [41, Theorem 2.4] provides the shown characterization of the barycenter. This concludes the proof.

The proofs for the symmetric PaWBar algorithm are very similar and thus omitted.

■

4.4.3 Proofs of results in Subsection 4.3.4

Proof: [Proof of Theorem 4.3.12] We first consider the directed PaWBar algorithm in case (i). Consider any $(i, j) \in E$ is selected at time t. From the definition of constant-

speed geodesics [151], it follows that,

$$\begin{aligned} W_2(\mu_i(t+1),\mu_j(t)) &= (1-a_{ij})W_2(\mu_i(t),\mu_j(t)),\\ W_2(\mu_i(t+1),\mu_i(t)) &= a_{ij}W_2(\mu_i(t),\mu_j(t)). \end{aligned} \tag{4.23}$$

If (i,j) is chosen τ times consecutively starting at time t, then $W_2(\mu_i(t+\tau),\mu_j(t)) = (1-a_{ij})^\tau W_2(\mu_i(t),\mu_j(t))$.

Now, set

$$U(t) = \sum_{(i,j)\in E} W_2(\mu_i(t),\mu_j(t)).$$

Assume any $(i^*,j^*)\in E$ is selected at time t, and let $(k^*,i^*)\in E$ (since G is a cycle). Then, setting $\mathcal{U}(t) = \sum_{(i,j)\in E\setminus\{(i^*,j^*),(k^*,i^*)\}} W_2(\mu_i(t),\mu_j(t))$,

$$\begin{aligned} U(t+1) &= W_2(\mu_{i^*}(t+1),\mu_{j^*}(t)) + W_2(\mu_{i^*}(t+1),\mu_{k^*}(t)) + \mathcal{U}(t)\\ &\le W_2(\mu_{i^*}(t+1),\mu_{j^*}(t)) + W_2(\mu_{i^*}(t+1),\mu_{i^*}(t)) + W_2(\mu_{i^*}(t),\mu_{k^*}(t)) + \mathcal{U}(t)\\ &= (1-a_{i^*j^*})W_2(\mu_{i^*}(t),\mu_{j^*}(t)) + a_{i^*j^*}W_2(\mu_{i^*}(t),\mu_{j^*}(t))\\ &\quad + W_2(\mu_{i^*}(t),\mu_{k^*}(t)) + \mathcal{U}(t)\\ &= W_2(\mu_{i^*}(t),\mu_{j^*}(t)) + W_2(\mu_{i^*}(t),\mu_{k^*}(t)) + \mathcal{U}(t) = U(t) \end{aligned}$$

where we used the triangle inequality and equation (4.23). Therefore, with probability one, $(U(t))_{t\ge 0}$ is a non-increasing sequence that is uniformly lower bounded by zero, which then implies that $U(t)$ converges to some lower bound which we need to prove to be zero. Consider the nontrivial case $U(t)\neq 0$. Consider again any $(i^*,j^*)\in E$. Since G is a cycle, there is a unique directed path $\mathcal{P}_{j^*\to i^*}$ from j^* to i^* of length $n-1$. Let $\mathcal{P}_{j^*\to i^*} = ((j^*,\ell_1),\dots,(\ell_{n-2},i^*))$. Consider (i^*,j^*) was selected at any time t. Now, pick positive numbers $\epsilon_1,\dots,\epsilon_{n-1}$ such that $\sum_{k=1}^{n-1}\epsilon_k < \frac{U(t)}{2}$. Then, from the sentence below (4.23), we

can first select T_1 times the edge (ℓ_{n-2}, i^*) such that $W_2(\mu_{\ell_{n-2}}(t+T_1), \mu_{i^*}(t)) < \epsilon_L$; then, we can select T_2 times the edge (ℓ_{n-3}, ℓ_{n-2}) such that $W_2(\mu_{\ell_{n-3}}(t+T_1+T_2), \mu_{\ell_{n-2}}(t+T_1)) < \epsilon_{n-2}$; and we can continue like this until finally selecting T_{n-1} times the edge (j^*, ℓ_1) such that $W_2(\mu_{j^*}(t+T), \mu_{\ell_1}(t+\sum_{k=1}^{n-2} T_k)) < \epsilon_1$, with $T = \sum_{k=1}^{n-1} T_k$. Then,

$$
\begin{aligned}
&\sum_{(i,j)\in\mathcal{P}_{j^*\to i^*}} W_2(\mu_i(t+T), \mu_j(t+T)) \\
&\qquad = W_2\Big(\mu_{j^*}(t+T), \mu_{\ell_1}\big(\sum_{k=1}^{n-2} T_k\big)\Big) \\
&\qquad\quad + \sum_{m=1}^{n-3} W_2\Big(\mu_{\ell_m}\big(t + \sum_{k=1}^{n-1-m} T_k\big), \mu_{\ell_{m+1}}\big(t + \sum_{k=1}^{n-1-(m+1)} T_k\big)\Big) \\
&\qquad\quad + W_2(\mu_{\ell_{n-2}}(t+T_1), \mu_{i^*}(t)) \\
&\qquad < \sum_{i=1}^{n-1} \epsilon_i < \frac{U(t)}{2}.
\end{aligned}
$$

Moreover, this result and the triangle inequality imply $W_2(\mu_{i^*}(t+T), \mu_{j^*}(t+T)) \leq \sum_{(i,j)\in\mathcal{P}_{j^*\to i^*}} W_2(\mu_i(t+T), \mu_j(t+T)) < \frac{U(t)}{2}$, and thus $U(t+T) = W_2(\mu_{i^*}(t+T), \mu_{j^*}(t+T)) + \sum_{(i,j)\in\mathcal{P}_{j^*\to i^*}} W_2(\mu_i(t+T), \mu_j(t+T)) < \frac{U(t)}{2} + \frac{U(t)}{2} = U(t)$. Since the event "$U(t+T) < U(t)$ for some finite $T > 0$" has positive probability of happening at any time t (because the finite sequence of edges described above has a positive probability of being selected sequentially at any time t), it can happen infinitely often with probability one. Therefore, we conclude that $U(t) \to 0$ as $t \to \infty$ with probability one. Then, G being a cycle implies $U(t) = 0$ iff $\mu_i(t) = \mu_j(t)$ for any $i, j \in V$, and the consensus result (4.11) follows. Note that the particular value of the converged measure μ_∞ may depend on the specific realization of the edge selection process. This finishes the convergence proof for case (i).

Now, consider the symmetric PaWBar algorithm in case (ii). Without loss of gener-

ality, let $E = \{(1,2), \ldots, (n-1,n)\}$ and set

$$U(t) = \sum_{i=1}^{n-1} W_2(\mu_i(t), \mu_{i+1}(t)). \tag{4.24}$$

Consider any $\{i, i+1\} \in E$ is selected at time t. In the following, consider this notation: for any $a, b \in \{t, t+1\}$ and $k \geq 1$, set $W_2(\mu_{1-k}(a), \mu_1(b)) = 0$ and $W_2(\mu_n(a), \mu_{n+k}(b)) = 0$. Then, setting $\mathcal{U}(t) = \sum_{\substack{j=1 \\ j \neq i-1,i,i+1}}^{n} W_2(\mu_j(t), \mu_{j+1}(t))$,

$$\begin{aligned}
U(t+1) &= W_2(\mu_{i-1}(t), \mu_i(t+1)) + W_2(\mu_{i+1}(t+1), \mu_{i+2}(t)) + \mathcal{U}(t) \\
&\leq W_2(\mu_{i-1}(t), \mu_i(t)) + W_2(\mu_i(t), \mu_i(t+1)) + W_2(\mu_{i+1}(t+1), \mu_{i+1}(t)) \\
&\quad + W_2(\mu_{i+1}(t), \mu_{i+2}(t)) + \mathcal{U}(t) \\
&= \frac{1}{2} W_2(\mu_i(t), \mu_{i+1}(t)) + \frac{1}{2} W_2(\mu_i(t), \mu_{i+1}(t)) + \sum_{\substack{j=1 \\ j \neq i}}^{n} W_2(\mu_j(t), \mu_{j+1}(t)) \\
&= U(t),
\end{aligned}$$

where we used the triangle inequality and equation (4.23). Then $U(t+1) \leq U(t)$ with probability one. Following a similar analysis to case (i), assume the nontrivial case $U(t) \neq 0$. If $W_2(\mu_1(t), \mu_2(t)) \neq 0$ or $W_2(\mu_{n-1}(t), \mu_n(t)) \neq 0$, then it follows from our previous derivation that choosing the edge $\{1,2\}$ or $\{n-1, n\}$ at time t implies $U(t+1) < U(t)$. Now if $W_2(\mu_1(t), \mu_2(t)) = W_2(\mu_{n-1}(t), \mu_n(t)) = 0$ (obviously we consider $n \geq 4$ since for $n = 2, 3$ there is nothing to prove), then, it is easy to prove that we can select a finite sequence of edges, say of some length T', such that $W_2(\mu_1(t+T'), \mu_2(t+T')) \neq 0$ or $W_2(\mu_{n-1}(t+T'), \mu_n(t+T')) \neq 0$. After such sequence is selected, we can select $\{1,2\}$ or $\{n-1, n\}$ so that $U(t+T'+1) < U(t+T') \leq U(t)$. Therefore, at any time t, the event "$U(t+T) < U(t)$ for some finite $T > 0$" has positive probability. Finally, following a similar analysis to case (i), we conclude that $U(t) \to 0$ as $t \to \infty$ with probability one

and conclude the convergence proof of case (ii).

We now focus on proving equation (4.12). We only consider the directed PaWBar algorithm, since the symmetric case is very similar. If $(i,j) \in E$ is selected at time t, then, from the definition of displacement interpolation, any $x \in \text{supp}(\mu_i(t+1))$ can be expressed as a convex combination of one element from $\text{supp}(\mu_i(t))$ and one from $\text{supp}(\mu_j(t))$, with their respective coefficients a_{ij} and $1 - a_{ij}$. Then,

$$\text{supp}(\mu_i(t+1)) \subseteq \{(1-a_{ij})x_i + a_{ij}x_j \,|\, x_i \in \text{supp}(\mu_i(t)), x_j \in \text{supp}(\mu_j(t))\}. \tag{4.25}$$

Now, consider any $i \in V$. With probability one, there is a time when all the sequence of edges in the unique cycle $\mathcal{P}_{i\to i}$ has been selected at least n times (with the order of the selection being arbitrary). After this event happens, say at time T, it is clear that if $(i,j) \in E$ is selected at time $t \geq T$, then

$$\text{supp}(\mu_i(t+1)) \subseteq \{\sum_{i=1}^{n} \lambda_i x_i \,|\, x_i \in \text{supp}(\mu_{i,0}), \lambda_i \geq 0, i \in V, \text{ and } \sum_{i=1}^{n} \lambda_i = 1\}. \tag{4.26}$$

Indeed, equation (4.26) follows from (4.25) and the following property of convex analysis: if x is a convex combination of numbers $u_1, \ldots, u_k$ and each u_i is a convex combination of numbers $v_1^i, \ldots, v_{\ell_i}^i$, $i \in \{1, \ldots, k\}$, then x is a convex combination of the numbers in $\cup_{i=1}^{k} \{v_1^i, \ldots, v_{\ell_i}^i\}$. Finally, since the right hand side of (4.26) is a set independent of time t, we can take the limit $t \to \infty$ and obtain (4.12). ■

4.5 The relevance of the PaWBar algorithm in opinion dynamics

In this section we discuss how the directed PaWBar algorithm generalizes a well-known opinion dynamics model with real-valued beliefs to a model with probability distributions as beliefs. Assume the strongly-connected weighted digraph $G = (V, E, A)$ describes a social network, whereby each agent is an individual and the weight $a_{ij} \in (0, 1)$, for each $(i, j) \in E$, indicates how much influence individual i accords to individual j. Traditionally in the field of opinion dynamics, the opinion or belief of any $i \in V$ at time t is modeled as a scalar $x_i(t) \in \mathbb{R}$. In the popular *asynchronous averaging model* (e.g., see [56, 5]) beliefs evolve as follows: if $(i, j) \in E$ is selected at time t, then $x_i(t + 1) = (1 - a_{ij})x_i(t) + a_{ij}x_j(t)$. Note that the PaWBar algorithm specializes to the asynchronous averaging model (as a consequence of Theorem 4.3.4 or Proposition 4.3.9) when each agent has a degenerate initial distribution with unit mass at a single scalar value.

It is easy to formulate a second generalization of the asynchronous averaging model. Let $\mu_i(t)$ and $\mu_j(t)$ denote the beliefs of individuals i and j, assume $(i, j) \in E$ is selected at time t, and consider the update $\mu_i(t+1) = (1-a_{ij})\mu_i(t)+a_{ij}\mu_j(t)$. This second model is a simple (weighted) averaging of the beliefs; we call it the *AoB model*. To understand the similarities and difference between the PaWBar and AoB models, assume the beliefs of individuals i and j at time t are Gaussian distributions $\mathcal{N}(x_i(t), \sigma)$ and $\mathcal{N}(x_j(t), \sigma)$ with equal variance. Under this assumption, one can see that both models predict that i's mean opinion evolves according to $x_i(t + 1) = (1 - a_{ij})x_i(t) + a_{ij}x_j(t)$. However, the

two models differ in the predicted overall belief and, specifically:

$$\text{PaWBar model:} \qquad \mu_i(t+1) := \mathcal{N}\big((1-a_{ij})x_i(t) + a_{ij}x_j(t), \sigma\big), \tag{4.27}$$

$$\text{AoB model:} \qquad \mu_i(t+1) := (1-a_{ij})\mathcal{N}(x_i(t), \sigma) + a_{ij}\mathcal{N}(x_j(t), \sigma). \tag{4.28}$$

In other words, the PaWBar model predicts a Gaussian belief and the AoB model predicts a Gaussian mixture belief. Even though both resulting beliefs have the same mean, they overall differ substantially.

Finally, we argue that the PaWBar algorithm is preferable over the AoB model for opinion evolution from a cognitive psychology viewpoint. In the case of initial Gaussian beliefs, the PaWBar algorithm dictates that i's belief is simply Gaussian at every time. Thus, as i continues her interactions in the social network, the memory cost associated to her belief at all times is constant: i remembers only two scalars, i.e., the mean opinion and its variance. Instead, if i updates her belief according to the AoB model, then her belief is a Gaussian mixture at every time and i is required to remember a more complicated belief structure. Thus, the AoB model implies that i requires more cognitive power and memory to process the information she gathers from her interactions. The problem with the AoB approach is that arguably individuals tend to simplify beliefs in order to both remember and process thoughts more economically. This simplification of beliefs has attributed humans the metaphor of being *cognitive misers* in cognitive psychology [68, 134]. Therefore, a model with more economic belief memory requirements, such as our PaWBar algorithm, is arguably more adequate.

4.6 Conclusion

We propose the PaWBar algorithm based on stochastic asynchronous pairwise interactions. For specific classes of discrete and absolutely continuous measures, we characterize the computation of both randomized and standard Wasserstein barycenters under arbitrary graphs. For the case of general measures, we prove a consensus result and a necessary condition for the existence of a barycenter, under specific graph structures. We also specialize our algorithm to the Gaussian case and establish a relationship with models of opinion dynamics.

We hope our paper elicits research on efficient numerical solvers for the distributed computation of Wasserstein barycenters based on pairwise computations. Given the plethora of applications for barycenters, we hope our paper elicits interest in the application of distributed randomized barycenters to engineering, economic or scientific domains. Finally, given the importance of Gaussian distributions, we envision theoretical progress in proving the conjecture proposed in our paper.

Chapter 5

Contraction Theory for Dynamical Systems on Hilbert Spaces

5.1 Introduction

Problem statement and motivation Contraction theory establishes the exponential incremental stability of ordinary differential equations. Its mature development can be traced back to the work by Coppel [48], where linear systems were studied, and to the textbook treatment by Vidyasagar [168]. Later, a reformulation was proposed in the seminal work by Slotine [112]. We refer to [11] for an introduction and a survey of applications on contraction theory, and to [157] for extensions to Riemannian manifolds. Generalizations of the classical contraction theory have been proposed in the literature. The notion of partial contraction, first introduced in [142], studies the exponential convergence of trajectories to invariant subspaces [142, 59]. Recently, [89] introduces the concept of semi-contraction, which establishes the exponential incremental semi-stability of trajectories. Contraction theory has also been used for control design [117].

To the best of our knowledge, a general contraction theory on Hilbert and Banach

spaces is missing. The importance of working with systems defined on such general space is, for example, the wide scope of possible applications of systems based on partial differential equations, delayed differential equations, functional differential equations, and integro-differential equations (e.g., see [125, Chapter 9]). The purpose of this note is to concisely present such a theory, with the hope that it will be relevant in both theoretical and applied work. We also provide an application example to illustrate the theory. This work builds a bridge between the abstract theory of differential equations developed in mathematics [53][99] and the widely-established contraction theory in the field of systems and control.

Literature review To the best of our knowledge, a first approach to contraction theory on general Banach spaces can be traced back to the 1972 book by Ladas & Lakshmikantham [99], in its Lemma 5.4.1 and 5.4.2. However, these results do not parallel much of the richer development of contraction theory in Euclidean spaces (see our Contributions below). Interestingly, the results in [99] seem to be unknown in the literature on contraction theory, which developed decades later. Applications of contraction theory have been proposed to specific classes of partial differential equations [11, 9, 10] and more recently to functional differential equations [131]. Besides these notable exceptions, the study of contraction theory on infinite dimensional systems has not received the same development as the Euclidean case, e.g., no concept of semi- or partially contractive systems on Hilbert spaces exists in the literature either.

In the controls community, the recent works [163, 98] have considered dynamical systems on Banach and Hilbert spaces and their applications to PDEs. Other recent interests in dynamical systems on these abstract spaces include controller design [143], event-triggered control [171], observability studies [71], optimal control [165], and stability characterizations [126].

Contributions First, we review two little-known results from [99] that establish a generalization of Coppel's inequality to Banach spaces and that provide some sufficient conditions for contraction using *operator measures* when the vector fields are continuously differentiable. Then, we prove that every time-invariant contractive system has a unique globally exponentially stable equilibrium poin. We also provide a sufficient condition using operator measures for when the norm of a time-invariant system has its vector field exponentially decreasing on trajectories of the system. In the case of time-invariant systems on Hilbert spaces, we introduce a simpler sufficient condition for contraction without the differentiability requirement on the vector field: the *integral contractivity condition*. Moreover, under the differentiability requirement, we prove for time-invariant systems: (i) that the condition using operator measures presented in [99] can be relaxed and still imply contraction (in particular, it is no longer needed for the Jacobian of the system to be uniformly bounded); (ii) the integral contractivity condition is implied by the one using operator measures.

Second, associated with a surjective linear operator $\mathcal{T}$, we introduce the concepts of $\mathcal{T}$*-seminorms* and $\mathcal{T}$*-operator semi-measures* which can be considered as generalization of recently introduced concepts in the study of the classical Euclidean setting [89]. Then, we introduce the concepts of partial and semi-contraction for systems on Hilbert spaces. Using the concepts of seminorms and semi-measures, we provide sufficient conditions for partial contraction for both time-varying and time-invariant systems, and semi-contraction for time-invariant systems. For time-invariant systems, we present a series of novel results. Firstly, we introduce the *integral partial contractivity condition*, a sufficient condition for partial contraction. Secondly, we introduce the *integral semi-contractivity condition*, a sufficient condition for semi-contraction. For continuous differentiable vector fields, we prove this condition is implied by another sufficient condition for semi-contraction using operator semi-measures. When there exists an invariant subspace

for the system, our conditions for semi-contraction imply partial contraction. We remark that, to the best of our knowledge, our characterization of partial and semi-contraction using integral conditions are new even in the classic Euclidean setting (with the usual inner product); e.g. as studied in the work [142] and in our previous work [89].

Finally, we present an example of a reaction-diffusion system and use partial contraction to prove the same result as [16]; moreover, we establish semi-contraction when the reaction term is linear in the state variable.

Paper organization Section 6.2 has preliminaries and notation. Sections 5.3 and 5.4 contain the main results on Banach and Hilbert spaces. Section 5.5 presents the application example and Section 6.7 is the conclusion.

5.2 Preliminaries and notation

5.2.1 Notation, definitions and useful results

A *Banach space* is a complete normed vector space $(\mathcal{X}, \|\cdot\|)$, where $\mathcal{X}$ is a vector space and $\|\cdot\|$ a norm over $\mathcal{X}$. A *Hilbert space* is a pair $(\mathcal{X}, \langle\cdot,\cdot\rangle)$, where $\mathcal{X}$ is a vector space and $\langle\cdot,\cdot\rangle$ is an inner product over $\mathcal{X}$, such that its induced norm $\|\cdot\| := \sqrt{\langle\cdot,\cdot\rangle}$ makes the space a Banach space. In what follows we assume $\mathcal{X}$ is a vector space over the field of real numbers.

Let $B(\mathcal{X})$ be the space of bounded linear operators with domain and codomain $\mathcal{X}$. Let 0 be the null element of $\mathcal{X}$, or the number zero, depending on the context. Let I be the identity operator. Given an operator $A \in B(\mathcal{X})$, $\|A\| = \max_{\substack{x\neq 0 \\ x\in\mathcal{X}}} \frac{\|Ax\|}{\|x\|}$ is its associated operator norm. Given an open set $\Omega \subseteq \mathcal{X}$, we say a function $H : \mathcal{X} \to \mathcal{X}$ is continuously Fréchet differentiable in Ω when H is Fréchet differentiable at each $x_o \in \Omega$ (i.e., $DH(x_o)$ exists) and $DH : \Omega \to B(\mathcal{X})$ is continuous. Finally, we say a subspace $\mathcal{V} \subset \mathcal{X}$ is invariant

for $A \in \mathcal{X}$ if for any $x \in \mathcal{V}$ then $Ax \in \mathcal{V}$.

Let I_n be the $n \times n$ identity matrix and $\mathbb{0}_n \in \mathbb{R}^n$ be the all-zeros column vector with n entries.

The following is a generalization of the definition of matrix measures, e.g., see [99, Definition 5.4.2], also known as the logarithmic norm.

Definition 5.2.1 (Operator measure) *Let $A \in B(\mathcal{X})$ and define the* operator measure *of A as:*

$$\mu(A) = \lim_{h \to 0^+} \frac{\|I + hA\| - 1}{h}.$$

5.2.2 Dynamical systems on Banach spaces

Given the Banach space $(\mathcal{X}, \|\cdot\|)$ and the vector field $F : \mathbb{R} \times \mathcal{X} \to \mathcal{X}$, consider the differential equation:

$$\dot{x} = F(t, x) \tag{5.1}$$

with $\frac{dx}{dt} = \dot{x}$. Following closely the setting in [125, Chapter 9], a continuous function $\phi : [t_0, t_0 + c) \to \mathcal{X}$, $c > 0$, is a solution of (5.1) if it is differentiable with respect to t for $t \in [t_0, t_0 + c)$ and if ϕ satisfies the equation $\dot{\phi} = F(t, \phi(t))$ for all $t \in [t_0, t_0 + c)$. When the system (5.1) is associated the initial condition $x(t_0) = x_0$, we have an initial value problem or Cauchy problem. In this paper we assume that for any $x_0 \in \mathcal{X}$, there exists at least one solution $\phi(t, t_0, x_0)$ to the initial value problem with $x(t_0) = x_0 = \phi(t_0, t_0, x_0)$ for all $t \geq t_0$, $t_0 \in \mathbb{R}_{\geq 0}$.

We say that a set $\mathcal{U}$ is (positively) invariant for the system (5.1) if $\phi(t', t_0, x_0) \in \mathcal{U}$ at some time $t' \geq t_0$ implies $\phi(t, t_0, x_0) \in \mathcal{U}$ for any $t \geq t'$.

The dynamical system (5.1) is time-invariant whenever the vector field F is time-invariant, i.e., F does not explicitly depend on t. If the system (5.1) is time-invariant, it has an equilibrium point x^* if $F(x^*) = 0$.

The system (5.1) has exponential incremental stability if, for any $x_0, y_0 \in \mathcal{X}$, the trajectories $\phi(t, t_0, x_0)$ and $\phi(t, t_0, y_0)$ for any $t \geq t_0$ satisfy $\|\phi(t, t_0, x_0) - \phi(t, t_0, y_0)\| \leq e^{-c(t-t_0)} \|x_0 - y_0\|$. A system is *contractive* with respect to norm $\|\cdot\|$ when $c > 0$, with c known as the *contraction rate*. In the Euclidean case, a central tool for studying contractivity is the matrix measure [11], i.e., the operator measure taking matrices as arguments.

5.3 Contraction on Banach and Hilbert spaces

The following Lemma 5.3.1 was proved in [99, Lemma 5.4.1],[1] and the next Theorem 5.3.2 is an application of [99, Lemma 5.4.2]. These are the only two existing results from the scarce literature on contraction on Banach spaces that we use.

Lemma 5.3.1 (Coppel's inequality for Banach spaces [99, Lemma 5.4.1]) *Let the linear time-varying dynamical system be*

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

on the Banach space $(\mathcal{X}, \|\cdot\|)$, with $A(t) \in B(\mathcal{X})$ and $t \mapsto A(t)$ being continuous for every $t \in \mathbb{R}_{\geq 0}$. Suppose that $\phi(t, t_0, x_0)$ is a solution of the Cauchy problem, then

$$\|x_0\| \exp\left(\int_{t_0}^{t} -\mu(A(\tau))d\tau\right) \leq \|\phi(t, t_0, x_0)\| \leq \|x_0\| \exp\left(\int_{t_0}^{t} \mu(A(\tau))d\tau\right). \quad (5.2)$$

Theorem 5.3.2 (Contraction on Banach Spaces. [99, Lemma 5.4.2]) *Consider the dynamical system (5.1) on the Banach space $(\mathcal{X}, \|\cdot\|)$ with $F(t, \cdot)$ continuously Fréchet differentiable for each t and such that $\|DF(t, u)\| \leq a$ for any $u \in \mathcal{X}$, $t \geq 0$, and some*

[1] The result [99, Lemma 5.4.1] does not prove the left inequality in equation (5.2), but this follows immediately from the same proof.

constant $a > 0$. Assume that there exists $c > 0$ such that $\mu(DF(t,x)) \leq -c$ for every $(t,x) \in \mathbb{R}_{\geq 0} \times \mathcal{X}$. Then the system (5.1) *is* contractive, *i.e.,*

$$\|\phi(t,t_0,x_0) - \phi(t,t_0,x_0')\| \leq e^{-c(t-t_0)}\|x_0 - x_0'\| \tag{5.3}$$

for all $t \geq t_0$ and any $x_0, x_0' \in \mathcal{X}$.

Proof: First, observe that, for any $u, h \in \mathcal{X}$, $\|DF(t,u)h\| \leq a\,\|h\|$. Moreover, observe that the differential equation $\dot{r} = ar$ satisfies that if it has a solution $r(t)$ such that $r(t_0) = 0$, then $r(t) = 0$ for any $t \geq t_0$. These two conditions satisfy the hypotheses of [99, Theorem 5.3.3], and thus we can use [99, Lemma 5.4.2] from which equation (5.3) follows. ■

Beginning now, all of the following results presented in this paper are novel. We present additional properties for contractive systems when the vector field is time-invariant.

Theorem 5.3.3 (Time-invariant contractive systems) *Consider the dynamical system* (5.1) *on the Banach space $(\mathcal{X}, \|\cdot\|)$ with F time-invariant.*

(i) If the system is contractive with contraction rate c, then there exists a unique globally exponentially stable equilibrium point x^* such that

$$\|\phi(t,t_0,x_0) - x^*\| \leq e^{-c(t-t_0)}\|x_0 - x^*\|,$$

for all $t \geq t_0$ and any $x_0 \in \mathcal{X}$.

(ii) If F is continuously Fréchet differentiable and $\mu(DF(x)) \leq -c$ for every $x \in \mathcal{X}$, then $\|F(\phi(t,t_0,x_0))\| \leq e^{-c(t-t_0)}\|F(x_0)\|$, for all $t \geq t_0$ and any $x_0 \in \mathcal{X}$.

Proof: We prove statement (i). Recalling that the system is contractive, we have $\|\phi(t,t_0,x_0) - \phi(t,t_0,x_0')\| \leq e^{-c(t-t_0)}\,\|x_0 - x_0'\|$ for any $t \geq t_0$, $x_0, x_0' \in \mathcal{X}$. Fix any

$t > t_0$. Since $e^{-c(t-t_0)} < 1$, we can use the Banach fixed point theorem to conclude there exists a unique fixed point x^* such that $\phi(t, t_0, x^*) = x^*$, which implies that x^* is either an equilibrium point or is a point which is revisited by the trajectory at time t. By contradiction, if we assume the latter, then any point y^* at time $t_0 \leq t' \leq t$ will be revisited at time $t + (t' - t_0)$ (since there is uniqueness of solutions from (5.3) and F is time invariant) and thus y^* is also a fixed point of $\phi(t, t_0, \cdot)$, which violates the uniqueness of x^* as a fixed point. Then, x^* must be the unique equilibrium of F. We just proved statement (i).

Finally, to prove statement (ii), observe that using the chain rule on Banach spaces [2, Theorem 2.4.3],

$$\frac{d}{dt}F(\phi(t,0,x_0)) = DF(\phi(t,t_0,x_0))\frac{d}{dt}\phi(t,t_0,x_0) = DF(\phi(t,t_0,x_0))F(\phi(t,t_0,x_0))$$

, i.e., $F(\phi(t, t_0, x_0)) \in \mathcal{X}$ satisfies a linear time-varying differential equation on Banach spaces. Now, using Lemma 5.3.1,

$$\|F(\phi(t,t_0,x_0))\| \leq \|F(x_0)\| \, e^{\int_{t_0}^{t} \mu\left(DF(\phi(\tau,t_0,x_0))\right) d\tau} \leq e^{-c(t-t_0)} \, \|F(x_0)\| \,, \tag{5.4}$$

where we used $\mu(DF(x)) \leq -c$, for every $x \in \mathcal{X}$. ■

Assume now that $\mathcal{X}$ is also a Hilbert space (over the field of real numbers) equipped with some inner product $\langle \cdot, \cdot \rangle$. Then, a weaker and simpler sufficient condition for contractivity than the one in Theorem 5.3.2 can be obtained if the dynamical system (5.1) is time-invariant.

Theorem 5.3.4 (Integral contractivity condition) *Consider the dynamical system* (5.1) *on the Hilbert space* $(\mathcal{X}, \langle \cdot, \cdot \rangle)$ *with* F *time-invariant.*

(i) If the following integral contractivity condition *holds*

$$\langle x-y, F(x)-F(y)\rangle \leq -c\left\|x-y\right\|^2 \tag{5.5}$$

for some $c>0$ and any $x,y\in\mathcal{X}$, then system (5.1) *is contractive.*

(ii) If F is continuously Fréchet differentiable, and $\mu(DF(x))\leq -c$ for any $x\in\mathcal{X}$, then condition (5.5) *holds.*

Proof: Consider (5.5) and define $e:=x-y$. Then, $\dot{e}=F(x)-F(y)$ and we obtain $\langle e,\dot{e}\rangle \leq -c\left\|e\right\|^2 \Rightarrow \frac{d\|e\|^2}{dt}=\frac{d\langle e,e\rangle}{dt}=(\langle e,\dot{e}\rangle+\langle \dot{e},e\rangle)\leq -2c\left\|e\right\|^2$, where we used the fact that the inner product is a bilinear function. Solving this differential inequality using the Grönwall's Lemma, we obtain $\|e(t)\|\leq e^{-c(t-t_0)}\left\|e(t_0)\right\|$ for any $t\geq t_0$, establishing that the system is contractive and proving statement (i).

Now, we prove statement (ii) of the theorem. First, let $A\in B(\mathcal{X})$, then

$$\begin{aligned}\mu(A)&=\lim_{h\to 0^+}\frac{\max\limits_{\substack{x,y\neq 0\\ x,y\in\mathcal{X}}}\frac{|\langle x,(I+hA)y\rangle|}{\|x\|\|y\|}-1}{h}\\ &\geq \lim_{h\to 0^+}\frac{\frac{\langle x,(I+hA)x\rangle}{\langle x,x\rangle}-1}{h}=\frac{\langle x,Ax\rangle}{\langle x,x\rangle}\end{aligned} \tag{5.6}$$

for any $x\in\mathcal{X}$ and $x\neq 0$; the first equality follows from [130, p. 187]. However, note that $\mu(A)\langle x,x\rangle\geq\langle x,Ax\rangle$ does hold for any $x\in\mathcal{X}$. Now, consider F to be continuously Fréchet differentiable, and consider any $x,y\in\mathcal{X}$. From the fundamental theorem of calculus for Fréchet derivatives [2, Proposition 2.4.7],

$$F(x)-F(y)=\left(\int_0^1 DF(s_\lambda(x,y))d\lambda\right)(x-y) \tag{5.7}$$

with $s_\lambda(x,y):=x+\lambda(x-y)$, and where the integral is the Riemann integral on Banach

spaces [57, Chapter 7][99, Section 1.3] ($B(\mathcal{X})$ is a Banach space with the operator norm). Then, $\langle x-y, F(x)-F(y)\rangle = \langle x-y, \int_0^1 DF(s_\lambda(x,y))d\lambda(x-y)\rangle$, and using (5.6),

$$\begin{aligned}
&\langle x-y, F(x)-F(y)\rangle \\
&\leq \mu\left(\int_0^1 DF(s_\lambda(x,y))d\lambda\right)\|x-y\|^2 \\
&\leq \int_0^1 \mu\left(DF(s_\lambda(x,y))\,d\lambda\,\|x-y\|^2 \leq -c\,\|x-y\|^2\right.\!.
\end{aligned}$$

We now justify the second inequality above. Let S_n be the nth partial sum of a Riemann integral $\mathcal{I}$, and set $q_n(h) = \frac{\|I+hS_n\|-1}{h}$. Observe that (i) $\lim_{h\to 0^+} q_n(h) = \mu(S_n)$ for each n; (ii) $\lim_{n\to\infty} q_n(h) = \frac{\|I+h\mathcal{I}\|-1}{h}$ uniformly over h. Then, the Moore-Osgood Theorem implies $\lim_{n\to\infty}\mu(S_n) = \mu(\mathcal{I})$. This and the sub-additive property of operator measures [99, Problem 5.4.1] prove the second inequality above. ■

Remark 5.3.5 (On the integral contractivity condition)

(i) The integral contractivity condition does not require the vector field F to be Fréchet differentiable.

(ii) When F is continuously Fréchet differentiable, the Jacobian is no longer required to be uniformly bounded as in Theorem 5.3.2 (which follows from [99, Lemma 5.4.2]). Then, for time-invariant systems, statement (ii) of Theorem 5.3.4 provides a more relaxed condition for contraction using operator measures.

(iii) The integral contractivity condition generalizes a known sufficient condition of contractivity (e.g., [44, Lemma 2.1]) that has been established in the Euclidean space and is related to the so-called QUAD condition for dynamical systems [58, 141].

Remark 5.3.6 (Uniqueness of solutions) *For any system satisfying the assumptions of Theorem 5.3.2 or Theorem 5.3.4, the existence of a solution implies its uniqueness.*

5.4 Semi- and partial contraction on Hilbert spaces

In this section, let $(\mathcal{X}, \langle\cdot,\cdot\rangle_X)$ and $(\mathcal{Y}, \langle\cdot,\cdot\rangle_Y)$ be Hilbert spaces and let $\mathcal{T} : \mathcal{X} \to \mathcal{Y}$ be linear, surjective and bounded.[2] A classic example is $\mathcal{X} = \mathbb{R}^n$, $\mathcal{Y} = \mathbb{R}^m$ with $m \leq n$, and $\mathcal{T} \in \mathbb{R}^{m\times n}$ being a full rank matrix.

Define the bilinear function $\langle\!\langle\cdot,\cdot\rangle\!\rangle_{\mathcal{T}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ by $\langle\!\langle x_1, x_2\rangle\!\rangle_{\mathcal{T}} = \langle \mathcal{T}x_1, \mathcal{T}x_2\rangle_Y$, and define the seminorm $\|x_1\|_{\mathcal{T}} := \sqrt{\langle\!\langle x_1, x_1\rangle\!\rangle_{\mathcal{T}}}$. Let $\mathcal{T}^\dagger$ be the Moore-Penrose (generalized) inverse of $\mathcal{T}$, which is a well-defined operator since $\mathcal{T}$ is surjective (and trivially has closed range) [172, Corollary 11.1.1].

Definition 5.4.1 (Partial and semi-contraction) *The system* (5.1) *is*

(i) partially contractive *with respect to* $\|\cdot\|_{\mathcal{T}}$ *if there exists* $c > 0$ *such that, for any* $x_0 \in \mathcal{X}$ *and* $t \geq t_0$,

$$\|\phi(t, t_0, x_0)\|_{\mathcal{T}} \leq e^{-c(t-t_0)} \|x_0\|_{\mathcal{T}}; \tag{5.8}$$

(ii) semi-contractive *with respect to* $\|\cdot\|_{\mathcal{T}}$ *if there exists* $c > 0$ *such that, for any* $x_0, y_0 \in \mathcal{X}$ *and* $t \geq t_0$,

$$\|\phi(t, t_0, x_0) - \phi(t, t_0, y_0)\|_{\mathcal{T}} \leq e^{-c(t-t_0)} \|x_0 - y_0\|_{\mathcal{T}}. \tag{5.9}$$

We remark that the concept of partial and semi-contraction have not been formalized before on Hilbert spaces. Indeed, in the Euclidean space (with the usual inner product), our formalization becomes the classic cases studied in [89] and [142] respectively, where $\mathcal{T}$ becomes an $n \times m$, $n < m$, full-row rank matrix.

We introduce the following useful concepts.

[2] In this case, the operator norm of $\mathcal{T}$ is $\|\mathcal{T}\| = \sup_{\substack{x\in\mathcal{X}\\ x\neq 0}} \frac{\|\mathcal{T}x\|_{\mathcal{Y}}}{\|x\|_{\mathcal{X}}}$.

Definition 5.4.2 ($\mathcal{T}$-seminorms and $\mathcal{T}$-operator semi-measures) *Consider a linear, surjective, bounded operator $\mathcal{T} : \mathcal{X} \to \mathcal{Y}$ and let $A \in B(\mathcal{X})$. The associated* $\mathcal{T}$-seminorm *of A as*

$$\|A\|_{\mathcal{T}} = \max_{\substack{x \in \ker(\mathcal{T})^{\perp} \\ x \in \mathcal{X}, x \neq 0}} \frac{\|Ax\|_{\mathcal{T}}}{\|x\|_{\mathcal{T}}}$$

and the $\mathcal{T}$-operator semi-measure *of A as*

$$\mu_{\mathcal{T}}(A) = \lim_{h \to 0^+} \frac{\|I + hA\|_{\mathcal{T}} - 1}{h}.$$

The definition of $\mathcal{T}$-operator semi-measure is well-posed, since $\|\cdot\|_{\mathcal{T}}$ (with the argument in $B(\mathcal{X})$) is a seminorm, and one can easily follow the steps in [57, Example 7.7.] to show the existence of directional derivatives.

Theorem 5.4.1 (Partial contraction on Hilbert spaces) *Let $(\mathcal{X}, \langle \cdot, \cdot \rangle_X)$ and $(\mathcal{Y}, \langle \cdot, \cdot \rangle_Y)$ be Hilbert spaces and $\mathcal{T} : \mathcal{X} \to \mathcal{Y}$ be linear, surjective and bounded. Consider the dynamical system (5.1) on $(\mathcal{X}, \langle \cdot, \cdot \rangle_X)$ with $F(t, \cdot)$ continuously Fréchet differentiable for each t and such that $\left\| \mathcal{T} DF(t,u) \mathcal{T}^{\dagger} \right\| \leq a$ for any $u \in \mathcal{X}$, $t \geq 0$, and some constant $a > 0$. Assume that*

(i) there exists $c > 0$ such that $\mu_{\mathcal{T}}(DF(t,x)) = \mu(\mathcal{T} DF(t,x) \mathcal{T}^{\dagger}) \leq -c$ for every $(t,x) \in \mathbb{R}_{\geq 0} \times \mathcal{X}$ (with the operator measure μ associated to $\|\cdot\|_Y$),

(ii) the subspace $\ker(\mathcal{T})$ is positively invariant.

Then the system (5.1) is partially contractive with respect $\|\cdot\|_{\mathcal{T}}$.

Proof: We first observe that

$$\begin{aligned}\|A\|_{\mathcal{T}} &= \max_{\substack{x\in\ker(\mathcal{T})^{\perp}\\ x\in\mathcal{X},x\neq 0}} \frac{\left\|\mathcal{T}A\mathcal{T}^{\dagger}\mathcal{T}x\right\|_{\mathcal{X}}}{\|\mathcal{T}x\|_{\mathcal{X}}} \\ &= \max_{\substack{y\neq 0\\ y\in\mathcal{Y}}} \frac{\left\|\mathcal{T}A\mathcal{T}^{\dagger}y\right\|_{\mathcal{Y}}}{\|y\|_{\mathcal{Y}}} = \left\|\mathcal{T}A\mathcal{T}^{\dagger}\right\|_{\mathcal{Y}}\end{aligned} \tag{5.10}$$

where the first equality follows from the fact that $\mathcal{T}^{\dagger}\mathcal{T}$ is a projection operator on $\ker(\mathcal{T})^{\perp}$ [85, Theorem 3.5.8]. Then, using the fact that $\mathcal{T}\mathcal{T}^{\dagger} = I$, which follows from $\mathcal{T}$ being surjective [172, Definition 11.1.3], it follows that

$$\begin{aligned}\mu_{\mathcal{T}}(A) &= \lim_{h\to 0^{+}} \frac{\left\|\mathcal{T}(I+hA)\mathcal{T}^{\dagger}\right\|_{\mathcal{Y}} - 1}{h} \\ &= \lim_{h\to 0^{+}} \frac{\left\|I+h\mathcal{T}A\mathcal{T}^{\dagger}\right\|_{\mathcal{Y}} - 1}{h} = \mu(\mathcal{T}A\mathcal{T}^{\dagger}).\end{aligned} \tag{5.11}$$

Now, set $y = \mathcal{T}x$, with x being the state of the system, and since x is (Fréchet) differentiable with respect to time, by the chain rule, y is differentiable with respect to time and $\dot{y} = \mathcal{T}\dot{x} = \mathcal{T}F(t,x)$. Now, since $\mathcal{T}$ is a bounded linear operator, $\ker(\mathcal{T})$ is a closed linear subspace of $\mathcal{X}$, and so, we have the following decomposition $\mathcal{X} = \ker(\mathcal{T}) \oplus \ker(\mathcal{T})^{\perp}$ [114, Theorem 1, Section 3.4]. Set $U := I - \mathcal{T}^{\dagger}\mathcal{T}$. Then, for any trajectory $t \mapsto x(t)$, we have $x(t) = \mathcal{T}^{\dagger}\mathcal{T}x(t) + Ux(t) = \mathcal{T}^{\dagger}y(t) + Ux(t)$, with $\mathcal{T}^{\dagger}y(t) \in \ker(\mathcal{T})^{\perp}$ and $Ux(t) \in \ker(\mathcal{T})$. Then,

$$\dot{y} = \mathcal{T}F(t, \mathcal{T}^{\dagger}y + Ux(t)) \tag{5.12}$$

is a time-varying dynamical system on the Hilbert space $(\mathcal{Y}, \langle\cdot,\cdot\rangle_Y)$, and so the Fréchet derivative, using the chain rule, of the right-hand side of (5.12) (with respect to y) is $\mathcal{T}DF(t, \mathcal{T}^{\dagger}y + Ux(t))\mathcal{T}^{\dagger}$. Then, from (5.11), it easily follows from Theorem 5.3.2 that: if $\mu(\mathcal{T}DF(t,x)\mathcal{T}^{\dagger}) \leq -c$ as in the theorem statement and assumption (A1), then the dynamical system (5.12) is contracting with respect to the norm $\|\cdot\|_Y$. Now, we make the

following observation: let us consider a solution with initial condition $x_o \in \ker(\mathcal{T})$ at time t_0, then, $\phi(t, t_0, x_o) \in \ker(\mathcal{T})$ and so $\mathcal{T}\phi(t, t_0, x_o) = 0$ for any $t \geq t_0$ (by assumption (A2)). Differentiating, we obtain $\mathcal{T}F(t, \phi(t, t_0, x_o)) = 0$, which let us conclude that if $u \in \ker(\mathcal{T})$, then $\mathcal{T}F(t, u) = 0$. In conclusion, there are two solutions known for the system (5.12): $y = 0$ (because if $y = 0$, then $\dot{y} = \mathcal{T}F(t, Ux) = 0$ follows from $Ux \in \ker(\mathcal{T})$ as we just showed) and $t \mapsto y(t) = \mathcal{T}x(t)$, and these two solutions should exponentially converge to each other due to contraction. Then, equation (5.8) follows from $\|y\|_Y = \|\mathcal{T}x\|_X = \|x\|_{\mathcal{T}}$. ■

Theorem 5.4.2 (Integral partial contractivity condition) *Consider the dynamical system* (5.1) *on the Hilbert space* $(\mathcal{X}, \langle\cdot,\cdot\rangle_X)$ *with* F *time-invariant, and the linear, surjective, bounded operator* $\mathcal{T} : \mathcal{X} \to \mathcal{Y}$. *If the following* integral partial contractivity condition *holds*

$$\langle\!\langle x, F(x)\rangle\!\rangle_{\mathcal{T}} \leq -c \|x\|_{\mathcal{T}}^2 \tag{5.13}$$

for some $c > 0$ *and any* $x \in \mathcal{X}$, *then the system is partially contractive with respect to* $\|\cdot\|_{\mathcal{T}}$, *and, as a consequence,* $\ker(\mathcal{T})$ *is a positively invariant subspace.*

Proof: The proof is very similar to the first part of Theorem 5.4.3 for proving its respective integral condition, and thus is omitted. ■

We now introduce the counterpart of Theorem 5.3.4 for semi-contracting systems.

Theorem 5.4.3 (Integral semi-contractivity condition) *Consider the dynamical system* (5.1) *on the Hilbert space* $(\mathcal{X}, \langle\cdot,\cdot\rangle_X)$ *with* F *time-invariant, and the linear, surjective, bounded operator* $\mathcal{T} : \mathcal{X} \to \mathcal{Y}$.

(i) If the following integral semi-contractivity condition *holds*

$$\langle\!\langle x - y, F(x) - F(y)\rangle\!\rangle_{\mathcal{T}} \leq -c \|x - y\|_{\mathcal{T}}^2 \tag{5.14}$$

for some $c > 0$ and any $x, y \in \mathcal{X}$, then the system (5.1) *is semi-contractive with respect to $\|\cdot\|_{\mathcal{T}}$.*

(ii) If F is continuously Fréchet differentiable, $\mu(\mathcal{T}DF(x)\mathcal{T}^\dagger) \leq -c$, and $\ker(\mathcal{T})$ *is an invariant subspace for $DF(x)$, for every $x \in \mathcal{X}$, then condition* (5.14) *holds.*

Proof: First, consider the inequality (5.14) and define $e := x - y$ and follow the same procedure as in the proof of Theorem 5.3.4 to show that $\|e(t)\|_{\mathcal{T}} \leq e^{-c(t-t_0)} \|e(t_0)\|_{\mathcal{T}}$ for any $t \geq t_0$, thus establishing the system is semi-contractive and proving statement (i).

Now, we prove statement (ii) of the theorem. Consider F to be continuously Fréchet differentiable, and consider any $x, y \in \mathcal{X}$. Then, using the fundamental theorem of calculus for Fréchet derivatives [2, Proposition 2.4.7], we obtain, $F(x) - F(y) = B(x, y)(x - y)$ with $B(x, y) := \int_0^1 DF(y + \lambda(x - y))d\lambda$.

Then,

$$\begin{aligned}
\langle \mathcal{T}(x-y), \mathcal{T}(F(x)-F(y))\rangle_Y &= \langle \mathcal{T}(x-y), \mathcal{T}B(x,y)(I - \mathcal{T}^\dagger\mathcal{T} + \mathcal{T}^\dagger\mathcal{T})(x-y)\rangle_Y \\
&= \langle \mathcal{T}(x-y), \mathcal{T}B(x,y)\mathcal{T}^\dagger\mathcal{T}(x-y)\rangle_Y \\
&\leq \mu(\mathcal{T}B(x,y)\mathcal{T}^\dagger)\langle \mathcal{T}(x-y), \mathcal{T}(x-y)\rangle_Y \\
&= \mu(\mathcal{T}B(x,y)\mathcal{T}^\dagger)\, \|x-y\|_{\mathcal{T}}^2 . \qquad (5.15)
\end{aligned}$$

We now justify the second equality in (5.15). First, the invariance assumption implies that $DF(u)v \in \ker(\mathcal{T})$ for any $v \in \ker(\mathcal{T})$ and $u \in \mathcal{X}$, and so: $\mathcal{T}B(x, y)v = \mathcal{T}\int_0^1 DF(y + \lambda(x - y))d\lambda v = \int_0^1 \mathcal{T}DF(y + \lambda(x - y))vd\lambda = 0$. Then, we use this to obtain the second equality: $(I - \mathcal{T}^\dagger\mathcal{T})(x - y) \in \ker(\mathcal{T})$, and so $B(x, y)(I - \mathcal{T}^\dagger\mathcal{T})(x - y) \in \ker(\mathcal{T})$ and so $\mathcal{T}B(x, y)(I - \mathcal{T}^\dagger\mathcal{T})(x - y) = 0$.

Now, observe that

$$\mu(\mathcal{T}B(x,y)\mathcal{T}^\dagger) = \mu(\int_0^1 \mathcal{T}DF(y+\lambda(x-y)\mathcal{T}^\dagger d\lambda) \\ \leq \int_0^1 \mu(\mathcal{T}DF(y+\lambda(x-y)\mathcal{T}^\dagger)d\lambda \leq -c,$$

where the first inequality is justified in the same way as in the last part of the proof of Theorem 5.3.4. Then, using this relationship in (5.15), we obtain $\langle \mathcal{T}(x-y), \mathcal{T}(F(x) - F(y))\rangle_Y \leq -c\,\|x-y\|_{\mathcal{T}}^2$, which is condition (5.14). ■

Remark 5.4.4 (About partial and semi-contraction)

(i) The integral semi- and partial contractivity conditions do not require F to be continuously Frechét differentiable.

(ii) If $\ker(\mathcal{T})$ is positively invariant for the system, then the integral condition in Theorem 5.4.3 implies partial contractivity.

(iii) For continuous differentiable vector fields on Euclidean spaces, the semi-contraction condition in Theorem 5.4.3 was first introduced in [89].

5.5 Application to reaction-diffusion systems

Reaction-diffusion PDEs have a long history of study due to their importance in chemistry and biology [129]. Of particular interest are conditions under which the system does not present the phenomenon of pattern formation, which occurs from diffusion-driven instabilities [16]. Particular instances of these systems have been studied using analysis related to contraction [9, 10, 11].

Consider a bounded and convex domain $\Omega \subset \mathbb{R}^m$ with smooth boundary $\partial\Omega$. For any function $h : \mathbb{R}^m \to \mathbb{R}^n$, define the vector Laplacian operator ∇^2 by

$\nabla^2 h = (\nabla^2 h_1, \ldots, \nabla^2 h_n)^\top$ and $(\nabla^2 h(x))_i = \sum_{j=1}^n \frac{\partial^2 h_i(x)}{\partial x_j^2}$. Let $\mathcal{L}^2(\Omega)$ be the space of squared-integrable functions $h : \Omega \subset \mathbb{R}^m \to \mathbb{R}^n$ with $\int_\Omega h_i^2 dx < \infty$, $i \in \{1, \ldots, n\}$, endowed with inner product $\langle u, v \rangle = \int_\Omega u^\top v dx$ for any $u, v \in \mathcal{L}^2(\Omega)$ and induced norm $\|u\| = \sqrt{\langle u, u \rangle}$. It is known that $(\mathcal{L}^2(\Omega), \langle \cdot, \cdot \rangle)$ is a Hilbert space.

Given a continuously differentiable *reaction function* $f : \mathbb{R}^n \to \mathbb{R}^n$ and a nonnegative matrix of *diffusion rates* $\Gamma \in \mathbb{R}^{n \times n}$, the *reaction-diffusion system with Neumann boundary conditions* is

$$\begin{aligned} \frac{\partial u}{\partial t} &= f(u) + \Gamma \nabla^2 u \\ \nabla u_i(x) \cdot \widehat{n}(x) &= 0 \quad \text{for all } x \in \partial\Omega,\ i\{1, \ldots, n\}, \end{aligned} \tag{5.16}$$

for $u \in \mathcal{L}^2(\Omega)$ and $(t, x) \in \mathbb{R}_{\geq 0} \times \Omega$. We refer to [16] and references therein for the system's well-posedness and existence of classical solutions $u = u(x, t)$ such that $u(t, \cdot)$ is twice continuously differentiable for each fixed $t \in \mathbb{R}_{\geq 0}$, and that $t \mapsto u(t, \cdot)$ is a twice continuously differentiable function on Ω. We assume that classical solutions exist.

A Neumann eigenvalue $\lambda \in \mathbb{R}$ for the Laplacian operator ∇^2 on Ω is defined by

$$\begin{aligned} -\nabla^2 u &= \lambda u \\ \nabla u_i(x) \cdot \widehat{n}(x) &= 0 \text{ for all } x \in \partial\Omega,\ i \in \{1, \ldots, n\}. \end{aligned}$$

The set of Neumann eigenvalues of the Laplacian operator consists of countably many nonnegative values with no finite accumulation point [79, Section 7.1]. For our Ω, the eigenspace associated with the lowest eigenvalue $\lambda = 0$ is

$$\begin{aligned} \mathcal{S} = \{h \in \mathcal{L}^2(\Omega) \mid h(x) = c \text{ for all } x \in \Omega \\ \text{and some constant vector } c\}. \end{aligned} \tag{5.17}$$

The volume of Ω is $|\Omega| = \int_\Omega dx$ and the spatial average of $h \in \mathcal{L}^2(\Omega)$ over Ω is $\overline{h} =$

$\frac{1}{|\Omega|}\int_\Omega h(x)dx \in \mathbb{R}^n$. Define the operator $\Pi_{\mathcal{S}} : \mathcal{L}^2(\Omega) \to \mathcal{S}^\perp$ by

$$\Pi_{\mathcal{S}}(u) = u - \overline{u}. \tag{5.18}$$

One can easily check that $\Pi_{\mathcal{S}}$ is the orthogonal projection onto $\mathcal{S}^\perp$. Since $\Pi_{\mathcal{S}}$ is surjective, i.e., $\mathrm{im}(\Pi_{\mathcal{S}}) = \mathcal{S}^\perp$, it follows from the definition of the Moore-Penrose pseudoinverse and its uniqueness that $\Pi_{\mathcal{S}}^\dagger : \mathcal{S}^\perp \to \mathcal{L}^2(\Omega)$ is given by $\Pi_{\mathcal{S}}^\dagger(u) = u$.

Given a matrix $A \in \mathbb{R}^{n\times n}$, the matrix measure associated to the standard Euclidean 2-norm, $\mu_2(A)$, has the following property [11]: $\mu_2(A) \le c$ if and only if $\frac{A+A^\top}{2} \preceq cI_n$, i.e., $cI_n - \frac{A+A^\top}{2}$ is positive semi-definite.

Theorem 5.5.1 (Partial and semi-contraction of reaction-diffusion systems) *Let the reaction-diffusion system* (5.16) *with the standard assumptions on f, Γ, and over a bounded and convex set domain $\Omega \subset \mathbb{R}^m$. Suppose that there exists a positive definite matrix $P \in \mathbb{R}^{n\times n}$ such that $\mu_2(P\Gamma) \ge 0$ and $\mu_2(P(Df(x) - \lambda_2\Gamma)) \le -c$ for all $x \in \Omega$ and some constant $c > 0$. Define $u \mapsto \|u\|_{\Pi_{\mathcal{S}},P^{1/2}} := \left\|\Pi_{\mathcal{S}}(P^{1/2}u)\right\|$ with the set $\mathcal{S}$ as in* (5.17)*, and let $\lambda_{\max}(P)$ be the largest eigenvalue of P. Then,*

(i) system (5.16) *is partially contractive with respect to $\|\cdot\|_{\Pi_{\mathcal{S}},P^{1/2}}$, that is, for every solution $u : \mathbb{R}_{\ge 0} \times \Omega \to \mathbb{R}^n$,*

$$\|u(t,\cdot)\|_{\Pi_{\mathcal{S}},P^{1/2}} \le e^{-\frac{c}{\lambda_{\max}(P)}} \|u(0,\cdot)\|_{\Pi_{\mathcal{S}},P^{1/2}},$$

(ii) $\ker(\Pi_{\mathcal{S}}) = \mathcal{S}$ is an invariant subspace and all trajectories exponentially converge to it; and

(iii) if additionally $f(u) = Au$, $A \in \mathbb{R}^{n\times n}$, then (5.16) *is semi-contractive with respect*

to $\|\cdot\|_{\Pi_{\mathcal{S}},P^{1/2}}$*, that is, for every solution* $u, v : \mathbb{R}_{\geq 0} \times \Omega \to \mathbb{R}^n$*,*

$$\|u(t,\cdot) - v(t,\cdot)\|_{\Pi_{\mathcal{S}},P^{1/2}} \leq e^{-\frac{c}{\lambda_{\max}(P)}} \|u(0,\cdot) - v(0,\cdot)\|_{\Pi_{\mathcal{S}},P^{1/2}},$$

Proof: Note that the reaction-diffusion system we are analyzing can be written as $\frac{\partial u}{\partial t} = F(u)$ where $F : \mathcal{L}^2(\Omega) \to \mathcal{L}^2(\Omega)$ is defined by $F(u) := f(u) + \Gamma\nabla^2 u$. Let $\langle\cdot,\cdot\rangle_{\Pi_{\mathcal{S}},P^{1/2}} := \langle \Pi_{\mathcal{S}}(P^{1/2}\cdot), \Pi_{\mathcal{S}}(P^{1/2}\cdot)\rangle$.

We start by proving statement (i). Consider any $u \in \mathcal{L}^2(\Omega)$, and set $\tilde{u} = u - \bar{u}$, so that $\tilde{u} \in \mathcal{S}^{\perp}$. Then,

$$\langle u, F(u)\rangle_{\Pi_{\mathcal{S}},P^{1/2}} = \langle u, f(u)\rangle_{\Pi_{\mathcal{S}},P^{1/2}} + \langle u, \Gamma\nabla^2(u)\rangle_{\Pi_{\mathcal{S}},P^{1/2}} \tag{5.19}$$

First, from the first term in the right-hand side of (5.19),

$$\begin{aligned} \langle u, f(u)\rangle_{\Pi_{\mathcal{S}},P^{1/2}} &= \langle \Pi_{\mathcal{S}}(P^{1/2}(u)), \Pi_{\mathcal{S}}(P^{1/2}(f(u)))\rangle \\ &= \int_{\Omega} \tilde{u}^{\top} P(f(u) - f(\bar{u}))dx \\ &= \int_0^1 \int_{\Omega} \tilde{u}^{\top} P Df(s(\lambda))\tilde{u}dxd\lambda. \end{aligned} \tag{5.20}$$

where for the second equality we repeatedly used $\int_{\Omega} \tilde{u}^{\top} a dx = 0$ for any constant vector $a \in \mathbb{R}^n$ in Ω, and, since Ω is convex, the mean-value theorem: $f(u) - f(\bar{u}) = \int_0^1 Df(s(\lambda))\tilde{u}d\lambda$ with $s(\lambda) = u + \lambda(\bar{u} - u)$ for the last inequality. Now, from the second term in the right-hand side of (5.19),

$$\langle u, \Gamma\nabla^2 u\rangle_{\Pi_{\mathcal{S}},P^{1/2}} = \int_{\Omega} \tilde{u}^{\top} P\Gamma\nabla^2\tilde{u}dx, \tag{5.21}$$

where we used $\Pi_{\mathcal{S}}(\nabla^2 u) = \nabla^2 u - \frac{1}{|\Omega|}\int_{\Omega} \nabla^2 u dx = \nabla^2\tilde{u}$, since $\int_{\Omega} \nabla^2 u_i dx = \int_{\partial\Omega} \nabla u_i \cdot \widehat{n} dS = 0$ (from the divergence theorem and the boundary condition in (5.16)). Note that, for

every $i \in \{1, \dots, n\}$, we have $\nabla^2 \tilde{u}_i(x) = \nabla \cdot (\nabla \tilde{u}_i(x))$. Now, by the product rule, we have that for every $i \in \{1, \dots, n\}$, $\nabla \cdot (\sum_{j=1}^n \tilde{u}_i (P\Gamma)_{ij} \nabla \tilde{u}_j) = \sum_{j=1}^n \tilde{u}_i (P\Gamma)_{ij} \nabla^2 \tilde{u}_j + \sum_{j=1}^n (\nabla \tilde{u}_i)^\top (P\Gamma)_{ij} \nabla \tilde{u}_j$. Now, by the divergence theorem, we obtain for every $i \in \{1, \dots, n\}$, $\int_\Omega \nabla \cdot (\sum_{j=1}^n \tilde{u}_i (P\Gamma)_{ij} \nabla \tilde{u}_j) dx = \int_{\partial\Omega} \big(\sum_{j=1}^n \tilde{u}_i (P\Gamma)_{ij} \nabla \tilde{u}_j \cdot \widehat{n} \big) dS = 0$ where the last equality follows from the boundary condition in (5.16). Then, from the identity $\tilde{u}^\top P\Gamma \nabla^2 \tilde{u} = \sum_{i=1}^n \sum_{j=1}^n \tilde{u}_i (P\Gamma)_{ij} \nabla^2 \tilde{u}_j$ we get

$$\int_\Omega \tilde{u}^\top P\Gamma \nabla^2 \tilde{u} dx = - \int_\Omega \sum_{i=1}^n \sum_{j=1}^n (\nabla \tilde{u}_i)^\top (P\Gamma)_{ij} \nabla \tilde{u}_j dx.$$

Moreover, one can check that

$$\sum_{i=1}^n \sum_{j=1}^n (\nabla \tilde{u}_i)^\top (P\Gamma)_{ij} \nabla \tilde{u}_j = \sum_{k=1}^n (\frac{\partial \tilde{u}}{\partial x_k})^\top (P\Gamma) \frac{\partial \tilde{u}}{\partial x_k}.$$

Since $\mu_2(P\Gamma) \geq 0$, there exists a positive semi-definite matrix $Q \in \mathbb{R}^{n \times n}$ such that $Q^\top Q = \frac{1}{2}(P\Gamma + \Gamma^\top P^\top)$. This implies that $\sum_{k=1}^n (\frac{\partial \tilde{u}}{\partial x_k})^\top P\Gamma \frac{\partial \tilde{u}}{\partial x_k} = \sum_{k=1}^n (\frac{\partial \tilde{u}}{\partial x_k})^\top Q^\top Q \frac{\partial \tilde{u}}{\partial x_k} = \sum_{k=1}^n (\frac{\partial Q\tilde{u}}{\partial x_k})^\top \frac{\partial Q\tilde{u}}{\partial x_k} = \sum_{i=1}^n (\nabla((Q\tilde{u})_i))^\top \nabla((Q\tilde{u})_i)$.

Combining all of these results, we finally obtain $\int_\Omega \tilde{u}^\top P\Gamma \nabla^2 \tilde{u} dx = - \sum_{i=1}^n \|\nabla (Q\tilde{u})_i\|^2$. Now, since $\int_\Omega Q\tilde{u} dx = Q \int_\Omega \tilde{u} dx = \mathbb{0}_n$, we can use the Poincaré inequality on simply connected domains [79, Section 1.3] (since our domain is convex), and obtain $\|\nabla (Q\tilde{u})_i\|^2 \geq \lambda_2 \|(Q\tilde{u})_i\|^2$. As a result, we get

$$\int_\Omega \tilde{u}^\top P\Gamma \nabla^2 \tilde{u} dx \leq -\lambda_2 \sum_{i=1}^n \|(Q\tilde{u})_i\|^2 = \int_\Omega \tilde{u}^\top \left(-\lambda_2 P\Gamma\right) \tilde{u} dx.$$

Replacing this result in (5.21), and then replacing the resulting expression with the one

in (5.20) back in (5.19):

$$\begin{aligned}\langle u, F(u)\rangle_{\Pi_{\mathcal{S}},P^{1/2}} &= \int_0^1 \int_\Omega \tilde{u}^\top P\left(Df(s(\lambda)) - \lambda_2\Gamma\right)\tilde{u}dxd\lambda \\ &\leq -c\int_0^1 \int_\Omega \tilde{u}^\top\tilde{u}dxd\lambda \leq -\frac{c}{\lambda_{\max}(P)}\left\|P^{1/2}\tilde{u}\right\| \\ &= -\frac{c}{\lambda_{\max}(P)}\left\|\tilde{u}\right\|_{\Pi_{\mathbb{S}},P^{1/2}} = -\frac{c}{\lambda_{\max}(P)}\left\|u\right\|_{\Pi_{\mathbb{S}},P^{1/2}}\end{aligned}$$

where the inequality comes from the assumption $\mu_2(P(Df(x)-\lambda_2\Gamma)) \leq -c$ for any $x \in \Omega$. This expression has the form of the integral partial contractivity condition. Although the set of classical solutions endowed with $\langle\cdot,\cdot\rangle$ is not a Hilbert space, we follow the proof of Theorem 5.4.2 using the Leibniz rule to differentiate the inner product and obtain partial contraction as in statement (i). Statement (ii) follows by noting that $\|u\|_{\Pi_{\mathcal{S}},P^{1/2}} = 0 \implies P^{1/2}u \in \ker(\Pi_{\mathcal{S}}) \implies P^{1/2}u \in \mathcal{S} \implies u \in \mathcal{S}$. Finally, statement (iii) is proved in a similar way to statement (i), using the difference of any two solutions as a new state variable. ■

Remark 5.5.2 *Statements (i) and (ii) of Theorem 5.5.1 are essentially the same result as [16, Theorem 1]; however, these statements and statement (iii) are now consequences of a general contraction theory.*

5.6 Conclusion

This paper presents a general contraction theory for dynamical systems on Hilbert spaces. We provide sufficient conditions for contraction, semi-contraction and partial contraction based on operator measures or operator semi-measures, and on the differentiability of the vector field. Moreover, when the system is time-invariant, we present weaker conditions that do not require differentiability. Finally, we present an example of

reaction-diffusion systems.

Our work brings the machinery of contraction theory, so far mainly applied to ODEs, to other possible application domains related to a variety of systems that can be expressed as dynamical systems on functional spaces.

Chapter 6

Distributed and time-varying primal-dual dynamics via contraction analysis

6.1 Introduction

Problem statement and motivation Primal-dual (PD) dynamics are dynamical systems that solve constrained optimization problems. Their study can be traced back to many decades ago [18] and has regained interest since the last decade [66]. PD dynamics have been made popular due to their scalability and simplicity. They have been widely adopted in engineering applications such as resource allocation problems in power networks [158], frequency control in micro-grids [116], solvers for linear equations [174], etc. In this note, we study optimization problems with linear equality constraints. In general, PD dynamics seek to find a saddle point of the associated Lagrangian function to the constrained problem, which is characterized by the equilibria of the dynamics. For a general treatise of asymptotic stability of saddle points, we refer to [42] and refer-

ences therein. However, despite the long history of study and application, there are very recent studies on PD dynamics related to linear equality constraints further studying different dynamic properties such as: exponential convergence under different convexity assumptions [145, 40] and contractivity properties [132].

We are particularly interested in studying primal-dual dynamics in distributed and time-varying optimization problems. We refer to the recent survey [176] for an overview of the long-standing interest on distributed optimization. Of particular interest is to provide strong convergence guarantees such as global (and exponential) convergence for the distributed solvers. We aim to provide them using contraction theory. Time-varying optimization has found applications in system identification, signal detection, robotics, traffic management, etc. [65, 160]. The goal is to employ a dynamical system able to track the time-varying optimal solution up to some bounded error in real time. Although different dynamics have been proposed to both time-varying centralized [65] and distributed problems [160, 148], to the best of our knowledge, there has not been a characterization of the PD dynamics in such application contexts. The importance of PD algorithms is their simplicity of implementation, i.e., they do not require more complex information structures like the inverse of the Hessian of the system at all times, as in [65] and [148] for the centralized and distributed cases respectively. However, simplicity may come with a possible trade-off in the tracking error.

Contraction is valuable in practice because it introduces strong stability and robustness guarantees. For example, it implies input-to-state stability for systems subject to state-independent disturbances. It also guarantees fast correction after transient perturbations to the trajectory of the solution, since initial conditions are forgotten. Moreover, a contractive system may be robust towards structural perturbations on the vector field, e.g., when a non-convex term is added to the objective function. Finally, contraction guarantees stable numerical discretizations with geometric convergence rates, an ideal

situation for practical implementations. All these properties are transparent to whether the system is time-varying or not. All of this motivates a contraction analysis of PD algorithms in contrast to the prevalent Lyapunov or invariance analysis in the literature.

Literature review The recent works [145, 132, 40] study convergence properties of PD dynamics under different assumptions on the objective function. In distributed optimization, solvers based on PD dynamics are fairly recent, e.g., [173, 49, 176]. An application of distributed optimization of current interest - as seen in the recent survey [174] - is the *distributed least-squares problem* for solving an over-determined system of linear equations. To the best of our knowledge, solvers for this problem (in continuous-time) with exponential global convergence are still missing in the literature.

Finally, this paper is related to contraction theory, a mathematical tool to analyze incremental stability [112, 168]. An introduction and survey can be found in [11]. A variant of contraction theory, *partial contraction* [142, 59], analyzes the convergence to linear subspaces and has been used in the synchronization analysis of diffusively-coupled network systems [142, 12]; however, its application to distributed algorithms is still missing, and our paper provides such contribution.

Contributions In this paper we consider the PD dynamics associated to optimization problems with a twice differentiable and strongly convex objective function and linear equality constraints. We use contraction theory to perform an overarching study of PD dynamics in a variety of implementations and applications. In particular:

(i) We introduce new theoretical results of how *weak* and *partial contraction* can imply exponential convergence to a point in a subspace of equilibria.

(ii) For the standard and distributed PD dynamics, we prove: 1) convergence under weak contraction when the objective function is convex; 2) contraction for the standard

problem and partial contraction for the distributed one in the strongly convex case, with closed-form exponential global convergence rates. The analysis in result 1) is novel, since it uses the new results introduced in (i). Compared to the work [132] that also shows contraction for the standard PD, our proof method provides an explicit closed-form expression of the system's contraction rate. Our exponential convergence rate is different from the one by [145] via Lyapunov analysis, and both rates cannot be compared without extra assumptions on the numerical relationships among various parameters associated to the objective function or constraints. Moreover, we propose using the *augmented Langrangian* in order to achieve contraction when the objective function is only convex. In the case of distributed optimization, there exist other solvers that show exponential convergence, e.g., as in [96, 105], but none of these have contractivity.

(iii) We propose a new solver for the distributed least-squares problem based on PD dynamics, and use our results in (ii) to prove its convergence. Compared to the recent work [110], our new model exhibits global convergence; and compared to the recent work [111], ours exhibits exponential convergence and has a simpler structure.

(iv) We characterize the performance of PD dynamics associated to time-varying versions of both standard and distributed optimization problems in terms of the problems' parameters - to the best of our knowledge, this is the first characterization of time-varying PD dynamics in the literature. We prove the tracking error to the time-varying solutions is uniformly ultimately bounded (UUB) in either case and that the bound decreases as the contraction rate increases. Our analysis builds upon the contraction results in contribution (ii).

Paper organization Section 6.2 has notation and preliminary concepts. Section 6.3 has results on contraction theory. Section 6.4 analyzes contractive properties of the standard PD dynamics. The contractive analysis of distributed (with the least-squares

problem application) and time-varying versions of PD dynamics are in Sections V and VI respectively. Section 6.7 is the conclusion.

6.2 Preliminaries and notation

6.2.1 Notation, definitions and useful results

Consider $A \in \mathbb{R}^{n\times n}$, then $\sigma_{\min}(A)$ denote its minimum singular value and $\sigma_{\max}(A)$ its maximum one. If A has only real eigenvalues, let $\lambda_{\max}(A)$ be its maximum eigenvalue. A is an orthogonal projection if it is symmetric and $A^2 = A$. Let $\|\cdot\|$ denote any norm, and $\|\cdot\|_p$ denote the ℓ_p-norm. When the argument of a norm is a matrix, we refer to its respective induced norm. The matrix measure associated to $\|\cdot\|$ is $\mu(A) = \lim_{h\to 0^+} \frac{\|I+hA\|-1}{h}$; e.g., the one associated to the ℓ_2-norm is $\mu_2(A) = \lambda_{\max}((A + A^\top)/2)$ [11]. Given invertible $Q \in \mathbb{R}^{n\times n}$, let $\|\cdot\|_{2,Q}$ be the weighted ℓ_2-norm $\|x\|_{2,Q} = \|Qx\|_2$, $x \in \mathbb{R}^n$, and whose associated matrix measure is $\mu_{2,Q}(A) = \mu_2(QAQ^{-1})$ [11].

Let I_n be the $n \times n$ identity matrix, $\mathbb{1}_n$ and $\mathbb{0}_n$ be the all-ones and all-zeros column vector with n entries respectively. Let $\mathrm{diag}(X_1, \dots, X_N) \in \mathbb{R}^{\sum_{i=1}^N n_i \times \sum_{i=1}^N n_i}$ be the block-diagonal matrix with elements $X_i \in \mathbb{R}^{n_i\times n_i}$. Let $\mathbb{R}_{\geq 0}$ be the set of non-negative real numbers. Given $x_i \in \mathbb{R}^{k_i}$, let $(x_1, \dots, x_N) = \begin{bmatrix} x_1^\top & \dots & x_N^\top \end{bmatrix}$.

Consider a differentiable function $f : \mathbb{R}^n \to \mathbb{R}^n$. We say f is *Lipschitz smooth* with constant $K_1 > 0$ if $\|\nabla f(x) - \nabla f(y)\|_2 \leq K_1 \|x - y\|_2$ for any $x, y \in \mathbb{R}^n$; and *strongly convex* with constant $K_2 > 0$ if $K_2 \|x - y\|_2^2 \leq (\nabla f(x) - \nabla f(y))^\top (x - y)$ for any $x, y \in \mathbb{R}^n$. Assuming f is twice differentiable, these two conditions are equivalent to $\nabla^2 f(x) \preceq K_1 I_n$ and $K_2 I_n \preceq \nabla^2 f(x)$ for any $x \in R^n$, respectively.

The proof of the next proposition is found in the Appendix.

Proposition 6.2.1 *For a full-row rank matrix $A \in \mathbb{R}^{m\times n}$, $B = B^\top \in R^{n\times n}$, and $b_2 \geq b_1 > 0$ such that $b_2 I_n \succeq B \succeq b_1 I_n \succ 0$, the matrix* $\begin{bmatrix} -B & -A^\top \\ A & \mathbb{0}_{m\times m} \end{bmatrix}$ *is Hurwitz.*

6.2.2 Review of basic concepts on contraction theory

Consider the dynamical system $\dot{x} = f(x,t)$ with $x \in \mathbb{R}^n$. Let $t \mapsto \phi(t, t_0, x_0)$ be the trajectory of the system starting from $x_0 \in \mathbb{R}^n$ at time $t_0 \geq 0$. Consider the system satisfies $\|\phi(t,t_0,x_0) - \phi(t,t_0,y_0)\| \leq \|x_0 - y_0\| e^{-c(t-t_0)}$, for any $x_0, y_0 \in \mathbb{R}^n$ and any $t_0 \in \mathbb{R}_{\geq 0}$. We say it is *contractive* with respect to $\|\cdot\|$ when $c > 0$, and *weakly contractive* when $c = 0$. A time-invariant contractive system has a unique equilibrium point. Now, assume the Jacobian of the system, i.e., $Df(x,t)$, satisfies: $\mu(Df(x,t)) \leq -c$ for any $(x,t) \in \mathbb{R}^n \times \mathbb{R}_{\geq 0}$, with μ being the matrix measure associated to $\|\cdot\|$ and constant $c \geq 0$. Then, this system has contraction rate c with respect to $\|\cdot\|$. Now, assume the system has a flow-invariant linear subspace $\mathcal{M} = \{x \in \mathbb{R}^n \mid Vx = \mathbb{0}_k\}$ with $V \in \mathbb{R}^{k\times n}$ being full-row rank with orthonormal rows. Then the system is *partially contractive* with respect to $\|\cdot\|$ and $\mathcal{M}$ if there exists $c > 0$ such that, for any $x_0 \in \mathbb{R}^n$ and $t_0 \in \mathbb{R}_{\geq 0}$, the system satisfies $\|V\phi(t,t_0,x_0)\| \leq \|Vx_0\| e^{-c(t-t_0)}$. When $c = 0$, the system is *partially weakly contractive* with respect to $\mathcal{M}$ [142]. Consequently, a partially contractive system has any of its trajectories approaching $\mathcal{M}$ with exponential rate; and a partially weakly contractive one has any of its trajectories at a non-increasing distance from $\mathcal{M}$.

Pick a symmetric positive-definite $P \in \mathbb{R}^{n\times n}$ and a scalar $c > 0$, then $\mu_{2,P^{1/2}}(Df(x,t)) \leq -c$ for all $(x,t) \in \mathbb{R}^n \times \mathbb{R}_{\geq 0}$ is equivalent to f satisfying the *integral contractivity condition*, i.e., for every $x, y \in \mathbb{R}^n$ and $t \geq 0$, $(y-x)^\top P(f(x,t) - f(y,t)) \leq -c\|x-y\|^2_{2,P^{1/2}}$.

6.3 Theoretical contraction results

The next result will be used throughout the paper.

Theorem 6.3.1 (Results on partial contraction) *Consider the system $\dot{x} = f(x,t)$, $x \in \mathbb{R}^n$, with a flow-invariant $\mathcal{M} = \{x \in \mathbb{R}^n \mid Vx = \mathbb{0}_k\}$ with $V \in \mathbb{R}^{k\times n}$ being a full-row rank matrix with orthonormal rows. Assume $\mu(VDf(x,t)V^\top) \leq -c$ for any $(x,t) \in \mathbb{R}^n \times \mathbb{R}_{\geq 0}$, some constant $c \geq 0$ and some matrix measure μ.*

(i) If $c > 0$, then the system is partially contractive with respect to $\mathcal{M}$ and every trajectory exponentially converges to the subspace $\mathcal{M}$ with rate c.

(ii) If $c = 0$ and $\mu(VDf((I_n - V^\top V)x,t)V^\top) < 0$ for any $(x,t) \in \mathbb{R}^n \times \mathbb{R}_{\geq 0}$, then the system is partially weakly contractive with respect to $\mathcal{M}$ and every trajectory converges to the subspace $\mathcal{M}$.

Moreover, assume that one of the conditions in parts (i) and (ii) holds and $\mathcal{M}$ is a set of equilibrium points. If the system is weakly contractive, then

(iii) every trajectory of the system converges to an equilibrium point, and if $c > 0$, then it does it with exponential rate c.

Remark 6.3.2 *Statement (i) in Theorem 6.3.1 was proved in [142]. To the best of our knowledge, statements (ii) and (iii) are novel.*

Proof: [Proof of Theorem 6.3.1] It is easy to check that $V^\top V$ is an orthogonal projection matrix onto $\mathcal{M}^\perp$; and that $U := I_n - V^\top V$ is also an orthogonal projection matrix onto $\mathcal{M}$. Using these results, we can express the given system as $\dot{x} = f(Ux + V^\top Vx, t)$. Now, we set $z := Vx$, and observe that $x(t)$ converges to $\mathcal{M}$ if and only if $z(t)$ converges $\mathbb{0}_k$. Then, using this change of coordinates, we obtain the system:

$$\dot{z} = Vf(Ux + V^\top z, t). \tag{6.1}$$

It has been proved in [142, Theorem 3] that $z^* = \mathbb{0}_k$ is an equilibrium point for the system (6.1).

To prove (ii), assume that $\mu(VDf(x,t)V^\top) = 0$ for any $(x,t) \in \mathbb{R}^n \times \mathbb{R}_{\geq 0}$; i.e., that the system (6.1) is weakly contractive. Now, if we assume that $\mu(VDf(Ux,t)V^\top) < 0$ for any $(x,t) \in \mathbb{R}^n \times \mathbb{R}_{\geq 0}$, then by the Coppel's inequality [48], the fixed point $z^* = \mathbb{0}_k$ is locally exponentially stable. Now, we can use a generalization of [113, Lemma 6], namely Lemma 6.8.1 (proof found in the Appendix), to establish the convergence of $z(t)$ to z^*. This finishes the proof for (ii).

Now, we prove statement (iii). Let $t \mapsto x(t)$ be a trajectory of the dynamical system. For every $t \in \mathbb{R}_{\geq 0}$, $\big(I_n - V^\top V\big)x(t)$ is the orthogonal projection of $x(t)$ onto the subspace $\mathcal{M}$ and it is an equilibrium point. Since the dynamical system is weakly contractive, we have $\|x(s) - (I_n - V^\top V)x(t)\| \leq \|x(t) - (I_n - V^\top V)x(t)\| = \|V^\top V x(t)\|$, for all $s \geq t$. This implies that, for every $t \in \mathbb{R}_{\geq 0}$ and every $s \geq t$, the point $x(s)$ remains inside the closed ball $\overline{B}(x(t), \|V^\top V x(t)\|)$. Therefore, for every $t \geq 0$, the point $x(t)$ is inside the set C_t defined by $C_t = \mathrm{cl}\big(\bigcap_{\tau \in [0,t]} \overline{B}(x(\tau), \|V^\top V x(\tau)\|)\big)$. It is easy to see that, for $s \geq t$, we have $C_s \subseteq C_t$. This implies that the family $\{C_t\}_{t \in [0,\infty)}$ is a nested family of closed subsets of $\mathbb{R}^n$. Moreover, by parts (i) and (ii), we have that $\lim_{t\to\infty} \|V^\top V x(t)\| \to 0$ as $t \to \infty$, which in turn results in $\lim_{t\to\infty} \mathrm{diam}(C_t) = 0$, with convergence rate c for the case $c > 0$ because of $\mathrm{diam}(C_t) = \|V^\top V x(t)\| \leq 2e^{-ct}\|V^\top V x(0)\|$. Thus, by the Cantor Intersection Theorem [128, Lemma 48.3], there exists $x^* \in \mathbb{R}^n$ such that $\bigcap_{t \in [0,\infty)} C_t = \{x^*\}$. We first show that $\lim_{t\to\infty} x(t) = x^*$. Note that $x^*, x(t) \in C_t$, for every $t \in \mathbb{R}_{\geq 0}$. This implies that $\|x(t) - x^*\| \leq \mathrm{diam}(C_t)$. This in turn means that $\lim_{t\to\infty} \|x(t) - x^*\| = 0$ and $t \mapsto x(t)$ converges to x^*, with convergence rate c for the case $c > 0$. On the other hand, by part (i), the trajectory $t \mapsto x(t)$ converges to the subspace $\mathcal{M}$. Therefore, $x^* \in \mathcal{M}$ and x^* is an equilibrium point of the dynamical system. This completes the proof for statement (iii). ■

6.4 The standard optimization problem

We consider the constrained optimization problem:

$$\min_{x\in\mathbb{R}^n} \; f(x) \quad \text{subject to} \quad Ax = b \tag{6.2}$$

with the following standing assumptions: $A \in \mathbb{R}^{k\times n}$, $k < n$, $b \in \mathbb{R}^k$, A is full-row rank, and $f : \mathbb{R}^n \to \mathbb{R}$ is convex and twice differentiable.

Associated to the optimization problem (6.2) is the *Lagrangian function* $\mathcal{L}(x, \nu) = f(x) + \nu^\top (Ax - b)$ and the *primal-dual dynamics*

$$\begin{bmatrix} \dot{x} \\ \dot{\nu} \end{bmatrix} = \begin{bmatrix} -\frac{\partial \mathcal{L}(x,\nu)}{\partial x} \\ \frac{\partial \mathcal{L}(x,\nu)}{\partial \nu} \end{bmatrix} = \begin{bmatrix} -\nabla f(x) - A^\top \nu \\ Ax - b \end{bmatrix}. \tag{6.3}$$

We introduce two possible sets of assumptions:

(A1) the primal-dual dynamics (6.3) have an equilibrium (x^*, ν^*) and $\nabla^2 f(x^*) \succ \mathbb{0}_{n\times n}$;

(A2) the function f is strongly convex with constant $\ell_{\text{inf}} > 0$ and Lipschitz smooth with constant $\ell_{\text{sup}} > 0$, and, for $0 < \epsilon < 1$, we define

$$\begin{aligned} \alpha_\epsilon &:= \frac{\epsilon \ell_{\text{inf}}}{\sigma^2_{\max}(A) + \frac{3}{4}\sigma_{\max}(A)\sigma^2_{\min}(A) + \ell^2_{\text{sup}}} > 0 \\ P &:= \begin{bmatrix} I_n & \alpha_\epsilon\, A^\top \\ \alpha_\epsilon\, A & I_k \end{bmatrix} \in \mathbb{R}^{(n+k)\times(n+k)}. \end{aligned} \tag{6.4}$$

Theorem 6.4.1 (Contraction analysis of primal-dual dynamics) *Consider the constrained optimization problem* (6.2)*, its standing assumptions, and its associated primal-dual dynamics* (6.3)*.*

(i) The primal-dual dynamics is weakly contractive with respect to $\|\cdot\|_2$ and, if Assumption (A1) *holds, then (x^*, ν^*) is globally asymptotically stable.*

(ii) Under Assumption (A2)*,*

(a) the primal-dual dynamics are contractive with respect to $\|\cdot\|_{2,P^{1/2}}$ with contraction rate

$$\alpha_\epsilon \frac{3}{4} \frac{\sigma_{\max}(A)\sigma_{\min}^2(A)}{\sigma_{\max}(A)+1}, \quad \textit{and} \tag{6.5}$$

(b) there exists a unique globally exponentially stable equilibrium point (x^*, ν^*), and x^* is the unique solution to the optimization problem (6.2)*.*

Proof: Let $(\dot{x}, \dot{\nu})^\top := F_{\mathrm{PD}}(x, \nu)$. Then, $DF_{\mathrm{PD}}(x, \nu) = \begin{bmatrix} -\nabla^2 f(x) & -A^\top \\ A & 0 \end{bmatrix}$, and so $\mu_2(DF_{\mathrm{PD}}(x, \nu)) = \lambda_{\max}\left((DF_{\mathrm{PD}}(x, v) + DF_{\mathrm{PD}}(x, v)^\top)/2\right) = \lambda_{\max}\left(\mathrm{diag}(-\nabla^2 f(x), \mathbb{0}_{k\times k})\right) = 0$ for any $(x, \nu) \in \mathbb{R}^n \times \mathbb{R}^m$, because of convexity $\nabla^2 f(x) \succeq 0$, which implies the system is weakly contractive. For the second part of statement (i): Proposition 6.2.1 implies $DF_{\mathrm{PD}}(x^*, \nu^*)$ is Hurwitz since $\nabla^2 f(x^*) \succ 0$, and the proof follows from a simple generalization of [113, Lemma 6] (its proof can be found in the Appendix).

Now, we prove statement (ii). Define $P = \begin{bmatrix} I_n & \alpha A^\top \\ \alpha A & I_k \end{bmatrix}$ which is a positive-definite matrix when

$$0 < \alpha < \frac{1}{\sigma_{\max}(A)}. \tag{6.6}$$

We plan to use the integral contractivity condition to show that system (6.3) is contractive with respect to norm $\|\cdot\|_{2,P^{1/2}}$. Thus, we need to show

$$\eta := \begin{bmatrix} x_1 - x_2 \\ \nu_1 - \nu_2 \end{bmatrix}^\top P(F_{\mathrm{PD}}(x_1, \nu_1) - F_{\mathrm{PD}}(x_2, \nu_2)) + c \begin{bmatrix} x_1 - x_2 \\ \nu_1 - \nu_2 \end{bmatrix}^\top P \begin{bmatrix} x_1 - x_2 \\ \nu_1 - \nu_2 \end{bmatrix} \leq 0$$

for any $x_1, x_2 \in \mathbb{R}^n$ and $\nu_1, \nu_2 \in \mathbb{R}^m$, and some constant $c > 0$ which will be the contraction rate. After completing squares, using the strong convexity and Lipschitz smoothness of f, along with $\sigma^2_{\min}(A) I_k \preceq AA^\top$ and $A^\top A \preceq \sigma^2_{\max}(A) I_n$, we obtain

$$\begin{aligned}\eta \leq & -\left((3\alpha/4)\sigma^2_{\min}(A) - c - c\alpha\right) \|\nu_1 - \nu_2\|_2^2 - (\ell_{\inf} \\ & - \alpha\sigma^2_{\max}(A) - c - \alpha\ell^2_{\sup} - c\alpha\sigma^2_{\max}(A)) \|x_1 - x_2\|_2^2 \\ & - \alpha c \|(\nu_1 - \nu_2) - A(x_1 - x_2)\|_2^2 .\end{aligned}$$

Set $c = D\alpha$ for some $D > 0$. Then, to ensure that $\eta \leq 0$, we need to ensure

$$\frac{3\alpha}{4}\sigma^2_{\min}(A) - D\alpha - D\alpha^2 \geq 0, \tag{6.7}$$

$$\ell_{\inf} - \alpha\sigma^2_{\max}(A) - D\alpha - \alpha\ell^2_{\sup} - D\alpha^2\sigma^2_{\max}(A) \geq 0. \tag{6.8}$$

Now, to ensure inequality (6.7) holds, using the inequalities (6.6), it is easy to see that it suffices to ensure that

$$\frac{3\sigma_{\max}(A)\sigma^2_{\min}(A)}{4(\sigma_{\max}(A) + 1)} > D. \tag{6.9}$$

Now, using inequalities (6.6) and (6.9), we obtain: $\ell_{\inf} - \alpha\sigma^2_{\max}(A) - D\alpha - \alpha\ell^2_{\sup} - D\alpha^2\sigma^2_{\max}(A) > \ell_{\inf} - \alpha(\sigma^2_{\max}(A) + \frac{3}{4}\sigma_{\max}(A)\sigma^2_{\min}(A) + \ell^2_{\sup})$ and so, to ensure inequality (6.8) holds, it suffices that

$$\frac{\ell_{\inf}}{\sigma^2_{\max}(A) + \frac{3}{4}\sigma_{\max}(A)\sigma^2_{\min}(A) + \ell^2_{\sup}} > \alpha. \tag{6.10}$$

Now, the parameter α needs to satisfy inequalities (6.6) and (6.10); however, (6.10) implies (6.6) because the inequality $\pi_1^2 + \pi_2^2 \geq 2\pi_1\pi_2$ for $\pi_1, \pi_2 > 0$ let us conclude that $\frac{\ell_{\sup}}{\sigma^2_{\max}(A) + \ell^2_{\sup}} \leq \frac{1}{2\sigma_{\max}(A)}$. Finally, c must be less than the multiplication of the left-hand sides of the inequalities (6.9) and (6.10), which proves statement (ii)a.

Now, since the dynamics are contractive, there must exist a unique globally exponentially stable equilibrium point which also satisfies the (sufficient and necessary) KKT conditions of optimality for the optimization problem (6.2), thus proving statement (ii)b.

■

Remark 6.4.2 *Theorem 6.4.1 is a fundamental building block for the rest of results in the results in this paper and therefore, it was necessary to provide a comprehensive proof using the integral contractivity condition that could provide an explicit estimate of the contraction rate (as opposed to the different proof in [132]).*

For the case of convex f, Theorem 6.4.1 does not state convergence - nor contraction - without additional assumptions; indeed, oscillations may appear and convergence to the saddle points is not guaranteed [66]. In order to still be able to use Theorem 6.4.1 in this case, we consider a modification to the Lagrangian, known as the *augmented Lagrangian* [149]: $\mathcal{L}_{\text{aug}}(x,\nu) = \mathcal{L}(x,\nu) + \frac{\rho}{2}\|Ax-b\|_2^2$ with gain $\rho > 0$. Its associated *augmented primal-dual dynamics* become

$$\begin{bmatrix} \dot{x} \\ \dot{\nu} \end{bmatrix} = \begin{bmatrix} -\nabla f(x) - A^\top \nu - \rho A^\top A x + \rho A^\top b \\ Ax - b \end{bmatrix} \tag{6.11}$$

and have the same equilibria as the original one in (6.3). We introduce two possible sets of assumptions:

(A3) the primal-dual dynamics (6.3) have an equilibrium (x^*, ν^*), $\nabla^2 f(x^*) \succeq \mathbb{0}_{n\times n}$, and $\ker(\nabla^2 f(x^*)) \cap \ker(A) = \{\mathbb{0}_n\}$;

(A4) $\ker(\nabla^2 f(x)) \cap \ker(A) = \{\mathbb{0}_n\}$ for any $x \in \mathbb{R}^n$ and f is Lipschitz smooth with constant $\ell_{\text{sup}} > 0$, and, for $0 < \epsilon < 1$, we define

$$\bar{\alpha}_\epsilon := \frac{\epsilon\rho\sigma_{\min}^2(A)}{(1+\rho)\sigma_{\max}^2(A) + \frac{3}{4}\sigma_{\max}(A)\sigma_{\min}^2(A) + \ell_{\sup}^2}$$
$$\bar{P} := \begin{bmatrix} I_n & \bar{\alpha}_\epsilon\, A^\top \\ \bar{\alpha}_\epsilon\, A & I_k \end{bmatrix} \in \mathbb{R}^{(n+k)\times(n+k)}. \tag{6.12}$$

Corollary 6.4.3 (Contraction analysis of the augmented primal-dual dynamics) *Consider the constrained optimization problem* (6.2)*, its standing assumptions, and its associated augmented primal-dual dynamics* (6.11) *with* $\rho > 0$*.*

(i) Under Assumption (A3)*, the augmented primal-dual dynamics are weakly contractive with respect to* $\|\cdot\|_2$ *and* (x^*, ν^*) *is globally asymptotically stable.*

(ii) Under Assumption (A4)*,*

(a) the augmented primal-dual dynamics are contractive with respect to $\|\cdot\|_{2,\bar{P}^{1/2}}$ *with contraction rate*

$$\bar{\alpha}_\epsilon \frac{3}{4} \frac{\sigma_{\max}(A)\sigma_{\min}^2(A)}{\sigma_{\max}(A)+1}, \quad \textit{and} \tag{6.13}$$

(b) there exists a unique globally exponentially stable equilibrium point (x^*, ν^*) *for the augmented primal-dual dynamics and* x^* *is the unique solution to the constrained optimization problem* (6.2)*.*

Proof: The proof follows directly from Theorem 6.4.1. For statement (i), note that $\ker(\nabla^2 f(x^*)) \cap \ker(A) = \{\mathbb{0}_n\}$ implies that $\nabla^2 f(x^*) + \rho A^\top A \succ \mathbb{0}_{n\times n}$ for the Jacobian of the system

$$\begin{bmatrix} -\nabla^2 f(x) - \rho A^\top A & -A^\top \\ A & 0 \end{bmatrix}.$$

For statement (ii), note that $\ker(\nabla^2 f(x)) \cap \ker(A) = \{\mathbb{0}_n\}$ for any $x \in \mathbb{R}^n$ implies that $x \mapsto f(x) + \frac{\varrho}{2} x^\top A^\top A x$ is Lipschitz smooth with constant $\ell_{\sup}^2 + \rho\sigma_{\max}^2(A) > 0$ and strongly convex with constant $\rho\sigma_{\min}^2(A) > 0$. ■

Remark 6.4.4 (Augmented Lagrangian and contraction) *The benefit of using the augmented Lagrangian is that, unlike the conditions in Theorem 6.4.1, the primal-dual dynamics may be contractive despite f being only convex.*

6.5 Distributed algorithms

We study a popular distributed implementation for solving an unconstrained optimization problem [176]. We want to solve the problem $\min_{x\in\mathbb{R}^n} f(x) = \sum_{i=1}^N f_i(x)$ with $f_i : \mathbb{R}^n \to \mathbb{R}$ convex. Let $\mathcal{G}$ be an undirected connected interaction graph between N distinct agents. Let $\mathcal{N}_i$ be the neighborhood of node i and L be the Laplacian matrix of $\mathcal{G}$. Let $x^i \in \mathbb{R}^n$ be the state associated to agent i, and let $\mathbf{x} = (x^1, \dots, x^N)^\top$. Then, the problem becomes:

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^{nN}} \quad & \sum_{i=1}^N f_i(x^i) \\ & (L \otimes I_n)\mathbf{x} = \mathbb{0}_{nN} \end{aligned} \quad . \tag{6.14}$$

The associated *distributed primal-dual dynamics* are

$$\begin{aligned} \dot{x}^i &= -\nabla_{x^i} f_i(x^i) - \sum_{j\in\mathcal{N}_i} (\nu^j - \nu^i) \\ \dot{\nu}^i &= \sum_{j\in\mathcal{N}_i} (x^j - x^i) \end{aligned} \tag{6.15}$$

for $i \in \{1, \dots, N\}$. In system (6.15), any agent only uses information from herself and the set of her neighbors.

To study this system, we introduce two possible sets of assumptions:

(A5) $\min_{x\in\mathbb{R}^n} f(x)$ has a solution x^* and $\nabla^2 f_i(x^*) \succ \mathbb{0}_{n\times n}$ for any $i \in \{1,\dots,N\}$;

(A6) $\min_{x\in\mathbb{R}^n} f(x)$ has a solution x^* and the function f_i is strongly convex with constant $\ell_{\text{inf},i} > 0$ and Lipschitz smooth with constant $\ell_{\text{sup},i} > 0$ for any $i \in \{1,\dots,N\}$, with $\ell_{\text{inf}} = (\ell_{\text{inf},1},\dots,\ell_{\text{inf},N})$ and $\ell_{\text{sup}} = (\ell_{\text{sup},1},\dots,\ell_{\text{sup},N})$.

With either assumption, note that we cannot apply Theorem 6.4.1 directly since the linear equality constraint in (6.14) is not full-row rank. However, if we instead consider partial contraction, then Theorem 6.4.1 can be used to prove the next result.

Theorem 6.5.1 (Contraction analysis of distributed dynamics) *Consider the distributed primal-dual dynamics* (6.15).

(i) The distributed primal-dual dynamics are weakly contractive with respect to $\|\cdot\|_2$, and

(ii) under Assumption (A5)*, for any $(x^i(0),\nu^i(0)) \in \mathbb{R}^n \times \mathbb{R}^n$, $\lim_{t\to\infty} x^i(t) = x^*$ and $\lim_{t\to\infty} \nu^i(t) = \nu_i^*$, for some ν_i^* such that $\sum_{k=1}^N \nu_k^* = \sum_{k=1}^N \nu^k(0)$.*

(iii) Under Assumption (A6)*, the convergence results in statement (ii) hold and, for $0 < \epsilon < 1$, the convergence of $(\boldsymbol{x}(t),\nu(t))^\top$ has exponential rate*

$$\frac{3\epsilon}{4}\frac{\lambda_N\lambda_2^2}{\lambda_N+1}\frac{\min_{i\in\{1,\dots,N\}}\ell_{\text{inf},i}}{\lambda_N^2+\frac{3}{4}\lambda_N\lambda_2^2+\|\ell_{\text{sup}}\|_\infty^2}, \tag{6.16}$$

where λ_2 and λ_N are the smallest non-zero and the largest eigenvalues of L, respectively.

Proof: Set $f(\mathbf{x}) = \sum_{i=1}^N f_i(x^i)$ and $\nu = (\nu^1,\dots,\nu^N)$. Succinctly, the dynamics of

the system are

$$\begin{aligned} \dot{\mathbf{x}} &= -\nabla f(\mathbf{x}) - (L \otimes I_n)\nu \\ \dot{\nu} &= (L \otimes I_n)\mathbf{x} \end{aligned}. \tag{6.17}$$

Now, let $\bar{A} := (L \otimes I_n)$ and $(\dot{\mathbf{x}}, \dot{\nu}) := F_{\text{PD-d}}(\mathbf{x}, \nu)$, and so

$$DF_{\text{PD-d}}(\mathbf{x}, \nu) = \begin{bmatrix} -\nabla^2 f(\mathbf{x}) & -\bar{A}^\top \\ \bar{A} & \mathbb{0}_{m\times m} \end{bmatrix}.$$

Since $-\nabla^2 f(\mathbf{x}) \preceq 0_{nN\times nN}$ because of convex f_i, it follows that $\mu_2(DF_{\text{PD-d}}(\mathbf{x}, \nu)) = 0$ for any $\mathbf{x} \in \mathbb{R}^{nN}, \nu \in \mathbb{R}^{nN}$, and the system is weakly contractive, which proves (i).

Consider the equilibrium equations of (6.17), and let $(\mathbf{x}^*, \nu^*)$ be a (candidate) fixed point of the system. From the second equation in (6.17), $\mathbf{x}^* = \mathbb{1}_N \otimes v$ with $v \in \mathbb{R}^n$. Now, from the first equation in (6.17), we get $\mathbb{0}_{nN} = \nabla f(\mathbf{x}^*) + (L \otimes I_n)\nu^*$ and left multiplying by $\mathbb{1}_N^\top \otimes I_n$, we obtain $\mathbb{0}_n = \sum_{i=1}^N \nabla f_i(v)$. This is exactly the necessary and sufficient conditions of optimality for the problem $\min_{x\in\mathbb{R}^n} f(x) = \sum_{i=1}^N f_i(x)$, and so $v = x^*$ is an optimal solution to this problem. Moreover, ν^* is just some Lagrange multiplier for the constraint in (6.14).

Now, define the change of coordinates $(\mathbf{x}', \nu') = (\mathbf{x} - \mathbf{x}^*, \nu - \nu^*)$, then we get $(\dot{\mathbf{x}}', \dot{\nu}') = (\dot{\mathbf{x}}, \dot{\nu}) = F_{\text{PD-d}}(\mathbf{x}' + \mathbf{x}^*, \nu' + \nu^*)$, and whenever we refer to the word "system" for the rest of the proof, we refer to the dynamics after this coordinate change. Observe the system has an equilibrium point $(\mathbb{0}_{nN}, \mathbb{0}_{nN})$, but it is not unique; in fact, it is easy to verify that the following is a linear subspace of equilibria for the system: $\mathcal{M} = \left\{(\mathbf{x}', \nu') \in \mathbb{R}^{nN} \times \mathbb{R}^{nN} \mid \mathbf{x}' = \mathbb{0}_{nN}, \nu' = \mathbb{1}_N \otimes \alpha \text{ with } \alpha \in \mathbb{R}^n\right\}$. As a corollary, the subspace $\mathcal{M}$ is flow-invariant for the distributed system.

Now, since L has $N - 1$ strictly positive eigenvalues, we can write them as $0 =$

$\lambda_1 < \lambda_2 \leq \cdots \leq \lambda_N$, and, by eigendecomposition, we can obtain an orthogonal matrix $R' \in \mathbb{R}^{N\times N}$ such that $R'LR'^{\top} = \mathrm{diag}(0, \lambda_2, \ldots, \lambda_N)$. From here, we obtain the matrix $R \in \mathbb{R}^{N-1\times N}$ as a submatrix of R' such that $RLR^{\top} = \Lambda$ with $\Lambda = \mathrm{diag}(\lambda_2, \ldots, \lambda_N)$ with the properties: $RL = \Lambda R$, $RR^{\top} = I_{N-1}$ and $R^{\top}R \neq I_N$. Now, define $V = \mathrm{diag}(I_{nN}, (R \otimes I_n))$ which has orthonormal rows and expresses

$$\mathcal{M} = \left\{(\mathbf{x}', \nu') \in \mathbb{R}^{nN} \times \mathbb{R}^{nN} \mid V(\mathbf{x}', \nu')^{\top} = \mathbb{0}_{(2N-1)n}\right\}.$$

Then, we can use Theorem 6.3.1 for stating the convergence of trajectories of the system to $\mathcal{M}$ using partial contraction. First, note that

$$VDF_{\text{PD-d}}(\mathbf{x}', \nu')V^{\top} = \begin{bmatrix} \nabla^2 f(\mathbf{x}' + \mathbf{x}^*) & -((R^{\top}\Lambda) \otimes I_n) \\ ((\Lambda R) \otimes I_n) & \mathbb{0}_{n(N-1)\times n(N-1)} \end{bmatrix},$$

where we have used the fact that $(R \otimes I_n)(L \otimes I_n) = (\Lambda R) \otimes I_n$. Now, set $\bar{A}^* := (\Lambda R) \otimes I_n$ and note that $\sigma_{\min}(\bar{A}^*) = \lambda_2$ and $\sigma_{\max}(\bar{A}^*) = \lambda_N$.

Since $\nabla^2 f(\mathbf{x}' + \mathbf{x}^*) \preceq \mathbb{0}_{nN\times nN}$, it follows that $\mu_2(VDF_{\text{PD-d}}(\mathbf{x}' + \mathbf{x}^*, \nu' + \nu^*)V^{\top}) \leq 0$. Now, since $\nabla^2 f(\mathbf{x}^*) \prec \mathbb{0}_{nN\times nN}$, Proposition 6.2.1 implies that $VDF_{\text{PD-d}}(\mathbf{x}^*, \nu^*)V^{\top}$ is a Hurwitz matrix, which implies that $\mu_2(VDF_{\text{PD-d}}(\mathbf{x}' + \mathbf{x}^*, \nu' + \nu^*)V^{\top}) < 0$ for any $(\mathbf{x}', \nu') \in \mathcal{M}$, and thus result (ii) of Theorem 6.3.1 implies the convergence to the subspace $\mathcal{M}$ (and this implies convergence of the trajectories of the original system to the set $\mathcal{M}' = \left\{(\mathbf{x}, \nu) \in \mathbb{R}^{nN} \times \mathbb{R}^{nN} \mid \mathbf{x} = \mathbf{x}^*, \nu = \mathbb{1}_N \otimes \alpha \text{ with } \alpha \in \mathbb{R}^n\right\}$, i.e., $\mathcal{M}$ is simply the set $\mathcal{M}'$ translated or anchored to the origin). Since $\mathcal{M}$ is a set of equilibria for the system and the system is weakly contractive, result (iii) of Theorem 6.3.1 concludes that any trajectory of the system converges to some equilibrium point in $\mathcal{M}$.

Now, observe that $(\mathbb{1}_N^\top \otimes I_n)\dot{\nu} = (\mathbb{1}_N^\top L \otimes I_n)\mathbf{x} = \mathbb{0}_n$, and so the set

$$\left\{(\mathbf{x},\nu) \in \mathbb{R}^{nN} \times \mathbb{R}^{nN} \mid (\mathbb{1}_N^\top \otimes I_n)\nu = (\mathbb{1}_N^\top \otimes I_n)\nu(0)\right\}$$

is positively-invariant for (6.15). Then, it follows that $\sum_{k=1}^N \nu_i^k(t) = \sum_{k=1}^N \nu_i^k(0)$ for any $i \in \{1,\dots,n\}$ and any $t \geq 0$. Then, since $\lim_{t\to\infty} \nu^i(t) < \infty$, we conclude the proof for statement (ii).

We prove statement (iii). Observe that $\min_{i\in\{1,\dots,N\}} \ell_{\text{inf},i}\ I_{nN\times nN} \preceq \nabla^2 f(\mathbf{x}' + \mathbf{x}^*) \preceq \max_{i\in\{1,\dots,N\}} \ell_{\text{sup},i}\ I_{nN\times nN}$ and that $\bar{A}^*$ is full-row rank since it is easy to verify that $\text{rank}(\Lambda R) = N-1$ and so $\text{rank}((\Lambda R)\otimes I_n) = n(N-1)$. Then, for $0 < \epsilon < 1$, defining $P = \begin{bmatrix} I_{nN} & \alpha_\epsilon\, \bar{A}^{*\top} \\ \alpha_\epsilon\, \bar{A}^* & I_{n(N-1)} \end{bmatrix} \in \mathbb{R}^{nN\times nN}$ and $\alpha_\epsilon := \frac{\epsilon \min_{i\in\{1,\dots,N\}} \ell_{\text{inf},i}}{\sigma_{\max}^2(\bar{A}^*) + \frac{3}{4}\sigma_{\max}(\bar{A}^*)\sigma_{\min}^2(\bar{A}^*) + \|\ell_{\text{sup}}\|_\infty^2} > 0$, we can use Theorem 6.4.1 to conclude that $\mu_{2,P^{1/2}}(VDF_{\text{PD-d}}(\mathbf{x}'+\mathbf{x}^*, \nu'+\nu^*)V^\top) \leq -c$ with c as in equation (6.16) for any positive $\epsilon < 1$. So then, any trajectory $(\mathbf{x}'(t), \nu'(t))$ exponentially converges to the subspace $\mathcal{M}$ with rate c, due to statement (iii) from Theorem 6.3.1. Finally, the proof finishes by following a similar proof to statement (ii). ■

For the case of convex f_i, Theorem 6.5.1 does not state convergence - nor partial contraction - without additional assumptions. Similar to the analysis in Section 6.4, we present an example where augmenting the Lagrangian let us use Theorem 6.5.1. We consider the popular *distributed least-squares problem* [174]. Given a full-column rank matrix $H \in \mathbb{R}^{N\times n}$, $n < N$, it is known that $x^* = (H^\top H)^{-1}H^\top z$ is the unique solution to the least-squares problem $\min_{x\in\mathbb{R}^n} \|z - Hx\|_2^2$, for $z \in \mathbb{R}^N$. An equivalent distributed

version is

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^{nN}} \quad & \sum_{i=1}^{N}(h_i^\top x^i - z_i)^2 \\ & (L\otimes I_n)\mathbf{x} = \mathbb{0}_{nN} \end{aligned} \tag{6.18}$$

with $h_i^\top \in \mathbb{R}^{1\times n}$ being the ith row of the matrix H, $\mathbf{x} = (x^1,\dots,x^N)^\top$ and $z = (z_1,\dots,z_N)^\top$. Notice that $f(\mathbf{x}) = \sum_{i=1}^N |h_i^\top x^i - z_i|^2$ is convex, since

$$\nabla^2 f(\mathbf{x}) = \operatorname{diag}(h_1h_1^\top,\dots,h_Nh_N^\top) \succeq \mathbb{0}_{nN\times nN}.$$

We propose to augment the Lagrangian with the quadratic term $\frac{\rho}{2}\mathbf{x}^\top(L\otimes I_n)\mathbf{x}$ with $\rho>0$ (which does not alter the original saddle points) and obtain

$$\begin{aligned} \dot{x}^i &= -(h_i^\top x^i - z_i)h_i - \rho\sum_{j\in\mathcal{N}_i}(x^j - x^i) - \sum_{j\in\mathcal{N}_i}(\nu^j - \nu^i) \\ \dot{\nu}^i &= \sum_{j\in\mathcal{N}_i}(x^j - x^i) \end{aligned} \tag{6.19}$$

for $i\in\{1,\dots,N\}$. The new algorithm is distributed.

Observe that $\ker(\operatorname{diag}(h_1h_1^\top,\dots,h_Nh_N^\top)) \cap \ker(L\otimes I_n) = \{\mathbb{0}_{nN}\}$ implies $\ell^*_{\inf} I_{nN} \preceq \operatorname{diag}(h_1h_1^\top,\dots,h_Nh_N^\top) + (L\otimes I_n)$ for some constant $\ell^*_{\inf} > 0$. Then, the following follows from Theorem 6.5.1.

Corollary 6.5.2 (Contraction analysis of distributed least-squares) *Consider the system* (6.19)*, and let x^* be the unique solution to the least-squares problem. Then, for any $(x^i(0),\nu^i(0)) \in \mathbb{R}^n\times\mathbb{R}^n$, $\lim_{t\to\infty} x^i(t) = x^*$ and $\lim_{t\to\infty}\nu^i(t) = \nu_i^*$ for some ν_i^* such that $\sum_{k=1}^N \nu_k^* = \sum_{k=1}^N \nu^k(0)$; and, for $0<\epsilon<1$, the convergence of $(\boldsymbol{x}(t),\nu(t))$ has*

exponential rate

$$\epsilon \frac{3}{4} \frac{\lambda_N \lambda_2^2}{\lambda_N + 1} \frac{\ell_{\inf}^*}{\lambda_N^2 + \frac{3}{4}\lambda_N \lambda_2^2 + \left(\lambda_N + \rho \max_i \|h_i\|_2^2\right)^2}$$

where λ_2 and λ_N are the smallest non-zero and the largest eigenvalues of L, respectively.

6.6 Time-varying optimization

6.6.1 Time-varying standard optimization

Our results in Section 6.4 can be used to prove new results for the case where the associated optimization problem is time-varying. Consider

$$\min_{x \in \mathbb{R}^n} \; f(x,t) \quad \text{subject to} \quad Ax = b(t) \tag{6.20}$$

with the following standing assumptions: $A \in \mathbb{R}^{k \times n}$, $k < n$, $b \in \mathbb{R}^k$, A is full-row rank, and, for every $(x,t) \in \mathbb{R}^n \times \mathbb{R}_{\geq 0}$,

(i) $x \mapsto f(x,t)$ is twice continuously differentiable, uniformly strongly convex with constant $\ell_{\inf} > 0$, i.e., $\nabla^2 f(x,t) \succeq \ell_{\inf} I_n$; and uniformly Lipschitz smooth with constant $\ell_{\sup} > 0$, i.e., $\nabla^2 f(x,t) \preceq \ell_{\sup} I_n$;

(ii) $t \mapsto \nabla f(x,t)$ and $t \mapsto b(t)$ are continuously differentiable functions.

The associated *time-varying primal-dual dynamics* are

$$\begin{bmatrix} \dot{x} \\ \dot{\nu} \end{bmatrix} = \begin{bmatrix} -\nabla f(x,t) - A^\top \nu \\ Ax - b(t) \end{bmatrix}. \tag{6.21}$$

Given a fixed time t, let $x^*(t)$ be a solution to the program $\min_{x:Ax=b(t)} f(x,t)$ and $\nu^*(t)$ its associated Lagrange multiplier. From the standing assumptions and Theorem 6.4.1,

for any fixed t, there exists a unique optimizer $(x^*(t), \nu^*(t))$. Then, $(x^*(t), \nu^*(t))_{t\geq 0}$ defines the *optimizer trajectory* of the optimization problem (6.20). The following result establishes the performance of the primal-dual dynamics in tracking the optimizer trajectory.

Theorem 6.6.1 (Contraction analysis of time-varying primal-dual dynamics) *Consider the time-varying optimization problem* (6.20)*, its standing assumptions, and its associated primal-dual dynamics* (6.21)*.*

(i) The primal-dual dynamics are contractive with respect to $\|\cdot\|_{2,P^{1/2}}$ with contraction rate c, where P is the matrix defined in (6.4) *and c is the same contraction rate as in* (6.5) *of Theorem 6.4.1.*

Assume that, for any $t \geq 0$, $\left\|\dot{b}(t)\right\|_2 \leq \beta_1$ and $\left\|\dot{\nabla} f(x,t)\right\|_2 \leq \beta_2$ for some positive constants β_1, β_2, and let $z(t) := (x(t), \nu(t))^\top$ and $z^*(t) := (x^*(t), \nu^*(t))^\top$.

(ii) Then,

$$\|z(t) - z^*(t)\|_{2,P^{1/2}} \leq \left(\|z(0) - z^*(0)\|_{2,P^{1/2}} - \frac{\rho}{c}\right) e^{-ct} + \frac{\rho}{c}, \tag{6.22}$$

i.e., the tracking error is uniformly ultimately bounded by $\frac{\rho}{c}$ with

$$\rho = \lambda_{\max}(P)\left(\frac{\beta_2}{\ell_{\inf}} + \left(\frac{\sigma_{\max}(A)}{\ell_{\inf}} + 1\right)\frac{\ell_{\max}}{\sigma^2_{\min}(A)}\left(\beta_1 + \frac{\sigma_{\max}(A)}{\ell_{\inf}}\beta_2\right)\right).$$

Proof: Let $(\dot{x}, \dot{\nu}) := F_{\text{PD-tv}}(x, \nu, t)$, and so $DF_{\text{PD-tv}}(x, \nu, t) = \begin{bmatrix} -\nabla^2 f(x,t) & -A^\top \\ A & \mathbb{0}_{k\times k} \end{bmatrix}$. Since A is constant and considering item (i) of the standing assumptions, we can finish the proof for statement (i) by following the same proof as in Theorem 6.4.1. Now we prove

statement (ii). Let us fix any $t \geq 0$. Then, the KKT conditions that the optimizers $x^*(t)$ and $\nu^*(t)$ must satisfy (i.e., equivalent to the equilibrium equations of the system (6.21)) are

$$\mathbb{0}_n = -\nabla f(x^*(t), t) - A^\top \nu^*(t) \tag{6.23}$$

$$\mathbb{0}_k = Ax^*(t) - b(t), \tag{6.24}$$

We first show that the curves $t \mapsto x^*(t)$ and $t \mapsto \nu^*(t)$ are continuously differentiable. Define the function $g : \mathbb{R}^{k+n+1} \to \mathbb{R}^{k+n}$ as $g(t, x, \nu) = \begin{bmatrix} -\nabla f(x,t) - A^\top \nu \\ Ax - b(t) \end{bmatrix}$. Since $t \mapsto b(t)$ and $t \mapsto \nabla f(x, t)$ are continuously differentiable, the function g is continuously differentiable on $\mathbb{R}^{n+k+1}$. Moreover, note that $\nabla_{(x,\nu)} g(t, x, \nu) = DF_{\text{PD-tv}}(x, \nu, t)$. By item (i) of the standing assumptions, we know that $-\nabla^2 f(x, t) \preceq -\ell_{\text{inf}} I_n$ and A is full row rank. From Proposition 6.2.1, this implies that $\nabla_{(x,\nu)} g(t, x, \nu)$ is Hurwitz and therefore, nonsingular. Finally, the Implicit Function Theorem [2, Theorem 2.5.7] implies the solutions $t \mapsto x^*(t)$ and $t \mapsto \nu^*(t)$ of the algebraic equations (6.23) and (6.24) are continuously differentiable for any $t \in \mathbb{R}_{\geq 0}$.

Now, observe that equation (6.24) implies $\left\|A\dot{x^*}(t)\right\|_2 \leq \beta_1$. Then, from (6.23)

$$\begin{aligned}
&\implies \mathbb{0}_n = \nabla^2 f(x^*(t), t)\dot{x^*}(t) + \dot{\nabla} f(x^*(t), t) + A^\top \dot{\nu^*}(t) \\
&\implies \mathbb{0}_m = A\dot{x^*}(t) + A(\nabla^2 f(x^*(t), t))^{-1} \dot{\nabla} f(x^*(t), t) + A(\nabla^2 f(x^*(t), t))^{-1} A^\top \dot{\nu^*}(t) \\
&\implies \left\|\dot{\nu^*}(t)\right\|_2 \leq \frac{\ell_{\max}}{\sigma^2_{\min}(A)} \left(\beta_1 + \frac{\sigma_{\max}(A)}{\ell_{\text{inf}}} \beta_2 \right),
\end{aligned}$$

where the first implication follows from differentiation, the second one follows from the Hessian being invertible, and the third one is derived considering that A is full-row rank.

Similarly, we differentiate (6.23) again and obtain

$$\left\|\dot{x^*}(t)\right\|_2 \leq \frac{\beta_2}{\ell_{\inf}} + \frac{\sigma_{\max}(A)}{\ell_{\inf}} \left\|\dot{\nu^*}(t)\right\|_2 .$$

Now, considering the contraction result on item (i), we set

$$\Delta(t) := \left\| \begin{bmatrix} x(t) \\ \nu(t) \end{bmatrix} - \begin{bmatrix} x^*(t) \\ \nu^*(t) \end{bmatrix} \right\|_{2,P^{1/2}}$$

and use [132, Lemma 2] to obtain the following differential inequality $\dot{\Delta}(t) \leq -c\Delta(t) + \left\| \begin{bmatrix} \dot{x^*}(t) \\ \dot{\nu^*}(t) \end{bmatrix} \right\|_{2,P^{1/2}}$. Then, $\dot{\Delta}(t) \leq -c\Delta(t) + \lambda_{\max}(P)(\left\|\dot{x^*}(t)\right\|_2 + \left\|\dot{\nu^*}(t)\right\|_2)$ and using our previous results, we immediately obtain $\dot{\Delta}(t) \leq -c\Delta(t) + \rho$ with ρ as in the theorem statement. Now, observe the function $h(u) = -cu + \rho$ is Lipschitz (since it is a linear function), and we can use the Comparison Lemma [94] to upper bound $\Delta(t)$ by the solution to the differential equation $\dot{u}(t) = -cu(t) + \rho$ for all $t \geq 0$, from which (ii) follows. ■

Remark 6.6.2 *The bounds in the assumptions for statement (ii) in Theorem 6.6.1 ensure that the rate at which the time-varying optimization changes is bounded. Indeed, the right-hand side of equation* (6.22) *is consistent: the larger (lower) these bounds, the larger (lower) the asymptotic tracking error. Moreover, the tracking is better the larger the contraction rate.*

6.6.2 Time-varying distributed optimization

Our partial contraction analysis of Section 6.5 can be extended to prove new results of performance guarantees for the following *time-varying distributed optimization problem*

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^{nN}} \quad & \sum_{i=1}^{N} f_i(x^i,t) \\ & (L\otimes I_n)\mathbf{x} = \mathbb{0}_{nN}, \end{aligned} \tag{6.25}$$

where we consider a time-invariant connected undirected graph whose Laplacian matrix is L, and set $\mathbf{x} = (x^i,\dots,x^N)^\top$ with $x^i \in \mathbb{R}^n$, with the following standing assumptions: for every $(x,t) \in \mathbb{R}^n \times \mathbb{R}_{\geq 0}$, and for any $i \in \{1,\dots,N\}$

(i) $x \mapsto f_i(x,t)$ is twice continuously differentiable, uniformly strongly convex with constant $\ell_{\text{inf},i} > 0$, i.e., $\nabla^2 f_i(x,t) \succeq \ell_{\text{inf},i} I_n$; and uniformly Lipschitz smooth with constant $\ell_{\text{sup},i} > 0$, i.e., $\nabla^2 f_i(x,t) \preceq \ell_{\text{sup},i} I_n$;

(ii) $t \mapsto \nabla f_i(x,t)$ is continuously differentiable.

Then, the associated primal-dual dynamics are

$$\begin{aligned} \dot{x}^i &= -\nabla_{x^i} f_i(x^i,t) - \sum_{j\in\mathcal{N}_i} (\nu^j - \nu^i) \\ \dot{\nu}^i &= \sum_{j\in\mathcal{N}_i} (x^j - x^i) \end{aligned} \tag{6.26}$$

for $i \in \{1,\dots,N\}$. Given a fixed time t, let $\mathbf{x}^*(t) = \mathbb{1}_N \otimes x^*(t)$ with $x^*(t)$ being the unique solution to the program $\min_x \sum_{i=1}^N f_i(x,t)$. Then, $(x^*(t))_t$ is a unique trajectory; however, there may exist multiple trajectories of the dual variables associated to the constraint in (6.25). Let $\nu^*(t) = (\nu^{1*}(t),\dots,\nu^{N*}(t))^\top$ be any dual variable obtained by solving the problem (6.25) for a fixed t. Then, we define the *time-varying set of optimizers*

as:

$$\mathcal{M}(t) = \{(\mathbf{x}, \nu) \in \mathbb{R}^{nN} \times \mathbb{R}^{nN} | V(\mathbf{x} - \mathbb{1}_N \otimes x^*(t), \nu - \nu^*(t))^\top = (\mathbb{0}_{nN}, \mathbb{0}_{n(N-1)})^\top\}$$

where $V = \mathrm{diag}(I_{nN}, R \otimes I_n)$ with $R \in \mathbb{R}^{N-1 \times N}$ as in the proof of Theorem 6.5.1. For convenience, let $\ell_{\inf} = (\ell_{\inf,1}, \ldots, \ell_{\inf,N})$ and $\ell_{\sup} = (\ell_{\sup,1}, \ldots, \ell_{\sup,N})$; and for $0 < \epsilon < 1$, we define

$$\begin{aligned} \tilde{\alpha}_\epsilon &:= \frac{\epsilon \min_{i \in \{1,\ldots,N\}} \ell_{\inf,i}}{\lambda_N^2 + \frac{3}{4}\lambda_N \lambda_2^2 + \|\ell_{\sup}\|_\infty^2} > 0 \\ \tilde{P} &:= \begin{bmatrix} I_{nN} & \alpha_\epsilon \, \bar{A}^{*\top} \\ \alpha_\epsilon \, \bar{A}^* & I_{n(N-1)} \end{bmatrix} \in \mathbb{R}^{nN \times nN} \end{aligned} \tag{6.27}$$

where $\bar{A}^* = (\Lambda R) \otimes I_n$, with $\Lambda = \mathrm{diag}(\lambda_2, \ldots, \lambda_N)$ containing the nonzero eigenvalues of L in nondecreasing order. The following result establishes the perfomance of the primal-dual dynamics at tracking the time-varying set of optimizers.

Theorem 6.6.3 [Contraction analysis of time-varying distributed primal-dual dynamics] *Consider the time-varying optimization problem* (6.25)*, its standing assumptions, and its associated primal-dual dynamics* (6.26)*. Set* $z(t) := V(\boldsymbol{x}(t), \nu(t))^\top$ *and* $z^*(t) := V(\boldsymbol{x}^*(t), \nu^*(t))^\top$.

(i) The system associated to $\dot{z}$ *is contractive with respect to* $\|\cdot\|_{2,\tilde{P}^{1/2}}$ *with rate* $c := \alpha_\epsilon \frac{3}{4} \frac{\lambda_2^2}{\lambda_N + 1}$.

Moreover, for any $t \geq 0$*, if* $\left\| \dot{\nabla} f_i(x,t) \right\|_2 \leq \beta_{1,i}$ *for some positive constant* $\beta_{1,i}$ *and any* $i \in \{1, \ldots, N\}$*, then,*

(ii)

$$\|z(t) - z^*(t)\|_{2,\tilde{P}^{1/2}} \leq \left(\|z(0) - z^*(0)\|_{2,\tilde{P}^{1/2}} - \frac{\rho}{c}\right) e^{-ct} + \frac{\rho}{c}, \tag{6.28}$$

i.e., the tracking error is asymptotically bounded by $\frac{\rho}{c}$ with

$$\rho = \lambda_{\max}(P)\frac{\|\beta_1\|_1}{\|\ell_{\inf}\|_1}N + \lambda_{\max}(P)\frac{\|\beta_1\|_1}{\lambda_2}\left(\frac{\|\ell_{\sup}\|_\infty}{\|\ell_{\inf}\|_1} + 1\right). \tag{6.29}$$

Proof: Define $f(\mathbf{x}(t), t) := \sum_{i=1}^N f_i(x^i(t), t)$; then

$$\dot{z} = \begin{bmatrix} -\nabla f(\mathbf{x}(t), t) - (L \otimes I_n)\nu(t) \\ (\Lambda R \otimes I_n)\mathbf{x}(t) \end{bmatrix}.$$

Then, decomposing $(\mathbf{x}(t), \nu(t))^\top = U(\mathbf{x}(t), \nu(t))^\top + V^\top z$ where $U = I_{n(2N-1)} - V^\top V$ is a projection matrix, we use the chain rule and obtain that the Jacobian for this system is

$$\begin{bmatrix} -\nabla^2 f(\mathbf{x}(t), t) & -(R^\top \Lambda \otimes I_n) \\ (\Lambda R \otimes I_n) & \mathbb{0}_{n(N-1)\times n(N-1)} \end{bmatrix},$$

so then, based on our standing assumptions, using Proposition 6.2.1 and following a similar proof to Theorem 6.5.1, we obtain that this system is contractive as in item (i).

Now we prove statement (ii). The KKT conditions that the optimizers $\mathbf{x}^*(t)$ and $\nu^*(t)$ must satisfy (i.e., the equilibrium equation of the system (6.26)), for any t, are

$$\mathbb{0}_{nN} = -\nabla f(\mathbf{x}^*(t), t) - (L \otimes I_n)\nu^*(t) \tag{6.30}$$

$$\mathbb{0}_{nN} = (L \otimes I_n)\mathbf{x}^*(t). \tag{6.31}$$

Now, observe that (6.31) and (6.30) $\implies \mathbf{x}^*(t) = \mathbb{1}_N \otimes x^*(t)$ with $x^*(t)$ being the first nN coordinates of any element of $\mathcal{M}(t)$. Moreover, by left multiplying (6.31) with $\mathbb{1}_N^\top \otimes I_n$, we obtain that $\mathbb{0}_n = \sum_{i=1}^N \nabla_{x^i} f_i(x^*(t), t)$. Then, the Implicit Function Theorem [2, Theorem 2.5.7] (akin to its use in the proof of Theorem 6.6.1) implies the curve $t \mapsto x^*(t)$ is continuously differentiable for any $t \in \mathbb{R}_{\geq 0}$.

Now, from (6.30) we obtain that $\mathbb{0}_{n(N-1)} = (R \otimes I_n)\nabla f(\mathbf{x}^*(t), t) + (\Lambda \otimes I_n)(R \otimes I_n)\nu^*(t)$. Defining $y^*(t) := (R \otimes I_n)\nu^*(t)$, we get $\mathbb{0}_{n(N-1)} = (R \otimes I_n)\nabla f(\mathbf{x}^*(t), t) + (\Lambda \otimes I_n)y^*(t)$. Again, an application of the Implicit Function Theorem let us conclude that the solution $(\mathbf{x}^*, t) \mapsto y^*(\mathbf{x}^*, t)$ is continuously differentiable for any $(\mathbf{x}^*, t) \in \mathbb{R}^{nN} \times \mathbb{R}_{\geq 0}$; however, since $t \mapsto \mathbf{x}^*(t)$ is continuously differentiable for any $t \in \mathbb{R}_{\geq 0}$, then $t \mapsto y^*(t)$ is continuously differentiable too.

Then, we can differentiate equation (6.30) and left multiply it by $(\mathbb{1}_N^\top \otimes I_n)$ to obtain

$$\dot{x^*}(t) = -\left(\sum_{i=1}^N \nabla_{x_i}^2 f_i(x^*(t))\right)^{-1} \sum_{i=1}^N \dot{\nabla}_{x^i} f_i(x^*(t), t)).$$

Recall that $RL = \Lambda R$. Then, since y^* is continuously differentiable, we differentiate equation (6.30) and left multiply it by $(R \otimes I_n)$ to obtain

$$\dot{y^*}(t) = -(\Lambda^{-1} R \otimes I_n)(\nabla^2 f(\mathbf{x}^*(t), t)(\mathbb{1}_N \otimes h_1(x^*(t)) + \dot{\nabla} f(\mathbf{x}^*(t), t)).$$

Therefore, observe that $\left\|\dot{x^*}(t)\right\|_2 \leq \frac{1}{\sum_{i=1}^N \ell_{\inf,i}} \sum_{i=1}^N \beta_{1,i} = \frac{\|\beta_1\|_1}{\|\ell_{\inf}\|_1}$, and $\dot{\mathbf{x}^*}(t) = \mathbb{1}_N \otimes \dot{x^*}(t)$. Moreover, $\left\|\dot{y^*}(t)\right\|_2 \leq \frac{1}{|\lambda_2|}\left(\|\ell_{\sup}\|_\infty \left\|\dot{\mathbf{x}^*}(t)\right\|_2 + \|\beta_1\|_1\right)$, where we used: $\left\|\dot{\nabla} f(\mathbf{x}^*(t), t)\right\|_2 \leq \sum_{i=1}^N \left\|\dot{\nabla}_{x^i} f_i(x^*(t))\right\|_2$, and $\|(\Lambda^{-1}R) \otimes I_n\|_2 = \sqrt{\lambda_{\max}(\Lambda^{-2} \otimes I_n)} = \frac{1}{\lambda_2}$.

Now, for any t, let $(a_1(t), a_2(t)) \in \mathcal{M}(t)$. Note that, no matter which element of $\mathcal{M}$ we choose, $a_1(t) = \mathbb{1}_N \otimes x^*(t)$ and so it is uniquely defined for any t and we also know is differentiable. Now, note that $a_2(t) = \gamma(t) + \mathbb{1}_N \otimes \alpha$, with $\alpha \in \mathbb{R}^n$ and some uniquely

defined $\gamma(t)$; and note that $(R\otimes I_n)a_2(t) = (R\otimes I_n)\gamma(t)$ for any t. Therefore $(R\otimes I_n)a_2(t)$ is uniquely defined for any t and we also know is differentiable. In conclusion, the trajectory $((a_1(t), (R \otimes I_n)a_2(t)))_{t\geq 0} = \left(V(a_1(t), a_2(t))^\top\right)_{t\geq 0}$ is unique and $t \mapsto V(a_1(t), a_2(t))^\top$ is a continuously differentiable curve.

Since the system associated to $\dot{z}$ is contractive and the curve, as we just proved above, $t \mapsto z^*(t) := V(\mathbf{x}^*(t), \nu^*(t))^\top$ is unique and differentiable, we set $\Delta(t) := \|z(t) - z^*(t)\|_{2,P^{1/2}}$ and use the result in item (i) and [132, Lemma 2] to obtain the differential inequality

$$\begin{aligned}\dot{\Delta}(t) &\leq -c\Delta(t) + \left\| \begin{bmatrix} \dot{\mathbf{x}^*}(t) \\ \frac{d}{dt}\left((R \otimes I_n)\nu^*(t)\right) \end{bmatrix} \right\|_{2,P^{1/2}} \\ &\leq -c\Delta(t) + \lambda_{\max}(P)(N \left\|\dot{x^*}(t)\right\|_2 + \left\|\dot{z^*}(t)\right\|_2).\end{aligned}$$

Finally, replacing our previous results and using the Comparison Lemma [94] conclude the proof for (ii). ■

Remark 6.6.4 *As in Remark 6.6.2, there is consistency on the right-hand side of equation* (6.29).

6.7 Conclusion

Primal-dual (PD) dynamics associated to linear equality constrained optimization problems are studied in centralized, distributed and time-varying cases. Contraction theory provides an overarching analysis of the dynamical behavior and performance for all these cases of PD dynamics. As future work, we plan to design controllers that can improve the PD solver's tracking properties in the time-varying setting. We also plan to study distributed PD solvers for globally coupled linear equation constraints and PD

solvers in nonsmooth domains.

6.8 Appendix

6.8.1 Proof of Proposition 6.2.1

We remark that the proof of Proposition 6.2.1 is complementary to the one given (for a slightly different case) in [23, Theorem 3.6]). *Proof:* Let $P := \begin{bmatrix} -B & -A^\top \\ A & \mathbb{0}_{m\times m} \end{bmatrix}$ be the matrix in the proposition statement. First, note that $\Re(\lambda(P)) \leq \mu_2(P) = 0$. Therefore, every eigenvalue of P has non-positive real part. We first show that P has no eigenvalue equal to zero. Note that by the Schur complement determinant identity, we have $\det(P) = \det(-B))\det(-AB^{-1}A^\top)$. Note that $B \succeq b_1 I_n$, therefore $\det(-B) \neq 0$. Also, note that $B^{-1} \preceq b_2^{-1} I_n$; and thus $AB^{-1}A^\top \preceq A(b_2^{-1}I_n)A^\top = b_2^{-1}AA^\top \succ 0$, where the last inequality follows from $AA^\top$ being invertible. This implies that $\det(-AB^{-1}A^\top) \neq 0$. As a result, $\det(P) \neq 0$ and P has no zero eigenvalue. Now we show that P is Hurwitz. Assume that $\lambda = \mathrm{i}\eta$ is an eigenvalue of P with zero real part. This means that, there exists $u \in \mathbb{C}^n$ and $v \in \mathbb{C}^m$ such that

$$\begin{bmatrix} -B & -A^\top \\ A & \mathbb{0}_{m\times m} \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = \mathrm{i}\eta \begin{bmatrix} u \\ v \end{bmatrix}. \tag{6.32}$$

Multiplying this equation from the left by $[u^H, v^H]$, we get

$$\Re\left(\begin{bmatrix} u^H & v^H \end{bmatrix} \begin{bmatrix} -B & -A^\top \\ A & \mathbb{0}_{m\times m} \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} \right) = 0.$$

This implies that $\Re\left(u^H B u\right) = 0$. Assume that $u = \theta_1 + \mathrm{i}\theta_2$, where $\theta_1, \theta_2 \in \mathbb{R}^n$. Then $\Re\left(u^H B u\right) = 0$ is equivalent to $\theta_1^T B \theta_1 + \theta_2^\top B \theta_2 = 0$. Since $B \succeq b_1 I_n$, we get that $u = \mathbb{0}_n$. As a result, the equation (6.32) can be written as the system $A^\top v = \mathbb{0}_n$ and $v = \mathbb{0}_k$. This implies that if $\begin{bmatrix} u \\ v \end{bmatrix} \in \mathbb{C}^{n+m}$ is an eigenvector associated to the eigenvalue $\lambda = \mathrm{i}\eta$, then $\begin{bmatrix} u \\ v \end{bmatrix} = \mathbb{0}_{n+m}$. Thus, the matrix P has no eigenvalue with zero real part. Therefore, the real part of every eigenvalue of P is negative and the matrix P is Hurwitz. ■

6.8.2 A simple generalization of [113, Lemma 6]

Lemma 6.8.1 (Convergence of weakly contractive systems) *Consider the dynamical system $\dot{x} = f(x,t)$, $x \in \mathbb{R}^n$, where f is continuously-differentiable with respect to x and weakly contractive respect to some norm $\|\cdot\|$, and let x^* be an equilibrium for the system, i.e., $f(x^*,t) = \mathbb{0}_n$, for every $t \geq 0$. Then x^* is locally asymptotically stable if and only if it is globally asymptotically stable.*

Proof: We only prove the nontrivial implication: if x^* is locally asymptotically stable then it is globally asymptotically stable. Since x^* is a locally asymptotically stable equilibrium point for the dynamical system, then there exists $\epsilon > 0$, such that, for every $y \in \overline{B}(x^*, \epsilon)$ we have $\phi(t, 0, y) \to x^*$ as $t \to \infty$. Note that, for every $z \in \overline{B}(x^*, \epsilon)$, there exists T_z such that $\phi(T_z, 0, z) \in \overline{B}(x^*, \epsilon/2)$. Using the fact that the closed ball $\overline{B}(x^*, \epsilon)$ is compact, we get that, there exists T such that, for every $z \in \overline{B}(x^*, \epsilon)$, we have $\phi(T, 0, z) \in \overline{B}(x^*, \epsilon/2)$. Suppose that $t \mapsto x(t)$ is a trajectory of the dynamical system. Assume that $y \in \partial B(x^*, \epsilon)$ is a point on the straight line connecting $x(0)$ to the unique equilibrium point x^*. Then we have $\|x(T) - x^*\| \leq \|x(T) - \phi(T, 0, y)\| + \|\phi(T, 0, y) - x^*\| \leq \|x(0) - y\| + \epsilon/2 = \|x(0) - x^*\| - \epsilon/2$. Therefore, after time T,

$t \mapsto \|x(t) - x^*\|$ decreases by $\epsilon/2$. As a result, there exists a finite time T_{inf} such that, for every $t \geq T_{\text{inf}}$, we have $x(t) \in \overline{B}(x^*, \epsilon)$. Since $\overline{B}(x^*, \epsilon)$ is in the region of attraction of x^* the trajectory $t \mapsto x(t)$ converges to x^*. ■

Bibliography

[1] R. P. Abelson and J. C. Miller. Negative persuasion via personal insult. *Journal of Experimental Social Psychology*, 3(4):321–333, 1967. doi:10.1016/0022-1031(67)90001-7.

[2] R. Abraham, J. E. Marsden, and T. S. Ratiu. *Manifolds, Tensor Analysis, and Applications*, volume 75 of *Applied Mathematical Sciences*. Springer, 2 edition, 1988.

[3] P.-A. Absil, R. Mahony, and B. Andrews. Convergence of the iterates of descent methods for analytic cost functions. *SIAM Journal on Control and Optimization*, 6(2):531–547, 2005. doi:10.1137/040605266.

[4] D. Acemoglu, G. Como, F. Fagnani, and A. Ozdaglar. Opinion fluctuations and disagreement in social networks. *Mathematics of Operation Research*, 38(1):1–27, 2013. doi:10.1287/moor.1120.0570.

[5] D. Acemoglu and A. Ozdaglar. Opinion dynamics and learning in social networks. *Dynamic Games and Applications*, 1(1):3–49, 2011. doi:10.1007/s13235-010-0004-1.

[6] M. Agueh and G. Carlier. Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, 43:904–924, 2011. doi:10.1137/100805741.

[7] C. Altafini. Consensus problems on networks with antagonistic interactions. *IEEE Transactions on Automatic Control*, 58(4):935–946, 2013. doi:10.1109/TAC.2012.2224251.

[8] P. C. Álvarez Esteban, E. del Barrio, J. A. Cuesta-Albertos, and C. Matrán. A fixed-point approach to barycenters in Wasserstein space. *Journal of Mathematical Analysis and Applications*, 441(2):744–762, 2016. doi:10.1016/j.jmaa.2016.04.045.

[9] Z. Aminzare, Y. Shafi, M. Arcak, and E. D. Sontag. Guaranteeing spatial uniformity in reaction-diffusion systems using weighted L_2 norm contractions. In *A*

Systems Theoretic Approach to Systems and Synthetic Biology I: Models and System Characterizations, chapter 3, pages 73–101. Springer, 2014. doi:10.1007/978-94-017-9041-3_3.

[10] Z. Aminzare and E. D. Sontag. Logarithmic Lipschitz norms and diffusion-induced instability. *Nonlinear Analysis: Theory, Methods & Applications*, 83:31–49, 2013. doi:10.1016/j.na.2013.01.001.

[11] Z. Aminzare and E. D. Sontag. Contraction methods for nonlinear systems: A brief introduction and some open problems. In *IEEE Conf. on Decision and Control*, pages 3835–3847, December 2014. doi:10.1109/CDC.2014.7039986.

[12] Z. Aminzare and E. D. Sontag. Synchronization of diffusively-connected nonlinear systems: Results based on contractions with respect to general norms. *IEEE Transactions on Network Science and Engineering*, 1(2):91–106, 2014. doi:10.1109/TNSE.2015.2395075.

[13] E. Anderes, S. Borgwardt, and J. Miller. Discrete Wasserstein barycenters: Optimal transport for discrete data. *Mathematical Methods of Operations Research*, 85:389–409, 2016. doi:10.1007/s00186-016-0549-x.

[14] T. Antal, P. L. Krapivsky, and S. Redner. Dynamics of social balance on networks. *Physical Review E*, 72(3):036121, 2005. doi:10.1103/PhysRevE.72.036121.

[15] T. Antal, P. L. Krapivsky, and S. Redner. Social balance on networks: The dynamics of friendship and enmity. *Physica D: Nonlinear Phenomena*, 224(1):130–136, 2006. doi:10.1016/j.physd.2006.09.028.

[16] M. Arcak. Certifying spatially uniform behavior in reaction-diffusion PDE and compartmental ODE systems. *Automatica*, 47(6):1219–1229, 2011. doi:10.1016/j.automatica.2011.01.010.

[17] A. Aron and G. Lewandowski. Psychology of interpersonal attraction. In N. J. Smelser and P. B. Baltes, editors, *International Encyclopedia of the Social & Behavioral Sciences*, pages 7860–7862. Pergamon, 2001. doi:10.1016/B0-08-043076-7/01787-3.

[18] K. J. Arrow, L. Hurwicz, and H. Uzawa, editors. *Studies in Linear and Nonlinear Programming*. Standford University Press, 1958.

[19] F. Battiston, G. Cencetti, I. Iacopini, V. Latora, M. L., A. Patania, J.-G. Young, and G. Petri. Networks beyond pairwise interactions: Structure and dynamics. *Physics Reports*, 874:1–92, 2020. doi:10.1016/j.physrep.2020.05.004.

[20] M. Baum, P. K. Willett, and U. D. Hanebeck. On Wasserstein barycenters and MMOSPA estimation. *IEEE Signal Processing Letters*, 22(10):1511–1515, 2015. doi:10.1109/LSP.2015.2410217.

[21] J.-D. Benamou and Y. Brenier. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numerische Mathematik*, 84:375–393, 2000. doi:10.1007/s002110050002.

[22] A. R. Benson, R. Abebe, M. T. Schaub, A. Jadbabaie, and J. Kleinberg. Simplicial closure and higher-order link prediction. *Proceedings of the National Academy of Sciences*, 115(48):E11221–E11230, 2018. doi:10.1073/pnas.1800683115.

[23] M. Benzi, G. H. Golub, and J. Liesen. Numerical solution of saddle point problems. *Acta Numerica*, 14:1–137, 2005. doi:10.1017/S0962492904000212.

[24] A. N. Bishop. Information fusion via the Wasserstein barycenter in the space of probability measures: Direct fusion of empirical measures and Gaussian fusion with unknown correlation. In *International Conference on Information Fusion*, pages 1–7, 2014.

[25] A. N. Bishop and A. Doucet. Distributed nonlinear consensus in the space of probability measures. *IFAC Proceedings Volumes*, 47(3):8662–8668, 2014. doi:10.3182/20140824-6-ZA-1003.00341.

[26] F. Blanchini and S. Miani. *Set-Theoretic Methods in Control*. Springer, 2015.

[27] E. Boissard, T. Le Gouic, and J.-M. Loubes. Distribution's template estimate with Wasserstein, metrics. *Bernoulli*, 21(2):740–759, 2015. doi:10.3150/13-BEJ585.

[28] P. Bonacich, A. C. Holdren, and M. Johnston. Hyper-edges and multidimensional centrality. *Social Networks*, 26(3):189–203, 2004. doi:10.1016/j.socnet.2004.01.001.

[29] N. Bonneel, G. Peyré, and M. Cuturi. Wasserstein barycentric coordinates: Histogram regression using optimal transport. *ACM Transactions on Graphics*, 35(4), 2016. doi:10.1145/2897824.2925918.

[30] N. Bonneel, J. Rabin, G. Peyré, and H. Pfister. Sliced and radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45, 2015. doi:10.1007/s10851-014-0506-3.

[31] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah. Randomized gossip algorithms. *IEEE Transactions on Information Theory*, 52(6):2508–2530, 2006. doi:10.1109/TIT.2006.874516.

[32] F. Bullo. *Lectures on Network Systems*. Kindle Direct Publishing, 1.3 edition, July 2019. With contributions by J. Cortés, F. Dörfler, and S. Martínez. URL: http://motion.me.ucsb.edu/book-lns.

[33] F. Bullo. *Lectures on Network Systems*. Kindle Direct Publishing, 1.4 edition, July 2020. With contributions by J. Cortés, F. Dörfler, and S. Martínez. URL: http://motion.me.ucsb.edu/book-lns.

[34] G. Buttazzo, L. De Pascale, and P. Gori-Giorgi. Optimal-transport formulation of electronic density-functional theory. *Physical Review A*, 85, 2012. doi:10.1103/PhysRevA.85.062502.

[35] S. Byrne and P. Solomon Hart. The boomerang effect. A synthesis of findings and a preliminary theoretical framework. *Annals of the International Communication Association*, 33(1):3–37, 2009. doi:10.1080/23808985.2009.11679083.

[36] G. Carlier and I. Ekeland. Matching for teams. *Economic Theory*, 42:397–418, 2010. doi:10.1007/s00199-008-0415-z.

[37] G. Carlier, A. Oberman, and E. Oudet. Numerical methods for matching for teams and Wasserstein barycenters. *ESAIM: Mathematical Modelling and Numerical Analysis*, 49:1621–1642, 2015. doi:10.1051/m2an/2015033.

[38] D. Cartwright and F. Harary. Structural balance: A generalization of Heider's theory. *Psychological Review*, 63(5):277, 1956. doi:10.1037/h0046049.

[39] G. Chen, W. Su, W. Mei, and F. Bullo. Convergence properties of the heterogeneous Deffuant-Weisbuch model. *Automatica*, 114:108825, 2020. doi:10.1016/j.automatica.2020.108825.

[40] X. Chen and N. Li. Exponential stability of primal-dual gradient dynamics with non-strong convexity. In *American Control Conference*, pages 1612–1618, 2020. doi:10.23919/ACC45564.2020.9147393.

[41] Y. Chen, T. T. Georgiou, and A. Tannenbaum. Optimal transport for Gaussian mixture models. *IEEE Access*, 7:6269–6278, 2019. doi:10.1109/ACCESS.2018.2889838.

[42] A. Cherukuri, B. Gharesifard, and J. Cortes. Saddle-point dynamics: Conditions for asymptotic stability of saddle points. *SIAM Journal on Control and Optimization*, 55(1):486–511, 2017. doi:10.1137/15M1026924.

[43] P. Cisneros-Velarde and F. Bullo. Signed network formation games and clustering balance. *Dynamic Games and Applications*, 2020. doi:10.1007/s13235-019-00346-8.

[44] P. Cisneros-Velarde, S. Jafarpour, and F. Bullo. Distributed and time-varying primal-dual dynamics via contraction analysis. *IEEE Transactions on Automatic Control*, March 2020. Submitted. URL: https://arxiv.org/pdf/2003.12665.

[45] S. Claici, E. Chien, and J. Solomon. Stochastic Wasserstein barycenters. In *International Conference on Machine Learning*, volume 80, pages 999–1008, 2018.

[46] A. R. Cohen. A dissonance analysis of the boomerang effect. *Journal of Personality*, 30(1):75–88, 1962. doi:10.1111/j.1467-6494.1962.tb02306.x.

[47] S. Coogan. A contractive approach to separable Lyapunov functions for monotone systems. *Automatica*, 106:349–357, 2019. doi:10.1016/j.automatica.2019.05.001.

[48] W. A. Coppel. *Stability and Asymptotic Behavior Of Differential Equations*. Heath, 1965.

[49] J. Cortés and S. K. Niederländer. Distributed coordination for nonsmooth convex optimization via saddle-point dynamics. *Journal of Nonlinear Science*, 29(4):1247–1272, 2019. doi:10.1007/s00332-018-9516-4.

[50] N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1853–1865, 2016. doi:10.1109/TPAMI.2016.2615921.

[51] M. Cuturi and A. Doucet. Fast computation of Wasserstein barycenters. In *International Conference on Machine Learning*, volume 32, pages 685–693, 2014.

[52] M. Cuturi and G. Peyr. A smoothed dual approach for variational Wasserstein problems. *SIAM Journal on Imaging Sciences*, 9:320–343, 2016. doi:doi.org/10.1137/15M1032600.

[53] J. L. Daleckii and M. G. Krein. *Stability of Solutions of Differential Equations in Banach Space*. American Mathematical Society, 2002.

[54] J. A. Davis. Clustering and structural balance in graphs. *Human Relations*, 20(2):181–187, 1967. doi:10.1177/001872676702000206.

[55] G. F. de Arruda, G. Petri, and Y. Moreno. Social contagion models on hypergraphs. *Physical Review Research*, 2, 2020. doi:10.1103/PhysRevResearch.2.023032.

[56] G. Deffuant, D. Neau, F. Amblard, and G. Weisbuch. Mixing beliefs among interacting agents. *Advances in Complex Systems*, 3(1/4):87–98, 2000. doi:10.1142/S0219525900000078.

[57] K. Deimling. *Nonlinear Functional Analysis*. Springer, 1985.

[58] P. DeLellis, M. Di Bernardo, and G. Russo. On QUAD, Lipschitz, and contracting vector fields for consensus and synchronization of networks. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 58(3):576–583, 2011. doi:10.1109/TCSI.2010.2072270.

[59] M. Di Bernardo, D. Fiore, G. Russo, and F. Scafuti. Convergence, consensus and synchronization of complex networks via contraction theory. In J. Lü, X. Yu, G. Chen, and W. Yu, editors, *Complex Systems and Networks: Dynamics, Controls and Applications*, pages 313–339. Springer, 2016. doi:10.1007/978-3-662-47824-0_12.

[60] A. d'Onofrio. A note on the global behaviour of the network-based SIS epidemic model. *Nonlinear Analysis: Real World Applications*, 9(4):1567–1572, 2008. doi:10.1016/j.nonrwa.2007.04.001.

[61] P. Dvurechenskii, D. Dvinskikh, A. Gasnikov, C. Uribe, and A. Nedich. Decentralize and randomize: Faster algorithm for Wasserstein barycenters. In *Advances in Neural Information Processing Systems*, pages 10760–10770. 2018.

[62] D. Easley and J. Kleinberg. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, 2010.

[63] G. Facchetti, G. Iacono, and C. Altafini. Computing global structural balance in large-scale signed social networks. *Proceedings of the National Academy of Sciences*, 108(52):20953–20958, 2011. doi:10.1073/pnas.1109521108.

[64] A. Fall, A. Iggidr, G. Sallet, and J.-J. Tewa. Epidemiological models and Lyapunov functions. *Mathematical Modelling of Natural Phenomena*, 2(1):62–68, 2007. doi:10.1051/mmnp:2008011.

[65] M. Fazlyab, S. Paternain, V. M. Preciado, and A. Ribeiro. Prediction-correction interior-point method for time-varying convex optimization. *IEEE Transactions on Automatic Control*, 63(7):1973–1986, 2018. doi:10.1109/TAC.2017.2760256.

[66] D. Feijer and F. Paganini. Stability of primal–dual gradient dynamics and applications to network optimization. *Automatica*, 46(12):1974–1981, 2010. doi:10.1016/j.automatica.2010.08.011.

[67] L. Festinger. *A Theory of Cognitive Dissonance*. Stanford University Press, 1957.

[68] S. Fiske and S. E. Taylor. *Social Cognition: From Brains to Culture*. Sage, 3 edition, 2003.

[69] N. E. Friedkin. *A Structural Theory of Social Influence*. Cambridge University Press, 1998.

[70] N. E. Friedkin, A. V. Proskurnikov, and F. Bullo. Positive contagion and the macrostructures of generalized balance. *Network Science*, 7(4):445–458, 2019. doi:10.1017/nws.2019.19.

[71] D. Gallaun, C. Seifert, and M. Tautenhahn. Sufficient criteria and sharp geometric conditions for observability in Banach spaces. *SIAM Journal on Control and Optimization*, 58(4):26392657, 2020. doi:10.1137/19M1266769.

[72] S. Galln, J.-M. Loubes, and E. Maza. Statistical properties of the quantile normalization method for density curve alignment. *Mathematical Biosciences*, 242(2):129–142, 2013. doi:10.1016/j.mbs.2012.12.007.

[73] M. Granovetter. Threshold models of collective behavior. *The Americal Journal of Sociology*, 83(6):1420–1443, 1978. doi:10.1086/226707.

[74] F. Harary. On the notion of balance of a signed graph. *Michigan Mathematical Journal*, 2(2):143–146, 1953. doi:10.1307/mmj/1028989917.

[75] A. Hatcher. *Algebraic Topology*. Cambridge University Press, 2002.

[76] F. Heider. Attitudes and cognitive organization. *The Journal of Psychology*, 21(1):107–112, 1946. doi:10.1080/00223980.1946.9917275.

[77] U. Helmke and J. B. Moore. *Optimization and Dynamical Systems*. Springer, 2 edition, 1996.

[78] J. M. Hendrickx. A lifting approach to models of opinion dynamics with antagonisms. In *IEEE Conf. on Decision and Control*, pages 2118–2123, December 2014. doi:10.1109/CDC.2014.7039711.

[79] A. Henrot. *Extremum Problems for Eigenvalues of Elliptic Operators*. Springer, 2006.

[80] H. W. Hethcote. An immunization model for a heterogeneous population. *Theoretical Population Biology*, 14(3):338–349, 1978. doi:10.1016/0040-5809(78)90011-4.

[81] H. W. Hethcote. The mathematics of infectious diseases. *SIAM Review*, 42(4):599–653, 2000. doi:10.1137/S0036144500371907.

[82] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 2nd edition, 2012.

[83] A. R. Hota and S. Sundaram. Game-theoretic vaccination against networked SIS epidemics and impacts of human decision-making. *IEEE Transactions on Control of Network Systems*, 6(4):1461–1472, 2019. doi:10.1109/TCNS.2019.2897904.

[84] C. I. Hovland, O. J. Harvey, and M. Sherif. Assimilation and contrast effects in reactions to communication and attitude change. *The Journal of Abnormal and Social Psychology*, 55(2):244–252, 1957. doi:10.1037/h0048480.

[85] T. Hsing and R. Eubank. *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. John Wiley & Sons, 2015.

[86] H. Huang, J. Tang, L. Liu, J. Luo, and X. Fu. Triadic closure pattern analysis and prediction in social networks. *IEEE Transactions on Knowledge and Data Engineering*, 27(12):3374–3389, 2015. doi:10.1109/TKDE.2015.2453956.

[87] I. Iacopini, G. Petri, A. Barrat, and V. Latora. Simplicial models of social contagion. *Nature Communications*, 10(1):2485, 2019. doi:10.1038/s41467-019-10431-6.

[88] M. O. Jackson and S. Nei. Networks of military alliances, wars, and international trade. *Proceedings of the National Academy of Sciences*, 112(50):15277–15284, 2015. doi:10.1073/pnas.1520970112.

[89] S. Jafarpour, P. Cisneros-Velarde, and F. Bullo. Weak and semi-contraction for network systems and diffusively-coupled oscillators. *IEEE Transactions on Automatic Control*, May 2020. Submitted. URL: http://arxiv.org/pdf/2005.09774.

[90] I. M. James and N. J. Hitchin. *The Topology of Stiefel Manifolds*. Cambridge University Press, 1976.

[91] B. Jhun, M. Jo, and B. Kahng. Simplicial SIS model in scale-free uniform hypergraph. *Journal of Statistical Mechanics: Theory and Experiment*, 2019. doi:10.1088/1742-5468/ab5367.

[92] P. Jia, N. E. Friedkin, and F. Bullo. The coevolution of appraisal and influence networks leads to structural balance. *IEEE Transactions on Network Science and Engineering*, 3(4):286–298, 2016. doi:10.1109/TNSE.2016.2600058.

[93] K. F. Kee, L. Sparks, D. C. Struppa, and M. Mannucci. Social groups, social media, and higher dimensional social structures: A simplicial model of social aggregation for computational communication research. *Communication Quarterly*, 61(1):35–58, 2013. doi:10.1080/01463373.2012.719566.

[94] H. K. Khalil. *Nonlinear Systems*. Prentice Hall, 3 edition, 2002.

[95] A. Khanafer, T. Başar, and B. Gharesifard. Stability of epidemic models over directed graphs: A positive systems approach. *Automatica*, 74:126–134, 2016. doi:10.1016/j.automatica.2016.07.037.

[96] S. S. Kia, J. Cortes, and S. Martinez. Distributed convex optimization via continuous-time coordination algorithms with discrete-time communication. *Automatica*, 55:254–264, 2015. doi:10.1016/j.automatica.2015.03.001.

[97] K. Kułakowski, P. Gawroński, and P. Gronek. The Heider balance: A continuous approach. *International Journal of Modern Physics C*, 16(05):707–716, 2005. doi:10.1142/S012918310500742X.

[98] M. Kurula. Well-posedness of time-varying linear systems. *IEEE Transactions on Automatic Control*, 2019. doi:10.1109/TAC.2019.2954794.

[99] G. E. Ladas and V. Lakshmikantham. *Differential Equations in Abstract Spaces*. Academic Press, 1972.

[100] A. Lajmanovich and J. A. Yorke. A deterministic model for gonorrhea in a nonhomogeneous population. *Mathematical Biosciences*, 28(3):221–236, 1976. doi:10.1016/0025-5564(76)90125-5.

[101] T. Le Gouic and J.-M. Loubes. Existence and consistency of Wasserstein barycenters. *Probability Theory and Related Fields*, 168:901–917, 2017. doi:10.1007/s00440-016-0727-z.

[102] J. M. Lee. *Introduction to Smooth Manifolds*. Springer, 2003.

[103] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Signed networks in social media. In *Int. Conf. on Human Factors in Computing Systems*, pages 1361–1370, Atlanta, USA, 2010. doi:10.1145/1753326.1753532.

[104] Y. Li, W. Chen, Y. Wang, and Z.-L. Zhang. Voter model on signed social networks. *Internet Mathematics*, 11(2):93–133, 2015. doi:10.1080/15427951.2013.862884.

[105] S. Liang, L. Y. Wang, and G. Yin. Exponential convergence of distributed primal-dual convex optimization algorithm without strong convexity. *Automatica*, 105:298–306, 2019. doi:10.1016/j.automatica.2019.04.004.

[106] C.-C. Lin, C.-H. Lee, C.-S. Fuh, H.-F. Juan, and H.-C. Huang. Link clustering reveals structural characteristics and biological contexts in signed molecular networks. *PLoS One*, 8(6):1–9, 06 2013. doi:10.1371/journal.pone.0067089.

[107] X. Lin, Q. Jiao, and L. Wang. Opinion propagation over signed networks: Models and convergence analysis. *IEEE Transactions on Automatic Control*, 64(8):3431–3438, 2019. doi:10.1109/TAC.2018.2879568.

[108] J. Liu, X. Chen, T. Başar, and M.-A. Belabbas. Exponential convergence of the discrete- and continuous-time Altafini models. *IEEE Transactions on Automatic Control*, 62:6168–6182, 2017. doi:10.1109/TAC.2017.2700523.

[109] W.-M. Liu, H. W. Hethcote, and S. A. Levin. Dynamical behavior of epidemiological models with nonlinear incidence rates. *Journal of Mathematical Biology*, 25(4):359–380, 1987. doi:10.1007/BF00277162.

[110] Y. Liu, C. Lageman, B. D.O. Anderson, and G. Shi. An Arrow-Hurwicz-Uzawa type flow as least squares solver for network linear equations. *Automatica*, 100:187–193, 2019. doi:10.1016/j.automatica.2018.10.007.

[111] Y. Liu, Y. Lou, B. D. O. Anderson, and G. Shi. Network flows that solve least squares for linear equations, 2018. arXiv:1808.04140.

[112] W. Lohmiller and J.-J. E. Slotine. On contraction analysis for non-linear systems. *Automatica*, 34(6):683–696, 1998. doi:10.1016/S0005-1098(98)00019-3.

[113] E. Lovisari, G. Como, and K. Savla. Stability of monotone dynamical flow networks. In *IEEE Conf. on Decision and Control*, pages 2384–2389, Los Angeles, USA, December 2014. doi:10.1109/CDC.2014.7039752.

[114] D. G. Luenberger. *Optimization by Vector Space Methods*. John Wiley & Sons, 1969.

[115] M. Malekzadeh, M. Fazli, P. Jalaly Khalidabadi, H. R. Rabiee, and M. A. Safari. Social balance and signed network formation games. In *Proceedings of 5th KDD Workshop on Social Network Analysis (SNA-KDD)*, San Diego, USA, August 2011.

[116] E. Mallada, C. Zhao, and S. Low. Optimal load-side control for frequency regulation in smart grids. *IEEE Transactions on Automatic Control*, 62(12):6294–6309, 2017. doi:10.1109/TAC.2017.2713529.

[117] I. R. Manchester and J.-J. E. Slotine. Control contraction metrics: Convex and intrinsic criteria for nonlinear feedback design. *IEEE Transactions on Automatic Control*, 62(6):3046–3053, 2017. doi:10.1109/TAC.2017.2668380.

[118] M. Martcheva. *An Introduction to Mathematical Epidemiology*. Springer, 2015.

[119] S. A. Marvel, J. Kleinberg, R. D. Kleinberg, and S. H. Strogatz. Continuous-time model of structural balance. *Proceedings of the National Academy of Sciences*, 108(5):1771–1776, 2011. doi:10.1073/pnas.1013213108.

[120] S. A. Marvel, S. H. Strogatz, and J. M. Kleinberg. Energy landscape of social balance. *Physical Review Letters*, 103:198701, 2009. doi:10.1103/PhysRevLett.103.198701.

[121] J. T. Matamalas, S. Gómez, and A. Arenas. Abrupt phase transition of epidemic spreading in simplicial complexes. *Physical Review Research*, 2:012049, 2020. doi:10.1103/PhysRevResearch.2.012049.

[122] W. Mei, P. Cisneros-Velarde, G. Chen, N. E. Friedkin, and F. Bullo. Dynamic social balance and convergent appraisals via homophily and influence mechanisms. *Automatica*, 110:108580, 2019. doi:10.1016/j.automatica.2019.108580.

[123] W. Mei, S. Mohagheghi, S. Zampieri, and F. Bullo. On the dynamics of deterministic epidemic propagation over networks. *Annual Reviews in Control*, 44:116–128, 2017. doi:10.1016/j.arcontrol.2017.09.002.

[124] Ziyang Meng, Guodong Shi, Karl H. Johansson, Ming Cao, and Yiguang Hong. Behaviors of networks with antagonistic interactions and switching topologies. *Automatica*, 73:110 – 116, 2016. doi:https://doi.org/10.1016/j.automatica.2016.06.022.

[125] A. N. Michel, L. Hou, and D. Liu. *Stability of Dynamical Systems*. Springer, 2008.

[126] A. Mironchenko and F. Wirth. Characterizations of input-to-state stability for infinite-dimensional systems. *IEEE Transactions on Automatic Control*, 63(6):1692–1707, 2018. doi:10.1109/TAC.2017.2756341.

[127] Y. Mroueh. Wasserstein style transfer. In *International Conference on Artificial Intelligence and Statistics*, volume 108, pages 842–852, 2020.

[128] J. Munkres. *Topology*. Pearson, 2 edition, 2000.

[129] J. D. Murray. *Mathematical Biology I: An Introduction*. Springer, 3 edition, 2002.

[130] J. Muscat. *Functional Analysis: An Introduction to Metric Spaces, Hilbert Spaces, and Banach Algebras*. Springer, 2014.

[131] P. H. A. Ngoc and H. Trinh. On contraction of functional differential equations. *SIAM Journal on Control and Optimization*, 56(3):2377–2397, 2018. doi:10.1137/16M1092672.

[132] H. D. Nguyen, T. L. Vu, K. Turitsyn, and J.-J. E. Slotine. Contraction and robustness of continuous time primal-dual dynamics. *IEEE Control Systems Letters*, 2(4):755–760, 2018. doi:10.1109/LCSYS.2018.2847408.

[133] C. Nowzari, V. M. Preciado, and G. J. Pappas. Analysis and control of epidemics: A survey of spreading processes on complex networks. *IEEE Control Systems*, 36(1):26–46, 2016. doi:10.1109/MCS.2015.2495000.

[134] P. J. Oakes and J. C. Turner. Is limited information processing capacity the cause of social stereotyping? *European Review of Social Psychology*, 1(1):111–135, 1990. doi:10.1080/14792779108401859.

[135] M. Ogura, V. M. Preciado, and N. Masuda. Optimal containment of epidemics over temporal activity-driven networks. *SIAM Journal on Applied Mathematics*, 79(3), 2019. doi:10.1137/18M1172740.

[136] K. Paarporn, C. Eksin, J. S. Weitz, and J. S. Shamma. Networked SIS epidemics with awareness. *IEEE Transactions on Computational Social Systems*, 4(3):93–103, 2017. doi:10.1109/TCSS.2017.2719585.

[137] V. M. Panaretos and Y. Zemel. *An Invitation to Statistics in Wasserstein Space*. SpringerBriefs in Probability and Mathematical Statistics. Springer, 2020. doi:10.1007/978-3-030-38438-8.

[138] P. E. Paré, C. L. Beck, and A. Nedić. Epidemic processes over time-varying networks. *IEEE Transactions on Control of Network Systems*, 5(3), 2017. doi:10.1109/TCNS.2017.2706138.

[139] P. E. Paré, J. Liu, C. L. Beck, A. Nedić, and T. Başar. Multi-competitive viruses over static and time-varying networks. In *American Control Conference*, pages 1685–1690, 2017. doi:10.23919/ACC.2017.7963195.

[140] R. Pastor-Satorras and A. Vespignani. Epidemic spreading in scale-free networks. *Physical Review Letters*, 86(14):3200–3203, 2001. doi:10.1103/PhysRevLett.86.3200.

[141] A. Pavlov, A. Pogromsky, N. Van de Wouw, and H. Nijmeijer. Convergent dynamics, a tribute to Boris Pavlovich Demidovich. *Systems & Control Letters*, 52(3-4):257–261, 2004. doi:10.1016/j.sysconle.2004.02.003.

[142] Q. C. Pham and J.-J. E. Slotine. Stable concurrent synchronization in dynamic system networks. *Neural Networks*, 20(1):62–77, 2007. doi:10.1016/j.neunet.2006.07.008.

[143] A. Polyakov, J. Coron, and L. Rosier. On homogeneous finite-time control for linear evolution equation in Hilbert space. *IEEE Transactions on Automatic Control*, 63(9):3143–3150, 2018. doi:10.1109/TAC.2018.2797838.

[144] A. V. Proskurnikov and R. Tempo. A tutorial on modeling and analysis of dynamic social networks. Part I. *Annual Reviews in Control*, 43:65–79, 2017. doi:10.1016/j.arcontrol.2017.03.002.

[145] G. Qu and N. Li. On the exponential stability of primal-dual gradient dynamics. *IEEE Control Systems Letters*, 3(1):43–48, 2019. doi:10.1109/LCSYS.2018.2851375.

[146] J. Rabin, G. Peyré, J. Delon, and M. Bernot. Wasserstein barycenter and its application to texture mixing. In *Scale Space and Variational Methods in Computer Vision*, pages 435–446. Springer, 2012. doi:10.1007/978-3-642-24785-9_37.

[147] F. Radicchi, D. Vilone, S. Yoon, and H. Meyer-Ortmanns. Social balance as a satisfiability problem of computer science. *Physical Review E*, 75:026106, 2007. doi:10.1103/PhysRevE.75.026106.

[148] S. Rahili and W. Ren. Distributed continuous-time convex optimization with time-varying cost functions. *IEEE Transactions on Automatic Control*, 62(4):1590–1605, 2017. doi:10.1109/TAC.2016.2593899.

[149] R. T. Rockafellar. Augmented Lagrangians and applications of the proximal point algorithm in convex programming. *Mathematics of Operations Research*, 1(2):97–116, 1976. doi:10.1287/moor.1.2.97.

[150] F. D. Sahneh, C. Scoglio, and P. Van Mieghem. Generalized epidemic mean-field model for spreading processes over multilayer complex networks. *IEEE/ACM Transactions on Networking*, 21(5):1609–1620, 2013. doi:10.1109/TNET.2013.2239658.

[151] F. Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. Progress in Nonlinear Differential Equations and Their Applications. Birkhäuser, 2015. doi:10.1007/978-3-319-20828-2.

[152] M. T. Schaub, A. R. Benson, P. Horn, G. Lippner, and A. Jadbabaie. Random walks on simplicial complexes and the normalized Hodge 1-Laplacian. *SIAM Review*, 62(2):353–391, 2020. doi:10.1137/18M1201019.

[153] M. A. Schmitz, M. Heitz, N. Bonneel, F. Ngolé, D. Coeurjolly, M. Cuturi, G. Peyré, and J.-L. Starck. Wasserstein dictionary learning: Optimal transport-based unsupervised nonlinear dictionary learning. *SIAM Journal on Imaging Sciences*, 11:643–678, 2018. doi:10.1137/17M1140431.

[154] V. Seguy and M. Cuturi. Principal geodesic analysis for probability measures under the optimal transport metric. In *Advances in Neural Information Processing Systems*, pages 3312–3320. 2015.

[155] J. H. Shapiro. *A Fixed-Point Farrago*. Springer, 2016.

[156] G. Shi, C. Altafini, and J. Baras. Dynamics over signed networks. *SIAM Review*, 61(2):229–257, 2019. doi:10.1137/17M1134172.

[157] J. W. Simpson-Porco and F. Bullo. Contraction theory on Riemannian manifolds. *Systems & Control Letters*, 65:74–80, 2014. doi:10.1016/j.sysconle.2013.12.016.

[158] J. W. Simpson-Porco, B. K. Poolla, N. Monshizadeh, and F. Dörfler. Input-output performance of linear-quadratic saddle-point algorithms with application to distributed resource allocation problems. *IEEE Transactions on Automatic Control*, 65(5):2032–2045, 2019. doi:10.1109/TAC.2019.2927328.

[159] S. Srivastava, V. Cevher, Q. Dinh, and D. Dunson. WASP: scalable Bayes via barycenters of subset posteriors. In *International Conference on Artificial Intelligence and Statistics*, volume 38, pages 912–920, 2015.

[160] C. Sun, M. Ye, and G. Hu. Distributed time-varying quadratic optimization for multiple agents under undirected graphs. *IEEE Transactions on Automatic Control*, 62(7):3687–3694, 2017. doi:10.1109/TAC.2017.2673240.

[161] A. Tahbaz-Salehi and A. Jadbabaie. Consensus over ergodic stationary graph processes. *IEEE Transactions on Automatic Control*, 55(1):225–230, 2010. doi:10.1109/TAC.2009.2034054.

[162] R. Tempo, G. Calafiore, and F. Dabbene. *Randomized Algorithms for Analysis and Control of Uncertain Systems*. Springer, 2005.

[163] A. Terrand-Jeanne, V. Andrieu, V. Dos Santos Martins, and C. Xu. Adding integral action for open-loop exponentially stable semigroups and application to boundary control of PDE systems. *IEEE Transactions on Automatic Control*, 2019. doi:10.1109/TAC.2019.2957349.

[164] V. A. Traag, P. Van Dooren, and P. De Leenheer. Dynamical models explaining social balance and evolution of cooperation. *PLoS One*, 8(4):e60063, 2013. doi:10.1371/journal.pone.0060063.

[165] E. Trlat, C. Zhang, and E. Zuazua. Steady-state and periodic exponential turnpike property for optimal control problems in Hilbert spaces. *SIAM Journal on Control and Optimization*, 56(2):12221252, 2018. doi:10.1137/16M1097638.

[166] C. A. Uribe, D. Dvinskikh, P. Dvurechensky, A. Gasnikov, and A. Nedi. Distributed computation of Wasserstein barycenters over networks. In *IEEE Conf. on Decision and Control*, pages 6544–6549, 2018. doi:10.1109/CDC.2018.8619160.

[167] A. van de Rijt. The micro-macro link for the theory of structural balance. *Journal of Mathematical Sociology*, 35(1-3):94–113, 2011. doi:10.1080/0022250X.2010.532262.

[168] M. Vidyasagar. *Nonlinear Systems Analysis*. SIAM, 2002. doi:10.1137/1.9780898719185.

[169] C. Villani. *Topics in Optimal Transportation*. Graduate Studies in Mathematics. American Mathematical Society, 2003.

[170] C. Villani. *Optimal Transport: Old and New*. Springer, 2009. doi:10.1007/978-3-540-71050-9.

[171] M. Wakaiki and H. Sano. Event-triggered control of infinite-dimensional systems. *SIAM Journal on Control and Optimization*, 58(2):605–635, 2020. doi:10.1137/18M1179717.

[172] G. Wang, Y. Wei, and S. Qiao. *Generalized Inverses: Theory and Computations*. Springer, 2018.

[173] J. Wang and N. Elia. A control perspective for centralized and distributed convex optimization. In *IEEE Conf. on Decision and Control and European Control Conference*, pages 3800–3805, Orlando, USA, 2011. doi:10.1109/CDC.2011.6161503.

[174] P. Wang, S. Mou, J. Lian, and W. Ren. Solving a system of linear equations: From centralized to distributed algorithms. *Annual Reviews in Control*, 47:306–322, 2019. doi:10.1016/j.arcontrol.2019.04.008.

[175] N. J. Watkins, C. Nowzari, and G. J. Pappas. Robust economic model predictive control of continuous-time epidemic processes. *IEEE Transactions on Automatic Control*, 65(3):1116–1131, 2020. doi:10.1109/TAC.2019.2919136.

[176] T. Yang, X. Yi, J. Wu, Y. Yuan, D. Wu, Z. Meng, Y. Hong, H. Wang, Z. Lin, and K. H. Johansson. A survey of distributed optimization. *Annual Reviews in Control*, 47:278–305, 2019. doi:10.1016/j.arcontrol.2019.05.006.

[177] X. Zheng, D. Zeng, and F.-Y. Wang. Social balance in signed networks. *Information Systems Frontiers*, 17(5):1077–1095, 2015. doi:10.1007/s10796-014-9483-8.