TÉCNICO LISBOA

UNIVERSIDADE DE LISBOA

INSTITUTO SUPERIOR TÉCNICO

Distributed Banach-Picard Iteration with Applications to Inference Problems

Francisco de Lima Andrade

Supervisor: Doctor Mário Alexandre Teles de Figueiredo

Co-Supervisor: Doctor João Manuel de Freitas Xavier

Thesis approved in public session to obtain the PhD Degree in

Electrical and Computer Engineering

Jury final classification: Pass with Distinction

2022

TÉCNICO LISBOA

UNIVERSIDADE DE LISBOA
INSTITUTO SUPERIOR TÉCNICO

Distributed Banach-Picard Iteration with Applications to Inference Problems

Francisco de Lima Andrade

Supervisor: Doctor Mário Alexandre Teles de Figueiredo

Co-Supervisor: Doctor João Manuel de Freitas Xavier

Thesis approved in public session to obtain the PhD Degree in

Electrical and Computer Engineering

Jury final classification: Pass with Distinction

Jury

Chairperson: Doctor Leonel Augusto Pires Seabra de Sousa, Instituto Superior Técnico, Universidade de Lisboa

Members of the Committee:

Doctor Dušan Jakovetić, Faculty of Sciences, University of Novi Sad, Sérvia

Doctor Panagiotis Patrinos, Faculty of Engineering Science, KU Leuven, Bélgica

Doctor João Manuel de Freitas Xavier, Instituto Superior Técnico, Universidade de Lisboa

Doctor João Pedro Castilho Pereira Santos Gomes, Instituto Superior Técnico, Universidade de Lisboa

Doctor Pedro Tiago Martins Batista, Instituto Superior Técnico, Universidade de Lisboa

Funding Institutions

Fundação para a Ciência e Tecnologia, Grant PD/BD/135185/2017

Instituto de Telecomunicações, Grant O-0025-LX-15

2022

Acknowledgements

I would like to thank both my supervisors, João Xavier and Mário Figueiredo, for all the support and all the fruitful discussions.

Agradeço à minha companheira de vida, viagem e aventura, Léa Autier, sem quem nada disto teria sido possível. Um obrigado também muito especial ao meu pai, Cali, e à minha mãe, Eva, pelo apoio incondicional que sempre me deram. Finalmente, agradeço aos meus tios Zé Miguel e Mali, e à nossa grande amiga de família Norinha.

Resumo

A iteração de Banach-Picard é amplamente utilizada para encontrar pontos fixos de mapas localmente ou globalmente contractivos. O trabalho apresentado nesta tese estende a iteração de Banach-Picard a configurações distribuídas; específicamente, assumimos que o mapa cujo ponto fixo se pretende é uma média de mapas individuais (não necessariamente localmente ou globalmente contractivos) que pertencem a um conjunto de agentes ligados por uma rede de comunicações. Propomos um algoritmo distribuído, denominado *distributed Banach-Picard iteration* (DBPI), e provamos a sua convergência, de facto mostrando que se a média dos mapas individuais é um mapa localmente ou globalmente contractivo, então o mapa subjacente ao DBPI herda a propriedade correspondente. O desafio no caso de um mapa localmente contractivo (LC) é que não é assumido que este mapa emirja de um problema de optimização subjacente, o que impede a exploração de propriedades globais fortes como convexidade ou condições de Lipschitz.

A segunda parte desta tese parte do DBPI e das suas guarantias de convergência local linear para fazer várias contribuições. Mostramos que o algoritmo de Sanger para *análise de components principais* (ACP) corresponde à iteração de um mapa LC que pode ser escrito como uma média de mapas locais, cada mapa sendo conhecido por um agente que detém um subjconjunto dos dados. De forma semelhante, mostramos que uma variante do algoritmo de *expectativa-maximização* (EM) para a estimação de um parâmetro, a partir de medidas com ruído e/ou defeituosas obtidas por uma rede de sensores, pode ser escrita como uma média de mapas locais, cada um dos quais pertencendo a um único sensor. Consequentemente, partir do DBPI, obtemos dois algoritmos distribuídos – EM distribuído e ACP distribuído – cujas guarantias de convergência local linear seguem das guarantias provadas para o DBPI. A verificação da condição LC para a variante do algoritmo EM não é trivial, dado que o operador subjacente depende de amostras aleatórias, implicando, portanto, que a condição LC seja de natureza probabilística.

Palavras-chave: iteração de Banach-Picard, consenso, computação distribuída, estimação distribuída, ACP distribuído.

Abstract

The Banach-Picard iteration is widely used to find fixed points of locally or globally contractive maps. The work presented in this thesis extends the Banach-Picard iteration to distributed settings; specifically, we assume the map of which the fixed point is sought to be the average of individual (not necessarily locally or globally contractive) maps held by a set of agents linked by a communication network. We propose a distributed algorithm, termed *distributed Banach-Picard iteration* (DBPI), and prove its convergence, in fact showing that if the average map is locally or globally contractive, then the map underlying DBPI inherits the corresponding property. The challenge in the locally contractive (LC) case is that the map is not assumed to come from an underlying optimization problem, which prevents exploiting strong global properties such as convexity or Lipschitzianity.

The second part of this thesis builds upon the DBPI and its *local linear convergence* (LLC) guarantees to make several contributions. We show that Sanger's algorithm for *principal component analysis* (PCA) corresponds to the iteration of a LC map that can be written as the average of local maps, each map known to an agent holding a subset of the data. Similarly, we show that a variant of the *expectation-maximization* (EM) algorithm for parameter estimation from noisy and faulty measurements in a sensor network can be written as the iteration of a LC map that is the average of local maps, each available at just one node. Consequently, via the DBPI, we derive two distributed algorithms – distributed EM and distributed PCA – whose LLC guarantees follow from those that we proved for the DBPI. The verification of the LC condition for the variant of the EM algorithm is challenging, as the underlying map depends on random samples, thus the LC condition is of probabilistic nature.

Keywords: Banach-Picard iteration, consensus, distributed computation, distributed parameter estimation, distributed PCA.

Contents

1 Introduction **3**

1.1 Motivation . . . 3
1.2 Distributed Banach-Picard Iteration . . . 4
1.3 Distributed PCA and Distributed Estimation . . . 5
1.4 Contributions and Related Work . . . 5
1.4.1 Theoretical Contributions . . . 6
1.4.2 Remarks . . . 7
1.4.3 Related Work . . . 7
1.4.4 Generalizations . . . 8
1.4.5 Contributions to Applications and Related Work . . . 9
1.4.6 Generalizations . . . 10
1.5 Organization of this Thesis . . . 11

2 Distributed Average Consensus **13**

2.1 Introduction . . . 13
2.2 Problem Statement . . . 13
2.2.1 Basic Notions in Graphs Theory . . . 13
2.2.2 Problem Statement . . . 14
2.2.3 Applications . . . 16
2.3 A Solution to the DAC Problem . . . 17
2.3.1 Example . . . 18
2.3.2 Returning to the Solution . . . 20
2.3.3 The Final Solution . . . 21
2.3.3.1 The Outline of the Proof . . . 22
2.3.3.2 The Formal Proof . . . 22
2.3.4 The Metropolis Weight Matrix and the Consensus Matrices . . . 24
2.4 Comments and References . . . 25
2.4.1 Directed graphs . . . 25
2.4.2 Time-Varying Topologies . . . 26

2.4.3 Finite-Time Consensus 27
2.4.3.1 Flooding 27
2.4.3.2 Linear Finite-Time Consensus 27
2.4.4 Solutions Optimizing the Convergence Rate 28
2.4.5 The Virtues of the Metropolis Weight Matrix 29
2.4.5.1 Knowledge of the Graph Topology 29
2.4.5.2 Amount of Information Transmitted 30
2.4.5.3 Building Block 30

3 Basics of Fixed Point Theory and Problem Statement 31
3.1 Introduction 31
3.2 Basic Definitions and Results 31
3.2.1 Definition of a Fixed Point 31
3.2.2 Existence of a Fixed Point 33
3.2.3 Qualitative Character of Fixed Points 33
3.2.4 Metric Conditions on H and Fundamental Global Results 34
3.2.5 Local Conditions and Results 37
3.3 Problem Statement 41
3.4 Comments and References 42
3.4.1 The Notion of an Attractor 42
3.4.2 Convergence Rate 43
3.4.3 The Relevance of Theorem 3.2.4 44
3.4.4 Distributed Optimization 44

4 A Distributed Algorithm with a Shrinking Step-Size 47
4.1 Introduction 47
4.2 Preliminaries 48
4.3 The Step-Size 50
4.4 The Consensus and Off-Consensus Recursions 51
4.4.1 Consensus and Off-Consensus Recursions of (4.5) 51
4.5 The Global Contraction Case 55
4.6 The Local Contraction Case 57
4.7 Comments and References 58
4.7.1 Distributed Optimization 59
4.7.2 The Memory-Convergence Rate Trade-Off 60
4.7.3 On The Convergence Proof 61
4.7.3.1 The Stolz-Cesàro Theorem 62

5 Distributed Banach-Picard Iteration **65**
5.1 Introduction . . . 65
5.2 The Family of Algorithms . . . 66
5.2.1 "Distributed Description" of the Fixed Points of H . . . 67
5.2.2 Parametric Family of Algorithms . . . 68
5.2.3 Fixed Points of F and the Map $\tilde{F}$ . . . 69
5.2.3.1 The Map $\tilde{F}$ . . . 70
5.2.4 Connection Between F and $\tilde{F}$ . . . 72
5.3 Convergence Analysis . . . 73
5.3.1 The Linear Part of $\tilde{F}$ . . . 73
5.3.2 The Differential Local Contraction Case . . . 76
5.3.3 The Continuous Local and Global Contraction Cases . . . 78
5.3.3.1 The Local Contraction Case . . . 82
5.4 Distributed Implementations . . . 84
5.4.1 The EXTRA-Distributed Banach-Picard Iteration . . . 85
5.4.2 The DIGing-Distributed Banach-Picard Iteration . . . 87
5.5 Comments and References . . . 88
5.5.1 Intuition and Connection with Optimization . . . 88
5.5.2 The Local Contraction Case . . . 90
5.5.3 Distributed Implementation . . . 90
5.5.3.1 Communications Per Iteration . . . 91
5.5.4 Why EXTRA is "Natural" . . . 94

6 Distributed PCA **97**
6.1 Introduction . . . 97
6.2 Problem Statement: Distributed PCA . . . 97
6.3 Sanger's Algorithm . . . 98
6.3.1 The Case $m = 1$ . . . 99
6.3.2 The General Case $m \geq 1$ . . . 102
6.3.2.1 Fixed Points of H . . . 102
6.3.2.2 Stability Properties of the Fixed Points of H . . . 103
6.4 Comments and References . . . 109
6.4.1 Simulations . . . 109
6.4.2 Extensions of Our Previous Work . . . 109
6.4.3 ADSA Almost Surely Escapes the Unstable Fixed Points . . . 109

7 Distributed Parameter Estimation with Noisy and Faulty Measurements **111**
7.1 Introduction . . . 111
7.2 Preliminaries . . . 112
7.2.1 MLE . . . 112
7.2.2 Mixture Models . . . 113
7.2.3 The EM Algorithm . . . 114
7.3 Problem Statement . . . 115
7.3.1 Mixture Model Formulation . . . 116
7.3.2 Problem Statement in Terms of the MLE . . . 117
7.4 Roadmap . . . 117
7.4.1 Why this Makes Sense . . . 119
7.4.2 Probability of Having a Fixed Point . . . 120
7.4.3 The Missing Piece . . . 121
7.4.4 Summary . . . 121
7.5 Gradient of the Log-Likelihood . . . 122
7.5.1 Modified EM . . . 123
7.6 Convergence Analysis . . . 124
7.6.1 Infinite Sample Map . . . 124
7.6.2 Assumption on the Model . . . 125
7.6.3 Probability of Having a Fixed Point . . . 126
7.6.4 Probability of Having a Stable Fixed Point . . . 128
7.6.5 Putting Everything Together . . . 129
7.7 Simulations . . . 131
7.7.1 EM Algorithm for (7.7) . . . 131
7.7.2 Two Distributed Algorithms . . . 131
7.7.3 Simulation Results . . . 132
7.8 Comments and References . . . 134
7.8.1 Convergence Towards the Ground Truth . . . 135

8 Conclusion and Future Work **137**
8.1 Conclusion . . . 137
8.2 Remarks on the Drawbacks of the DBPI . . . 139
8.3 Future Work . . . 140
8.3.1 Asymptotically Stable but not Exponentially Stable . . . 140
8.3.2 Non-Differential Local Contraction . . . 141
8.3.3 The Local Diffeomorphism Condition for Distributed PCA . . . 141

8.3.4 Ground Truth of Variant of EM 142

A Proof of Theorem 5.3.2 **143**

A.1 Proof of Part 1) of Theorem 5.3.2 145

A.2 Proof of Part 2) of Theorem 5.3.2 146

B Proof of Remark 3.2.7 (Sublinear Convergence) **149**

Bibliography **151**

List of Notation

$\mathbb{R}^n_{>0}$	set of real n-dimensional vectors with positive components
$A, B, \ldots$	matrices are denoted by upper case letters
M_{st}	given a matrix M, M_{st} denotes the element on the sth line and tth column
M^T	transpose of a matrix M
L^+	Moore-Penrose (pseudo-inverse) of a matrix L
I_d	d-dimensional identity matrix
$\mathbf{0}_{m,n}$	$m \times n$ matrix of zeros
$M \succ 0$ ($M \prec 0$)	matrix M is positive (negative) definite
$M > 0$ ($M \geq 0$)	the entries of matrix M are positive (non-negative)
$\mathcal{U}(A)$	upper triangular matrix of the same dimension of the matrix A and whose upper triangular part coincides with that of A
$a, b, \ldots$	vectors are denoted by lower case letters
v_s	given a vector v, v_s denotes its sth component
$\mathbf{1}_d$	d-dimensional vector of ones
$\otimes$	Kronnecker product
$\rho(\cdot)$	spectral radius
$\|\cdot\|_F$	Frobenius norm
$\mathbf{J}_H(x)$	Jacobian of a map H at the point x
$\nabla_w f$	gradient of a function f with respect to w
$\mathrm{Fix}(H)$	set of fixed points of a map H
i	complex imaginary unit ($i^2 = -1$)
$\bar{B}(x, \delta)$	closed ball of center x and radius δ with respect to a distance that is clear from context
$f_Y(\cdot)$	probability density (or mass) of a random variable Y
$\mathcal{N}(\cdot\|\mu, \sigma^2)$	probability density of a Gaussian with mean μ and variance σ^2

Similar to matrices, random variables or vectors are denoted by upper case letters (the distinction should be clear from context). Whenever convenient, we will denote a vector with two stacked blocks $[v^T, u^T]^T$ simply as (u, v).

Chapter 1

Introduction

1.1 Motivation

The last decades have seen a surge in interest in distributed algorithms due to the ever increasing collection of data by spatially dispersed agents linked by a communication network; these network technologies take many forms ranging from social mobile media to the Internet of Things (IoT), passing through environment monitoring by sensors endowed with wireless communication, the so-called *wireless sensor networks.* The characteristic feature of distributed algorithms, distinguishing them from their centralized counterparts, is their non-isolated nature where communication between agents armed with computing power forms the backbone of the coordination towards a common goal.

Inferring a desired quantity $x^\star$ from data can often be naturally expressed as a fixed point equation $H(x^\star) = x^\star$, the map H relating the data to the desired quantity. Although in some rare cases this fixed point equation can be solved in closed-form, more often than not, $x^\star$ has to be numerically approximated using, *e.g.*, the so-called *Banach-Picard iteration*:

$$x^{k+1} = H(x^k). \tag{1.1}$$

For a recent comprehensive review of the fixed-point strategy to inference problems, see [1].

If the entire data, thus the map H, is available to some agent, that agent can perform the Banach-Picard iteration. In contrast, in the so-called *distributed* scenario, the data is acquired by spatially dispersed agents who only have access to local data. In such distributed setups, no single agent possesses the full data set, hence no single agent can compute the map H. Instead, each agent holds a local portion of the data and can communicate only with a subset of the other agents (its neighbours). Nevertheless, the goal

remains that of finding a fixed point of H, under the constraints of this distributed configuration: each agent can only engage in private/local computation and in communication with its neighbours.

1.2 Distributed Banach-Picard Iteration

The first goal of the work described in this thesis was to extend the Banach-Picard iteration to distributed scenarios, under the assumption that H is an average of individual maps held by a set of agents linked by a communication network. Formally, we consider a network of N agents, where the interconnection structure is represented by an undirected and connected graph: the nodes correspond to the agents and an edge between two agents indicates that they can directly communicate (are neighbors). Each agent $n \in \{1, \dots, N\}$ holds a map $H_n : \mathbb{R}^d \to \mathbb{R}^d$, and their common goal is to compute a fixed point of the average map

$$H = \frac{1}{N} \sum_{n=1}^{N} H_n. \tag{1.2}$$

Crucially, the extension of (1.1) to distributed setups should not only yield the fixed point, but also do so while preserving the convergence rate. Towards this end, we studied a parametric family of maps $F_{\eta,\beta,\alpha}$ on $\mathbb{R}^{dN} \times \mathbb{R}^{dN}$ built from $H_1, \dots, H_N$, whose corresponding Banach-Picard iteration, *i.e.,*

$$(z^{k+1}, w^{k+1}) = F_{\eta,\beta,\alpha}(z^k, w^k), \tag{1.3}$$

can be implemented in a distributed fashion and "lifts" the fixed points of H in the following sense: if $x^\star$ is a fixed point of H, then there exists $w^\star$ such that $(\mathbf{1}_N \otimes x^\star, w^\star)$[1] is a fixed point of $F_{\eta,\beta,\alpha}$. Moreover, the convergence properties, either local or global, of (1.1) are preserved by $F_{\eta,\beta,\alpha}$. Specifically, if (1.1) converges globally (locally) at a linear rate to $x^\star$, then (1.3) converges globally (locally) at a linear rate to $(\mathbf{1}_N \otimes x^\star, w^\star)$.

The second goal of this work was to build upon the theoretical results mentioned in the previous paragraph to obtain distributed algorithms for the two following problems: *principal component analysis* (PCA); coordinating N agents towards collectively estimating a parameter of which each agent has a noisy and possibly faulty measurement.

[1]The notation $\mathbf{1}_N$ indicates the N dimensional vector with all components equal to 1.

1.3 Distributed PCA and Distributed Estimation

Dimensionality reduction aims at representing high-dimensional data in a lower dimensional space, which can be crucial to reduce the computational complexity of manipulating and processing this data, and is a core task in modern data analysis, machine learning, and related areas. The standard linear dimensionality reduction tool is *principal component analysis* (PCA), which allows expressing a high-dimensional dataset on the basis formed by the top eigenvectors of its sample covariance matrix. PCA first appeared in the statistics community in the beginning of the 20th century [2] and became one of the workhorses of statistical data analysis, with *dimensionality reduction* being a notable application. Nowadays, as data is collected in multiple locations, developing algorithms for distributed PCA constitutes a relevant area of research; for a comprehensive review on the subject see, *e.g.*, [3].

Consider a collection of spatially distributed sensors monitoring the environment, a common scenario for information processing or decision making tasks see, *e.g.,* [4, 5, 6, 7, 8, 9, 10, 11]. Often, these sensors communicate wirelessly, maybe in a harsh environment, which may result in faulty communications or sensor malfunctions [12]. A decentralized algorithm, rather than one where each sensor sends its data to a central node, is potentially more robust to faulty wireless communications that may render a sensor useless. Moreover, a decentralized algorithm can yield considerable energy savings [4], a very desirable feature.

In line with the theoretical results, to arrive at the distributed algorithms we formulate both problems' goals as a fixed point of a map H that can be implicitly written as an average of local maps. Both distributed algorithms are then shown to enjoy local linear convergence towards the desired solution, as a corollary of the theoretical results. In fact, this comes "for free", once the corresponding property is verified for (1.1). In summary, these applications demonstrate the virtue of the theoretical results: we don't need to worry about proving the linear convergence of the distributed algorithm, since it follows from verifying the linear convergence of its centralized counterpart, *i.e.*, (1.1).

1.4 Contributions and Related Work

This work is mainly a presentation of our two articles [13] and [14] with minor generalizations, the contributions of the first being on the theoretical level, whereas those of the second have more of an applied flavor. In this section, both the theoretical and the applied contributions are separately discussed and, in both cases, the minor generalizations are highlighted along the way.

1.4.1 Theoretical Contributions

Our main contribution in [13] is to show the following. Let H be a map that is an average of local maps such as in (1.2) and suppose that H has a fixed point $x^\star$ satisfying[2]

$$\rho\big(\mathbf{J}_H(x^\star)\big) < 1. \tag{1.4}$$

Consider the parametric family of maps $F_{\eta,\beta,\alpha}$, whose correspondonding Banach-Picard iteration, *i.e.,* (1.3), has distributed implementation. Then, for particular choices of η and β, and for α sufficiently small, the map $F_{\eta,\beta,\alpha}$ satisfies

$$\rho\Big(\mathbf{J}_{F_{\eta,\beta,\alpha}}\big(\mathbf{1}_N \otimes x^\star, w^\star\big)\Big) < 1. \tag{1.5}$$

Although assuming a relatively weak set of conditions—essentially only local linear convergence of the centralized Banach-Picard iteration—and no global structure (*e.g.*, Lipschitzianity or coercivity), we propose a distributed algorithm, *i.e.,* (1.3), and prove that it inherits the local linear convergence of its centralized counterpart.

Even though the assumptions are rather weak, they nevertheless suffice to encapsulate relevant algorithms, namely some instances of the *expectation maximization* (EM, [15, 12]) algorithm and the one proposed in [16] for *principal component analysis* (PCA). The second work [14] is devoted to the application of the algorithmic framework proposed in [13] to obtain distributed versions of those algorithms, with local linear convergence guarantees.

As an additional contribution, we mention the proof technique, which, as far as we know, departs from the standard proof techniques used in distributed optimization. Specifically, we employ tools from perturbation theory of linear operators [17], which, to the best of our knowledge, are scarcely exploited in the context of distributed computation. Arguably, there are proof techniques that resemble a "perturbative argument" on the eigenvalues of a matrix (e.g., Proposition 2.8 in [18], and Theorem 2 in [19]). However, those techniques bypass the subtle issue of the differentiability of the eigenvalues, simply using the formula for the derivative of the determinant. In contrast, the theorem from perturbation theory (PT) of linear operators that we use simultaneously handles the differentiability issue and simplifies the computation of the derivative.

[2] $\mathbf{J}_H(x^\star)$ denotes the Jacobian of H at $x^\star$ and $\rho(\cdot)$ represents the spectral radius.

1.4.2 Remarks

The setup addressed in [13] departs from standard ones in two main aspects. First, it encompasses problems that are not naturally expressed as optimization problems. This last notion should be understood with a grain of salt, since a fixed point of H minimizes $\|H(x) - x\|^2$; however, in many cases, there is a more "natural" objective function than this one. For example, if the Jacobian $\mathbf{J}_H$ is symmetric in an open, convex set, then there exists a function f such that $H = \nabla f$ [20, Theorem 1.3.1], and the Banach-Picard iteration can be seen as method to find a stationary point of $f(x) - \frac{1}{2}\|x\|^2$.

Second, condition (1.4) is purely local *i.e.,* we consider only local guarantees. Many optimization problems benefit from global properties, such as Lipschitzianity or strong convexity. Such properties, however, are absent in many relevant algorithms, such as EM, for which only local guarantees can be given.

1.4.3 Related Work

In this section, we review relevant related work in distributed computation, highlighting how our contributions differ from that other work.

A setup that closely resembles ours is considered in [21] and [12]; in fact, the problems therein addressed are, respectively, distributed PCA and distributed EM. As shown in our second work [14], our setup encapsulates the problems addressed in [21] and [12]. However, the algorithm proposed in [12] uses a diminishing step-size, which, unlike our algorithm, results in a sacrifice of the convergence rate of the centralized EM. The algorithm in [21] is recovered by using our approach to build a distributed version of Sanger's algorithm [16]. Moreover, our approach in [14] has at least two advantages over that of [21]: we provide a proof of local linear convergence (which [21] does not) and our setup is not restricted to Sanger's algorithm [16].

The works presented in [22], [23], and [24] share a similarity with ours by addressing the distributed computation of fixed points. However, the setups therein considered have much more structure than ours: Lipschitzianity and quasi-nonexpansivness [23], non-expansiveness [24], and paracontractiveness [22]. Those are global properties that are absent in algorithms such as EM or the PCA algorithm proposed in [16].

A large body of work on distributed optimization has been produced in the last decade; see [25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41], and [42] for convex optimization, where the last reference considers a stochastic variant, and [43, 44, 45], as examples of work on distributed non-convex optimization. All the algorithms described in those publications can definitely be seen as distributed algorithms for finding fixed points. However, as their setups stem from optimization, they further assume conditions such as

coercivity, Lipschitzianity, or strong convexity. In our work, none of these properties are assumed, and only a basic local assumption is made.

1.4.4 Generalizations

This thesis generalizes the results presented in [13] in two ways. First, we show that $F_{\eta,\beta,\alpha}$ preserves a global property of H: instead of the local condition (1.4), which implicitly assumes differentiability, we consider the case where H is a continuous (not necessarily differentiable) global contraction with respect to a unique fixed point $x^\star$, and we show that $F_{\eta,\beta,\alpha}$ is also a global contraction with respect to $(\mathbf{1}_N \otimes x^\star, w^\star)$. This generalization, however, comes at the expense of assuming that each local map is globally Lipschitz. An interesting application of this result is the case where H is a gradient descent map, which, in the strongly convex and Lipschitz case, is, for a sufficiently small step-size, a global contraction; consequently, $F_{\eta,\beta,\alpha}$ is a distributed gradient descent algorithm. In fact, depending on the choices of η and β, (1.3) recovers the EXTRA algorithm (see [28]) and the DIGing algorithm (see [31, 32]), and, thus, the global contraction case proof can be seen as a "unifying proof " for these two well-known algorithms. The proof is loosely based on that presented in [31] for the strongly convex case, and we believe that it improves on it by identifying its key blocks. The second way the present work generalizes the results in [13] has to do with "unstable fixed points", that is, we show that $F_{\eta,\beta,\alpha}$ preserves a particular case of the inequality in (1.4) reversed. Specifically, we show that if $x^\star$ is a fixed point for which $\mathbf{J}_H(x^\star)$ as an eigenvalue with real part larger than one, then

$$\rho\Big(\mathbf{J}_{F_{\eta,\beta,\alpha}}\big(\mathbf{1}_N \otimes x^\star, w^\star\big)\Big) > 1, \tag{1.6}$$

for particular choices of η and β, and for α sufficiently small. Under certain conditions, (1.1) "almost surely" escapes a fixed point $x^\star$ for which $\mathbf{J}_H(x^\star)$ has an eigenvalue with real part larger than one, that is, the set of initial conditions for which (1.1) converges to such a point has Lebesgue measure zero. Ideally, a distributed extension of (1.1) should preserve this feature: if the probability that (1.1) initialized at x^0 converges to $x^\star$ is zero, then the probability that (1.3) initialized at z^0 converges to $(\mathbf{1}_N \otimes x^\star, w^\star)$ should also be zero. To see why this is desirable, suppose that the fixed points of H are the zeros of the gradient of a certain function f with multiple maxima and minima; suppose as well that we are interested in a minimum rather than a maximum; in that case, we would like to have (1.4) at minima and the reverse inequality at maxima. The preservation of instability, *i.e.,* (1.6), guarantees that (1.3) will not have a non-zero probability of converging to a maximum, an undesired outcome.

Finally, in the interest of self-containedness, rather than appealing to perturbation theory of linear operators to show that (1.4) implies (1.5), we establish the result via the *Geršhgorin circle theorem*, thereby obtaining a more elementary proof. In this way, we circumvent the background knowledge on complex analysis required to understand the perturbation argument used in [13].

1.4.5 Contributions to Applications and Related Work

In [14], we addressed the distributed PCA problem and a distributed estimation problem using two instantiations of the algorithmic framework proposed in [13]. More concretely, we obtained:

1. A distributed algorithm for PCA, which results from considering a map that can be implicitly written as an average of local maps and that has as a fixed point the solution to the PCA problem.

2. An algorithm that stems from formulating the estimation of a parameter from noisy and faulty measurements as a fixed point of a map induced by the stationary equations of the *maximum likelihood estimation* (MLE) criterion. This map corresponds to the iterations of a slightly modified EM algorithm for a *mixture of linear regressions* [46].

The guarantees of local linear convergence for these distributed algorithms involve verifying condition (1.4) for the centralized maps inducing them, which allows invoking the results from [13]. Consequently, a great portion of [14] is devoted to proving that (1.4) holds for these maps, which is far from trivial.

The distributed PCA problem was addressed in [21], where an algorithm termed *accelerated distributed Sanger's algorithm* (ADSA) was proposed. The authors consider a "mini-batch variant" of *Sanger's algorithm* (SA, see [16]) and, inspired by [28], arrive at ADSA. Although no proof of convergence was presented in [21], a very recent work by the same authors proves convergence of their algorithm [47]. Our contributions in this context are twofold: we show that ADSA is recovered by applying the distributed algorithmic framework of [13] to SA, and that condition (1.4) holds for SA, thus, the guarantees of local convergence follow directly as a consequence of the results in [13].

The parameter estimation under noisy and faulty measurements problem was addressed in [12], where it is modeled as finding the MLE of *finite mixture model* [48]. To arrive at the MLE, the authors proposed a distributed version of the EM algorithm, termed *diffusion-averaging distributed EM* (DA-DEM). However, DA-DEM, very much

in the spirit of [49, 25], uses a diminishing step-size to achieve convergence, leading to a sublinear convergence rate. In [14] we proposed an algorithm for this problem that extends a slightly modified version of the centralized EM algorithm to distributed settings. The key challenge is to show that we can "expect" condition (1.4) to hold, and we dedicate a considerable amount of effort to this endeavor. We use the term "expect", since the map underlying the centralized algorithm depends on the observed samples and, therefore, the existence of a fixed point satisfying (1.4) is a probabilistic question.

There is considerable work on the "probabilistic linear convergence" of EM [50, 51, 52]. However, neither the results in [50], nor those in [51] encompass the mixture model in [12]. The mixture of regressions presented in [52] bears some similarity with it, but it is not the same: with the mixture of regressions from [52], there would be no measurements with just noise, *i.e.,* there would be no faulty measurements. Furthermore, [52] is primarily concerned with statistical guarantees for the error with respect to the ground truth, while we address the goal of establishing (1.4).

As mentioned in [12], there are two other relevant works on distributed EM, namely, [53] and [54]. However (see [12]), both these works address a different problem of Gaussian mixture density estimation. Moreover, in the case of [53], the algorithm demands a cyclic network topology, and, in [54], the algorithm requires higher computational load on each node, since it is based on the *alternating direction method of multipliers* [55].

1.4.6 Generalizations

The present work improves upon [14] in two ways, both regarding the distributed PCA problem. First, proofs are more self-contained, no longer requiring *matrix differential calculus*, the tool used in [14] to prove that Sanger's algorithm satisfies (1.4) at the solution of PCA. Rules of matrix differential calculus are used in [14] to compute the differential of H at $x^\star$, which is then used to analyze the eigenvalues of $\mathbf{J}_H(x^\star)$, under the identification between differentials and Jacobians. In contrast, in this thesis, the Jacobian of H is studied by simply noting that it is the linear map satisfying

$$\mathbf{J}_H(x^\star)Z = \lim_{t\to 0} \frac{H(x^\star + tZ) - H(x^\star)}{t},$$

which, in the case of the Sanger's map, is straightforward to analyze. Second, and more importantly, we extend the results in [14] to the unstable fixed points of Sanger's map. Specifically, we show that its Jacobian has a real eigenvalue larger than one at the fixed points other than the solution to PCA, and we observe that if all eigenvalues are distinct (one dimensional eigenspaces), then Sanger's map has finitely many fixed points, the solutions to PCA corresponding to stable fixed points and the remaining being unstable.

As a consequence of the generalization of the results in [14] previously described, the distributed extension has a finite number of fixed points and their stability can be tuned to be that of the corresponding fixed points of its centralized counterpart.

1.5 Organization of this Thesis

We conclude this introductory chapter with a brief overview of the structure of the thesis.

The next chapter (Chapter 2) introduces the basics of *consensus matrices*, *i.e.,* $N \times N$ symmetric matrices W with the topology of an adjacency matrix of a connected graph on N nodes[3] that satisfy

$$\lim_{k\to\infty} W^k x^0 = \frac{1}{N}\sum_{n=1}^{N} x_n^0;$$

these matrices are a building block to all distributed algorithms studied in this thesis. Chapter 3 introduces the basics of fixed point theory and the problem addressed in this thesis. Chapter 4 proposes a distributed algorithm for finding fixed points that amounts to a generalization of both DA-DEM from [12] and the distributed gradient descent with a shrinking step-size [25]; we show that the shrinking step-size results in a sacrifice in the convergence rate. Chapter 5 is based on [13] and is the crucial theoretical chapter where the preservation of the stability properties and the global contraction case are analyzed. Chapters 6 and 7 are devoted to the two applications abovementioned and are based on [14]. Each chapter begins with its own introduction, includes a final section with comments, as well as references. Finally, Chapter 8 concludes this thesis and points at ongoing and future work.

[3] A matrix having zeros corresponding to non-neighboring nodes.

Chapter 2

Distributed Average Consensus

2.1 Introduction

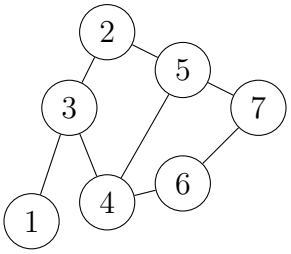
This chapter is devoted to a well-known and studied problem which we term the *distributed average consensus* (DAC) problem, involving, as part of a possible solution, a type of matrices, termed *consensus matrices*, that constitute a fundamental building block to all that follows. The necessary mathematical concepts for this section lie at the intersection of Matrix Analysis and Graph Theory, and are introduced/explained upon demand. In an effort to keep the presentation elementary and self-contained, we avoided the full generality of the DAC problem. In fact, we jumped over the analysis of non-symmetric consensus matrices to circumvent a digression into the Perron-Frobenius theory. This sacrifice in generality results, in any case, in the right degree of generality deemed necessary to understand what follows and has the benefit, we hope, of improving understanding and readability.

2.2 Problem Statement

2.2.1 Basic Notions in Graphs Theory

To introduce the problem, we start with the notion of an *undirected connected graph*, an elementary concept in mathematics. For our purposes, we see no virtue in going through a rigorous definition, but, nevertheless, point the interested reader to [56] for a formal treatment on Graph Theory.

An undirected graph $\mathcal{G}$ is a diagram consisting of a set of points joined by line segments; the set of points is denominated the *node set* and the set of line segments the *edge set*. A picture to have in mind looks like



In this case, the nodes are the (numbered) circles and the edges are the segments; for example, there is an edge joining node 4 to node 5.

Throughout this work, only undirected graphs are considered, *i.e.,* those with no oriented edges. By no orientation, what we mean is: suppose that, in the graph above, nodes represent villages and edges represent pedestrian roads; pedestrian roads are, by nature, not oriented in the sense that people can walk in both directions of the road, *e.g.,* one can walk from village 6 to village 4 and return to village 6 using the same road. In contrast, if edges represent one-way highways there is an orientation in the sense that cars are only allowed to drive in a determined direction.

With the villages example in mind, we mention that a graph is *connected* if one can walk from a given village to any other; for example, someone can travel from village 6 to village 2 by first going to village 4, then to village 5 and finally to village 2. This sequence of steps is called a *path* from node 6 to node 2 and will be denoted as

$$6 \to 4 \to 5 \to 2.$$

Furthermore, being formed by three edges, it is denominated a path of *length* three. The connected graph scenario contrasts with a two island scenario with no bridge connecting them; in each island there are villages connected by roads, but it is impossible to walk from a village of an island to a village of another.

Given an undirected and connected graph $\mathcal{G}$ with node set V and edge set E, the set of *neighbors* of a node $v \in V$, denoted by $\mathcal{N}_v$, are the nodes $w \in V$ for which there exists an edge joining v and w, denoted by $v \sim w$. In the example above, the neighbors of node 4, $\mathcal{N}_4$, are the nodes 6, 5, and 3. Finally, the degree of a node $v \in V$, denoted by $\deg(v)$, is the number of elements in $\mathcal{N}_v$.

2.2.2 Problem Statement

Let $\mathcal{G}$ be an undirected connected graph with node set V and edge set E, and suppose that V represents a collection of water tanks, with E representing a collection of water pipes attached to the bottom of the water tanks. In addition, suppose that each water tank is filled with a, possibly different, quantity of water, and that all the pipes have a

valve that is initially closed preventing the water to pass from a tank to another. If, at some point, all the valves are simultaneously open, intuition suggests that, after some time, all tanks will reach the same amount of water. Moreover, this common quantity will be roughly the average of the initial amounts, if each pipe is assumed to have a negligibly small volume when compared to each tank's initial volume. The problem addressed in this chapter can be motivated as a discrete-time analog of the water tanks example.

DAC Problem: Let $\mathcal{G}$ be an undirected and connected graph with node set V and edge set E. The *distributed average consensus problem* is that of defining $|V|$ maps $(F_v)_{v\in V}$, where $F_v : \mathbb{R}^{\deg(v)+1} \mapsto \mathbb{R}$, such that, for all $(x_v^0)_{v\in V}$, the $|V|$ sequences recursively given by

$$\begin{aligned} x_v^0 &\in \mathbb{R} \\ x_v^{k+1} &= F_v\big(x_v^k, (x_w^k)_{w\in\mathcal{N}_v}\big) \end{aligned} \tag{2.1}$$

satisfy

$$\lim_{k\to\infty} x_v^k = \frac{1}{|V|}\sum_{v\in V} x_v^0,$$

for all $v \in V$.

In connection with the water tanks example, imagine that each node v starts with an initial quantity x_v^0, then combines its quantity with that of its neighbors (F_v is a map whose arguments are the values of v and its neighbors), much in the same way that water is only exchanged with neighboring water tanks, and obtains a quantity x_v^1. The process is repeated resulting in a quantity x_v^2 and so on. If the process was repeated indefinitely, the quantity corresponding to each node should approach a common value and that common value should be the average of the initial quantities. We conclude this section with the following remarks.

Remark 2.2.1. *The order of the implicit quantifiers in the problem above is relevant. In fact, the maps $(F_v)_{v\in V}$ depend on the graph but do not depend on the quantities x_v^0. Otherwise, there would be a trivial solution with each F_v constantly equal to $1/|V|\sum_{v\in V} x_v^0$.*

Remark 2.2.2. *The connectedness condition is crucial and can be motivated by the water tanks example. Suppose that the graph is not connected; such a graph can be decomposed into connected components – think of a collection of tanks grouped as islands. In such scenario, when the valves are simultaneously open, the quantity of water in each "island" of tanks will reach the same quantity but this quantity need not be the same among "islands". This behavior will resurface in disguise in the solution that is later presented.*

Remark 2.2.3. *The name* distributed average consensus problem *should now be clear. The "average consensus" part refers to the fact that, in the limit, all nodes agree (are in consensus) on (over) a value and that value is the average of the initial values. Finally, the "distributed" part refers to the fact that the maps F_v depend on an underlying graph.*

Remark 2.2.4. *Even though in the formulation of the DAC problem, the sequences x_v^k are sequences of real numbers, the generalization, both of the formulation and of the solution, to sequences of vectors or matrices poses no serious challenge.*

2.2.3 Applications

An undirected graph is the natural mathematical model to describe real world scenarios where a set of objects are somehow related (connected) to each other. The list of such situations is endless and, thus, the following examples are far from exhaustive.

Example 1: The graph is a representation of a tube map, where nodes are tube stations and edges are tube lines between them.

Example 2: The nodes represent a collection of websites and the edges correspond to hyperlinks between them.

Example 3: The graph represents a social network, where the nodes correspond to people and the edges encode friendship.

Example 4: The graph represents a molecule, where the nodes are atoms and the edges are chemical bonds between them.

In many situations there is further structure associated to the interconnections. For instance, a number assigned to an edge might encode a property of the connection such as time travel between stations in example 1 or the strength of a chemical bond in example 2.

Concerning the DAC problem, a picture to have in mind is a collection of N sensors monitoring the temperature of the environment. Each sensor is endowed with wireless communications and can communicate with other sensors within a certain wireless range; suppose that all sensors have the same wireless range. This setup is naturally modeled by a graph where the nodes are the sensors ($|V| = N$) and an edge between two nodes indicates that they are within the wireless range of each other, *i.e.,* can communicate. Finally, suppose that sensor n has a noisy measurement θ_n of $\theta^\star$, the true temperature. Specifically, suppose that θ_n is a sample from a Gaussian with mean $\theta^\star$ and unit variance, and that each sample is independent.

The sensors seek to estimate $\theta^\star$ from $\theta_1, \ldots, \theta_N$ and, to this end, their goal is formulated as that of finding the *maximum likelihood estimate* (see Chapter 7 for further details on maximum likelihood estimation), *i.e.,* to compute

$$\arg\max_{\theta} \prod_{n=1}^{N} \mathcal{N}(\theta_n|\theta, 1). \tag{2.2}$$

By taking the logarithm, we can reformulate (2.2) as

$$\arg\max_{\theta} \frac{1}{2} \sum_{n=1}^{N} (\theta - \theta_n)^2. \tag{2.3}$$

To conclude, observe that the solution of (2.3) is $\bar{\theta} := 1/N \sum_{n=1}^{N} \theta_n$, that is, the sensors seek to compute $\bar{\theta}$, but, to arrive at $\bar{\theta}$, can only engage in communications with neighbors, *i.e.*, with other sensors within wireless range; this is exactly the DAC problem. As a final observation, we mention that Chapter 7 considers a much more general estimation problem in which the sensors might malfunction and obtain a faulty measurement.

2.3 A Solution to the DAC Problem

There are plenty solutions to the DAC problem and the one presented here relies on a type of matrices that is instrumental to all that will unfold. Instead of immediately presenting the solution we will gradually build towards it.

The naive approach is to take the functions $(F_v)_{v \in V}$ to be linear, *i.e.,* for each $v \in V$ define vectors $c_v \in \mathbb{R}^{\deg(v)+1}$ and take $F_v(x) = c_v^T x$. The functions $(F_v)_{v \in V}$ should combine the values of a node with those of its neighbors in such a way that the recursion (2.1) converges to the average of the initial values. Therefore, and with the water tanks in mind, it is natural to let the functions $(F_v)_{v \in V}$ themselves correspond to a local *weighted average* of the values of a node and those of its neighbors – by a weighted average of the components of $x \in \mathbb{R}^n$, we mean $p^T x$, where $p = (p_1, \ldots, p_n) \in \mathbb{R}^n$ is a *probability* (weight) vector, *i.e.,* $p_i \geq 0$ for all $i = 1, \ldots, n$, and $\sum_{i=1}^{n} p_i = 1$.

A rather natural naive way to proceed is to assign an equal weight to all neighbors, that is, to let

$$c_v = \frac{1}{\deg(v) + 1} \mathbf{1}.$$

However, as it will shortly be shown, this is only works for a specific class of graphs. In fact, unless the graph is *regular* – a regular graph is one where all nodes have the same

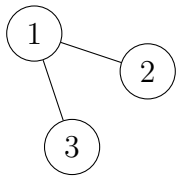
degree – this weight assignment only reaches average consensus for specific sets of initial quantities.

To proceed, it is useful to have a more compact form of writing (2.1) in the linear case, and, to this end, suppose that the graph is labeled, *i.e.,* assume that each node has been assigned a number[1]. The value of a node that is not a neighbor of v is not among arguments of the function F_v associated to node v, which, from an abstract point of view, is equivalent to setting a zero weight to the value of a non-neighboring node. Formally, this corresponds to regard c_v not as a vector in $\mathbb{R}^{\deg(v)+1}$ but rather as a vector in $\mathbb{R}^{|V|}$ with zeros in positions corresponding to non-neighboring nodes. The result is the following compact matrix representation of (2.1) in the linear case: let C be the $|V| \times |V|$ matrix whose ij entry, denoted by C_{ij}, corresponds to the weight that node i gives to node j (from the previous discussion the ij entry is zero if node i is not a neighbor of node j); the recursive process (2.1) can be compactly written as $x^{k+1} = Cx^k$, which, after unfolding, yields $x^k = C^k x^0$. Recall that the rows of C are weight vectors and, hence, $C_{ij} \geq 0$ for all i and j, and $\sum_{j=1}^{|V|} C_{ij} = 1$.

At this point it pays to see an example that illustrates what has been said so far and that additionally shows that assigning an equal weight to all nodes requires a specific set of initial quantities.

2.3.1 Example

Consider the graph that follows.



Nodes 2 and 3 are not joined by an edge, *i.e.*, are not neighbors, and, from the previous discussion, the weight node 2 assigns to node 3 is equal to the weight node 3 assigns to node 2 and is equal to zero, that is, $C_{23} = C_{32} = 0$. The remaining weights are, otherwise, arbitrary, hence, the general form of the matrix C for this graph is given by

$$C = \begin{bmatrix} C_{11} & C_{12} & C_{13} \\ C_{21} & C_{22} & 0 \\ C_{31} & 0 & C_{33} \end{bmatrix}.$$

[1]We have bypassed a rigorous discussion on graph isomorphism and labeling since we do not see any value in doing so for the purpose of presentation.

The equal weights assignment corresponds to the following specification of the matrix

$$C = \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/2 & 1/2 & 0 \\ 1/2 & 0 & 1/2 \end{bmatrix} = \frac{1}{6}\begin{bmatrix} 2 & 2 & 2 \\ 3 & 3 & 0 \\ 3 & 0 & 3 \end{bmatrix},$$

and, to analyze the limiting behavior of the recursion (2.1), we look at the eigenvalues of C or, for simplicity, of $6C$, whose characteristic polynomial is given by

$$\begin{aligned} p(\lambda) &= (3-\lambda)\big((2-\lambda)(3-\lambda)-6\big) - 6(3-\lambda) \\ &= -\lambda^3 + 8\lambda^2 - 9\lambda - 18. \end{aligned}$$

From $C\mathbf{1} = \mathbf{1}$ it follows that 1 is an eigenvalue of C and, hence, that 6 is an eigenvalue of $6C$, allowing the remaining roots of the characteristic polynomial to be found by polynomial division. The resulting factorization is given by

$$p(\lambda) = -(\lambda-6)(\lambda^2 - 2\lambda - 3) = -(\lambda-6)(\lambda+1)(\lambda-3),$$

and, thus, the eigenvalues of C are $1, 1/2$, and $-1/6$.

We are already a position to conclude that the recursion $x^k = C^k x^0$ converges to consensus, *i.e.,* converges to a vector with equal components: it is known ([57]) that a matrix with distinct eigenvalues is diagonalizable, *i.e.,* there are vectors v_1 and v_2 satisfying $Cv_1 = 1/2v_1$ and $Cv_2 = -1/6v_2$, such that $\{\mathbf{1}, v_1, v_2\}$ is a (not necessarily orthogonal) basis of $\mathbb{R}^3$, and, thus, let $x^0 = \alpha_0\mathbf{1} + \alpha_1 v_1 + \alpha_2 v_2$ be an expansion of x^0 in that basis. Observe that $x^k = \alpha_0\mathbf{1} + \alpha_1 1/2^k v_1 + \alpha_2(-1)^k 1/6^k v_2$, which shows that $\lim_{k\to\infty} x^k = \alpha_0\mathbf{1}$.

To conclude, we show that α_0 need not be the average of the components of x^0, *i.e.,* (2.1) converges to a consensus, but not necessarily to one over the average of the initial values. Given that $\det(A) = \det(A^T)$, the eigenvalues of C^T coincide with those of C, and, hence, let $w_1 \neq 0$ be an eigenvector of C^T associated to the eigenvalue 1. Observe that $w_1^T v_1 = \big(C^T w_1\big)^T v_1 = w_1^T\big(Cv_1\big) = 1/2 w_1^T v_1$, which implies that $w_1^T v_1 = 0$ (similarly, $w_1^T v_2 = 0$). We conclude that $w_1^T x^0 = \alpha_0 w_1^T\mathbf{1}$. Moreover, $w_1^T\mathbf{1} \neq 0$ [2] and, hence, dividing by $w_1^T\mathbf{1}$ yields

$$\lim_{k\to\infty} x^k = \frac{w_1^T x^0}{w_1^T\mathbf{1}}\mathbf{1}.$$

[2] In fact, if $w_1^T\mathbf{1} = 0$, then w_1 would be orthogonal to all elements of a basis. This happens if only if $w_1 = 0$.

Instead of computing w_1, assume that

$$\frac{w_1^T x^0}{w_1^T \mathbf{1}} = \frac{1}{3}\mathbf{1}^T x^0,$$

i.e., x^k converges to a consensus over the average of the components of x^0, which, after rearrangement, is equivalent to

$$\left(x^0\right)^T \left(\frac{w_1^T \mathbf{1}}{3}\mathbf{1} - w_1\right) = 0. \tag{2.4}$$

Now just observe $\mathbf{1}$ is not an eigenvector of C^T, and, hence, $(w_1^T\mathbf{1}/3)\mathbf{1} - w_1 \neq 0$ (recall that $C^T w_1 = w_1$). To finish, note that any choice of x^0 for which (2.4) does not hold results in a limiting consensus over a value other than the average of the initial quantities.

2.3.2 Returning to the Solution

The example in the previous section shows that averaging with equal weights will generally fail to be a solution to the DAC problem. However, there are some takeaways from the derivation: the first is that the limiting behavior of (2.1) is intimately connected with the eigenvalues of C, and the second is that the existence of an orthonormal basis of eigenvectors likely results in a considerably simpler analysis – these two considerations naturally suggest restricting C to be symmetric. Note, however, that allowing C to be symmetric is only "legal" because $\mathcal{G}$ was assumed to be undirected. The restrictions on C imposed so far are summarized below.

1) C is *weakly compatible with the graph structure*, *i.e.*, $C_{ij} = 0$ if i is not a neighbor of j;

2) $C \geq 0$;

3) $C\mathbf{1} = \mathbf{1}$;

4) $C = C^T$.

Suppose that all four conditions hold and let's inspect when $\lim_{k\to\infty} x^k = \bar{x}^0\mathbf{1}$, where $\bar{x}^0 = 1/|V|\mathbf{1}^T x_0$. Let $\mathcal{B} = \{1/\sqrt{|V|}\mathbf{1}, w_2, \ldots, w_{|V|}\}$ be an orthonormal basis of $\mathbb{R}^{|V|}$ such that $Cw_i = \lambda_i w_i$, where $\lambda_i \in \mathbb{R}$, for $i = 2, \ldots, |V|$ (the existence of such a basis follows from the symmetry of C). The first thing we show is that $|\lambda_i| \leq 1$, for all $i = 2, \ldots, |V|$.

To this end, consider the norm $\|\cdot\|_1$ on $\mathbb{R}^{|V|}$ and observe that

$$|\lambda_i|\|w_i\|_1 = \|Cw_i\|_1 = \Big\| \sum_{j=1}^{|V|} C_j w_{i_j} \Big\| \leq \sum_{j=1}^{|V|} |w_{i_j}| \|C_j\|_1$$

where C_j, $j = 1, \ldots, |V|$, denote the columns of C. From $C\mathbf{1} = \mathbf{1}$, $C \geq 0$ and $C = C^T$ follows that $\|C_j\|_1 = 1$ for all $j = 1, \ldots |V|$, and, hence,

$$|\lambda_i|\|w_i\|_1 \leq \sum_{j=1}^{|V|} |w_{i_j}| = \|w_i\|_1.$$

Given that $\|w_i\|_1 \neq 0$, both sides can be divided by $\|w_i\|_1$ resulting in $|\lambda_i| \leq 1$. From the symmetry of C, its eigenvalues are guaranteed to be real, and, hence, either $\lambda_i = 1$, $\lambda_i = -1$, or $\lambda_i \in (-1, 1)$. Let $\mathcal{C}, \mathcal{D}$, and $\mathcal{E}$ be the subsets of $\{2, \ldots, |V|\}$ corresponding, respectively, to these three cases. The expansion of vector x^k in the basis $\mathcal{B}$ is

$$x^k = \bar{x}^0 \mathbf{1} + \sum_{r \in \mathcal{C}} (w_r^T x^0) w_r + \sum_{s \in \mathcal{D}} (-1)^k (w_r^T x_0) w_r + \sum_{t \in \mathcal{E}} \lambda_t^k (w_t^T x_0) w_t,$$

and from $\lim_{k\to\infty} \sum_{t\in\mathcal{E}} \lambda_t^k (w_t^T x_0) w_t = 0$, it is clear that $\mathcal{C}$ and $\mathcal{D}$ must be empty if we wish to ensure that $\lim_{k\to\infty} x_k = \bar{x}^0 \mathbf{1}$. In other words, $\mathbf{1}$ (in fact span$\{\mathbf{1}\}$) needs to be the only eigenvector associated to the eigenvalue 1, and the remaining eigenvectors must be associated to eigenvalues with magnitude less than 1.

2.3.3 The Final Solution

The four conditions on C considered in Section 2.2 are too weak to ensure that: 1) $\mathbf{1}$ is the only eigenvector associated to the eigenvalue 1; 2) the remaining eigenvectors are associated to eigenvalues with magnitude less than 1 (observe that the identity matrix satisfies all four conditions). Nevertheless, there is room for improvement in condition 1), since the full power of the graph connectedness has not yet been used. The weak compatibility can be strengthened by requiring $C_{ij} > 0$ if i is a neighbor of j or if $i = j$, leading to the following stronger version of condition 1)

1*) C is *compatible with the graph structure*, *i.e.*, $C_{ij} = 0$ if i is not a neighbor of j and $C_{ij} > 0$ if either i is a neighbor of j or $i = j$.

In this section we show that 1*), together with conditions 2), 3), and 4) from Section 2.2, is enough to solve the DAC problem.

2.3.3.1 The Outline of the Proof

The proof works as follows. First we show that, under $1^\star$), 2), 3), and 4), there is a power of C with only positive entries. Second, we note that this power of C inherits from C the property of having eigenvalues with magnitude less or equal than 1. Third, we show that a positive, symmetric matrix, having $\mathbf{1}$ has an eigenvector associated to the eigenvalue 1 satisfies: a) the eigenspace associated to the eigenvalue 1 is the span$\{\mathbf{1}\}$; b) the remaining eigenvalues have magnitude less than 1. We conclude by showing that this property is inherited by C.

Remark 2.3.1. *The connectedness of the graph encoded in C via $1^\star$) is crucial to establish the existence of a power of C with only positive entries. In fact, under the weaker version of $1^\star$), namely 1), this would not necessarily be the case, as it is evident by taking C to be the identity.*

2.3.3.2 The Formal Proof

Lemma 2.3.1. *There exists m such that $C^m > 0$.*

Proof. Note that the non-negativity of the entries of C, together with the positivity of its diagonal entries, implies that if $(C^k)_{ij} > 0$, then $(C^t)_{ij} > 0$ for all $t \geq k$. This observation reduces the proof to showing that the existence of a path from i to j of length t implies $(C^t)_{ij} \neq 0$. The case $t = 4$ suffices to illustrate the idea and the general case is easily established with a proof by induction: the ij entry of C^4 satisfies

$$(C^4)_{ij} = \sum_{s=1}^{|V|} (C^3)_{is} C_{sj} = \sum_{s=1}^{|V|} \Big(\sum_{r=1}^{|V|} (C^2)_{ir} C_{rs} \Big) C_{sj} = \sum_{s=1}^{|V|} \Big(\sum_{r=1}^{|V|} \big(\sum_{m=1}^{|V|} C_{im} C_{mr} \big) C_{rs} \Big) C_{sj},$$

and, from $C \geq 0$, it follows that

$$(C^4)_{ij} \geq C_{im} C_{mr} C_{rs} C_{sj}, \tag{2.5}$$

for all i, j, m, r and s in $\{1, \ldots, |V|\}$.

The existence of a path of length 4 between the nodes i and j implies that there are three nodes m, r, and s such that $i \to m \to r \to s \to j$, and, from $1^\star$), this implies that C_{im}, C_{mr}, C_{rs}, and C_{sj} are all positive numbers. From (2.5), we conclude that $(C^4)_{ij} > 0$. To finish the proof, observe that, since the graph is connected, there is a path from any node to any other. Moreover, $(C^t)_{ij} > 0$ implies that $(C^s)_{ij} > 0$ for all $s \geq t$, thus, combining these two observations establishes the result. □

For the proof of the following result, it is crucial to note that $\|Mv\| \leq \|v\|$, for every v, if M is a square and symmetric matrix with eigenvalues having magnitude less or equal than 1 (this simple observation can be proven by expanding v in an orthonormal basis of eigenvectors).

Lemma 2.3.2. *Let M be a $|V| \times |V|$ matrix satisfying $M > 0$, $M = M^T$ and $M\mathbf{1} = \mathbf{1}$. If $Mz = z$ for some z, then $z \in$ span$\{\mathbf{1}\}$.*

Proof. Suppose that $Mz = z$ for a non-zero vector z, and let $|z|$ denote the $|V|$-dimensional vector satisfying $|z|_i = |z_i|$. From $M > 0$, it follows that $|z| = |Mz| \leq M|z|$, and if strict inequality holds, namely if $M|z| > |z|$, then, the non-negativity of all the quantities implies that $\big\|M|z|\big\| > \big\||z|\big\|$; this contradicts the observation preceding the lemma, *i.e.,* that $\|Mv\| \leq \|v\|$ for all v. Because strict inequality does not hold, we conclude that $M|z| = |z|$, and the symmetry of M further implies that $|z| \in$ span$\{\mathbf{1}\}$ (in fact, if $|z| \notin$ span$\{\mathbf{1}\}$, then $|z|$ would be a non-zero and non-negative vector orthogonal to $\mathbf{1}$; it is clear that there are no such vectors).

We wish to conclude that z, not $|z|$, is in the linear space generated $\mathbf{1}$. If $z = |z|$, there is nothing to show, and, hence, suppose that $z \neq |z|$. By multiplying by a suitable constant, we can assume that $z_i = \pm 1$ for all i, and that at least two distinct components have opposite signs. Let $\mathcal{N}$ be the set of indices j such that $z_j = -1$ and let $\mathcal{M}$ be the set of indices j such that $z_j = 1$, which, from the previous assumption, are both non-empty. The eigenvector equation $Mz = z$, implies that, for all $j \in \mathcal{N}$,

$$-1 = \sum_{t \in \mathcal{M}} M_{jt} - \sum_{s \in \mathcal{N}} M_{js}$$

Moreover, the corresponding eigenvector equation for $\mathbf{1}$, that is, $M\mathbf{1} = \mathbf{1}$, similarly implies

$$1 = \sum_{t \in \mathcal{M}} M_{jt} + \sum_{s \in \mathcal{N}} M_{js}.$$

The addition of both equalities results in

$$0 = \sum_{t \in \mathcal{M}} M_{jt},$$

contradicting the positiveness of the entries of M. □

The final result can now be stated and the proof is an easy consequence of the two preceding lemmas.

Theorem 2.3.1. *Let C be a $|V| \times |V|$ matrix such that*

1*) *C is compatible with the graph structure;*

2) $C \geq 0$*;*

3) $C\mathbf{1} = \mathbf{1}$*;*

4) $C = C^T$*.*

Then, $\mathbf{1}$ *is the only eigenvector associated to the eigenvalue* 1*, and all the remaining eigenvalues have magnitude less than* 1*.*

Proof. Let $\mathcal{B} = \{1/|V|\mathbf{1}, w_2, \ldots, w_{|V|}\}$ be an orthonormal basis of $\mathbb{R}^{|V|}$ such that $Cw_i = \lambda_i$, where $\lambda_i \in \mathbb{R}$ for $i = 2, \ldots, |V|$. From Lemma 2.1, there exists m such that $C^m > 0$ and, since the diagonal entries of C are all positive, we may assume that m is even. Observe that $\mathcal{B}$ is an orthonormal basis of eigenvectors of C^m that satisfies $C^m w_i = \lambda_i^m w_i$ and $C^m\mathbf{1} = \mathbf{1}$. Moreover, $0 \leq \lambda_i^m \leq 1$. From Lemma 2.2, we can further say that $0 \leq \lambda_i^m < 1$, and, hence, $|\lambda_i| < 1$ as desired. □

2.3.4 The Metropolis Weight Matrix and the Consensus Matrices

Theorem 2.3.1 provides sufficient conditions under which the DAC problem is solved with linear maps. However, it still remains to show that it is possible to construct a matrix satisfying the aforementioned conditions. There are many possible constructions and we mention one termed the *Metropolis Weight Matrix*: For neighboring nodes i and j, define

$$C_{ij} = \frac{1}{1 + \max\big(d(i), d(j)\big)}$$

and define

$$C_{ii} = 1 - \sum_{j \sim i} C_{ij}.$$

The only condition that needs to be checked is $C_{ii} > 0$, and this is equivalent to showing that

$$\sum_{j \sim i} C_{ij} < 1.$$

To see this, just observe that

$$\sum_{i\sim j} C_{ij} \leq \sum_{i\sim j} \frac{1}{1+d(i)} < \sum_{i\sim j} \frac{1}{d(i)} = 1.$$

We conclude this section with the definition of a *consensus matrix*, for which we first introduce the concept of *spectral radius.*

Definition 2.3.1. *Given an $n \times n$ square (not necessarily symmetric) matrix M, the* spectral radius *of M, denoted by $\rho(M)$, is the maximum of the absolute values of the eigenvalues of M,* i.e.,

$$\rho(M) = \max\{|\lambda_1|, \ldots, |\lambda_n|\},$$

where $\lambda_1, \ldots, \lambda_n$ are the complex roots of the characteristic polynomial of M, i.e.*, the polynomial $p(\lambda) = det(M - \lambda I)$.*

Definition 2.3.2. *Let $\mathcal{G}$ be an undirected and connected graph with node set $|V|$ and edge set E. A* consensus matrix *is an $|V| \times |V|$ square matrix C that is weekly compatible with the graph structure and that satisfies*

1) $\mathbf{1}^T C = \mathbf{1}^T$;

2) $C\mathbf{1} = \mathbf{1}$,

3) $\rho\big(C - 1/|V|\mathbf{1}\mathbf{1}^T\big) < 1$.

Remark 2.3.2. *It is elementary, and, hence, omitted, to show that any consensus matrix solves the DAC problem. What was shown is that the Metropolis Weight Matrix is a consensus matrix.*

2.4 Comments and References

This section discusses some of the additional variations of the DAC problem together with different solutions. For a comprehensive review work on the subject, we point the reader to [58].

2.4.1 Directed graphs

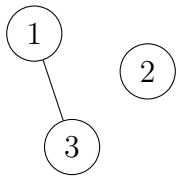
A generalization of this problem consists in allowing the graph to be directed, which corresponds to restricting the values transmitted between the nodes to respect the edge

orientation. From a mathematical point of view, a solution with linear maps does not give rise to a symmetric matrix, and hence, the analysis is more involved and the proofs rely on the Perron Frobenius theory [57]. It can be seen that, in the non-symmetric case, the notion of right eigenvectors is crucial and the nodes have to keep track of a second variable to achieve consensus.

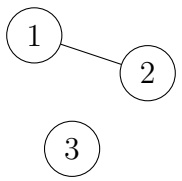
2.4.2 Time-Varying Topologies

In the time-varying topology scenario the graph $\mathcal{G}$ is not fixed but changes over time; there is a sequence of graphs $(\mathcal{G}_k)_{k\in\mathbb{N}}$ that are not necessarily equal nor connected, and, in the linear case, it corresponds to studying the properties of a recursion of the form $x^{k+1} = C_k x^k$, where C_k is compatible with the graph topology of the graph at time k, *i.e.,* $\mathcal{G}_k$.

The analysis is challenging for several reasons. In fact, to ensure that $x^k \to \bar{x}^0\mathbf{1}$, it is necessary to have a notion of "connectedness in the long run"; this notion won't be made precise, but we illustrate what can go wrong. Suppose, for example, that $\mathcal{G}_k$ alternates between two graphs $\mathcal{A}$ and $\mathcal{B}$, where $\mathcal{A}$ is



and $\mathcal{B}$ is



There are various types of behaviors that can occur. If for $k \geq K$, $\mathcal{G}_k = \mathcal{A}$, it is intuitive that we can expect

$$\lim_{k\to\infty} x_1^k = \lim_{k\to\infty} x_3^k = \frac{x_1^K + x_3^K}{2}$$
$$x_2^k = x_2^K, \text{ for } k \geq K.$$

This example seems to indicate that, to achieve consensus between the three nodes, both $\mathcal{A}$ and $\mathcal{B}$ should occur infinitely often. However, even in this case the behavior can

be quite different depending on "how often" $\mathcal{A}$ and $\mathcal{B}$ occur. For instance, suppose that

$$\mathcal{G}_k = \begin{cases} \mathcal{A} \text{ if } k \text{ is even} \\ \mathcal{B} \text{ otherwise} \end{cases} .$$

It seems plausible that the "rate" at which the nodes achieve consensus is considerably faster in this case than in the case in which

$$\mathcal{G}_k = \begin{cases} \mathcal{A} \text{ if } k \text{ is a power of two} \\ \mathcal{B} \text{ otherwise} \end{cases} .$$

To finish, observe as well that if the sequence $\mathcal{G}_k$ is not "deterministic" but "random", the analysis must be much more subtle. As an example, suppose that, instead of knowing that the sequence is deterministic (for example $\mathcal{G}_k = \mathcal{A}$ if k is even and $\mathcal{G}_k = \mathcal{B}$ otherwise), we only know that, at time k, $\mathcal{G}_k = \mathcal{A}$ with probability p, and $\mathcal{G}_k = \mathcal{B}$ with probability $1 - p$. It is clear that any convergence results will be probabilistic, and hence, not as straightforward to establish.

2.4.3 Finite-Time Consensus

Consider, for simplicity, the time-invariant topology case and an undirected and connected graph $\mathcal{G}$. In this section we describe two different ways in which the nodes can reach consensus in finite-time, *i.e.,* $x^k = \bar{x}^0\mathbf{1}$, for $k \geq K$.

2.4.3.1 Flooding

Each node has a unique identifier and maintains a table of pairs (a_v, x_v^0), where a_v is the identifier of node v and x_v^0 is its initial value. At time 0 the table of node v is initialized with only the pair (a_v, x_v^0), and at each step, the nodes exchange their table with their neighbors. It is clear that at time k equal to the diameter (largest distance between any two nodes) of the graph, all nodes will have obtained all initial values and can, therefore, compute their average (in fact, any function of the initial values).

2.4.3.2 Linear Finite-Time Consensus

Suppose the graph $\mathcal{G}$ is a *complete graph*, *i.e.,* one in which any two nodes are connected by an edge. In this setting, finite consensus is achieved in one step by taking $C = 1/|V|\mathbf{1}\mathbf{1}^T$. Now if the graph $\mathcal{G}$ is not complete, the matrix C above is not compatible with the graph topology. However, a reasonable question is whether it can be factorized into a product

of matrices compatible with the graph topology, *i.e.,* are there matrices $C_1, \ldots, C_N$ such that

$$\frac{1}{|V|}\mathbf{1}\mathbf{1}^T = C_1 \cdots C_N,$$

and where each C_i is compatible with the graph topology? The answer is yes and, given its elegance, we cannot resist to explain the general idea: Let C be the Metropolis Weight Matrix associated to an undirected and connected graph $\mathcal{G}$, and let $p(\lambda)$ be the minimal polynomial of C, *i.e.*, the smallest (in terms of degree) monic (leading coefficient equal to 1) polynomial $m(x)$ that satisfies $m(C) = 0$. From the *Cayley-Hamilton theorem* (see [57]), it is known that $p(\lambda)$ divides the characteristic polynomial of C, and, thus, $p(\lambda) = (\lambda - 1)q(\lambda)$ for some polynomial $q(\lambda)$ (this follows from $C\mathbf{1} = \mathbf{1}$). We conclude that $0 = p(C) = Cq(C) - q(C)$, which further implies that

$$Cq(C) = q(C). \tag{2.6}$$

In the previous pages we showed that the eigenspace associated to the eigenvalue 1 is span$\{\mathbf{1}\}$, and, hence, (2.6) shows that the columns of $q(C)$ are either 0 or eigenvectors of C associated to the eigenvalue 1, *i.e.,* $q(C)$ is of the form $q(C) = \begin{bmatrix}\alpha_1\mathbf{1} \ldots \alpha_{|V|}\mathbf{1}\end{bmatrix}$. The symmetry of C implies that $q(C)$, being a polynomial in C, is also symmetric, and this further implies that $q(C) = \alpha_1/|V|\mathbf{1}\mathbf{1}^T$. Now note that $q(C)\mathbf{1} = q(1)\mathbf{1}$ and, hence, $\alpha_1 = q(1)$. Finally, observe that, from the minimality of $p(\lambda)$, $q(C) \neq 0$, and, hence, $q(1) = \alpha_1 \neq 0$. We conclude that

$$\frac{1}{|V|}\mathbf{1}\mathbf{1}^T = \frac{1}{q(1)}q(C) = \frac{1}{q(1)}(C - \lambda_2)\cdots(C - \lambda_{|V|}), \tag{2.7}$$

where $\lambda_2, \ldots, \lambda_{|V|}$ are the eigenvalues of C other than 1.

2.4.4 Solutions Optimizing the Convergence Rate

Let C be a symmetric consensus matrix and let $\mathcal{B} = \{1/\sqrt{|V|}\mathbf{1}, w_2, \ldots, w_{|V|}\}$ be an orthonormal basis of eigenvectors of C. Suppose that $Cw_i = \lambda_i w_i$ and that the eigenvalues are in decreasing order, *i.e.*, $\lambda_2 \geq \ldots \geq \lambda_{|V|}$. Observe that the recursion $x^{k+1} = C^k x^k$ yields, after unfolding,

$$x^k = \bar{x}^0\mathbf{1} + \sum_{i=2}^{|V|}\alpha_i(\lambda_i)^k w_i,$$

where $x^0 = \bar{x}^0\mathbf{1} + \sum_{i=2}^{|V|} \alpha_i w_i$ is the expansion of x^0 in the basis $\mathcal{B}$. The error $\|x^k - \bar{x}^0\mathbf{1}\|$ satisfies

$$\|x^k - \bar{x}^0\mathbf{1}\| = \sqrt{\sum_{i=2}^{|V|} \alpha_i^2 \lambda_i^{2k}} \leq \sqrt{\lambda_2^{2k} \sum_{i=2}^{|V|} \alpha_i^2} = |\lambda_2^k| \|x^0 - \bar{x}^0\mathbf{1}\|,$$

and this indicates that the velocity at which consensus is achieved is governed by λ_2. In fact, the lower the magnitude of λ_2, the faster the values of x^k approach average consensus; this suggests designing C to have a low $|\lambda_2|$. For symmetric consensus matrices it is possible to formulate a convex optimization whose solution yields a consensus matrix with the lowest possible value of $|\lambda_2|$ – see [59] for further details on the formulation of the optimization problem and for heuristics for constructing consensus matrices other than the Metropolis Weight Matrix.

2.4.5 The Virtues of the Metropolis Weight Matrix

After all the previous discussion, a natural question arises: why would one opt for a Metropolis Weight Matrix solution and not for a solution that arrives at consensus in finite time or a solution that is based on a consensus matrix with minimal $|\lambda_2|$? In this section we give some partial answers to this question.

2.4.5.1 Knowledge of the Graph Topology

Imagine a scenario where a set of robots endowed with wireless communications is dropped from a plane. Upon landing, they stay in the same position and can only communicate with the robots within a distance depending on the wireless range; this is naturally modeled by a graph $\mathcal{G}$ where each node corresponds to a robot and an edge between robot v and robot w indicates that they are within the wireless range of each other. Suppose that the resulting graph is connected, that each robot measures some quantity of the environment (*e.g.* the temperature), and that their goal is to compute the average of all measured quantities (*e.g.* average temperature). Assume whoever is responsible for programming the robots to preform this task prior to them being dropped from the plane faces the issue of choosing between a Metropolis Weight Matrix solution and a linear finite-time consensus solution. It seems tempting to say that a finite-time solution is certainly superior, and, hence, that this is not a real debate. However, the subtlety is that this implicitly assumes that the graph $\mathcal{G}$ modeling the scenario is known prior to the robots being dropped. To see this, note that the factorization (2.7) depends on the eigenvalues of a matrix that is compatible with the graph topology, and, hence, the graph

topology would have to be known in advance if such a solution was to be implemented.

As the previous paragraph illustrates, the lack of perfect control over where the robots will land implies that the $\mathcal{G}$ is cannot be known with absolute accuracy, and, hence, the linear finite-time consensus solution is not a viable approach. On the other side, the Metropolis Weight Matrix solution can still be implemented, since, to execute this solution, the robots only need to communicate their degrees (how many robots are within their wireless range) to their neighbors prior to the beginning of the computation.

The takeaway from this section is that the question of choosing between the Metropolis Weight Matrix solution and the linear finite-time consensus solution is really a question of feasibility of implementation. The Metropolis Weight Matrix implementation only relies local knowledge, *i.e.,* node degrees, whereas the linear finite-time consensus solution relies on global knowledge, encoded in the spectral decomposition of matrices dependent on the graph topology. To finish, observe as well that a solution based on optimizing $|\lambda_2|$ also implicitly assumes global knowledge. The solution requires finding a matrix with minimal $|\lambda_2|$ among the matrices compatible with the graph topology, and hence, global knowledge is implicit in the constraint set of the optimization problem.

2.4.5.2 Amount of Information Transmitted

The sharp reader will quickly realize that the absence of global knowledge is not an issue in a flooding solution. However, in this case another issue pops up: the amount of information transmitted between any two neighboring nodes. In fact, in a flooding solution, a node transmits a table at each iteration, whereas in the Metropolis Weight Solution it transmits a number, and from the implementation point of view, this might constitute an issue.

2.4.5.3 Building Block

The problem described in this section is certainly applicable in a variety of situations, and, hence, it is interesting on its own. However, as the following chapters show, average consensus is not the endgame. In fact, this work looks at the problem of arriving at a consensus over a fixed point of a map, the computation of which requires an average at each iteration, and, as it will be seen, the linear consensus matrices constitute a building block of the suggested algorithms.

Chapter 3

Basics of Fixed Point Theory and Problem Statement

3.1 Introduction

The previous chapter introduced the problem of coordinating a group of agents linked by a communication network towards the computation of the average of individual values held by said agents. Although relevant in itself, the average problem is but a simple operation and the coordination of a group of agents towards more complex inference problems is desired and constitutes the goal of this work. Specifically, the later chapters focus on inference problems that are naturally cast as fixed point equations, and, to this end, this chapter begins with a review of the basics of fixed point theory and concludes by introducing the general problem addressed in this work.

3.2 Basic Definitions and Results

3.2.1 Definition of a Fixed Point

A *fixed point* of a map $H : \mathcal{D} \subseteq \mathbb{R}^d \to \mathbb{R}^d$ is a point $x^\star \in \mathbb{R}^d$ that is not "moved" by H, *i.e.*, that satisfies

$$H(x^\star) = x^\star.$$

Whenever $H(\mathcal{D}) \subseteq \mathcal{D}$, picturing H as "moving points" is quite natural: a point $x^0 \in \mathbb{R}^d$ is "moved" by H to $x^1 = H(x^0)$, which is then moved to $x^2 = H(x^1)$, and so on,

producing a trajectory induced by x^0, formally defined as the recursive sequence

$$\begin{aligned} x^0 &\in \mathcal{D}, \\ x^{k+1} &= H(x^k), \end{aligned} \tag{3.1}$$

which we will call the *orbit* of x^0. A fixed point $x^\star$ is, thus, a point whose orbit coincides with itself.

In the following plot, the blue line represents a roller coaster track and the vertical axis corresponds to the height relative to the ground (the horizontal axis). The green square corresponds to a train that can be placed anywhere with a particular initial velocity, the initial position/velocity pair being denoted by (x^0, v^0). Assume that the only forces acting on the train are gravity and the force exerted by the track, with (x^k, v^k) denoting the position/velocity pair after k seconds.

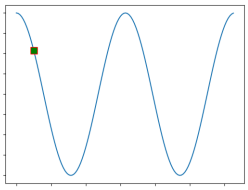


Figure 3.1: Roller Coaster Track.

Intuition suggests the existence of a map, the particular form of which is beyond of the scope of this discussion, $H : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}^2 \times \mathbb{R}^2$ that describes the motion of the train, *i.e.*, the sequence (x^k, v^k) satisfies

$$(x^{k+1}, v^{k+1}) = H(x^k, v^k),$$

for a map H modeling the physics of the roller coaster. As far as fixed points are concerned, it is intuitive that, with zero initial velocity, a train placed on the bottom and top points of the roller coaster (see the next figure) remains at rest. These points thus correspond to fixed points of H.

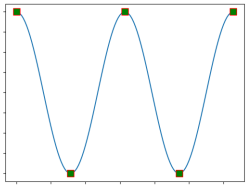


Figure 3.2: Multiple Fixed Points of the Roller Coaster Track.

3.2.2 Existence of a Fixed Point

Not all maps have fixed points and coming up with one that does not is elementary: Let $\mathbb{S}^1$ denote the collection of points distancing one unit to the origin, and let H be the map on $\mathbb{S}^1$ that rotates a point x by a non-trivial angle (*i.e.*, an angle different from a multiple of 2π). All points are moved, and, thus, H has no fixed points.

It should not come as a surprise that conditions on H and $\mathcal{D}$ have to be imposed if the existence of fixed points is to be established. A celebrated result with rather minimal conditions on both H and $\mathcal{D}$ is *Brouwer's fixed point theorem*. Given its importance for a later result, it is now stated without proof (see [60]).

Theorem 3.2.1 (Brouwer's fixed point theorem)**.** *Let $H : \mathcal{D} \to \mathcal{D}$ be a continuous map, where $\mathcal{D}$ is a closed, bounded and convex subset of $\mathbb{R}^d$. Then H has a fixed point.*

The proof of this result is non-trivial and, thus, as abovementioned, it is omitted. Nevertheless, in the one dimensional case ($d = 1$ above), the proof is straightforward, essentially reducing to the *Intermediate Value Theorem*: a closed, bounded, and convex set of $\mathbb{R}$ is a closed and bounded interval, that is, an interval of the form $[a, b]$. Let $H : [a, b] \to [a, b]$ be a continuous function. If $H(a) = a$ or $H(b) = b$ we are done, so suppose that $H(a) \neq a$, that $H(b) \neq b$, and let $g(x) = H(x) - x$. The assumptions imply that $g(a) > 0$ and $g(b) < 0$. From the continuity of H, and, consequently, that of g, it follows that g has a zero, and such a zero is a fixed point of H.

3.2.3 Qualitative Character of Fixed Points

Let us take a closer look at the roller coaster example and consider the two fixed points colored red and green in the figure below.

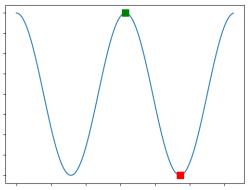


Figure 3.3: Qualitatively Different Fixed Points of the Roller Coaster Track: the green is unstable and the red is stable.

Our physical intuition suggests that if the green train is slightly perturbed from its position at rest, it will move away from the top position. On the other side, a small perturbation of the red train's position results in a different behavior, as it will eventually

return to the bottom position. This indicates that the red and the green fixed points are qualitatively different, one being "robust" to slight perturbations and the other not so.

Let $x^\star$ be a fixed point of a map H; we say that $x^\star$ is an *attractor* if there exists an open neighborhood $\mathcal{U}$ of $x^\star$ such that, for all $x^0 \in \mathcal{U}$, the orbit of x^0 tends to $x^\star$, that is, sequence (3.1) satisfies

$$\lim_{k\to\infty} x^k = x^\star.$$

Moreover, if $\mathcal{U}$ can be taken to be the whole domain of H, then $x^\star$ is a *global attractor*. As an example of a roller coaster track with a global attractor[1], imagine that its shape is that of the function $x \to x^2$.

3.2.4 Metric Conditions on H and Fundamental Global Results

Let $\|\cdot\|$ be a norm on $\mathbb{R}^d$ and μ a non-negative real number. A map $H : \mathcal{D} \subseteq \mathbb{R}^d \to \mathbb{R}^d$ satisfying, for all x and y,

$$\|H(x) - H(y)\| \leq \mu\|x - y\|,$$

is called a *μ-Lipschitz map*. If $\mu < 1$, we say that H is a *μ-contraction* or that H is a *contractive map*. A few remarks are due:

Remark 3.2.1. *Lipschitz maps are uniformly continuous.*

Remark 3.2.2. *Continuously differentiable maps are locally Lipschitz. This is an easy consequence of the* mean value theorem *and maximum value attainment by continuous maps on compact sets.*

Remark 3.2.3. *A consequence of norm equivalence is that Lipschitzianity is norm independent. However, to distinct norms are usually associated distinct Lipschitz constants, hence, contractiveness is a norm-dependent notion.*[2]

Contractive maps play a relevant role in mathematics due to their fixed point properties, with distance decrease towards the fixed point being a simple example. More relevant is that all orbits converge linearly to $x^\star$: let x^0 be any point and observe that substituting recursion (3.1) in the contractive condition yields a recursive inequality whose unfolding

[1] A map H as at most one global attractor.

[2] To see this, suppose that H is a μ-contraction with respect to $\|\cdot\|$ and define the map $\|x\|_\mu := \frac{1}{\mu}\|x\|$; an elementary argument shows that $\|\cdot\|_\mu$ is a norm with respect to which H is not contractive.

results in

$$\|x^k - x^\star\| \leq \mu^k \|x^0 - x^\star\|$$

for all k (note the validity of this expression for $k = 0$).

Remark 3.2.4. *The derivation above assumes the existence of a fixed point. However, contractive maps "usually" have fixed points; this is the content of the celebrated* Banach-Picard fixed point Theorem*: a contraction on a closed subset $\mathcal{D} \subseteq \mathbb{R}^d$ has a unique fixed point in $\mathcal{D}$. When compared to* Brouwer's fixed point theorem*, this theorem has a less restrictive assumption on $\mathcal{D}$ (merely closed, not necessarily bounded) at the cost of a much more restrictive condition on H (continuity is a much weaker condition than contractiveness).*

The following lemma regarding a slight generalization of the contractive property is instrumental for a later result.

Lemma 3.2.1. *Let $\|\cdot\|_{a_i}$ be norms on $\mathbb{R}^{n_i}$, for $i = 1, \ldots, m$, and let*

$$\begin{aligned} H : \mathbb{R}^{n_1} \times \ldots \times \mathbb{R}^{n_m} &\to \mathbb{R}^{n_1} \times \ldots \times \mathbb{R}^{n_m} \\ \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix} &\to \begin{bmatrix} H_1(x_1, \ldots, x_m) \\ \vdots \\ H_m(x_1, \ldots, x_m) \end{bmatrix} \end{aligned}$$

be a map satisfying, for all $x = (x_1, \ldots, x_m), y = (y_1, \ldots, y_m) \in \mathbb{R}^{n_1} \times \ldots \times \mathbb{R}^{n_m}$,

$$\begin{bmatrix} \|H_1(x) - H_1(y)\|_{a_1} \\ \vdots \\ \|H_m(x) - H_m(y)\|_{a_m} \end{bmatrix} \leq P \begin{bmatrix} \|x_1 - y_1\|_{a_1} \\ \vdots \\ \|x_m - y_m\|_{a_m} \end{bmatrix}, \tag{3.2}$$

where P is a non-negative $m \times m$ matrix with a spectral radius less than one. Then, if $x^\star \in \mathbb{R}^{n_1} \times \ldots \times \mathbb{R}^{n_m}$ is a fixed point of H, any orbit converges to $x^\star$ at least linearly.

Proof. Let $x^0 \in \mathbb{R}^{n_1} \times \ldots \times \mathbb{R}^{n_m}$ and consider the orbit of x^0, *i.e.*, the sequence recursively defined as

$$\begin{aligned} x^0 &\in \mathbb{R}^{n_1} \times \ldots \times \mathbb{R}^{n_m} \\ x^{k+1} &= H(x^k). \end{aligned}$$

Inequality (3.2) applied to the orbit of x^0 and to $x^\star$ yields the recursive inequality

$$\begin{bmatrix} \|x_1^{k+1} - x_1^\star\|_{a_1} \\ \vdots \\ \|x_m^{k+1} - x_m^\star\|_{a_m} \end{bmatrix} \leq P \begin{bmatrix} \|x_1^{k} - x_1^\star\|_{a_1} \\ \vdots \\ \|x_m^{k} - x_m^\star\|_{a_m} \end{bmatrix},$$

whose unfolding, permitted by the non-negativity of the entries of P, results in

$$\begin{bmatrix} \|x_1^{k} - x_1^\star\|_{a_1} \\ \vdots \\ \|x_m^{k} - x_m^\star\|_{a_m} \end{bmatrix} \leq P^k \begin{bmatrix} \|x_1^{0} - x_1^\star\|_{a_1} \\ \vdots \\ \|x_m^{0} - x_m^\star\|_{a_m} \end{bmatrix}.$$

Let $\|\cdot\|$ be a matrix norm satisfying $\|P\| < 1$, the existence of which follows from $\rho(P) < 1$ and Lemma 5.6.10 in [57]. Seen as a vector norm on $\mathbb{R}^{m^2}$, $\|\cdot\|$ is equivalent to the vector norm $\|\cdot\|_\infty$ on $\mathbb{R}^{m^2}$ defined as

$$\|x\|_\infty := \max_{i=1,\dots,m} |x_i|,$$

that is, there exists a constant $\beta > 0$ for which

$$\|P^k\|_\infty \leq \beta \|P^k\| \leq \beta \|P\|^k,$$

where the last inequality is the multiplicative property of matrix norms (see [57]). The entries of P^k are, thus, upper bounded by $\beta\|P\|^k$, and, therefore,

$$\|x_i^k - x_i^\star\|_{a_i} \leq \beta \|P\|^k \sum_{j=1}^{m} \|x_j^0 - x_j^\star\|_{a_j}.$$

From $\|P\| < 1$, the conclusion that x^k converges to $x^\star$ at least linearly follows. □

Remark 3.2.5. *The statement of this lemma is enough for our purposes. However, similar to the Banach-Picard fixed point theorem, the conditions of this lemma suffice to establish the existence of a fixed point (see [61]). Not surprisingly, for $m = 1$, it amounts to the Banach-Picard fixed point theorem: matrix P is just a number, thus reducing the spectral radius condition to a contractive one. For this reason, these maps are commonly called P-contractions (see, for example, [62]).*

3.2.5 Local Conditions and Results

Let $H : \mathcal{D} \subseteq \mathbb{R}^d \to \mathbb{R}^d$ be a map having at least two distinct fixed points $x^\star$ and $y^\star$, and observe that H cannot be contractive (if it was, H would move the two fixed points closer to each other, contradicting their nature as fixed points). Nevertheless, it can happen that H is a local contraction at $x^\star$, the contractive condition holding in a neighborhood of $x^\star$. Such local behavior naturally suggests looking at differential properties and, thus, in this section maps will be assumed to be differentiable at least in a neighborhood of a fixed point.

Theorem 3.2.2 (Ostrowski's theorem)**.** *Suppose $H : \mathcal{D} \subseteq \mathbb{R}^d \to \mathbb{R}^d$ has a fixed point $x^\star$ in the interior of $\mathcal{D}$. Suppose as well that it is differentiable at $x^\star$ and that*

$$\rho\big(\mathbf{J}_H(x^\star)\big) < 1. \tag{3.3}$$

Then, there exists a norm $\|\cdot\|$, a ball $\mathcal{B} = \{x \in \mathbb{R}^d : \|x - x^\star\| < \delta\}$, and a number $0 \leq \sigma < 1$ such that, for all $x \in \mathcal{B}$,

$$\|H(x) - x^\star\| \leq \sigma \|x - x^\star\|.$$

Proof. The proof is a relatively straightforward consequence of the existence of a norm satisfying $\|\mathbf{J}_H(x^\star)\| < 1$ (see [61]). □

Remark 3.2.6. *There is a subtlety regarding the conclusion of Theorem 3.2.2: H is not necessarily a local contraction. The map contracts distances with respect to $x^\star$, but not necessarily between any two points in a neighborhood of $x^\star$, a property, by definition, satisfied by a local contraction. Yet, if H is assumed to be not just differentiable at $x^\star$, but also* continuously differentiable *at $x^\star$, then H is a local contraction. An example explored in [63] is that of the function*

$$\begin{aligned} f : \mathbb{R} &\to \mathbb{R} \\ x &\to \begin{cases} x^2 \sin(\frac{1}{x}), & x \neq 0 \\ 0, & x = 0 \end{cases}, \end{aligned}$$

which satisfies: 1) it is differentiable everywhere; 2) $f'(0) = 0$ (and, hence, $\rho(f'(0)) < 1$); 3) f is not continuously differentiable at zero; 4) f is not a local contraction in any neighborhood of zero.

Remark 3.2.7. *In section 3.2.3, we looked at two distinct fixed points and highlighted their qualitative difference regarding robustness to small perturbations. An attractor was*

then defined to be a fixed point $x^\star$ to which orbits initialized nearby converged to. While it is certainly true that the conclusion of Ostrowski's theorem implies that $x^\star$ is an attractor of H, note that it is a stronger one: orbits initialized nearby not only tend to $x^\star$, but do so with at least linear rate. However, as the following examples shows, a point $x^\star$ can be an attractor with orbits tending slower than linearly to $x^\star$.

Let $g : \mathbb{R} \to \mathbb{R}$ be defined by $g(x) = x - x^3$ and observe that g has no fixed points other than zero. Moreover, $g'(0) = 1$, and, thus, the conditions of Ostrwoski's theorem are not met; still zero is an attractor, and, to see this, observe that

$$|g(x)| = |x||1 - x^2| < |x|$$

for $0 \neq x \in (-1, 1)$. This shows that, for any $0 \neq x^0 \in (-1, 1)$, the sequence $y_k = |x^k|$, with x^k being the orbit generated by x^0, is a positive and strictly decreasing sequence, thus a convergent sequence with a non-negative limit. An elementary argument establishes that this limit is zero, i.e., *that zero is an attractor of g. However, nearby points are attracted at a rate slower than linear (for an elementary proof see Appendix B).*

Informally, the conclusion of Ostrowski's theorem is that the behavior of orbits of H initialized sufficiently close to a fixed point $x^\star$ satisfying (3.3) mimics that of the linear approximation of H at $x^\star$, *i.e.,* the approximation

$$H(x) \approx x^\star + \mathbf{J}_H(x^\star)(x - x^\star) := \tilde{H}(x).$$

In fact, it is known that (3.3) implies that $\mathbf{J}_H(x^\star)^k \to 0$, as k goes to infinity, and that all orbits of $\tilde{H}$ tend to $x^\star$ with at least linear rate. On the other side, the nearby behavior of orbits of $g(x) = x - x^3$ initialized close to zero does not mimic that of the identity function (the linear approximation of g at $x^\star$).

A natural question is whether this mimicking behavior holds if (3.3) is replaced by

$$\rho\big(\mathbf{J}_H(x^\star)\big) > 1. \tag{3.4}$$

Suppose that $H : \mathcal{D} \subseteq \mathbb{R}^d \to \mathbb{R}^d$ has a fixed point $x^\star$ in the interior of $\mathcal{D}$, that H is differentiable at $x^\star$ and that (3.4) holds. Let v be an eigenvector of $\mathbf{J}_H(x^\star)$ associated to an eigenvalue λ satisfying $|\lambda| > 1$, and let x_ϵ^k be the orbits of the linear approximation $\tilde{H}$ intitialized at $x_\epsilon^0 = x^\star + \epsilon v$.

The sequence $y_\epsilon^k = x_\epsilon^k - x^\star$ satisfies

$$y_\epsilon^{k+1} = x_\epsilon^{k+1} - x^\star = \mathbf{J}_H(x^\star)(x_\epsilon^k - x^\star) = \mathbf{J}_H(x^\star) y_\epsilon^k,$$

which, after unfolding, yields

$$y_\epsilon^k = \mathbf{J}_H(x^\star)^k y_\epsilon^0 = \mathbf{J}_H(x^\star)^k \epsilon v = \lambda^k \epsilon v,$$

and, since $|\lambda| > 1$, the magnitude of y_ϵ^k becomes arbitrarily large, that is, x_ϵ^k becomes arbitrarily far from $x^\star$. The key observation is that since $\epsilon > 0$ is arbitrary, x_ϵ^0 is as close to $x^\star$ as we want, and, nevertheless, the orbit of $\tilde{H}$ initialized at x_ϵ^0 will "diverge" from $x^\star$. Contrary to the attractor scenario, $x^\star$ "repels" rather than attract – as an illustrative example look at the green fixed point in Section 3.2.3.

Having only looked at the behavior of the linear approximation and not to the local behavior of H, we still have the question unanswered. We are not aware of a result ruling out $x^\star$ to be an attractor when H is C^1 in a neighborhood of $x^\star$ and satisfies (3.4). There is, nevertheless, a result showing that $x^\star$ is *unstable*, a property satisfied as well by the linear approximation. Stability, a stronger notion than continuity, informally means that orbits initialized sufficiently close to a fixed point remain bounded to any desired accuracy.

Formally, given a (not necessarily differentiable) continuous map $H : \mathcal{D} \subseteq \mathbb{R}^d \to \mathbb{R}^d$, together with a fixed point $x^\star$, we say that $x^\star$ is a *stable* fixed point if for every $\epsilon > 0$, there exists $\delta > 0$ such that if $\|x^0 - x^\star\| < \delta$ then, for all $k \geq 1$,

$$\|H^k(x^0) - x^\star\| < \epsilon.$$

Remark 3.2.8. *It is trivial to prove that Ostrowski's theorem implies that $x^\star$ is a stable fixed point.*

Remark 3.2.9. *Stability is a norm-independent notion.*

As an illustrative example, observe that the green point and the red point in the figure of section 3.2.3 are, respectively, unstable and stable fixed points.

The extent to which the behavior of local orbits of H is comparable to that of the linear approximation is captured in the next result

Theorem 3.2.3. *Suppose $H : \mathcal{D} \subseteq \mathbb{R}^d \to \mathbb{R}^d$ has a fixed point $x^\star$ in the interior of $\mathcal{D}$. Suppose as well that it is differentiable at $x^\star$, and let*

$$\tilde{H}(x) = x^\star + \mathbf{J}_H(x^\star)(x - x^\star)$$

be the linear approximation of H near $x^\star$.

1) *If*

$$\rho\big(\mathbf{J}_H(x^\star)\big) < 1,$$

then $x^\star$ is a stable fixed point of both H and $\tilde{H}$. Moreover, $x^\star$ is an attractor of H and a global attractor of $\tilde{H}$. Finally, orbits of H initialized sufficiently close to $x^\star$, converge to $x^\star$ with at least linear rate, and orbits of $\tilde{H}$ initialized not necessarily close to $x^\star$ converge to $x^\star$ with at least linear rate.

2) *If*

$$\rho\big(\mathbf{J}_H(x^\star)\big) > 1,$$

then $x^\star$ is an unstable fixed point of both H and $\tilde{H}$. Moreover, $x^\star$ is not an attractor of $\tilde{H}$.

Proof. For the proof of the first assertion, just note that it corresponds to Ostrowski's theorem. For the second see [64]. □

Remark 3.2.10. *The stability of the "frontier" case $\rho\big(\mathbf{J}_H(x^\star)\big) = 1$ cannot be inspected by "only looking at first derivatives". This case won't be explored, but note that $g_1(x) = x - x^3$ and $g_2(x) = x + x^3$ are two maps belonging to the "frontier case" having zero as stable and an unstable fixed point, respectively.*

Although the question of whether $x^\star$ can be an attractor of H when (3.4) holds was left rather unanswered, this section finishes with a result ruling this out when H is a local diffeomorphism.

Theorem 3.2.4. *Let $H : \mathcal{D} \to \mathcal{D}$ be a C^1 map on an open set $\mathcal{D}$ of $\mathbb{R}^d$. Suppose that*

$$det\big(\mathbf{J}_H(x)\big) \neq 0$$

for all $x \in \mathcal{D}$. Then, the set of points in $\mathcal{D}$ whose orbit tends to a fixed point $x^\star$ satisfying

$$\rho\big(\mathbf{J}_H(x^\star)\big) > 1$$

has Lebesgue-measure zero.

Proof. See [65]. □

Remark 3.2.11. *The determinant condition implies that H is a local diffeomorphism; this is the content of the celebrated* Inverse Function Theorem *(see, for example [66]).*

As a final observation note that, according to this result, $x^\star$ cannot be an attractor, since the open neighborhood "attracted" to $x^\star$ has non-zero measure. Additional comments on this result and its relevance are provided at the end of the chapter.

3.3 Problem Statement

We are now in a position to set the goals that guide the rest of this work. Consider a network of N agents, where the interconnection structure is represented by an undirected and connected graph $\mathcal{G}$: the nodes correspond to the agents and an edge between two agents indicates they can communicate (are neighbors). Each agent $n \in \{1, \dots, N\}$ holds a map $H_n : \mathbb{R}^d \to \mathbb{R}^d$, and the goal is to compute a fixed point $x^\star$ of the average map

$$H = \frac{1}{N} \sum_{n=1}^{N} H_n.$$

Crucially, each agent n is restricted to performing computations involving H_n and communicating with its neighbors.

In essence, we will show how, given $H_1, \dots, H_N$, and $\mathcal{G}$, to construct maps[3]

$$\begin{aligned} F : \mathbb{R}^{dN} \times \mathbb{R}^{dN} &\to \mathbb{R}^{dN} \times \mathbb{R}^{dN} \\ (z, w) &\to \begin{bmatrix} \big(F_{11}(z,w), F_{12}(z,w)\big) \\ \vdots \\ \big(F_{N1}(z,w), F_{N2}(z,w)\big) \end{bmatrix} \end{aligned}$$

with the following properties:

1) The iteration $(z^{k+1}, w^{k+1}) = F(z^k, w^k)$ can be implemented in a distributed manner respecting $\mathcal{G}$, that is,

$$\frac{\partial F_{i1}}{\partial z_j} = \frac{\partial F_{i2}}{\partial z_j} = \frac{\partial F_{i1}}{\partial w_j} = \frac{\partial F_{i2}}{\partial w_j} = 0,$$

if, in $\mathcal{G}$, nodes i and j are not neighbors. Note that z_j and w_j are vectors in $\mathbb{R}^d$ and hence, we are overloading the notation of partial derivative; this should be interpreted as all partial derivatives with respect to the components of z_j and w_j.

2) The map F "lifts" the fixed points of H from $\mathbb{R}^d$ to $\mathbb{R}^{dN} \times \mathbb{R}^{dN}$, that is, if $x^\star$ is a fixed point of H, then $(\mathbf{1} \otimes x^\star, w^\star)$ is a fixed point of F for some $w^\star$. The converse should

[3]In reality, we should probably write $F_{H_1,\dots,H_N,\mathcal{G}}$ to be precise. However this would lead to an extremely heavy notation, hence the omission of subscripts.

also hold in the sense that F should not have new fixed points, *i.e.* if $(z^\star, w^\star)$ is a fixed point of F, then $z^\star$ should be of the form $z^\star = \mathbf{1} \otimes x^\star$, where $x^\star$ is a fixed point of H.

3) The local properties of fixed points of H are also "lifted" by F. This will require some further explanation later on, but the idea is that a fixed point $x^\star$ of H that satisfies the condition in Ostrowki's theorem, *i.e.,* (3.3), should be "lifted" to a fixed point $(\mathbf{1} \otimes x^\star, w^\star)$ of F that is also stable and enjoys local linear convergence. The same should hold for the reverse strict inequality (3.4), which informally means that, once "lifted by F", $x^\star$ will remain unstable. In fact, this will be established not for F, but for a map $\tilde{F}$ that is related to F by a quotient operation.

4) If H is a global contraction then F should also be. To be precise, similar to the point above, not F, but the still undefined map $\tilde{F}$ will be shown to be a P-contraction, whenever H is a contraction.

3.4 Comments and References

In the previous sections, we presented the basics of fixed points that are relevant for the rest of this work. In doing so, we bypassed many interesting topics, and put together notions/results that typically arise in apparently distinct areas.

Given a map H, the recursion (3.1) is typically called an *iterative process* in the *numerical analysis and optimization* community, and an *initial value problem of a first order difference equation* in the *dynamical systems* community, where it is regarded as a discrete time analog of an ordinary differential equation. In the latter community, a fixed point is most usually called an *equilibrium point*, and the notion of stability, a central one in dynamical systems, is often termed *Lyapunov stability.* The overlap between the two areas is discussed in [67].

As far as references are concerned, we mention: [60] for a theoretical approach on fixed points results in which geometric conditions on $\mathcal{D}$ and/or H play a crucial role, and where the *Banach-Picard fixed point theorem* and *Brouwer's fixed point theorem* are presented; [68] and [61] for a more applied approach; [64] and [69] for a dynamical systems approach, the latter reference focusing essentially on differential equations.

3.4.1 The Notion of an Attractor

For a while, the concept of an attractor was the subject of debate (see [70]). The author of [70] proposes a definition that does not coincide with ours and is based on the informal

notion that an attractor is

> "the set of points to which most points evolve under iterates ..."

We will not go over the formal definition, but we mention that it is a "probabilistic" one. In fact, we can informally say that, according to [70], a fixed point $x^\star$ of H is an attractor if an orbit initialized at a "random" x^0 has a non-zero probability of converging to $x^\star$. For this definition, there are examples of maps that are C^1 and that have an attractor satisfying (3.4).

As far as the definition of this chapter is concerned, we point the reader to [67] for an example of a continuous but not differentiable map with an unstable attractor.

3.4.2 Convergence Rate

Throughout this work, a sequence of non-negative real numbers $\{a_k\}_{k\in\mathbb{N}}$ is said to *converge at a linear rate to zero* if there exists a non-negative constant L and a non-negative real number $\mu < 1$ such that

$$a_k \leq L\mu^k. \tag{3.5}$$

This definition is framed only in terms of an upper bound, hence, a sequence converging "faster than linearly" to zero, *e.g.*, $a_k = 0$ for all k, is a sequence converging linearly to zero. While, to be completely precise, "the best" that can be said about a sequence satisfying (3.5) is that it converges with *at least linear rate*, the phrasing "converges linearly" will sometimes be employed without the risk of confusion.

A sequence $x^k \in \mathbb{R}^d$ converges (tends) linearly to $x^\star$ whenever, for a given norm $\|\cdot\|$, the non-negative real sequence $\|x^k - x^\star\|$ converges linearly to zero. This is a norm-independent notion although the constant L in (3.5) changes when the norm changes.

A sequence x^k satisfying the recursive inequality

$$\|x^{k+1} - x^\star\| \leq \mu\|x^k - x^\star\|$$

is, after unfolding, easily seen to converge to $x^\star$ at a linear rate, the constant L being $\|x^0 - x^\star\|$. It has an additional property, that of *monotonic convergence*: the iterates become successively closer to $x^\star$. Contrary to linear convergence, monotonic convergence is a norm-dependent notion.[4] We will not focus on particular norms, and, hence, norm-

[4]Let H be the map on $\mathbb{R}^2$ that first rotates the plane counterclockwise by a angle of $\frac{\pi}{2}$ and then shrinks it by a factor of $\frac{1}{2}$. The orbit of H generated by $(1,0)$ is easily seen to converge linearly and monotonically to $(0,0)$ with respect to the Euclidean distance. However, with respect to $\|\cdot\|_D$, where $\|(x_1,x_2)\|_D = \sqrt{x_1^2 + 16x_2^2}$, it converges linearly but not monotonically.

dependent notions are disregarded.

3.4.3 The Relevance of Theorem 3.2.4

The relevance of Theorem 3.2.4 is of practical nature. We are interested in "computing stuff" and, to us, an orbit of H generated by x^0 is just an algorithm initialized at x^0. With this in mind, note that if H has many fixed points, some of which are not of interest, initialization is an issue that must be addressed. To see this, observe that if an algorithm is initialized at an undesired fixed point, it will not move away from it, thus rendering the computation useless.

To make things slightly more concrete, suppose that (3.1) is an algorithm for finding a minimum of some wiggly function f with multiple minima and maxima, all of which are fixed points of H. Depending on x^0, the algorithm can converge to a minimum, to a maximum, or not converge at all. However, if H satisfies the conditions of Theorem 3.2.4, there is margin to be careless regarding the choice of x^0; in fact, Theorem 3.2.4 essentially guarantees that with a "blind" choice of x^0, the algorithm will not converge to a maximum. While it does not promise that the algorithm converges to a minimum, it does, nevertheless, insure that it almost surely does so, provided it converges.

Such a margin of carelessness regarding initialization should not be destroyed once H is "extended" to a distributed scenario, and, hence, this feature should be inherited by the maps F mentioned in Section 3.3.

3.4.4 Distributed Optimization

In the last two decades, a large body of work has been produced in distributed computation, the largest portion of it in distributed optimization. A commonly studied problem in distributed optimization is that of the coordination of a network of agents towards the computation of the minimum and/or the minimizer of the average of the functions held by said agents. Formally, each agent $n \in \{1, \dots, N\}$ holds a function $f_n : \mathbb{R}^d \to \mathbb{R}$, and the goal is to solve

$$\underset{x}{\text{minimize}} \quad f(x) := \frac{1}{N} \sum_{n=1}^{N} f_n(x),$$

where, similarly to the scenario described in section 3.3, each agent n is restricted to performing computations using only f_n and communicating with its neighbors[5]. Computations using only f_n should be understood in rather general sense that include, for

[5]Similar to section 3.3, there is an underlying graph $\mathcal{G}$ representing the communication links.

example, the evaluation at a chosen point of the gradient of f_n or of the proximal operator[6] associated to f_n.

Research works on this problem usually begin with assumptions (such as convexity, gradient Lipschitzianity, coercivity, etc) on the functions f_n (sometimes on the whole average function f) and then propose algorithms whose convergence properties rely on said assumptions.

Despite the great affinity between the problem in Section 3.3 and distributed optimization problems just mentioned, the former departs from the latter in two main aspects. First, it encompasses problems that are not naturally expressed as optimization problems. Second, many of the assumptions made in distributed optimization are absent in relevant algorithms, *e.g.* the *expectation maximization algorithm*, whose extension to distributed scenarios is sought.

It is safe to say that there is a distinction in the starting point of view. Whereas in distributed optimization it is

> "here is an average function, give me a method for computing its minimum in a distributed fashion, leveraging on the underlying function properties,"

we believe that, in our case, it is more generally,

> "here is a centralized algorithm, give me a general method to extend it to a distributed one while preserving its relevant features."

Nevertheless, given the similarities between the two, the parallel between the problem in Section 3.3 and problems in distributed optimization will be recurrent throughout this work.

[6]The proximal operator associated to f_n is the map $z \to \arg\min_x f_n(x) + \frac{\rho}{2}\|z - x\|^2$ (see, for example, [71]).

Chapter 4

A Distributed Algorithm with a Shrinking Step-Size

4.1 Introduction

Section 3.3 of Chapter 3 introduced the general problem studied in this work. This Chapter loosely addresses that problem by presenting an algorithm for the distributed computation of fixed points, though, unlike in Section 3.3, not by constructing a map F satisfying the properties therein described.

The distributed algorithm in this section can be written as an iteration of the form

$$z^{k+1} = F_k(z^k), \tag{4.1}$$

for a collection of maps $F_k : \mathbb{R}^{dN} \to \mathbb{R}^{dN}$, $k = 1, \ldots$, that respect the graph topology. Contrary to the implementation of the distributed algorithms arising from the maps F, which requires agents to have in memory a $2d$-dimensional vector, the implementation of (4.1) merely requires each agent to have in memory a d-dimensional vector (the dimension of the domain of the average map). This, however, comes with a double cost. First, the maps F_k depend on k, hence, the agents need to know which iteration they are executing. Second, and more importantly, whenever the average map is a global or a local contraction, the rate of convergence of (4.1) is sub-linear, hence, qualitatively slower than that of $x^{k+1} = H(x^k)$.

The contents of this Chapter are an original and unpublished contribution that builds upon [12], a work that suggests an algorithm that extends to a distributed scenario a particular application of the expectation-maximization (EM) algorithm. As a first contribution, we mention that our setup is more general than that of [12], the distributed EM of [12] being a particular instance of (4.1). Secondly, our proof of convergence is

considerably simpler, and, contrary to that of [12], which relies heavily on the particular form of a step-size sequence, our proof shows that their result holds provided the step-size sequence is merely vanishing and non-summable; furthermore, we give a bound on the convergence rate, a gap left by [12]. Finally, by focusing on a general setup, we address the case in which the average map H is a global contraction, a property not enjoyed by the particular map H underlying the "non-distributed" EM algorithm.

4.2 Preliminaries

Consider, as in Section 3.3 of Chapter 3, a network of N agents, where the interconnection structure is represented by an undirected and connected graph. Recall that each agent holds a map $H_n : \mathbb{R}^d \to \mathbb{R}^d$, and the goal is to compute a fixed point of the average map

$$H = \frac{1}{N} \sum_{n=1}^{N} H_n.$$

Moreover, each agent is restricted to performing computations using H_n and communicating with its neighbors. Throughout this Chapter, each H_n is assumed to be bounded and Lipschitz; without loss of generality, we may assume that all H_n are β-Lipschitz and bounded by the same constant M.

In a non-distributed scenario where, for example, an agent holds all the maps H_n, a common algorithm to approximate a fixed point of H is the *Banach-Picard iteration* of H,

$$x^{k+1} = H(x^k). \tag{4.2}$$

Associated to a map H, there is a family of map H_α that correspond to averaging H with the identity,

$$H_\alpha = (1 - \alpha)I + \alpha H,$$

where $\alpha \in]0, 1]$; note that H_1 coincides with H and, that, for $\alpha \neq 0$, the fixed point set of H_α coincides with that of H. In addition to (4.2), a common method to approximate a fixed point of H is the Banach-Picard iteration of H_α, typically called the *Krasnoselskij iteration* (see [68]). More generally, allowing the averaging weight α to vary along the iterations, that is,

$$x^{k+1} = (1 - \alpha^k)x^k + \alpha^k H(x^k),$$

leads to the so-called *normal Mann iteration* (see [68]).

Let $\tilde{H} : \mathbb{R}^{dN} \to \mathbb{R}^{dN}$ be defined by

$$\tilde{H}(z_1, \ldots, z_N) = \big(H_1(z_1), \ldots, H_N(z_N)\big). \tag{4.3}$$

In a distributed setting with a network represented by a complete graph, a natural distributed algorithm is to let each agent execute the Mann iteration of H. Up to initialization, this can essentially be written as

$$z^{k+1} = \frac{1}{N}\Big(\mathbf{1}\mathbf{1}^T \otimes I_d\Big)\big((1-\alpha^k)z^k + \alpha^k \tilde{H}(z^k)\big), \tag{4.4}$$

with $z^k \in \mathbb{R}^{dN}$. Recursion (4.4) can be seen as a local computation, $(1-\alpha^k)z^k + \alpha^k \tilde{H}(z^k)$, followed by a communication (combination) step, represented by the multiplication by $1/N\mathbf{1}\mathbf{1}^T \otimes I_d$. In contrast, if the communication network is not a complete graph, the local computation step can be carried out, but the multiplication step cannot. In this case, a naïve approach is to perform an "imperfect average", *i.e.,* to substitute the matrix $1/N\mathbf{1}\mathbf{1}^T \otimes I_d$ in (4.4) by a consensus matrix respecting the graph topology (see Chapter 2),

$$z^{k+1} = W\big((1-\alpha^k)z^k + \alpha^k \tilde{H}(z^k)\big), \tag{4.5}$$

where $W = \tilde{W} \otimes I_d$, for a consensus matrix $\tilde{W}$. This chapter studies the properties of (4.5). The following remarks are due:

Remark 4.2.1. *The stepsizes α^k in* (4.5) *will be non-constant. For this reason,* (4.5) *is an iteration of the form*

$$z^{k+1} = F_k(z^k),$$

where the maps F_k depend on k. As a result, the agents need to "know" which iteration they are executing. Rather than resulting in an imperfect Krasnoselskij iteration, the non-constant nature of the step-sizes renders (4.5) *a "true" imperfect Mann iteration.*

Remark 4.2.2. *If each map H_n is of the form $H_n = x - \nabla f_n$ for a differentiable function f_n, the Krasnoselskij and Mann iterations correspond to gradient descent with constant and non-constant stepsizes, respectively. Moreover, with vanishing setpsizes,* (4.5) *corresponds to the* distributed gradient descent with shrinking stepsize*, a well known algorithm in the distributed optimization community (see [25]).*

Remark 4.2.3. *If the algorithm is, by miracle, initialized at a fixed point of interest,* i.e., $z^0 = \mathbf{1} \otimes x^\star$, *then, unless the maps* H_n *are very specific, the next iterate,* z^1 *will be different from* $\mathbf{1} \otimes x^\star$. *This, for obvious reasons, is undesirable.*

4.3 The Step-Size

Recall that all agents seek to compute a fixed point of H, thus, the iteration (4.5) should converge to $\mathbf{1} \otimes x^\star$, where $x^\star \in \mathbb{R}^d$ is a fixed point of H. Let us assume that it does so, for a, possibly constant, convergent step-size sequence α^k with limit $\alpha^\star$; then,

$$\mathbf{1} \otimes x^\star = W\big((1-\alpha^\star)\mathbf{1} \otimes x^\star + \alpha^\star \tilde{H}(\mathbf{1} \otimes x^\star)\big),$$

which implies that

$$0 = \alpha^\star\big(\mathbf{1} \otimes x^\star - W\tilde{H}(\mathbf{1} \otimes x^\star)\big),$$

and for a non-zero $\alpha^\star$, the component corresponding to agent i satisfies

$$x^\star = \sum_{j=1}^{N} \tilde{W}_{ij} H_j(x^\star).$$

The details are omitted, but this imposes a rather strong structure on the maps H_n: $x^\star$ ought to be, not only a fixed point of the average map H, but also of all weighted averages arising from multiplication by $\tilde{W}$.[1] Avoiding this requires $\alpha^\star$ to be zero.

From now onward, α^k is a non-negative sequence that converges to zero. Moreover, we assume that $\alpha^0 = 1$, that $\alpha^k \in (0, 1]$, and that α^k is non-summable, *i.e.,*

$$\sum_{n=0}^{\infty} \alpha^k = \infty,$$

which is a technical condition. Informally, it prevents α^k from converging to zero "too fast". To motivate why this is instrumental, note that, if α^k converged to zero "too fast", then the recursion (4.5) would "very quickly approach" the distributed average consensus algorithm, thus preventing "enough contribution" of the maps H_n to the computation.

[1] A simple situation where this holds is $H_n(x^\star) = x^\star$ for all n, a prohibitively restrictive condition.

4.4 The Consensus and Off-Consensus Recursions

A good practice to gain insight into the working of a distributed algorithm is to look at the recursions satisfied by the consensus and the off-consensus sequences. Let $z = (z_1, \ldots, z_N) \in (\mathbb{R}^d)^N$; the *consensus component of* z, denoted by $\bar{z}$, is the vector in $\mathbb{R}^d$ defined by

$$\bar{z} := \frac{1}{N}\sum_{n=1}^{N} z_n.$$

The *off-consensus component*, denoted by $\hat{z}$ is the vector in $(\mathbb{R}^d)^N$ defined by

$$\hat{z} := z - \mathbf{1} \otimes \bar{z}.$$

Given a sequence of vectors $z^k \in (\mathbb{R}^d)^N$, the corresponding consensus and the off-consensus sequences will be denoted by $\bar{z}^k$ and $\hat{z}^k$.[2]

4.4.1 Consensus and Off-Consensus Recursions of (4.5)

From the properties of consensus matrices, it follows that

$$\begin{aligned}\bar{z}^{k+1} &= (1-\alpha^k)\bar{z}^k + \alpha^k \bar{H}(\hat{z}^k + \mathbf{1}\otimes\bar{z}^k)\\ &= (1-\alpha^k)\bar{z}^k + \alpha^k H(\bar{z}^k) + \alpha^k\big(\bar{H}(\hat{z}^k + \mathbf{1}\otimes\bar{z}^k) - (H(\bar{z}^k)\big),\end{aligned}$$

where $\bar{H}(z_1, \ldots, z_N) = \frac{1}{N}\sum_{n=1}^{N} H_n(z_n)$, and where z^k was written as $\hat{z}^k + \mathbf{1}\otimes\bar{z}^k$. Similarly, the properties of $\tilde{W}$ (recall that $W = \tilde{W}\otimes I_d$, where $\tilde{W}$ is a consensus matrix), imply that

$$\tilde{W} - \frac{1}{N}\mathbf{1}\mathbf{1}^T z = (\tilde{W} - \frac{1}{N}\mathbf{1}\mathbf{1}^T)(z - \frac{1}{N}\mathbf{1}\mathbf{1}^T z),$$

hence, the off-consensus sequence satisfies the recursion

$$\hat{z}^{k+1} = (W - \frac{1}{N}\mathbf{1}\mathbf{1}^T \otimes I_d)\big((1-\alpha^k)\hat{z}^k + \alpha^k \tilde{H}(\hat{z}^k + \mathbf{1}\otimes\bar{z}^k)\big).$$

Remark 4.4.1. *The consensus recursion corresponds to an "imperfect" Mann iteration,*

[2]The consensus component of $z \in \mathbb{R}^{dN}$ is ambiguous (it is not clear whether it lives in $\mathbb{R}^d$ or $\mathbb{R}^N$); however, the number of agents will always be N and H will always be a map in $\mathbb{R}^d$, hence, the consensus component of $z \in \mathbb{R}^{dN}$ is that of $z \in (\mathbb{R}^d)^N$.

the Mann iteration of H plus an error, $\alpha^k\epsilon^k$, where

$$\epsilon^k = \big(\bar{H}(\hat{z}^k + \mathbf{1}\otimes\bar{z}^k) - H(\bar{z}^k)\big).$$

From the β-Lipschitzianity of each H_n, it follows that

$$\|\epsilon^k\| \leq \beta_1\|\hat{z}^k\|,$$

for a constant β_1 depending on β.

Remark 4.4.2. *We can now motivate the need for the boundedness assumption on each H_n. To have convergence of (4.5) to $\mathbf{1}\otimes x^\star$ is to have convergence to $x^\star$ of the consensus sequence and a vanishing off-consensus sequence,* i.e., *$\bar{z}^k \to x^\star$ and $\hat{z}^k \to 0$; the boundedness of each H_n implies that the recursion satisfied by $\hat{z}^k$ is of the form*

$$\hat{z}^{k+1} = B\big((1-\alpha^k)\hat{z}^k + \alpha^k\delta^k\big),$$

where $\rho(B) < 1$ and δ^k is a bounded sequence. From this observation, a straightforward argument establishes that $\hat{z}^k$ converges to zero, and, thus, as a result of the previous remark, the consensus component satisfies

$$\bar{z}^{k+1} = (1-\alpha^k)\bar{z}^k + \alpha^k H(\bar{z}^k) + \alpha^k\epsilon^k,$$

with $\epsilon^k \leq \beta_1\|\hat{z}^k\| \to 0$. Consequently, the recursion of the consensus sequence is the Mann iteration of H plus a vanishing error.

If, on the other side, at least one map H_n was unbounded, then δ^k could "escape" to infinity "faster" than α^k converges to zero, thus forcing the product $\alpha^k\delta^k$ to "escape" to infinity as well, possibly preventing $\hat{z}^k$ from vanishing. It seems tempting to choose a sequence α^k with a "very very fast" convergence to zero, in order to cancel the speed at which δ^k "escapes to infinity"; however, as previously noted, this "very quickly" leads to $\bar{z}^{k+1} \approx \bar{z}^k$ and $\hat{z}^{k+1} \approx B\hat{z}^k$.

The observations made in the two previous remarks are made precise in Lemma 4.4.1, the proof of which relies on the *Stolz-Cesàro theorem* [72].

Theorem 4.4.1 (Stolz-Cesaro)**.** *Let c^k and b^k be two real sequences and suppose that b_k is strictly monotone and divergent. If*

$$\lim_k \frac{c^{k+1}-c^k}{b^{k+1}-b^k} = l \in [-\infty,\infty],$$

then

$$\lim_k \frac{c^k}{b^k} = l.$$

Proof. See [72] □

Lemma 4.4.1. *Consider the iteration* (4.5)*, where* $\tilde{H}$ *is as defined in* (4.3)*. Suppose that each* H_n *is a* β*-Lipschitz map bounded by* M*. Then,*

1) *The off-consensus sequence* $\hat{z}^k$ *converges to zero;*

2) *The recursion satisfied by consensus sequence is of the form*

$$\bar{z}^{k+1} = (1-\alpha^k)\bar{z}^k + \alpha^k H(\bar{z}^k) + \alpha^k \epsilon^k,$$

where ϵ^k *satisfies* $\|\epsilon^k\| \leq \beta_1 \|\hat{z}^k\|$*, for a non-negative constant* β_1 *depending on* β*. Hence, from 1),* ϵ^k *converges to zero.*

Moreover, if α^k *is of the form* $\alpha^k = \frac{1}{(k+1)^s}$*, with* $0 < s < 1$*, then*

a) *The off-consensus sequence satisfies* $\|\hat{z}^k\| \leq \frac{T}{k^s}$*, for a non-negative constant* T*;*

b) *The error* ϵ^k *satisfies* $\|\epsilon^k\| \leq \frac{\beta_1 T}{k^s}$.

Proof. Part 2) is a straightforward consequence of Lipschitzianity and b) follows directly from 2) and a).

To prove 1), recall that W is symmetric and that $\rho(W - \frac{1}{N}\mathbf{1}\mathbf{1}^T \otimes I_d) < 1$, hence, up to a similarity transformation, the recursion satisfied by $\hat{z}^k$ is of the form

$$\hat{z}^{k+1} = D\big((1-\alpha^k)\hat{z}^k + \alpha^k \delta^k\big),$$

for a bounded sequence δ^k and a diagonal matrix D whose entries are, in magnitude, less than one.[3] Let λ be the largest (in magnitude) diagonal entry of D and observe that the magnitude of each entry of $\hat{z}^k$, that is, each $|\hat{z}_i^k|$, denoted as $a^k = |\hat{z}_i^k|$, satisfies the recursive inequality

$$a^{k+1} \leq |\lambda|\big((1-\alpha^k)a^k + \alpha^k C\big), \tag{4.6}$$

[3]This is a standard argument: choose V satisfying $V^T(W - \frac{1}{N}\mathbf{1}\mathbf{1}^T \otimes I_d)V = D$, and, instead of $\hat{z}^k$, consider the change of coordinates given by $\hat{z}^k \to V\hat{z}^k$.

where $|\lambda| < 1$ and C is a non-negative constant. The fact that $\alpha^0 = 1$ implies that $a^1 \leq |\lambda|C < C$; moreover, if $a^k \leq C$, then

$$a^{k+1} \leq |\lambda|\big((1-\alpha^k)a^k + \alpha^k C\big) \leq |\lambda|\big((1-\alpha^k)C + \alpha^k C\big) = |\lambda|C < C,$$

showing inductively that a^k is a non-negative bounded sequence.

Let $a^\star = \limsup a^k$; from the boundedness of a^k, it follows that $0 \leq a^\star < \infty$. Statement 1) now follows from taking the limit suppremum on both sides of (4.6), leading to

$$a^\star \leq |\lambda| a^\star,$$

and, since $|\lambda| < 1$, we conclude that $a^\star = 0$, thus proving 1).

The proof of a) is based on unfolding the recursive inequality (4.6), yielding

$$a^{k+1} \leq C\sum_{n=0}^{k} |\lambda|^{k-n+1}\alpha^n = C\sum_{n=0}^{k} \frac{|\lambda|^{k-n+1}}{(n+1)^s} = C|\lambda|^{k+2}\sum_{n=1}^{k+1}\frac{\mu^n}{n^s}, \tag{4.7}$$

where $\mu = \frac{1}{|\lambda|} > 1$. We show that the right-hand side of (4.7) is upper bounded by $\frac{T}{k^s}$, for a non-negative constant T. In fact, we show something stronger, namely that the sequence

$$\frac{|\lambda|^{k+2}\sum_{n=1}^{k+1}\frac{\mu^n}{n^s}}{\frac{1}{k^s}} = \frac{\sum_{n=1}^{k+1}\frac{\mu^n}{n^s}}{\frac{\mu^{k+2}}{k^s}}$$

converges: if this sequence converges, it must bounded, hence the right-hand side of (4.7) is upper bounded by $\frac{T}{k^s}$ for a non-negative constant T.

Define $c^k = \sum_{n=1}^{k+1}\frac{\mu^n}{n^s}$ and $b^k = \frac{\mu^{k+2}}{k^s}$ and the idea is to use the Stolz-Cesaro theorem (see Theorem 4.4.1). From $\mu > 1$, it follows that b^k is divergent; however, it is not necessarily strictly monotone: note that $b^{k+1} > b^k$ is equivalent to

$$\mu > \frac{(k+1)^s}{k^s}. \tag{4.8}$$

Nevertheless, the sequence $\frac{(k+1)^s}{k^s}$ tends to one, hence, there exists k_0 such that (4.8) holds for $k \geq k_0$. Redefine c^k and b^k to be, respectively, c^{k+k_0} and b^{k+k_0}, and the conditions of Theorem 4.4.1 hold; note as well that the convergence of shifted sequences implies that of the non-shifted ones.

A straightforward manipulation shows that

$$\Big(\frac{c^{k+1}-c^k}{b^{k+1}-b^k}\Big)^{-1} = \mu\Big(\frac{(k+k_0)(k+2+k_0)}{(k+1+k_0)(k+k_0)}\Big)^s - \Big(\frac{(k+1+k_0)(k+2+k_0)}{(k+1+k_0)(k+k_0)}\Big)^s.$$

Each of the terms inside the parentheses is a quotient of monic polynomials in k of the same degree and the function $x \to x^s$ is continuous. Consequently, as $k \to \infty$,

$$\Big(\frac{c^{k+1}-c^k}{b^{k+1}-b^k}\Big)^{-1} \to \mu - 1 > 0,$$

thus, from Theorem 4.4.1, we conclude that $\frac{c^k}{b^k}$ converges to $\frac{1}{\mu-1}$, proving the result. □

While the results of this section establish that the off-consensus sequence vanishes, no claim is made regarding the convergence of the consensus sequence. This should not be surprising, since the conditions on H are rather weak (Lipschizianity and boundedness of each H_n). In the next two sections we look at what more can be said, if H is further assumed to be a global contraction (Section 4.5) and a local contraction (Section 4.6).

4.5 The Global Contraction Case

Let each H_n satisfy the assumptions of the previous section, that is, each H_n is β-Lipschitz and bounded by M. Throughout this section we further assume that the average map H is a (global) μ- contraction, that is, for all x and y,

$$\|H(x) - H(y)\| \leq \mu\|x - y\|,$$

where $0 \leq \mu < 1$.

From the triangle inequality, it follows that

$$\begin{aligned}\|\bar{z}^{k+1} - x^\star\| &= \Big\|(1-\alpha^k)(\bar{z}^k - x^\star) + \alpha^k\big(H(\bar{z}^k) - H(x^\star)\big) + \alpha^k\epsilon^k\Big\| \\ &\leq \big(1 - \alpha^k(1-\mu)\big)\|\bar{z}^k - x^\star\| + \alpha^k\|\epsilon^k\|,\end{aligned} \tag{4.9}$$

and, from Lemma 4.4.1, $\|\epsilon^k\|$ vanishes. The non-summability of α^k has not yet played a role, and it is here that it will do so. To motivate its relevance, suppose that ϵ^k is identically zero and observe that, in this case, (4.9) implies that $\|\bar{z}^{k+1} - x^\star\| < \|\bar{z}^k - x^\star\|$, so let $m^\star$ be the limit of $\|\bar{z}^k - x^\star\|$. The recursive inequality (4.9) with $\epsilon^k = 0$ can be equivalently written as

$$\|\bar{z}^k - x^\star\| - \|\bar{z}^{k+1} - x^\star\| \geq \alpha^k(1-\mu)\|\bar{z}^k - x^\star\|,$$

and summing both sides from 0 to K yields

$$\|\bar{z}^0 - x^\star\| \geq \|\bar{z}^0 - x^\star\| - \|\bar{z}^{K+1} - x^\star\| \geq (1-\mu)\sum_{k=0}^{K} \alpha^k \|\bar{z}^k - x^\star\| \geq (1-\mu)m^\star \sum_{k=0}^{K} \alpha^k.$$

We conclude that α^k is summable if $m^\star$ is non-zero. Since α^k was assumed to be non-summable, $m^\star$ must be zero, *i.e.*, $\bar{z}^k$ must converge to $x^\star$.

Recursion (4.9) is a particular instance of the recursive inequality

$$b^{k+1} \leq (1-\delta^k)b^k + \delta^k \epsilon^k, \tag{4.10}$$

where b^k is a non-negative real sequence. There is an extensive literature, particularly on stochastic approximation algorithms, with results on under which conditions this recursive inequality implies the vanishment of b^k (for example, see [68] and the references therein; [73]; [74]); the following Lemma summarizes two of these results that are enough for our purposes.

Lemma 4.5.1. *Let b^k and ϵ^k be two non-negative sequences and suppose that ϵ^k converges to zero. Let δ^k be a non-summable sequence in $[0,1]$, that is,*

$$\sum_k \delta^k = \infty.$$

If there exists k_0 such that, for all $k \geq k_0$, inequality (4.10) is satisfied, then,

1) *The sequence b^k converges to zero;*

2) *If both δ^k and ϵ^k are of the form $\frac{A}{(k+1)^s}$, with $0 < s < 1$ and $A > 0$ (with the constant A associated to δ^k possibly different from that associated to ϵ^k), then*

$$b^k \leq \frac{L}{k^s},$$

for a non-negative constant L.

Proof. See [74] for 1). For 2), see Lemma 4 in [73] with $t = 2s$. □

The proof of the following theorem follows directly from Lemmas 4.4.1 and 4.5.1.

Theorem 4.5.1. *Suppose that: 1) each H_n is a β-Lipschitz map bounded by M; 2) the average map H is a global contraction. Let α^k be a non-summable sequence in $[0,1]$ that converges to zero. Then, the recursion (4.5) converges to $\mathbf{1} \otimes x^\star$. Moreover, if α^k is of the form $\frac{A}{(k+1)^s}$, with $0 < s < 1$ and $A > 0$, then,*

a) *The off-consensus sequence satisfies* $\|\hat{z}^k\| \leq \frac{T_1}{k^s}$, *for a non-negative constant* T_1*;*

b) *The consensus sequence satisfies* $\|\bar{z}^k - x^\star\| \leq \frac{T_2}{k^s}$, *for a non-negative constant* T_2.

From both a) and b) and the fact that $z^k = \mathbf{1} \otimes \bar{z}^k + \hat{z}^k$, *it follows that*

$$\|z^k - \mathbf{1} \otimes x^\star\| \leq \frac{T}{k^s},$$

for a non-negative constant T.

4.6 The Local Contraction Case

Similar to the global contraction case, in this section, each H_n is assumed to be a β-Lipschitz map bounded by a constant M. Contrary to the previous Section, the average map H is not assumed to be a global contraction, but merely a local contraction with respect to $x^\star$, that is, there exist constants $\delta > 0$ and $0 \leq \mu < 1$ such that

$$\|H(x) - x^\star\| \leq \mu \|x - x^\star\|,$$

for all x satisfying $\|x - x^\star\| < \delta$. Recall that, if H is differentiable at $x^\star$, a sufficient condition for H to be a local contraction with respect to $x^\star$ is that $\rho\big(\mathbf{J}_H(x^\star)\big) < 1$ (see Ostrowski's theorem in Chapter 3).

Informally, in this section we establish that if $\bar{z}^k$ gets sufficiently close to $x^\star$ for a sufficiently large k, then the sequence $\bar{z}^k$ converges to $x^\star$. Additionally, if α^k is of the form

$$\alpha^k = \frac{A}{(k+1)^s},$$

for a non-negative constant A and for $0 < s < 1$, then, the convergence of the consensus sequence satisfies

$$\|\bar{z}^k - x^\star\| \leq \frac{T}{k^s},$$

for a non-negative constant T.

As in the global contraction case, the idea is to use Lemma 4.5.1 to prove the result. However, contrary to the global contraction case, we cannot use this result immediately since it is not necessarily true that the whole sequence satisfies (4.9). In fact, while

$$\|\bar{z}^{k+1} - x^\star\| \leq \big(1 - \alpha^k(1-\mu)\big)\|\bar{z}^k - x^\star\| + \alpha^k \|\epsilon^k\|,$$

provided that $\|\bar{z}^k - x^\star\| < \delta$, this does not necessarily imply that $\|\bar{z}^{k+1} - x^\star\| < \delta$, hence, $\bar{z}^{k+1}$ does not necessarily satisfy (4.9), that is, it does not necessarily follow that

$$\|\bar{z}^{k+2} - x^\star\| \leq \big(1 - \alpha^{k+1}(1-\mu)\big)\|\bar{z}^{k+1} - x^\star\| + \alpha^{k+1}\|\epsilon^{k+1}\|.$$

To overcome this issue, let k_0 be such that

$$\|\epsilon^k\| \leq (1-\mu)\delta$$

for $k \geq k_0$, and observe that if $\bar{z}^k$ satisfies $\|\bar{z}^k - x^\star\| < \delta$, for $k \geq k_0$, then,

$$\begin{aligned}\|\bar{z}^{k+1} - x^\star\| &\leq \big(1 - \alpha^{k+1}(1-\mu)\big)\|\bar{z}^k - x^\star\| + \alpha^k\|\epsilon^{k+1}\| \\ &\leq \big(1 - \alpha^k(1-\mu)\big)\delta + \alpha^k(1-\mu)\delta = \delta.\end{aligned}$$

By induction, this shows that if $\|\bar{z}^{k_1} - x^\star\| < \delta$, for $k_1 \geq k_0$, then, for all $k \geq k_1$,

$$\|\bar{z}^{k+1} - x^\star\| \leq \big(1 - \alpha^k(1-\mu)\big)\|\bar{z}^k - x^\star\| + \alpha^k\|\epsilon^k\|.$$

This observation, together with Lemmas 4.4.1 and 4.5.1, yields the following theorem.

Theorem 4.6.1. *Suppose that: 1) each H_n is a β-Lipschitz map bounded by M; 2) the average map H is a local contraction with respect to $x^\star$. Let α^k be a non-summable sequence in $[0,1]$ that converges to zero. Then, there exists k_0 such that if, $\|\bar{z}^{k_1} - x^\star\|$ is sufficiently small, for $k_1 \geq k_0$, then recursion (4.5) converges to $\mathbf{1} \otimes x^\star$. Moreover, if $\alpha^k = \frac{A}{(k+1)^s}$, with $0 < s < 1$ and $A > 0$, then*

a) *The off-consensus sequence satisfies $\|\hat{z}^k\| \leq \frac{T_1}{k^s}$, for a non-negative constant T_1;*

b) *The consensus sequence satisfies $\|\bar{z}^k - x^\star\| \leq \frac{T_2}{k^s}$, for a non-negative constant T_2.*

From both a) and b) and the fact that $z^k = \mathbf{1} \otimes \bar{z}^k + \hat{z}^k$, it follows that

$$\|z^k - \mathbf{1} \otimes x^\star\| \leq \frac{T}{k^s},$$

for a non-negative constant T.

4.7 Comments and References

In this chapter, we presented a distributed algorithm for finding fixed points $x^\star$ of an average map H, and showed two things:

1) If H is a global contraction, then the corresponding distributed algorithm converges to $\mathbf{1} \otimes x^\star$;

2) If H is a local contraction, then, if, for a sufficiently large k, z^k is sufficiently close to $\mathbf{1} \otimes x^\star$, then the distributed algorithm converges to $\mathbf{1} \otimes x^\star$.

While, similar to its centralized counterpart, $x^{k+1} = H(x^k)$, the distributed algorithm (4.5) requires only the agents to store a d-dimensional vector at each iteration, it has at least two big downsides, as a consequence of relying on a vanishing step-size. First, the agents need to know which iteration they are executing. Secondly, the rate of convergence of $x^{k+1} = H(x^k)$ is sacrificed. To be fair, we point out that we only provided an upper bound on the convergence rate that seems to indicate its sub-linearity; however, as the simulations in Chapter 7 demonstrate, the rate is in fact sub-linear.

4.7.1 Distributed Optimization

As noted in Remark 4.2.2, if each map H_n is of the form $H_n = I - \nabla f_n$, for a differentiable function f_n, algorithm (4.5) is nothing but the distributed gradient descent with shrinking step-size (see [25]). Historically speaking, this algorithm is of great importance, being one of the first algorithms suggested to solve the distributed optimization problem described in Chapter 3. A corollary of Theorem 4.5.1 is that the distributed gradient descent with vanishing step-size converges, whenever each f_n is a Lipschitz function and the average function, $\frac{1}{N}\sum_{n=1}^N f_n$, is strongly convex. To prove this via Theorem 4.5.1, we only need to appeal to the following standard result in convex optimization which shows that "centralized" gradient descent is a global contraction.

Lemma 4.7.1. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable strongly convex function with β-Lipschitz gradient. Then, the gradient map, that is, the map on $\mathbb{R}^d$ defined by*

$$x \to x - \alpha \nabla f(x),$$

is a contraction for a sufficiently small non-negative constant α.

Proof. See any standard book on optimization; for example [75]. □

Suppose that: 1) each f_n is a differentiable function on $\mathbb{R}^d$ with a Lipschitz gradient; 2) the gradient maps, ∇f_n, are all bounded by the same constant M; 3) the average map,

$$f = \frac{1}{N}\sum_{n=1}^N f_n$$

is strongly convex. Consider the maps H_n on $\mathbb{R}^d$ given by

$$H_n(x) = x - \alpha \nabla f_n(x)$$

and, from the lemma above, let α be such that the average map H,

$$H(x) = \frac{1}{N} \sum_{n=1}^{N} H_n(x) = x - \alpha \nabla f(x),$$

is a μ-contraction. Recursion (4.5) with the maps H_n so-defined reduces to the distributed gradient descent with shrinking step-size, whose convergence guarantees follow from Theorem 4.5.1, the contractiveness of H, the Lipschitzianity of each H_n, and the boundedness of each ∇f_n.

Remark 4.7.1. *Regarding the boundedness condition there is a subtlety. Our results require the maps H_n to be bounded and, in this case, even though the gradients are assumed to be bounded, the maps H_n are not. However, a careful inspection of the convergence proofs reveals that the only place where this condition is used is in the proof of Theorem 4.4.1, the modification of which presents no serious challenge. Observe that the distributed gradient descent recursion is given by*

$$z^{k+1} = W\big(z^k - \alpha^k \nabla(z^k)\big), \tag{4.11}$$

where $\nabla(z^k) = \big(H_1(z_1^k), \ldots, H_N(z_N^k)\big)$. Hence, (4.6) *in Theorem 4.4.1 should be replaced by*

$$a^{k+1} \leq |\lambda|(a^k + \alpha^k C), \tag{4.12}$$

which, from Lemma 1.6 in [68], converges to zero. Unfolding (4.12) *results in an inequality that can be analyzed in the same way as* (4.7)*, that is, using the Stolz-Cesaro theorem.*

4.7.2 The Memory-Convergence Rate Trade-Off

The distributed algorithm (4.5) is of the form

$$z^{k+1} = F_k(z^k),$$

where F_k is a map defined on $\mathbb{R}^{dN}$. As a result, the implementation of (4.5) only requires each agent to store a d-dimensional vector, the dimension of the domain of H. Whenever

H is a global contraction, (4.5) converges to $\mathbf{1} \otimes x^\star$, although it does so at a sub-linear rate, a qualitatively slower rate than that exhibited by the convergence of $x^{k+1} = H(x^k)$ towards $x^\star$; this "downgrade" in convergence rate is a consequence of the vanishing step-size. With this in mind, a natural and important question is whether there exists a distributed algorithm that can be written as

$$z^{k+1} = F(z^k),$$

for a map F defined on $\mathbb{R}^{dN}$ and that preserves the linear convergence rate of $x^{k+1} = H(x^k)$, whenever H is a global contraction; or, by contrast, whether insisting that each agent should merely store a d-dimensional vector at each iteration, necessarily comes with the price of a sacrifice in convergence rate.

Two interesting works, [76] and [77], touch upon this issue. The first, [76], shows that if α^k is constant in (4.11), rather than vanishing, then, instead of convergence to $\mathbf{1} \otimes x^\star$, convergence is only guaranteed to a neighborhood of $\mathbf{1} \otimes x^\star$; specifically, in the strongly convex setup, the algorithm converges linearly to a $O(\alpha)$-radius neighborhood of $\mathbf{1} \otimes x^\star$ (see [76] for further details and simulations). This, however, could be a specific feature of (4.11), and there could be another algorithm that preserved the convergence guarantees of the centralized gradient, requiring agents to only store a d-dimensional vector; this last point leads us to the second work, [77]. In [77], the authors show that for a sufficiently general class of distributed algorithms with constant step-sizes, no algorithm where the agents merely store a d-dimensional vector can solve the distributed optimization problem, for all graph topologies and functions $f_1, \ldots, f_N$ (see [77] for the formal treatment); as the authors put it

> "This provides an explanation as to why the distributed gradient descent algorithm must use a diminishing stepsize..."

4.7.3 On The Convergence Proof

In the global contractive case, the high-level structure of the convergence proof is as follows. First, the recursion

$$z^{k+1} = F_k(z^k)$$

is "broken" into two recursions

$$\bar{z}^{k+1} = F_{1,k}(\bar{z}^k, \hat{z}^k)$$
$$\hat{z}^{k+1} = F_{2,k}(\bar{z}^k, \hat{z}^k),$$

and it is noted that the convergence of z^k to $\mathbf{1} \otimes x^\star$ is equivalent to the convergence of $\bar{z}^k$ to $x^\star$ and the vanishment of $\hat{z}^k$. Secondly, it is shown that $\hat{z}^k$ vanishes by ignoring the contribution of $\bar{z}^k$ and by "replacing this" by a bounded sequence. Finally, Lipschitzianity is used to show that,

$$\|\bar{z}^{k+1} - x^\star\| \leq \big(1 - \alpha_k(1-\mu)\big)\|\bar{z}^k - x^\star\| + \alpha^k\beta_1\|\hat{z}^k\|,$$

which, after having established that $\hat{z}^k$ vanishes, leads to a relatively straightforward proof via Lemma 4.5.1.

This is a rather sequential proof that can be broken into the following two logical statements:

(A) The sequence $\hat{z}^k$ vanishes.

(B) If the sequence $\hat{z}^k$ vanishes, then the sequence $\bar{z}^k$ converges to $x^\star$.

The proof of **(A)** is achieved by an "exogenous" argument (the boundedness of each H_n), given that it ignores how $\bar{z}^k$ "helps" $\hat{z}^k$ to vanish. By contrast, the analysis of the algorithms in the next chapter replaces this "hierarchical reasoning" by a more "circular" one. In fact, logically speaking, it looks more like proving:

(A') If the sequence $\hat{z}^k$ vanishes, then $\bar{z}^k$ converges to $x^\star$.

(B') If the sequence $\bar{z}^k$ converges to $x^\star$, then $\hat{z}^k$ vanishes.

4.7.3.1 The Stolz-Cesàro Theorem

We finish this Chapter with some remarks on the Stolz-Cesàro theorem(Theorem 4.4.1). While its proof is beyond the scope of this work (we refer the reader to [72]), we mention that it can be seen as a discrete-time analog of the much more familiar *L'Hôpital's rule*. To see this, we introduce the extremely useful notation used in [64].

A sequence x^k in $\mathbb{R}$ is essentially a function $f : \mathbb{N}_0 \to \mathbb{R}$, the identification being $x^k = f(k)$. With this in mind we can define the "derivative" of f, denoted by f' as the function $f'(k) = f(k+1) - f(k)$, since this measures the change in f at time k. Under this notation, an algorithm of the form $x^{k+1} = H(x^k)$ is the discrete-time analog of a

differential equation, since it can be written as

$$f' = \tilde{H}(f), \tag{4.13}$$

where $\tilde{H}(x) = H(x) - x$. To see this, note that

$$f(k+1) - f(k) = f'(k) = \tilde{H} \circ f(k) = H\big(f(k)\big) - f(k),$$

which is equivalent to $f(k+1) = H\big(f(k)\big)$; the identification $x^k = f(k)$ leads to the algorithm $x^{k+1} = H(x^k)$. Observe that, unlike in the theory of differential equations, the existence and uniqueness of the solution of a *discrete-time differential equation*, *i.e.* (4.13), is immediate, that solution being simply $f(k) = H^k\big(f(0)\big)$.

Another theorem, whose proof in the discrete-time version poses no serious challenge, is the *fundamental theorem of calculus.* In fact, if $f : \mathbb{N} \to \mathbb{R}$ is a function and we define the "integral" of f from $k = 0$ to K by

$$\int_{k=0}^{K} f := \sum_{k=0}^{K} f(k),$$

we easily see that

$$\int_{k=0}^{K} f' = \sum_{k=0}^{K} f'(k) = \sum_{k=0}^{K} f(k+1) - f(k) = f(K+1) - f(0).$$

Concerning the Stolz-Cesàro theorem, note that if the sequences c^k and b^k in the statement of Theorem 4.4.1 are identified, respectively, with $f(k)$ and $g(k)$, then, the Stolz-Cesàro theorem can be reformulated as: if

$$\lim_{k\to\infty} \frac{f'(k)}{g'(k)} = l$$

then

$$\lim_{k\to\infty} \frac{f(k)}{g(k)} = l;$$

this is the familiar L'Hôpital's rule when f and g are functions in $\mathbb{R}$ rather than $\mathbb{N}$.

This connection between "discrete-time calculus" and standard differential calculus sheds some light on why it is "natural" to use the Stolz-Cesàro theorem when approaching

the problem faced in the proof of Lemma 4.4.1. Recall that, in (4.7), we showed that

$$C|\lambda|^{k+2}\sum_{n=1}^{k+1}\frac{\mu^n}{n^s}\leq\frac{T}{k^s},$$

for a non-negative constant T. We did this by observing that

$$\frac{|\lambda|^{k+2}\sum_{n=1}^{k+1}\frac{\mu^n}{n^s}}{\frac{1}{k^s}}=\frac{\sum_{n=1}^{k+1}\frac{\mu^n}{n^s}}{\frac{\mu^{k+2}}{k^s}}$$

converged, due to the Stolz-Cesàro theorem. If we were not "aware" of the existence of the Stolz-Cesàro, a natural way to approach this problem would be to see how its continuous-time "twin" looks like; that amounts to the inspection of the limit

$$\lim_{k\to\infty}\frac{\int_1^{k+1}\frac{\mu^x}{x^s}dx}{\frac{\mu^{k+2}}{k^s}}.$$

An elementary calculus approach is to note that the L'Hôpital's rule, together with the Fundamental Theorem of Calculus applied to the numerator, easily yields the limit. This illustrates why it is natural to look for a discrete-time version of the L'Hôpital's rule.

To conclude, we remark that, incidentally, this good practice of inspecting the "continuous-time twin" of a problem in discrete-time is behind the derivation in Appendix B corresponding to Remark 3.2.7 of Chapter 3. There, the goal was to show that the iteration $x^{k+1}=x^k-(x^k)^3$ converges to zero at a sub-linear rate, if initialized sufficiently close to zero. The continuous-time analog is the differential equation

$$f'=-f^3,$$

which is easily solved by integrating both sides of

$$-\frac{f'}{f^3}=1.$$

The details are omitted but each step of the derivation can be motivated by a corresponding continuous-time step.

Chapter 5

Distributed Banach-Picard Iteration

5.1 Introduction

This chapter addresses the problem described in Section 3.3 of Chapter 3, that, is, it shows how, given N maps $H_1, \ldots, H_N$ on $\mathbb{R}^d$ and an undirected connected graph $\mathcal{G}$, to construct maps $F : \mathbb{R}^{dN} \times \mathbb{R}^{dN} \to \mathbb{R}^{dN} \times \mathbb{R}^{dN}$ such that:

1) the Banach-Picard iteration of F has distributed implementation respecting $\mathcal{G}$;

2) the fixed points of F are of the form $(\mathbf{1} \otimes x^\star, w^\star)$, where $x^\star$ is a fixed point of $H = 1/N \sum_{n=1}^{N} H_n$;

3) the convergence properties of the Banach-Picard iteration of F with respect to a fixed point $(\mathbf{1} \otimes x^\star, w^\star)$ are those of the Banach-Picard iteration of H with respect to $x^\star$.

Chapter 4 introduced an algorithm for the distributed computation of fixed points that essentially (for the precise meaning see the results therein) satisfies 1) and 2), although it fails to meet 3); in fact, as a consequence of relying on a diminishing step-size, the convergence properties are lost. As anticipated in Section 4.7.2 of Chapter 4, the price paid for 3) is that the implementation of the Banach-Picard iteration of F requires twice the memory of the algorithm described therein.

The contents of this chapter are an original contribution and build upon our work, [13], published in the *IEEE Transactions on Automatic Control*. [13] addresses 3), showing that $\rho\big(\mathbf{J}_H(x^\star)\big) < 1$ implies that $\rho\big(\mathbf{J}_F\big(\mathbf{1} \otimes x^\star, w^\star\big)\big) < 1$, by relying on an result ([78, 17]) from *perturbation theory* (PT) of linear operators that establishes the differentiability of semi-simple eigenvalues. The results of this chapter extend those of [13] in three ways. First, the proof in [13] is shown to be robust enough to handle a particular case of $\rho\big(\mathbf{J}_H(x^\star)\big) > 1$; specifically, we show that if $\mathbf{J}_H(x^\star)$ has an eigenvalue with real part

larger than one, then $\rho\big(\mathbf{J}_F\big(\mathbf{1}\otimes x^\star, w^\star\big)\big) > 1$. Second, if H is a continuous (not necessarily differentiable) global contraction with respect to $x^\star$, then the same is true for F with respect to $(\mathbf{1}\otimes x^\star, w^\star)$. Third, if H is a continuous local contraction with respect to $x^\star$, then the same is true for F with respect to $(\mathbf{1}\otimes x^\star, w^\star)$. The local contraction result appears to be a a considerable generalization of the result in [13], in the sense that, from *Ostrowski's theorem* (see Chapter 3), a map satisfying $\rho\big(\mathbf{J}_H(x^\star)\big) < 1$ is, in particular, a local contraction. However, this is not the case, because, to arrive at the result, we assume that each map H_n is Lipschitz, an assumption absent in [13].

The three extensions are relevant on their own. The first, is relevant in the light of Theorem 3.2.4 of Chapter 3 if both H and F are local diffeomorphisms, then, with a random initialization, the probability that H and F converge, respectively, towards $x^\star$ and $(\mathbf{1}\otimes x^\star, w^\star)$ is zero. The second contains, as a particular case, the EXTRA (see [28]) and DIGing (see [32, 31]) algorithms for gradient descent, hence it is interesting as a "unifying proof"; the proof is loosely based on that in [31] for the strongly convex case, and we believe that improves on it by identifying its key blocks. Finally, the third, is relevant as a step towards the generalization of the result in [13], that is, the continuous (not necessarily differentiable) local contraction case without the Lipschitzianity assumption.

To finish, Appendix A presents an "elementary" proof of a simplified version (enough for our needs) of the PT result. In fact, as far as we know, the known proofs of this result (see [78] and [17]) rely on tools from Complex Analysis. Our proof, by contrast, only relies on results from matrix analysis (*e.g. Geršgorin's Theorem*) and is inspired on the proof for *simple eigenvalue* case that is presented in [57], constituting, therefore, a generalization of the latter.

5.2 The Family of Algorithms

Consider, as in Section 3.3 of Chapter 3, a network of N agents, where the interconnection structure is represented by an undirected and connected graph. Recall that agent n holds a map $H_n : \mathbb{R}^d \to \mathbb{R}^d$, and the goal is to compute a fixed point of the average map,

$$H = \frac{1}{N}\sum_{n=1}^{N} H_n.$$

Moreover, each agent is restricted to performing computations using H_n and communicating with its neighbors.

5.2.1 "Distributed Description" of the Fixed Points of H

Let $\mathcal{C}$ be the *consensus space* of $\mathbb{R}^{dN}$, that is, $\mathcal{C} := \{(z_1, \ldots, z_N) \in \mathbb{R}^{dN} : z_1 = \cdots = z_N\}$, and let $\pi : \mathbb{R}^d \to \mathcal{C}$ be the map that lifts the points in $\mathbb{R}^d$ to the consensus space of $\mathbb{R}^{dN}$, *i.e.,* $\pi(x) = \mathbf{1} \otimes x$. This initial section provides a "distributed description" of the fixed points of H through the construction of a map $G : \mathbb{R}^{dN} \times \mathbb{R}^{dN} \to \mathbb{R}^{dN} \times \mathbb{R}^{dN}$ for which $\pi\big(\mathrm{Fix}(H)\big)$ is the set of $z \in \mathbb{R}^{dN}$ such that $G(z, w) = 0$, for a $w \in \mathbb{R}^{dN}$. Crucially, computing the map G has distributed implementation, hence the phrasing *"distributed description"*.

The motivation for searching G should be quite evident: once G is found, a naïve distributed algorithm to consider is

$$\begin{aligned} z^{k+1} &= z^k + \alpha G_1(z^k, w^k) \\ w^{k+1} &= w^k + \alpha G_2(z^k, w^k), \end{aligned} \tag{5.1}$$

for a non-zero constant α. In fact, provided that G is continuous, if z^k and w^k converge, respectively, to $z^\star$ and $w^\star$, then, $G(z^\star, w^\star) = 0$, which in turn implies that $z^\star \in \pi\big(\mathrm{Fix}(H)\big)$.

Let L be a symmetric $dN \times dN$ matrix with $\ker(L) = \mathcal{C}$, the consensus space. A point $z \in \mathbb{R}^{dN}$ is the lift of a fixed point of H, *i.e.,* $z \in \pi\big(\mathrm{Fix}(H)\big)$, if and only if

$$\begin{cases} H(\bar{z}) - \bar{z} & = 0 \\ Lz & = 0, \end{cases}$$

where recall that $\bar{z} = 1/N \sum_{n=1}^N z_n$ (see Chapter 4). Equivalently,

$$\begin{cases} \big(\frac{1}{N}\mathbf{1}^T \otimes I_d\big)\big(\tilde{H}(\mathbf{1} \otimes \bar{z}) - \mathbf{1} \otimes \bar{z}\big) & = 0 \\ Lz & = 0, \end{cases}$$

where $\tilde{H}(z_1, \ldots, z_N) = \big(H_1(z_1), \ldots, H_N(z_N)\big)$ (see Chapter 4). Now, if $Lz = 0$, then all the components of z are equal, *i.e.,* $z = \mathbf{1} \otimes \bar{z}$. Therefore,

$$\begin{cases} \big(\frac{1}{N}\mathbf{1}^T \otimes I_d\big)\big(\tilde{H}(z) - z\big) & = 0 \\ Lz & = 0. \end{cases}$$

The first of these equations is equivalent to $\tilde{H}(z) - z \in \ker(L)^\perp$, since the rows of $\mathbf{1}^T \otimes I_d$ form a basis of this space. From $\ker(L)^\perp = \mathrm{range}(L)$, it follows that $z \in \pi\big(\mathrm{Fix}(H)\big)$ if

and only if

$$\begin{cases} \tilde{H}(z) - z & \in \text{range}(L) \\ z & \in \ker(L). \end{cases} \tag{5.2}$$

Let $G(z, w) = \big(\tilde{H}(z) - z + Lw, -Lz\big)$ and, by further assuming L to be compatible with the graph structure[1], we obtain a "distributed description" of the fixed points of H, that is, $z \in \pi\big(\text{Fix}(H)\big)$ if and only if there exists $w \in \mathbb{R}^{dN}$ such that $G(z, w) = 0$. Furthermore, since products by L respect the network topology and computing $\tilde{H}$ can be carried out locally, the computation of the map G has distributed implementation.

5.2.2 Parametric Family of Algorithms

A natural distributed algorithm to consider is (5.1): if (5.1) converges, then the agents succeed in agreeing on a fixed point of H. However, we are concerned with ensuring convergence, thus, to this end, we consider a parametric family of algorithms, of which (5.1) is an instance (take $\alpha = -\beta$ and $\eta = 0$ below). The rest of this chapter is devoted to studying the parametric family defined by

$$\begin{aligned} F : \mathbb{R}^{dN} \times \mathbb{R}^{dN} &\to \mathbb{R}^{dN} \times \mathbb{R}^{dN} \\ (z, w) &\to \big(z + \alpha R(z) + \beta L^s w - \eta L z, w - \beta L^s z\big), \end{aligned} \tag{5.3}$$

where $R(z) = \tilde{H}(z) - z$, and

1) α, β, and η are positive;

2) s is either 1 or 1/2;

3) L is a $dN \times dN$ matrix such that

- **a)** L is symmetric and positive semidefinite;
- **b)** $\rho(L) < 2$;
- **c)** $\ker(L) = \mathcal{C} = \{(z_1, \ldots, z_N) \in \mathbb{R}^{dN} : z_1 = \cdots = z_N\}$; and
- **d)** $L = \tilde{L} \otimes I_d$, where $\tilde{L}$ is $N \times N$ and has the property that $\tilde{L}_{ij} = 0$ if and only if agents i and j are not neighbors (thus establishing the compatibility with the network structure).

The following remarks are due:

[1]The existence of L satisfying these conditions follows from the results in Chapter 2. For example, take $L = I - W$, where $W = \tilde{W} \otimes I_d$ and $\tilde{W}$ is a consensus matrix.

Remark 5.2.1. *The existence of L with the abovementioned conditions is ensured by the results in Chapter 2. For example, let $W = \tilde{W} \otimes I_d$, for a consensus matrix $\tilde{W}$, and define $L = I - W$.*

Remark 5.2.2. *Because L is assumed to be positive semidefinite, the square root of L, $L^{\frac{1}{2}}$ (corresponding to $s = 1/2$), is well defined.*

Remark 5.2.3. *For $s = 1$, the iteration $(z^{k+1}, w^{k+1}) = F(z^k, w^k)$ has distributed implementation, and, in contrast, for $s = 1/2$, given the presence of $L^{\frac{1}{2}}$, it does not (whereas products by L only require each node to communicate with its neighbors, the same is not true with $L^{\frac{1}{2}}$, given that $L^{\frac{1}{2}}$ need not be compatible with the graph topology). Nevertheless, as shown in Section 5.4, the elimination of the second variable yields an algorithm having distributed implementation.*

5.2.3 Fixed Points of F and the Map $\tilde{F}$

Consider the map $G(z, w) = \big(\tilde{H}(z) - z - Lw, Lz\big)$ from Section 5.2.1. We saw that $x^\star$ is a fixed point of H if and only if there exists $w^\star$ such that $G(\mathbf{1} \otimes x^\star, w^\star) = 0$. To have a "distributed description", L must be compatible with the network structure. However, to have a, not necessarily distributed, description of the form $G(\mathbf{1} \otimes x^\star, w^\star) = 0$, it is enough to have a symmetric L with $\ker L = \mathcal{C}$. Consequently, if L is replaced by $\alpha/\beta L^s$ in G, it still holds that $x^\star$ is a fixed point of H if and only there exists $w^\star$ such that

$$\tilde{H}(\mathbf{1} \otimes x^\star) - \mathbf{1} \otimes x^\star + \frac{\beta}{\alpha} L^s w^\star = 0.$$

This establishes the following simple lemma.

Lemma 5.2.1. *If $x^\star \in \mathbb{R}^d$ is a fixed point of H, then there exists $w^\star \in \mathbb{R}^{dN}$ such that $(\mathbf{1} \otimes x^\star, w^\star)$ is a fixed point of F. Conversely, if $(z^\star, w^\star)$ is a fixed point of F, then $z^\star = \mathbf{1} \otimes \bar{z}^\star$ and $\bar{z}^\star$ is a fixed point of H.*

The following caveat must be addressed before proceeding: $w^\star$ is not unique. In fact, if $w^\star$ is such that $(\mathbf{1} \otimes x^\star, w^\star)$ is a fixed point of F, then any point in the set $w^\star + \ker(L)$ is also a fixed point of F. Informally, to each fixed point $x^\star$ of H, there is an associated d-dimensional (the dimension of $\ker(L)$) affine subspace of fixed points of F. As a result, regardless of the properties of H (global contraction, local contraction, etc), no fixed point of F is isolated.[2] This is a crucial observation, because it shows to be pointless to

[2] A fixed point $x^\star$ of a map is isolated if there exists a neighborhood of $x^\star$ in which the map has no fixed points other than $x^\star$.

attempt to prove the local contractivness of F, since, otherwise, its fixed points would be isolated.

To address this issue, we begin by singling out a natural $w^\star$ to play the role of a "reference" fixed point. In fact, every $w \in \mathbb{R}^{dN}$ has a unique orthogonal decomposition in $\ker(L) + \ker(L)^\perp$, $w = \mathbf{1} \otimes \bar{w} + \hat{w}$, where $\bar{w}$ and $\hat{w}$ are, respectively, the consensus and the off-consensus components (see Chapter 4). Given a fixed point $x^\star$ of H, the set of $w^\star$ such that $(\mathbf{1} \otimes x^\star, w^\star)$ is a fixed point of F is the set of $w^\star$ satisfying $-\alpha/\beta R(\mathbf{1} \otimes x^\star) = L^s w^\star$. Therefore, a natural $w^\star$ to take is the one (it is unique) with zero consensus component, *i.e.,* the unique $w^\star$ such that

$$\begin{cases} -\frac{\alpha}{\beta} R(\mathbf{1} \otimes x^\star) & = L^s w^\star \\ \bar{w}^\star & = 0. \end{cases}$$

The unique $w^\star$ is easily found using the *Moore-Penrose inverse*[3] of L^s, denoted by $(L^s)^+$, *i.e.,* let

$$w^\star = -\frac{\alpha}{\beta}(L^s)^+ R(\mathbf{1} \otimes x^\star).$$

All of this leads to the following refined version of Lemma 5.2.1.

Lemma 5.2.2. *Let $\psi : \mathbb{R}^d \to \mathbb{R}^{dN}$ be the map defined by*

$$\psi(x) = \Big(\mathbf{1} \otimes x, -\frac{\alpha}{\beta}(L^s)^+ R(\mathbf{1} \otimes x)\Big).$$

If $x^\star$ is a fixed point of H, then $\psi(x)$ is a fixed point of F. Conversely, if $(z^\star, w^\star)$ is a fixed point of F, then $\bar{z}^\star$ is a fixed point of H, and

$$(z^\star, w^\star) = \psi(\bar{z}^\star) + (0, \bar{w}^\star).$$

5.2.3.1 The Map $\tilde{F}$

We are ultimately interested in proving statements such as: if $x^\star$ is a fixed point of H satisfying $\rho\big(\mathbf{J}_H(x^\star)\big) < 1$, then $\psi(x^\star)$ is a fixed point of F satisfying $\rho\big(\mathbf{J}_F(\psi(x^\star)\big) < 1$. However, this quest is doomed to fail: $\psi(x^\star)$ is not an isolated fixed point, thus preventing F from being a local contraction with respect to $\psi(x^\star)$. To overcome this

[3]The Moore-Penrose inverse of a symmetric matrix A can be defined in the following way: let V be an orthogonal matrix such that $V^T D V = A$ for a diagonal matrix D, the Moore-Penrose inverse of A is the matrix $V^T \tilde{D} V$, where $\tilde{D}$ is a diagonal matrix with $\tilde{D}_{ii} = 0$ if $D_{ii} = 0$ and $\tilde{D}_{ii} = D_{ii}^{-1}$ otherwise. It can be proved that this construction is independent of V, that is, we obtain the same matrix if we replace V by any other $\tilde{V}$ satisfying $\tilde{V}^T \hat{D} V = A$, where $\hat{D}$ is diagonal.

issue, this section introduces a map $\tilde{F}$ that does not distinguish between fixed points points in $\psi(x^\star) + \big(0, \ker(L)\big)$. The construction of $\tilde{F}$ (see below) stems from the simple observation that, if (z_1, w_1) and (z_2, w_2) satisfy $(z_1, w_1) - (z_2, w_2) \in \big(0, \ker(L)\big)$, then $F(z_1, w_1) - F(z_2, w_2) \in \big(0, \ker(L)\big)$. This shows that F "descends to a map" $\tilde{F}$ on the quotient $\mathbb{R}^{dN} \times \mathbb{R}^{dN} / \big(\ker(L)\big)$, defined by $(z, w) + \big(0, \ker(L)\big) \to F(z, w) + \big(0, \ker(L)\big)$. The notion of quotient, in the case of vectors spaces, can be easily expressed in coordinates, by introducing a matrix $\tilde{U}$ with columns forming an orthonormal basis of range(L), because $\mathbb{R}^{dN} / \big(\ker(L)\big)$ is isomorphic to range(L). To this end, let

$$\begin{aligned} \tilde{F} : \mathbb{R}^{dN} \times \mathbb{R}^{d(N-1)} &\to \mathbb{R}^{dN} \times \mathbb{R}^{d(N-1)} \\ (z, \tilde{w}) &\to \big(z + \alpha R(z) + \beta L^s \tilde{U} \tilde{w} - \eta L z, \tilde{w} - \beta \tilde{U}^T L^s z\big), \end{aligned} \tag{5.4}$$

where $\tilde{U}$ is a matrix with columns forming an orthonormal basis of range(L), and, without loss of generality, assume that the columns of $\tilde{U}$ are eigenvectors of L associated to non-zero eigenvalues of L. Moreover, let

$$\begin{aligned} \tilde{\psi} : \mathbb{R}^d &\to \mathbb{R}^{dN} \times \mathbb{R}^{d(N-1)} \\ x &\to \big(\mathbf{1} \otimes x, -\frac{\alpha}{\beta} \tilde{U}^T (L^s)^+ R(\mathbf{1} \otimes x)\big). \end{aligned}$$

Given that $\tilde{F}$ "ignores" the consensus component of w, a straightforward calculation shows that, "under $\tilde{\psi}$", the fixed points of H are in one-to-one correspondence with those of $\tilde{F}$; this is the content of the next lemma.

Lemma 5.2.3. *If $x^\star$ is a fixed point of H, then $\tilde{\psi}(x^\star)$ is a fixed point of $\tilde{F}$. Conversely, if $(z^\star, w^\star)$ is a fixed point of $\tilde{F}$, then $(z^\star, w^\star) = \psi(\bar{z}^\star)$.*

This chapter establishes three types of results:

1) If $\rho\big(\mathbf{J}_H(x^\star)\big) < 1$ or $\mathbf{J}_H(x^\star)$ has an eigenvalue with real part larger than one, then, for particular choices of η and β depending on s, there exists α sufficiently small such that

$$\operatorname{sign}\Big(1 - \rho\big(\mathbf{J}_{\tilde{F}}(\tilde{\psi}(x^\star))\big)\Big) = \operatorname{sign}\Big(1 - \rho\big(\mathbf{J}_H(x^\star)\big)\Big). \tag{5.5}$$

2) If H is not necessarily differentiable, each H_n is a Lipschitz map, and H is a local contraction with respect to a fixed point $x^\star$, then, for particular choices of η and β depending on s, there exists α sufficiently small such that $\tilde{F}$ is a local contraction with respect to the fixed point $\psi(x^\star)$;

3) If H is a global contraction with respect to a fixed point $x^\star$ and each H_n is a Lipschitz map, then, for particular choices of η and β depending on s, there exists α sufficiently small such that $\tilde{F}$ is a global contraction with respect to the fixed point $\psi(x^\star)$.

Before proceeding, note that the presence of $\tilde{U}$ prevents the Banach-Picard iteration of $\tilde{F}$, *i.e.,* $(z^{k+1}, \tilde{w}^{k+1}) = \tilde{F}(z^k, \tilde{w}^k)$, from having distributed implementation. However, as the next section shows, to understand the Banach-Picard iteration of F it is enough to understand that of $\tilde{F}$.

5.2.4 Connection Between F and $\tilde{F}$

Consider the trajectory induced by the Banach-Picard iteration of F with initialization in (z^0, w^0), that is, the sequence recursively defined as

$$\begin{aligned} (z^0, w^0) &\in \mathbb{R}^{dN} \times \mathbb{R}^{dN} \\ (z^{k+1}, w^{k+1}) &= F(z^k, w^k). \end{aligned} \tag{5.6}$$

Consider as well the orthogonal decomposition of w^0 in its consensus and off-consensus components $w^0 = \mathbf{1} \otimes \bar{w}^0 + \hat{w}^0$ (see Chapter 4). Since $\tilde{U}$ is a matrix with columns forming an orthonormal basis of $\text{range}(L) = \ker(L)^\perp$, and $\ker(L)$ is the consensus space, the orthogonal decomposition can be rewritten as $w^0 = \mathbf{1} \otimes \bar{w}^0 + \tilde{U}\tilde{U}^T w^0$. From the properties of F, it follows that (z^k, w^k) is easily recovered from (u^k, v^k), where

$$\begin{aligned} (u^0, v^0) &= (z^0, \tilde{U}\tilde{U}^T w^0) \\ (u^{k+1}, v^{k+1}) &= F(u^k, v^k). \end{aligned}$$

In fact, note that $z^k = u^k$, and $w^k = \mathbf{1} \otimes \bar{w}^0 + v^k$, where $v^k \in \text{range}(L)$ for all k. To make the connection with $\tilde{F}$, express v^k in the basis formed by the columns of $\tilde{U}$, *i.e.,* let $\tilde{w}^k = \tilde{U}^T v^k$. The fact that, for all k, $v^k \in \text{range}(L)$ implies that $\tilde{U}\tilde{w}^k = \tilde{U}\tilde{U}^T v^k = v^k$. Consequently, $z^k = u^k$ and $w^k = \mathbf{1} \otimes \bar{w}^0 + \tilde{U}\tilde{w}^k$, where $(u^k, \tilde{w}^k)$ is the trajectory given by

$$\begin{aligned} (u^0, \tilde{w}^0) &= (z^0, \tilde{U}^T\tilde{U}\tilde{U}^T w_0) = (z^0, \tilde{U}^T w^0) \\ (u^{k+1}, \tilde{w}^{k+1}) &= \tilde{F}(u^k, \tilde{w}^k). \end{aligned}$$

We summarize this in the following lemma.

Lemma 5.2.4. *Consider the sequence* $(z^k, w^k) \in \mathbb{R}^{dN} \times \mathbb{R}^{dN}$ *recursively given by*

$$(z^0, w^0) \in \mathbb{R}^{dN} \times \mathbb{R}^{dN}$$
$$(z^{k+1}, w^{k+1}) = F(z^k, w^k),$$

and define the sequence $(u^k, \tilde{w}^k) \in \mathbb{R}^{dN} \times \mathbb{R}^{d(N-1)}$ *by*

$$(u^0, \tilde{w}^0) = (z^0, \tilde{U}^T w^0)$$
$$(u^{k+1}, \tilde{w}^{k+1}) = \tilde{F}(u^k, \tilde{w}^k).$$

Then, (z^k, w^k) *and* $(u^k, \tilde{w}^k)$ *are related by* $z^k = u^k$ *and* $w^k = \mathbf{1} \otimes \bar{w}^0 + \tilde{U}\tilde{w}^k$.

This result is shows that the properties of $(z^{k+1}, w^{k+1}) = F(z^k, w^k)$ are completely characterized by those of $(z^{k+1}, \tilde{w}^{k+1}) = \tilde{F}(z^k, \tilde{w}^k)$.

5.3 Convergence Analysis

5.3.1 The Linear Part of $\tilde{F}$

The map $\tilde{F}$, defined in (5.4), is of the form "linear map + non-linear map". In fact, $\tilde{F}$ can be written as $(z, \tilde{w}) \to M(\eta, \beta)[z^T, \tilde{w}^T]^T + \alpha T(z, \tilde{w})$, for a matrix $M(\eta, \beta)$ and a non-linear map $T(z, w)$. If, for sufficiently small α, there is any hope for $\tilde{F}$ to be a local contraction, then it seems plausible to look for η and β such that $\rho\big(M(\eta, \beta)\big) \leq 1$. To give an intuition on the plausibility of this statement, observe that if α is "very very small", then $\tilde{F}(z, \tilde{w}) \approx M(\eta, \beta)[z^T, \tilde{w}^T]^T$ and the convergence of $M^k(\eta, \beta)[z_0^T, \tilde{w}_0^T]^T$ requires $\rho\big(M(\eta, \beta)\big) \leq 1$. Otherwise, there would be an "expansive" invariant direction of $M(\eta, \beta)$ and in that direction the iteration diverges.

The "linear part" of $\tilde{F}$, denoted above by $M(\eta, \beta)$, is given by $I + \hat{A}(\eta, \beta)$, where

$$\hat{A}(\eta, \beta) = \begin{bmatrix} -\eta L & \beta L^s \tilde{U} \\ -\beta \tilde{U}^T L^s & 0 \end{bmatrix}. \tag{5.7}$$

The eigenvalues of $I + \hat{A}(\eta, \beta)$ are of the form $1 + \mu$, where μ is an eigenvalue of $\hat{A}(\eta, \beta)$, hence, it is enough to focus on these. Recall that $\tilde{U}$ is a matrix with columns forming an orthonormal basis of eigenvectors of L associated to non-zero eigenvalues, *i.e.,* an orthonormal basis of range(L). Consider the matrix $U = [\tilde{U}, \hat{U}]$, where

$$\hat{U} = \frac{1}{\sqrt{N}} \mathbf{1}_N \otimes I_d,$$

that is, $\hat{U}$ is a matrix with columns forming an orthonormal basis of $\ker(L)$.

The matrix $A(\eta, \beta)$ obtained by replacing $\tilde{U}$ by U in $\hat{A}(\eta, \beta)$, *i.e.,*

$$A(\eta, \beta) = \begin{bmatrix} -\eta L & \beta L^s U \\ -\beta U^T L^s & 0 \end{bmatrix}$$

corresponds to "appending" to $\hat{A}(\eta, \beta)$ d rows and columns of zeros, *i.e.*,

$$A(\eta, \beta) = \begin{bmatrix} \hat{A}(\eta, \beta) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix},$$

because we are enlarging $\tilde{U}$ by vectors in $\ker(L)$; we, thus, focus on the eigenvalues of $A(\eta, \beta)$.

Since $U^T L U = \Lambda$, where Λ is a diagonal matrix with the elements in the diagonal being the eigenvalues of L, and eigenvalues are preserved by similarity, consider the similarity transformation

$$\begin{bmatrix} U^T & \mathbf{0} \\ \mathbf{0} & I \end{bmatrix} A(\eta, \beta) \begin{bmatrix} U & \mathbf{0} \\ \mathbf{0} & I \end{bmatrix} = \begin{bmatrix} -\eta\Lambda & \beta\Lambda^s \\ -\beta\Lambda^s & \mathbf{0} \end{bmatrix},$$

which implies the following lemma.

Lemma 5.3.1. *The non-zero eigenvalues of $\hat{A}(\eta, \beta)$ are those of*

$$\begin{bmatrix} -\eta\Lambda & \beta\Lambda^s \\ -\beta\Lambda^s & \mathbf{0} \end{bmatrix}. \tag{5.8}$$

Let ξ be an eigenvalue of (5.8) and let $(u, v) \neq (0, 0)$ be an associated eigenvector. There must exist i such that $(u_i, v_i) \neq 0$ and, from the eigenvalue equation,

$$\begin{bmatrix} -\eta\Lambda & \beta\Lambda^s \\ -\beta\Lambda^s & \mathbf{0} \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = \xi \begin{bmatrix} u \\ v \end{bmatrix},$$

we conclude that

$$\begin{bmatrix} -\eta\lambda_i & \beta\lambda_i^s \\ -\beta\lambda_i^s & 0 \end{bmatrix} \begin{bmatrix} u_i \\ v_i \end{bmatrix} = \xi \begin{bmatrix} u_i \\ v_i \end{bmatrix}.$$

Therefore, if ξ is an eigenvalue of (5.8), there exists $i \in \{1, \ldots, dN\}$ such that ξ is an

eigenvalue of

$$\begin{bmatrix} -\eta\lambda_i & \beta\lambda_i^s \\ -\beta\lambda_i^s & 0 \end{bmatrix}. \tag{5.9}$$

Conversely, if ξ is an eigenvalue of (5.9) associated to $(u_i, v_i) \neq 0$, then taking (u, v) with $(u_j, v_j) = 0$, for $j \neq i$, yields an eigenvector of (5.8) associated to ξ. We conclude that:

Lemma 5.3.2. *Let $\lambda_1, \ldots, \lambda_{dN}$ be the eigenvalues of L. The set of non-zero eigenvalues of $\hat{A}(\eta, \beta)$ is the set*

$$\bigcup_{i=1}^{dN} \Big\{ x \in \mathbb{C} - \{0\} : x^2 + \eta\lambda_i x + \lambda_i^{2s}\beta^2 = 0 \Big\}.$$

Moreover, since, by assumption, $\beta \neq 0$, the set of non-zero eigenvalues of $\hat{A}(\eta, \beta)$ is the set

$$\bigcup_{i=1,\ldots,dN:\lambda_i \neq 0} \Big\{ x \in \mathbb{C} : x^2 + \eta\lambda_i x + \lambda_i^{2s}\beta^2 = 0 \Big\}.$$

Recall that the goal is to choose η and β such that $\rho\big(I + \hat{A}(\eta, \beta)\big) \leq 1$, and, thus, consider $\big(I + \hat{A}(\eta, \beta)\big)v = v$, or, equivalently, consider $\ker\big(\hat{A}(\eta, \beta)\big)$. From the properties of $\tilde{U}$ it follows that, for $\eta, \beta \neq 0$, $\ker\big(\hat{A}(\eta, \beta)\big) = \ker(L) \times \{0\}$. Observe that

$$\Big(\ker(L) \times \{0\}\Big)^{\perp} = \ker(L)^{\perp} \times \{0\}^{\perp} = \mathrm{span}(L) \times \mathbb{R}^{d(N-1)},$$

and note that both $\ker(L) \times \{0\}$ and $\mathrm{span}(L) \times \mathbb{R}^{d(N-1)}$ are invariant under $\hat{A}(\eta, \beta)$ (this follows from $\mathrm{span}(L) = \mathrm{span}(L^s)$). Let

$$\hat{\mathbf{U}} = \begin{bmatrix} \hat{U} \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{N}}\mathbf{1}_N \otimes I_d \\ \mathbf{0} \end{bmatrix}, \tag{5.10}$$

i.e., $\hat{\mathbf{U}}$ is a matrix with columns forming an orthonormal basis of $\ker(L) \times \{0\}$, and, let $\tilde{\mathbf{U}}$ be a matrix with columns forming an orthonormal basis of $\mathrm{span}(L) \times \mathbb{R}^{d(N-1)}$. We conclude that

Lemma 5.3.3. *Let $\lambda_1, \ldots \lambda_{dN}$ be the eigenvalues of L. The matrix $I + \hat{A}(\eta, \beta)$ is unitarily similar to*

$$\begin{bmatrix} \hat{\mathbf{U}}^T \\ \tilde{\mathbf{U}}^T \end{bmatrix} \Big(I + \hat{A}(\eta, \beta)\Big) \Big[\hat{\mathbf{U}}, \tilde{\mathbf{U}}\Big] = \begin{bmatrix} I_d & \mathbf{0} \\ \mathbf{0} & I + \hat{\mathcal{A}}(\eta, \beta) \end{bmatrix}, \tag{5.11}$$

where the eigenvalues of $\hat{\mathcal{A}}(\eta,\beta)$ *are the non-zero eigenvalues of* $\hat{A}(\eta,\beta)$, i.e., *the set*

$$\bigcup_{i=1,\ldots,dN:\lambda_i\neq 0}\Big\{x\in\mathbb{C}: x^2+\eta\lambda_i x+\lambda_i^{2s}\beta^2=0\Big\}.$$

To finish this section, we show that for both cases, $s\in\{1/2,1\}$, there are positive η and β such that $I+\hat{A}(\eta,\beta)$ is unitarily similar to (5.11) with $\rho\big(I+\hat{\mathcal{A}}(\eta,\beta)\big)<1$. At this point, we make use of $\rho(L)<2$ (see Section 5.2.2).

Remark 5.3.1. *Choose* $\eta(s)=2s$ *and* $\beta^2(s)=s$. *The polynomials in Lemma 5.3.3 reduce to*

$$\begin{cases} x^2+\lambda_i x+\frac{\lambda_i}{2}=0, & s=\frac{1}{2}\\ (x+\lambda_i)^2=0, & s=1\end{cases}.$$

with roots given by

$$\begin{cases} x=\frac{-\lambda_i\pm i\sqrt{\lambda_i}\sqrt{2-\lambda_i}}{2}, & s=\frac{1}{2}\\ x=-\lambda_i, & s=1\end{cases}.$$

Now, since $0<\lambda_i<2$, *it follows that*

$$\begin{cases} |1+x|^2=\Big(1-\frac{\lambda_i}{2}\Big)^2+\frac{2\lambda_i-\lambda_i^2}{4}=1-\frac{\lambda_i}{2}<1, & s=\frac{1}{2}\\ |1+x|^2=(1-\lambda_i)^2<1, & s=1\end{cases}.$$

5.3.2 The Differential Local Contraction Case

Throughout this section, the average map, $H=1/N\sum_{n=1}^{N}H_n$, is assumed to be differentiable at $x^\star$. The main result is the following theorem.

Theorem 5.3.1. *Let* $\lambda_1,\ldots,\lambda_{dN}$ *be the eigenvalues of* L. *Choose* $\eta>0$ *and* $\beta>0$ *such that* $|1+y|<1$, *for every*

$$y\in\bigcup_{i=1,\ldots,dN:\lambda_i\neq 0}\Big\{x\in\mathbb{C}: x^2+\eta\lambda_i x+\lambda_i^{2s}\beta^2=0\Big\}.$$

If $\rho\big(\mathbf{J}_H(x^\star)\big)<1$ *or there exists an eigenvalue* μ *of* $\mathbf{J}_H(x^\star)$ *such that* $Re(\mu)>1$, *then, there exists* $\alpha^\star$ *such that, for* $0<\alpha<\alpha^\star$,

$$sign\Big(1-\rho\big(\mathbf{J}_{\tilde{F}}(\tilde{\psi}(x^\star))\big)\Big)=sign\Big(1-\rho\big(\mathbf{J}_H(x^\star)\big)\Big).$$

Remark 5.3.2. *An interesting consequence of this theorem is that if H has a finite number of fixed points, we can choose α sufficiently small such that*

$$sign\Big(1-\rho\big(\mathbf{J}_{\tilde{F}}(\tilde{\psi}(x^\star))\big)\Big) = sign\Big(1-\rho\big(\mathbf{J}_H(x^\star)\big)\Big),$$

for all $x^\star \in Fix(H)$ such that either $\rho\big(\mathbf{J}_H(x^\star)\big) < 1$ or $\mathbf{J}_H(x^\star)$ has an eigenvalue with real part larger than one.

The proof of Theorem 5.3.1 reduces to understanding how the eigenvalues of a matrix change under a linear perturbation. Observe that

$$\mathbf{J}_{\tilde{F}}\big(\tilde{\psi}(x^\star)\big) = I + \hat{A}(\eta,\beta) + \alpha \begin{bmatrix} \mathbf{J}_R\big(\mathbf{1}_N \otimes x^\star\big) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix},$$

where $\hat{A}(\eta,\beta)$ is the linear part of $\tilde{F}$ (see (5.7)). From Lemma 5.3.3, $\mathbf{J}_{\tilde{F}}\big(\tilde{\psi}(x^\star)\big)$ is unitarily similar to

$$\begin{bmatrix} \hat{\mathbf{U}}^T \\ \tilde{\mathbf{U}}^T \end{bmatrix} \mathbf{J}_{\tilde{F}}\big(\tilde{\psi}(x^\star)\big) \begin{bmatrix} \hat{\mathbf{U}}, \tilde{\mathbf{U}} \end{bmatrix} = \begin{bmatrix} I_d & \mathbf{0} \\ \mathbf{0} & I + \mathcal{A}(\eta,\beta) \end{bmatrix} + \alpha \begin{bmatrix} \hat{\mathbf{U}}^T \\ \tilde{\mathbf{U}}^T \end{bmatrix} \begin{bmatrix} \mathbf{J}_R\big(\mathbf{1}_N \otimes x^\star\big) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{U}}, \tilde{\mathbf{U}} \end{bmatrix},$$

where $\hat{\mathbf{U}}$ is defined in (5.10). Moreover, if η and β are chosen according to the statement of the theorem, then, $\rho\big(I + \mathcal{A}(\eta,\beta)\big) < 1$. The matrix multiplied by α can be partitioned according to the blocks of $I + \mathcal{A}(\eta,\beta)$ yielding

$$\begin{bmatrix} \hat{\mathbf{U}}^T \\ \tilde{\mathbf{U}}^T \end{bmatrix} \begin{bmatrix} \mathbf{J}_R\big(\mathbf{1}_N \otimes x^\star\big) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{U}}, \tilde{\mathbf{U}} \end{bmatrix} = \begin{bmatrix} \hat{U}^T \mathbf{J}_R\big(\mathbf{1}_N \otimes x^\star\big)\hat{U} & * \\ * & * \end{bmatrix},$$

where $\hat{U}^T \mathbf{J}_R\big(\mathbf{1}_N \otimes x^\star\big)\hat{U}$ is as $d \times d$ block, and the symbols $*$ correspond to other blocks that do not concern us. Finally, note that

$$\hat{U}^T \mathbf{J}_R\big(\mathbf{1}_N \otimes x^\star\big)\hat{U} = \Big(\frac{1}{\sqrt{N}}\mathbf{1}_N^T \otimes I_d\Big)\mathbf{J}_R(\mathbf{1}_N \otimes x^\star)\Big(\frac{1}{\sqrt{N}}\mathbf{1}_N \otimes I_d\Big),$$

and, from $R(z) = \tilde{H}(z) - z$,

$$\mathbf{J}_R(\mathbf{1}_N \otimes x^\star) = \begin{bmatrix} \mathbf{J}_{H_1}(x^\star) - I_d & & \\ & \ddots & \\ & & \mathbf{J}_{H_N}(x^\star) - I_d \end{bmatrix}.$$

Consequently,

$$\hat{U}^T \mathbf{J}_R\big(\mathbf{1}_N \otimes x^\star\big)\hat{U} = \mathbf{J}_H(x^\star) - I_d.$$

We summarize these observations in the following lemma.

Lemma 5.3.4. *Let $\lambda_1, \ldots, \lambda_{dN}$ be the eigenvalues of L. Choose $\eta > 0$ and $\beta > 0$ such that $|1 + y| < 1$, for every*

$$y \in \bigcup_{i=1,\ldots,dN:\lambda_i \neq 0} \Big\{ x \in \mathbb{C} : x^2 + \eta\lambda_i x + \lambda_i^{2s}\beta^2 = 0 \Big\}.$$

Then, $\mathbf{J}_{\tilde{F}}\big(\tilde{\psi}(x^\star)\big)$ is unitarily similar to

$$\begin{bmatrix} I_d & \mathbf{0} \\ \mathbf{0} & I + \mathcal{A}(\eta,\beta) \end{bmatrix} + \alpha \begin{bmatrix} \mathbf{J}_H(x^\star) - I_d & * \\ * & * \end{bmatrix},$$

where $\rho\big(I + \mathcal{A}(\eta,\beta)\big) < 1$.

The proof of Theorem 5.3.1 follows immediately from the following non-trivial theorem on linear perturbations, the proof of which can be found in Appendix A.

Theorem 5.3.2. *Let A and B be two block matrices of the form*

$$A = \begin{bmatrix} I_d & \mathbf{0}_{d,k} \\ \mathbf{0}_{d,n} & A_{22}, \end{bmatrix}, B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix},$$

where A_{22} is a $k \times k$ matrix satisfying $\rho(A_{22}) < 1$, and where the block partition of B is compatible with that of A. Let α be a real parameter and consider the curve of matrices $A(\alpha)$ defined by $A(\alpha) = A + \alpha B$. Then,

1) *If all eigenvalues of B_{11} have negative real part, then, there exists $\alpha^\star$ such that, for $0 < \alpha < \alpha^\star$, $\rho\big(A(\alpha)\big) < 1$;*

2) *Conversely, if B_{11} has at least one eigenvalue with positive real part, then there exists $\alpha^\star$ such that, for $0 < \alpha < \alpha^\star$, $\rho\big(A(\alpha)\big) > 1$.*

5.3.3 The Continuous Local and Global Contraction Cases

Throughout this section the average map, $H = 1/N \sum_{n=1}^{N} H_n$, is not assumed to be differentiable at $x^\star$, but a global contraction relative to $x^\star$, that is, for $0 \leq \mu < 1$,

$$\|H(x) - x^\star\| \leq \mu \|x - x^\star\|, \tag{5.12}$$

for all $x \in \mathbb{R}^d$. The local contraction case, where (5.12) holds with $x \in \mathbb{R}^d$ replaced by

$$x \in \bar{B}(x^\star, \delta) = \big\{ y \in \mathbb{R}^d : \|y - x^\star\| \leq \delta \big\}, \tag{5.13}$$

is also analyzed. Furthermore, in both cases (local and global), each H_n is assumed to be globally γ-Lipschitz. We begin with the global case.

Recall that $\bar{z} = 1/N \sum_{n=1}^N z_n$ and that each $z \in \mathbb{R}^{dN}$ can be written as $z = \mathbf{1} \otimes \bar{z} + \big(z - \mathbf{1} \otimes \bar{z}\big)$. Moreover, recall that $\tilde{U}$ is a matrix with columns forming an orthonormal basis of eigenvectors of L associated to non-zero eigenvalues, *i.e.,* a basis of range(L), and, thus, $z = \mathbf{1} \otimes \bar{z} + \tilde{U}\tilde{U}^T z$. Consider iteration $(z^{k+1}, w^{k+1}) = \tilde{F}(z^k, w^k)$ and let $\bar{z}^k = 1/N \sum_{n=1}^N z_n^k$, $y^k = \tilde{U}^T z^k$. By construction, $z^k = \mathbf{1} \otimes \bar{z}^k + \tilde{U} y^k$. A straightforward manipulation shows that

$$\begin{aligned}
\bar{z}^{k+1} &= (1-\alpha)\bar{z}^k + \alpha \bar{H}\big(\tilde{U} y^k + \mathbf{1} \otimes \bar{z}^k\big) \\
y^{k+1} &= \big(I - \eta \tilde{U}^T L \tilde{U}\big) y^k + \beta \tilde{U}^T L^s \tilde{U} w^k + \alpha \tilde{U}^T R\big(\tilde{U} y^k + \mathbf{1} \otimes \bar{z}^k\big) \\
w^{k+1} &= w^k - \beta \tilde{U}^T L^s \tilde{U} y^k,
\end{aligned}$$

where $\bar{H}(z_1, \ldots, z_N) = \frac{1}{N} \sum_{n=1}^N H_n(z_n)$. Consider the linear part of the recursions followed by y^k and w^k: note that $\tilde{U}^T L \tilde{U} = \tilde{\Lambda}$, where $\tilde{\Lambda}$ is a diagonal matrix with the elements in the diagonal being the non-zero eigenvalues of L, and, thus, the linear part of the recursions satisfied by y^k and w^k is given by

$$\mathcal{B}(\eta, \beta) := I + \begin{bmatrix} -\eta\tilde{\Lambda} & \beta\tilde{\Lambda}^s \\ \beta\tilde{\Lambda}^s & \mathbf{0} \end{bmatrix}.$$

Compare $\mathcal{B}(\eta, \beta)$ with (5.8) from Lemma 5.3.1 and it is straightforward to see that $\mathcal{B}(\eta, \beta)$ is unitarily similar to the matrix $I + \mathcal{A}(\eta, \beta)$ from Lemma 5.3.3. This leads to the following lemma.

Lemma 5.3.5. *Let $\lambda_1, \ldots, \lambda_{dN}$ be the eigenvalues of L. Choose $\eta > 0$ and $\beta > 0$ such that $|1 + y| < 1$, for every*

$$y \in \bigcup_{i=1,\ldots,dN : \lambda_i \neq 0} \Big\{ x \in \mathbb{C} : x^2 + \eta\lambda_i x + \lambda_i^{2s}\beta^2 = 0 \Big\}.$$

Given the sequence (z^k, w^k) recursively defined as $(z^{k+1}, w^{k+1}) = \tilde{F}(z^k, w^k)$, define the sequences $\bar{z}^k = 1/N \sum_{n=1}^N z_n^k$ and $y^k = \tilde{U}^T z^k$. Then, the sequence $(\bar{z}^k, y^k, w^k)$ satisfies the

recursions

$$\begin{aligned} \bar{z}^{k+1} &= (1-\alpha)\bar{z}^k + \alpha\bar{H}\big(\tilde{U}y^k + \mathbf{1}\otimes\bar{z}^k\big) \\ \begin{bmatrix} y^{k+1} \\ w^{k+1} \end{bmatrix} &= \mathcal{B}(\eta,\beta)\begin{bmatrix} y^k \\ w^k \end{bmatrix} + \alpha\begin{bmatrix} \tilde{U}^T R\big(\tilde{U}y^k + \mathbf{1}\otimes\bar{z}^k\big) \\ \mathbf{0} \end{bmatrix}, \end{aligned} \tag{5.14}$$

where $\rho\big(\mathcal{B}(\eta,\beta)\big) < 1$.

The recursions given by (5.14) are rather ugly and, to analyze them, it makes more sense to abstract their general features. Let

$$\begin{aligned} G_{1,\alpha}(a,b) &= (1-\alpha)a + \alpha T_1(a,b) \\ G_{2,\alpha}(a,b) &= Mb + \alpha T_2(a,b), \end{aligned} \tag{5.15}$$

where $\rho(M) < 1$, and, instead of (5.14), consider $(a^{k+1}, b^{k+1}) = \big(G_{1,\alpha}(a^k,b^k), G_{2,\alpha}(a^k,b^k)\big)$. The proof that (5.14) converges at least linearly will follow easily from the following theorem. In what follows, given the equivalence between all norms, we "move freely" between norms, whenever Lipschitzianity is concerned.

Theorem 5.3.3. *Consider a map* $G_\alpha(a,b) = \big(G_{1,\alpha}(a,b), G_{2,\alpha}(a,b)\big)$, *where* $G_{1,\alpha}$ *and* $G_{2,\alpha}$ *are given by* (5.15). *Suppose that, for each* $\alpha \neq 0$, G_α *has a unique fixed point* $\big(a^\star(\alpha), b^\star(\alpha)\big)$. *Additionally, suppose that*

1) $T_1\big(\cdot, b^\star(\alpha)\big)$ *is a* $\mu-$*contraction with respect to* $a^\star(\alpha)$ *and a norm* $\|\cdot\|_1$;

2) $\|M\|_2 < 1$ *for some matrix norm* $\|\cdot\|_2$ *induced by a vector norm* $\|\cdot\|_2$;

3) *Both* T_1 *and* T_2 *are* $\tilde{\gamma}$*-Lipschitz, and, without loss of generality, assume that*

$$\|T_i(a,b) - T_i(\tilde{a},\tilde{b})\|_i \leq \tilde{\gamma}\Big(\|a-\tilde{a}\|_1 + \|b-\tilde{b}\|_2\Big), i = 1,2.$$

Then, there exists $\alpha^\star$ *such that, for* $0 < \alpha < \alpha^\star$, *there exists a matrix* $P(\alpha)$, *satisfying* $\rho\big(P(\alpha)\big) < 1$, *such that*

$$\begin{bmatrix} \|G_{1,\alpha}(a,b) - a^\star(\alpha)\|_1 \\ \|G_{2,\alpha}(a,b) - b^\star(\alpha)\|_2 \end{bmatrix} \leq P(\alpha)\begin{bmatrix} \|a - a^\star(\alpha)\|_1 \\ \|b - b^\star(\alpha)\|_2 \end{bmatrix},$$

that is, G_α *is a* $P(\alpha)$*-contraction with respect to* $\big(a^\star(\alpha), b^\star(\alpha)\big)$.

Proof. The triangle inequality implies that

$$\begin{aligned}
&\|G_{1,\alpha}(a,b) - a^\star(\alpha)\|_1 = \Big\|G_{1,\alpha}(a,b) - G_{1,\alpha}\big(a^\star(\alpha), b^\star(\alpha)\big)\Big\|_1 \\
\leq &\Big\|G_{1,\alpha}(a,b) - G_{1,\alpha}\big(a, b^\star(\alpha)\big)\Big\|_1 + \Big\|G_{1,\alpha}\big(a, b^\star(\alpha)\big) - G_{1,\alpha}\big(a^\star(\alpha), b^\star(\alpha)\big)\Big\|_1 \\
\leq &\alpha\Big\|T_1(a,b) - T_1\big(a, b^\star(\alpha)\big)\Big\|_1 + \big(1 - \alpha(1-\mu)\big)\|a - a^\star(\alpha)\|_1 \\
\leq &\alpha\tilde{\gamma}\big\|b - b^\star(\alpha)\big\|_2 + \big(1 - \alpha(1-\mu)\big)\|a - a^\star(\alpha)\|_1.
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\|G_{2,\alpha}(a,b) - b^\star(\alpha)\|_2 \leq \Big\|G_{2,\alpha}(a,b) - G_{2,\alpha}\big(a^\star(\alpha), b\big)\Big\|_2 + \Big\|G_{2,\alpha}\big(a^\star(\alpha), b\big) - b^\star(\alpha)\Big\|_2 \\
\leq &\alpha\Big\|T_2(a,b) - T_2\big(a^\star(\alpha), b\big)\Big\|_2 + \|M\|_2\|b - b^\star(\alpha)\|_2 + \alpha\Big\|T_2\big(a^\star(\alpha), b^\star(\alpha)\big) - T_2\big(a^\star(\alpha), b\big)\Big\|_2,
\end{aligned}$$

where we used the fact that $b^\star(\alpha) = Mb^\star(\alpha) + \alpha T_2\big(a^\star(\alpha), b^\star(\alpha)\big)$. The Lipschitzianity of T_2 implies that

$$\|G_{2,\alpha}(a,b) - b^\star(\alpha)\|_2 \leq \alpha\tilde{\gamma}\big\|a - a^\star(\alpha)\big\|_1 + \|M\|_2\|b - b^\star(\alpha)\|_2 + \alpha\tilde{\gamma}\big\|b - b^\star(\alpha)\big\|_2.$$

Therefore, we conclude that

$$\begin{bmatrix} \|G_{1,\alpha}(a,b) - a^\star(\alpha)\|_1 \\ \|G_{2,\alpha}(a,b) - b^\star(\alpha)\|_2 \end{bmatrix} \leq \begin{bmatrix} 1 - \alpha(1-\mu) & \alpha\tilde{\gamma} \\ \alpha\tilde{\gamma} & \|M\|_2 + \alpha\tilde{\gamma} \end{bmatrix} \begin{bmatrix} \|a - a^\star(\alpha)\|_1 \\ \|b - b^\star(\alpha)\|_2 \end{bmatrix}.$$

Define

$$P(\alpha) = \begin{bmatrix} 1 & 0 \\ 0 & \|M\|_2 \end{bmatrix} + \alpha \begin{bmatrix} \mu - 1 & \tilde{\gamma} \\ \tilde{\gamma} & \tilde{\gamma} \end{bmatrix}.$$

All that is left to prove is that there exists $\alpha^\star$ such that, for $0 < \alpha < \alpha^\star$, $\rho\big(P(\alpha)\big) < 1$. $P(\alpha)$ is symmetric 2×2 matrix, hence its roots are easy to study. However, to avoid a messy computation, just note that $P(0)$ is a diagonal matrix with two distinct elements in the diagonal, namely, 1 and $\|M\|_2 < 1$. Since $P(\alpha)$ is symmetric, there exist two continuous real valued functions $x_1(\alpha)$ and $x_2(\alpha)$ such that correspond to the eigenvalues of $P(\alpha)$. Assume that $x_1(0) = 1$ and that $x_2(0) = \|M\|_2$. It is known that both of these functions are differentiable at zero with $x_1'(0) = \mu - 1$ and $x_2'(0) = \tilde{\gamma}$ (see [57]). From $0 \leq \mu < 1$, we conclude that $x_1'(0) < 0$. Moreover, since $\|M\|_2 < 1$, we can define $\alpha^\star$. $\square$

Observe that with the identifications $a^k \to \bar{z}^k$, $b^k \to (y^k, w^k)$, and defining

$$T_1(\bar{z}, y, w) = \bar{H}(\tilde{U} + \mathbf{1} \otimes \bar{z})$$
$$T_2(\bar{z}, y, w) = \begin{bmatrix} \tilde{U}^T R(\tilde{U} y + \mathbf{1} \otimes \bar{z}) \\ \mathbf{0} \end{bmatrix},$$

the linear convergence (5.14) follows from Theorem 5.3.3 and the linear convergence of P-contractions (see Lemma 3.2.1 of Chapter 3). However, conditions of Theorem 5.3.3 still need to be checked. The map $\tilde{F}$ has a unique fixed point given by $\tilde{\psi}(x^\star)$, and, hence, for $\alpha \neq 0$, (5.14) has a unique fixed point given by

$$\big(x^\star, 0, -\frac{\alpha}{\beta}\tilde{U}^T (L^s)^+ R(\mathbf{1} \otimes x^\star)\big).$$

Therefore $x^\star$ and $\big(0, -\frac{\alpha}{\beta}\tilde{U}^T (L^s)^+ R(\mathbf{1} \otimes x^\star)\big)$ play, the role of $a^\star(\alpha)$ and $b^\star(\alpha)$, respectively. Note as well that

$$T_1\big(x, 0, -\frac{\alpha}{\beta}\tilde{U}^T (L^s)^+ R(\mathbf{1} \otimes x^\star)\big) = \bar{H}(\mathbf{1} \otimes x) = H(x),$$

and, thus, $T_1\big(\cdot, b^\star(\alpha)\big)$ is μ-contractive where $\|\cdot\|_1$ is the norm $\|\cdot\|$ with respect to which H is contractive. To finish, note that, although tedious to verify, the Lipschitzianity of T_1 and T_2 follows from that of each $H_1, \ldots, H_N$. In fact, T_1 and T_2 only involve operations (e.g. linear combinations, compositions with linear maps, etc) that preserve Lipschitzianty.

We conclude that:

Theorem 5.3.4. *Suppose that H is a global μ-contraction with respect to a fixed point $x^\star$ and that each $H_1, \ldots, H_N$ is globally Lipschitz. Let $\lambda_1, \ldots, \lambda_{dN}$ be the eigenvalues of L. Choose $\eta > 0$ and $\beta > 0$ such that $|1 + y| < 1$, for every*

$$y \in \bigcup_{i=1,\ldots,dN : \lambda_i \neq 0} \Big\{ x \in \mathbb{C} : x^2 + \eta \lambda_i x + \lambda_i^{2s} \beta^2 = 0 \Big\}.$$

Then, there exists $\alpha^\star$ such that, for $0 < \alpha < \alpha^\star$, the iteration $(z^{k+1}, \tilde{w}^{k+1}) = \tilde{F}(z^k, \tilde{w}^k)$ converges (globally) at least linearly to $\tilde{\psi}(x^\star)$.

5.3.3.1 The Local Contraction Case

Although it might sound surprising, the local contraction case, that is if (5.12) only holds in the ball defined in (5.13), is not considerably more challenging than the global case.

In fact, observe that the global contraction proof amounted to showing that: 1)

$$\begin{bmatrix} \|\bar{z}^{k+1} - x^\star\| \\ \left\| y^{k+1}, \tilde{w}^{k+1} + \frac{\alpha}{\beta}\tilde{U}^T(L^s)^+R(\mathbf{1}\otimes x^\star)\right)\right\|_2 \end{bmatrix} \leq P(\alpha) \begin{bmatrix} \|\bar{z}^{k} - x^\star\| \\ \left\| y^{k}, \tilde{w}^{k} + \frac{\alpha}{\beta}\tilde{U}^T(L^s)^+R(\mathbf{1}\otimes x^\star)\right)\right\|_2 \end{bmatrix}, \tag{5.16}$$

where

$$P(\alpha) = \begin{bmatrix} 1-\alpha(1-\mu) & \alpha\tilde{\gamma} \\ \alpha\tilde{\gamma} & \|M\|_2 + \alpha\tilde{\gamma} \end{bmatrix},$$

with $\|M\|_2 < 1$; 2) there exists $\alpha^\star$ such that, for $0 < \alpha < \alpha^\star$, $\rho\big(P(\alpha)\big) < 1$.

If H is a global contraction, then (5.16) holds for every $\bar{z}^k \in \mathbb{R}^d$. In contrast, if H is merely a local, rather than a global, contraction, (5.16) only holds provided that $\bar{z}^k \in \bar{B}(x^\star, \delta)$. However, it is not necessarily true that $\bar{z}^k \in \bar{B}(x^\star, \delta)$ implies that $\bar{z}^{k+1} \in \bar{B}(x^\star, \delta)$, a condition that should hold if (5.16) is to be "unfolded" to establish the linear convergence (see the proof of Theorem 3.2.1 of Chapter 3). To fix this, let $\alpha^\star$ be such that $\rho\big(P(\alpha)\big) < 1$, for $0 < \alpha < \alpha^\star$, and, by reducing $\alpha^\star$, if necessary, we may assume that the entries of $P(\alpha)$ are all positive for $0 < \alpha < \alpha^\star$. From *Perron's Theorem* (see [57]) and $P(\alpha) > 0$, there exists $v(\alpha) = \big(v_1(\alpha), v_2(\alpha)\big)$ with only positive entries such that

$$P(\alpha)v(\alpha) = \rho\big(P(\alpha)\big)v(\alpha).$$

By multiplying by a suitable positive constant, we may assume that $v_1(\alpha) = \delta$. Let $b^\star(\alpha) = \big(0, -\alpha/\beta\tilde{U}^T(L^s)^+R(\mathbf{1}\otimes x^\star)\big)$ and define an open neighborhood $\mathcal{Y}(\alpha)$ of $(x^\star, b^\star(\alpha)\big)$ by

$$\mathcal{Y}(\alpha) = \bar{B}(x^\star, \delta) \times \bar{B}_{\|\cdot\|_2}\big(b^\star(\alpha), v_2(\alpha)\big).$$

Note that if $(\bar{z}^k, y^k, \tilde{w}^k) \in \mathcal{Y}(\alpha)$, then, in particular $\bar{z}^k \in \bar{B}(x^\star, \delta)$ and, hence, (5.16) holds. Moreover, since $P(\alpha)$ has only positive entries, if $(\bar{z}^k, y^k, \tilde{w}^k) \in \mathcal{Y}(\alpha)$, then, by construction,

$$\begin{aligned}\begin{bmatrix} \|\bar{z}^{k+1} - x^\star\| \\ \left\| y^{k+1}, \tilde{w}^{k+1} + \frac{\alpha}{\beta}\tilde{U}^T(L^s)^+ R(\mathbf{1}\otimes x^\star)\right) \right\|_2 \end{bmatrix} &\leq P(\alpha) \begin{bmatrix} \|\bar{z}^{k} - x^\star\| \\ \left\| y^{k}, \tilde{w}^{k} + \frac{\alpha}{\beta}\tilde{U}^T(L^s)^+ R(\mathbf{1}\otimes x^\star)\right) \right\|_2 \end{bmatrix} \\ &\leq P(\alpha) \begin{bmatrix} \delta \\ v_2(\alpha) \end{bmatrix} \\ &= \rho\big(P(\alpha)\big) \begin{bmatrix} \delta \\ v_2(\alpha) \end{bmatrix} \\ &< \begin{bmatrix} \delta \\ v_2(\alpha) \end{bmatrix},\end{aligned}$$

which implies that $(\bar{z}^{k+1}, y^{k+1}, \tilde{w}^{k+1}) \in \mathcal{Y}(\alpha)$. We conclude that

1) If $(\bar{z}^k, y^k, \tilde{w}^k) \in \mathcal{Y}(\alpha)$, then $(\bar{z}^{k+1}, y^{k+1}, \tilde{w}^{k+1}) \in \mathcal{Y}(\alpha)$;

2) If $(\bar{z}^k, y^k, \tilde{w}^k) \in \mathcal{Y}(\alpha)$, then (5.16) holds.

Therefore, if $(\bar{z}^{k_0}, y^{k_0}, \tilde{w}^{k_0}) \in \mathcal{Y}(\alpha)$ for some k_0, then $(\bar{z}^k, y^k, \tilde{w}^k) \in \mathcal{Y}(\alpha)$ converges to $\big(x^\star, b^\star(\alpha)\big)$ at least linearly. We summarize this result in the following theorem.

Theorem 5.3.5. *Suppose that H is a local μ-contraction with respect to a fixed point $x^\star$, that is* (5.12) *holds for $x \in \bar{B}(x^\star, \delta)$. Furthermore, assume that each $H_1, \ldots, H_N$ is globally Lipschitz. Let $\lambda_1, \ldots, \lambda_{dN}$ be the eigenvalues of L. Choose $\eta > 0$ and $\beta > 0$ such that $|1 + y| < 1$, for every*

$$y \in \bigcup_{i=1,\ldots,dN:\lambda_i \neq 0} \Big\{ x \in \mathbb{C} : x^2 + \eta\lambda_i x + \lambda_i^{2s}\beta^2 = 0 \Big\}.$$

Then, there exists $\alpha^\star$ such that, for $0 < \alpha < \alpha^\star$, there exists a neighborhood $\mathcal{Y}(\alpha)$ of $\psi(x^\star)$ such that if $(z^{k_0}, \tilde{w}^{k_0}) \in \mathcal{Y}(\alpha)$, then $(z^{k+1}, \tilde{w}^{k+1}) = \tilde{F}(z^k, \tilde{w}^k)$ converges at least linearly to $\tilde{\psi}(x^\star)$.

5.4 Distributed Implementations

As it was observed, the iteration $(z^{k+1}, \tilde{w}^{k+1}) = \tilde{F}(z^k, \tilde{w}^k)$ does not have distributed implementation, because of the presence of the matrix $\tilde{U}$. In contrast, for $s = 1$, the iteration $(z^{k+1}, w^{k+1}) = F(z^k, w^k)$ has and its convergence guarantees follow from those of $(z^{k+1}, \tilde{w}^{k+1}) = \tilde{F}(z^k, \tilde{w}^k)$ and Lemma 5.2.4. If $H_n(x) = x - t\nabla f_n$, then $(z^{k+1}, w^{k+1}) = F(z^k, w^k)$ amounts to a distributed gradient descent algorithm; this section shows that

the choices $\eta(s) = 2s$ and $\beta^2(s) = s$ from Remark 5.3.1 lead to the well-known distributed gradient descent algorithms termed EXTRA and DIGing (see [28, 32]).

5.4.1 The EXTRA-Distributed Banach-Picard Iteration

If $s = 1/2$, the Banach-Picard iteration of F does not have distributed implementation. However, the elimination of the second variable leads to an algorithm having distributed implementation: consider two consecutive z-updates

$$\begin{aligned} z^{k+2} &= z^{k+1} + \alpha R(z^{k+1}) + \beta L^{\frac{1}{2}} w^{k+1} - \eta L z^{k+1} \\ z^{k+1} &= z^{k} + \alpha R(z^{k}) + \beta L^{\frac{1}{2}} w^{k} - \eta L z^{k} \end{aligned}$$

and consider their difference

$$z^{k+2} = 2z^{k+1} - z^k + \beta L^{\frac{1}{2}}(w^{k+1} - w^k) - \eta L(z^{k+1} - z^k) + \alpha\big(R(z^{k+1} - R(z^k)\big).$$

Observe that $w^{k+1} - w^k = -\beta L^{\frac{1}{2}} z^k$, and, hence, the variable w^k can be eliminated, yielding

$$z^{k+2} = (2I - \eta L)z^{k+1} - (I + \beta^2 L - \eta L)z^k + \alpha\big(R(z^{k+1}) - R(z^k)\big). \tag{5.17}$$

The elimination of the second variable results in a "second order" recursion, *i.e.,* a recursion of the form $q^{k+2} = J(q^k, q^{k+1})$, essentially a discrete-time version of a second order differential equation. Naturally, initialization must be specified at two points and, to have the trajectory produced by (5.17) to be exactly the one produced prior to the elimination of w^k, z^1 must be initialized according to

$$z^1 = z^0 + \alpha R(z^0) + \beta L^{\frac{1}{2}} w^0 - \eta L z^0.$$

Consequently, if z^k is initialized at $\big(z^0, z^0 + \alpha R(z^0) + \beta L^{\frac{1}{2}} w^0 - \eta L z^0\big)$, the trajectory followed by z^k is the that of u^k, for u^k recursively defined by

$$\begin{aligned} (u^0, v^0) &= (z^0, w^0) \\ (u^{k+1}, v^{k+1}) &= F(u^k, v^k). \end{aligned}$$

However, we have once again the problem that z^1 requires a product by $L^{\frac{1}{2}}$. Clearly, the only feasible "distributed initialization" corresponds to setting $w^0 = 0$, *i.e.,*

$$\begin{aligned} z^0 &\in \mathbb{R}^{dN} \\ z^1 &= z^0 + \alpha R(z^0) - \eta L z^0 \\ z^{k+2} &= (2I - \eta L)z^{k+1} - (I + \beta^2 L - \eta L)z^k + \alpha\big(R(z^{k+1}) - R(z^k)\big). \end{aligned} \tag{5.18}$$

From Lemma 5.2.4, we obtain that z^k recursively defined by (5.18) satisfies $z^k = u^k$, where u^k is recursively defined by

$$\begin{aligned} u^0 &= z^0 \in \mathbb{R}^{dN} \\ \tilde{v}^0 &= 0 \in \mathbb{R}^{d(N-1)} \\ u^{k+1} &= u^k + \alpha R(u^k) + \beta L^{\frac{1}{2}} \tilde{U} \tilde{v}^k - \eta L u^k \\ \tilde{v}^{k+1} &= \tilde{v}^k - \beta \tilde{U}^T L^{\frac{1}{2}} u^k. \end{aligned}$$

Moreover, the convergence guarantees for $(u^k, \tilde{v}^k)$, hence for z^k, are given in Theorems 5.3.1, 5.3.4, and 5.3.5, for the three conditions on H that were analyzed.

Finally, recall that Remark 5.3.1 shows that the choices $\eta(s) = 2s$ and $\beta^2(s) = s$ satisfy the conditions of Theorems 5.3.1, 5.3.4, and 5.3.5. If we additionally let $L = I - W$, where $W = \tilde{W} \otimes I_d$ and $\tilde{W}$ is a consensus matrix, (5.18) reduces to

$$\begin{aligned} z^0 &\in \mathbb{R}^{dN} \\ z^1 &= Wz^0 + \alpha R(z^0) \\ z^{k+2} &= (I + W)z^{k+1} - \frac{1}{2}(I + W)z^k + \alpha\big(R(z^{k+1}) - R(z^k)\big). \end{aligned} \tag{5.19}$$

If $H_n(x) = x - t\nabla f_n(x)$, that is, if

$$R(z) = -t \begin{bmatrix} \nabla f_1(z_1) \\ \vdots \\ \nabla f_N(z_N) \end{bmatrix},$$

then (5.19) reduces to the EXTRA algorithm (see [28]) for distributed gradient descent. Moreover, a sufficient condition for the existence of $t > 0$ such that

$$H(x) = \frac{1}{N}\sum_{n=1}^{N} x - t\nabla f_n(x) = x - t\frac{1}{N}\sum_{n=1}^{N} \nabla f_n(x)$$

is globally contractive with respect to the unique minimum of $f = 1/N \sum_{n=1}^{N} f_n$ is that f

is Lipschitz and strongly convex (see [75]). Therefore, we obtain the following corollary of Theorem 5.3.4.

Corollary 5.4.1. *Let $f_1, \dots, f_N$ be real valued differentiable functions such that ∇f_n are Lipschitz maps. Suppose that*

$$f := \frac{1}{N} \sum_{n=1}^{N} f_n,$$

is a strongly convex function. (The Lipschitzianity of each ∇f_n implies the Lipschitzianity of ∇f, hence f is a Lipschitz and strongly convex function.) Let $x^\star$ be the unique minimum of f.

Then, there exists $\alpha^\star$ such that for $0 < \alpha < \alpha^\star$, (5.19) converges to $\mathbf{1} \otimes x^\star$ at least linearly.

5.4.2 The DIGing-Distributed Banach-Picard Iteration

This section considers the case $s = 1$ and we observe that the elimination of the second variable recovers the form of DIGing (see [31, 32]). The procedure is very similar to the previous section, and, hence, the details are omitted. Choose η and β according to Remark 5.3.1 and let $L = I - W$, where $W = \tilde{W} \otimes I_d$ and $\tilde{W}$ is a consensus matrix. The elimination of the second variable leads to

$$\begin{aligned} z^0 &\in \mathbb{R}^{dN} \\ z^1 &= (2W - I)z^0 + \alpha R(z^0) + (I - W)w^0 \\ z^{k+2} &= 2Wz^{k+1} - W^2 z^k + \alpha\big(R(z^{k+1}) - R(z^k)\big). \end{aligned}$$

If we initialize at $w^0 = z^0$, we obtain

$$\begin{aligned} z^0 &\in \mathbb{R}^{dN} \\ z^1 &= Wz^0 + \alpha R(z^0) \\ z^{k+2} &= 2Wz^{k+1} - W^2 z^k + \alpha\big(R(z^{k+1}) - R(z^k)\big), \end{aligned}$$

which, if $H_n(x) = x - t\nabla f_n$, reduces to the algorithm we obtain if we eliminate the second variable in DIGing and assume that the second variable is initialized at

$$\begin{bmatrix} \nabla f_1(z_1^0) \\ \vdots \\ \nabla f_N(z_N^0) \end{bmatrix},$$

the initialization suggested both in [31, 32].

Finally, similar to Corollary 5.4.1, the convergence guarantees for DIGing in the strongly convex and Lipschitz case follow from Theorem 5.3.4.

5.5 Comments and References

5.5.1 Intuition and Connection with Optimization

The parametric family of maps F defined in (5.3) stems from (5.2) and these are a generalization of the KKT conditions (see *e.g.* [79]) that typically arise in the study of constrained optimization. A common approach (see *e.g.* [34, 35, 36, 38, 39]) to the distributed optimization problem described in Section3.4.4 of Chapter 3 is to reformulate

$$\underset{x}{\text{minimize}} \quad f(x) := \frac{1}{N}\sum_{n=1}^{N} f_n(x) \tag{5.20}$$

as a constrained optimization problem that "forces" the network structure into the problem. Afterwards, (5.20) is numerically approximated by a primal-dual method having distributed implementation. Typically, to "force" the network structure into (5.20) is reformulated as

$$\begin{aligned} \underset{z=(z_1,\ldots,z_N)}{\text{minimize}} \quad & \sum_{n=1}^{N} f_n(z_n), \\ \text{subject to} \quad & Lz = 0 \end{aligned} \tag{5.21}$$

where $L = \tilde{L} \otimes I$ and $\tilde{L}$ is compatible with the network structure. Moreover,

$$\ker\big(L \otimes I\big) = \{(z_1, \ldots, z_N) \in \mathbb{R}^{dN} : z_1 = \cdots = z_N\}.$$

Assuming L to be symmetric and positive semi-definite, (5.21) can be reformulated by introducing two positive parameters β' and η' and writing

$$\begin{array}{ll} \underset{z=(z_1,\ldots,z_N)}{\text{minimize}} & \displaystyle\sum_{n=1}^{N} f_n(z_n) + \frac{\eta'}{2}\|L^{\frac{1}{2}}z\|^2 \\ \text{subject to} & \beta' L^{\frac{1}{2}} z = 0 \end{array}, \tag{5.22}$$

with the *Lagrangian* (see [80]) defined by

$$\mathcal{L}(z,w) = \sum_{n=1}^{N} f_n(z_n) + \beta' w^T L^{\frac{1}{2}} z + \frac{\eta'}{2}\|L^{\frac{1}{2}}z\|^2.$$

Suppose that each $f_1,\ldots,f_N$ is assumed to be differentiable and consider the KKT conditions, that is, the set of equations characterizing stationary points of the Lagrangian, *i.e.,* $\nabla\mathcal{L}(z,w)=0$. Specifically,

$$\begin{cases} 0 &= \nabla_z \mathcal{L} = \begin{bmatrix} \nabla f_1(z_1) \\ \vdots \\ \nabla f_N(z_N) \end{bmatrix} + \beta' L^{\frac{1}{2}} w + \eta' L z \\ 0 &= -\nabla_w \mathcal{L} = -\beta' L^{\frac{1}{2}} z \end{cases},$$

where, in the last equation, the negative sign is introduced for convenience. Let $\alpha<0$, $\beta := \alpha\beta'$ and $\eta := -\alpha\eta'$, and the KKT conditions can be rewritten as a fixed point equation

$$\begin{cases} z &= z + \alpha \begin{bmatrix} \nabla f_1(z_1) \\ \vdots \\ \nabla f_N(z_N) \end{bmatrix} + \beta L^{\frac{1}{2}} w - \eta L z \\ w &= w - \beta L^{\frac{1}{2}} z \end{cases}.$$

Replacing the gradient vector by $R(z)$ leads to the fixed point equation of the parametric family of maps F ($s=1/2$) defined in Section 5.2.2.

Finally, if $\sum_{n=1}^{N} f_n$ is a convex function, then (5.22) is a constrained convex optimization problem. A primal dual algorithm that can be used to solve (5.22) is the *Arrow-Hurwitz-Uzawa* method (see [39]) which performs a gradient descent in the z-component

and a gradient ascent in the w-component, that is, for $\alpha < 0$, let

$$z^{k+1} = z^k + \alpha \nabla_z \mathcal{L}(z^k, w^k)$$
$$w^{k+1} = w^k - \alpha \nabla_w \mathcal{L}(z^k, w^k).$$

Renaming $\beta := \alpha\beta'$ and $\eta := -\alpha\eta'$ recovers the Banach-Picard iteration of F, once the gradient vector is substituted by $R(z)$.

5.5.2 The Local Contraction Case

If H is differentiable at $x^\star$ and $\rho\big(\mathbf{J}_H(x^\star)\big) < 1$, then, by Ostrowski's theorem, H is a local contraction with respect to $x^\star$. Conversely, if H is a local contraction with respect to $x^\star$ and H is differentiable at $x^\star$, then $\rho\big(\mathbf{J}_H(x^\star)\big) < 1$. This suggests that local contractiveness with respect to $x^\star$ is the continuous analog to $\rho\big(\mathbf{J}_H(x^\star)\big) < 1$. Theorem 5.3.1 shows that if H is differentiable at $x^\star$ and $\rho\big(\mathbf{J}_H(x^\star)\big) < 1$, then there exists $\alpha^\star$ such that, for $0 < \alpha < \alpha^\star$,

$$\rho\Big(\mathbf{J}_{\tilde{F}}\big(\tilde{\psi}(x^\star)\big)\Big) < 1.$$

In contrast, Theorem 5.3.5 shows that if H is a local contraction with respect to $x^\star$ and $H_1, \ldots, H_N$ are globally Lipschitz, then there exists $\alpha^\star$ such that, for $0 < \alpha < \alpha^\star$, $\tilde{F}$ is a local contraction with respect to $\tilde{\psi}(x^\star)$.

The fact that Theorem 5.3.1 does not require global Lipschitzianity, neither differentiability at points other than $x^\star$, suggests that it should be possible to prove a stronger version of Theorem 5.3.5. In fact, it seems plausible that the global Lipschitzianity is superfluous and that only local contractivenes of H with respect to $x^\star$ should be enough to conclude the same property for $\tilde{F}$ with respect to $\tilde{\psi}(x^\star)$. We were not yet able to prove this stronger version. Nevertheless, this is work in progress.

5.5.3 Distributed Implementation

Section 4.7.2 of Chapter 4 commented on the memory-convergence rate trade-off, observing that the insistence that, at each iteration, each agent merely stores a d-dimensional vector, leads to a sacrifice in convergence rate. The proofs therein suggest that the diminishing step-size has a role in ensuring that the agents are successively in agreement (the iteration is "increasingly" closer to a distributed average consensus iteration). This, however, comes with a price: the iteration inherits the convergence rate of the step-size sequence. The Banach-Picard iteration of F also has a "force" driving the agents towards

consensus, the variable w^k, which, for reasons explained in Section 5.5.1, we term the *dual variable*. To see this, consider the iterations of $\bar{z}^k$, y^k, and w^k defined as in Section 5.3.3; there are maps $\tilde{T}_1$ and $\tilde{T}_2$ such that

$$\begin{aligned}\bar{z}^{k+1} &= \tilde{T}_1(\bar{z}^k, y^k)\\ \begin{bmatrix} y^k \\ w^k \end{bmatrix} &= \mathcal{B}(\eta,\beta)\begin{bmatrix} y^k \\ w^k \end{bmatrix} + \tilde{T}_2(\bar{z}^k, y^k),\end{aligned}$$

where $\rho\big(\mathcal{B}(\eta,\beta)\big) < 1$. Clearly, w^k "influences" $\bar{z}^k$ (indirectly "via" y^k) and the way w^k "interacts" with the iteration is via the matrix product

$$\mathcal{B}(\eta,\beta)\begin{bmatrix} y^k \\ w^k \end{bmatrix}.$$

Since $\rho\big(\mathcal{B}(\eta,\beta)\big) < 1$, any multiplication by $\mathcal{B}(\eta,\beta)$ contracts a vector towards zero, and, thus, the role of w^k seems to be that of driving y^k to zero. Now, recall that $y^k = \tilde{U}^T z^k$, *i.e.,* y^k is the off-consensus component of z^k expressed in the basis formed by the columns of $\tilde{U}$. Therefore, the role of w^k is to drive the off-consensus component to zero, that is, to drive the agents towards consensus.

As it was pointed out several times, the price paid for using w^k, instead of a diminishing step-size, as the "driving force" is to have each agent maintaining two variables (z_n^k, w_n^k) in memory. We stress out that this price is still paid even if the variable w^k is eliminated, leading to algorithms such as the EXTRA-distributed Banach-Picard iteration or the DIGing-distributed Banach-Picard iteration. In fact, in both those cases the algorithm is of the form $z^{k+2} = J(z^k, z^{k+1})$ and, hence, the agents need to have in memory z_n^k and z_n^{k+1} to update.

5.5.3.1 Communications Per Iteration

Up until now, the issue of the number of communications required by an update of a distributed algorithm was disregarded. The reason to avoid this topic is that its formalization is beyond the scope of this work. To hint at its non-trivial nature, consider the meaning of an "update" of F, that is,

$$\begin{aligned}z^{k+1} &= z^{k+1} + \alpha R(z^k) + \beta L w^k - \eta L z^k\\ w^{k+1} &= w^k - \beta L z^k.\end{aligned}$$

From the implementation point of view, this update could be carried out as two updates, with the time between instant k and instant $k+1$ being broken into two smaller time intervals. It is as if there was a time instant $k+1/2$ between k and $k+1$, and the update was performed by first updating

$$\begin{aligned} z^{k+\frac{1}{2}} &= z^k + \alpha R(z^k) + L(\beta w^k - \eta z^k) \\ w^{k+\frac{1}{2}} &= w^k \end{aligned}$$

and then

$$\begin{aligned} z^{k+1} &= z^{k+\frac{1}{2}} \\ w^{k+1} &= w^{k+\frac{1}{2}} - \beta L z^k. \end{aligned}$$

This corresponds to regarding $(z^0, w^0) \to (z^1, w^1) \to \dots (z^k, w^k) \to \cdots$ as has having intermediate steps

$$(z^0, w^0) \to (z^{\frac{1}{2}}, w^{\frac{1}{2}}) \to (z^1, w^1) \cdots (z^k, w^k) \to (z^{k+\frac{1}{2}}, w^{k+\frac{1}{2}}) \cdots$$

Counting the number of communications by the "number of products by L", it seems that, at each iteration, only one round of communications is required: from time k to time $k+1/2$ agent n only needs to communicate $\beta w_n^k - \eta z_n^k$ to its neighbors, and from time $k+1/2$ to time k it only z_n^k. Consequently, it seems tempting to say that, at each iteration, agent n only communicates a d-dimensional vector. However, suddenly agent n has to keep in memory more than a $2d$-dimensional vector: at time k it needs to have z_n^k and w_n^k, but, once it updates to obtain $z_n^{k+1/2}$, it cannot erase z_n^k, since this will be required at time $k+1/2$ to update w_n^k. This is as much as we will say about what an update means from the implementation point of view.

For the sake of the discussion, let's leave aside issues such as the time it takes z_n^k to arrive at neighbors of agent n, and assume that once agent n "broadcasts" its value at time k, all of its neighbors receive it immediately and at the same time. Even though we are appealing to intuition when we speak of "an update", it is seems reasonable to say that the algorithm $z^{k+1} = F_k(z^k)$ from Chapter 4 can be implemented in such a way that agent n only keeps in memory a d-dimensional vector that communicates at each iteration. Similarly, the EXTRA-distributed Banach-Picard iteration, that is, (5.19), has the rearrangement

$$z^{k+2} = (I+W)(z^{k+1} - \frac{1}{2}z^k) + \alpha\big(R(z^{k+1}) - R(z^k)\big),$$

and, hence, agent n only needs to communicate a d-dimensional vector to its neighbors, *i.e.,* $z_n^{k+1} - 1/2 z_n^k$. Moreover, it has to maintain a $2d$-dimensional variable (z_n^{k+1}, z_n^k) in memory. We conclude that, when compared with the algorithms from Chapter 4, the EXTRA-distributed Banach-Picard iteration has the same "communication per iteration cost", at the expense of more memory. In contrast, the DIGing-distributed Banach-Picard iteration does not have a similar a rearrangement, because of the presence of W^2, *i.e.,* it is not obvious how to rearrange

$$z^{k+2} = 2Wz^{k+1} - W^2 z^k + \alpha\big(R(z^{k+1}) - R(z^k)\big)$$

in the form

$$z^{k+2} = \mathcal{W} g(z^{k+1}, z^k) + \alpha\big(R(z^{k+1}) - R(z^k)\big),$$

where $\mathcal{W}$ is a matrix compatible with the graph structure and where g is a map of the form

$$g(z^{k+1}, z^k) = \big(g_1(z_1^{k+1}, z_1^k), \ldots, g_N(z_N^{k+1}, z_N^k)\big).$$

Consequently, a tempting conclusion is: EXTRA is clearly superior to DIGing, because the latter has more communication per iteration cost for the same memory requirements. Not surprisingly, this comes with a price, although subtle: before the particular choice made for the initialization of w^k, *i.e.,* w^0, the elimination of the dual variable in $(z^{k+1}, w^{k+1}) = F(z^k, w^k)$ leads to

$$\begin{aligned} z^0 &\in \mathbb{R}^{dN} \\ z^1 &= z^0 + \alpha R(z^0) + \beta L^{\frac{1}{2}} w^0 - \eta L z^0 \\ z^{k+2} &= (I + W)(z^{k+1} - \frac{1}{2} z^k) + \alpha\big(R(z^{k+1}) - R(z^k)\big) \end{aligned}$$

for EXTRA, and

$$\begin{aligned} z^0 &\in \mathbb{R}^{dN} \\ z^1 &= w^0 - z^0 + W(2z^0 - w^0) + \alpha R(z^0) \\ z^{k+2} &= 2Wz^{k+1} - W^2 z^k + \alpha\big(R(z^{k+1}) - R(z^k)\big) \end{aligned} \tag{5.23}$$

for DIGing. Now note, that the only feasible "distributed initialization" of z^1 in EXTRA is $w^0 = 0$, because of the presence of $L^{\frac{1}{2}}$. However, even though, to coincide exactly with the elimination of the second variable of DIGing, we consider an initialization of z^1 with

$w^0 = z^0$, any choice of w^0 leads to a "distributed initialization" of z^1 in (5.23).

To conclude, at the expense of being less flexible, EXTRA is superior, as far as the communication per iteration cost is considered. To imagine why this rigidity is not necessarily irrelevant, suppose that H has multiple fixed points, one of which sought by the agents. It could happen that the agents had some "insight" that allowed them to "wisely" choose w^0 in order to ensure the convergence of DIGing to the desired fixed point. In contrast, by being flexible only in the choice of z^0, EXTRA could be bound to "escape" the desired fixed point.

5.5.4 Why EXTRA is "Natural"

Arriving at EXTRA from

$$z^{k+2} = (2I - \eta L)z^{k+1} - (I + \beta^2 L - \eta L)z^k + \alpha\big(R(z^{k+1}) - R(z^k)\big) \tag{5.24}$$

is rather natural if one is "aware" of EXTRA: let $L = I - W$ to obtain

$$z^{k+2} = \big(I(2-\eta) + \eta W\big)z^{k+1} - \big(I(1 - \eta + \beta^2) + (\eta - \beta^2)W\big)z^k + \alpha\big(R(z^{k+1}) - R(z^k)\big),$$

and "knowing" EXTRA leads to

$$\begin{cases} I(2-\eta) + \eta W & = I + W \\ I(1 - \eta + \beta^2) + (\eta - \beta^2)W & = \big(I + W\big)\frac{1}{2} \end{cases}.$$

The first equation leads to $\eta = 1$ which leads to $\beta^2 = 1/2$ in the second equation. Afterwards, we argue, as we did in Remark 5.10, that this choice of η and β satisfies the hypothesis of Theorems 5.3.1, 5.3.4, and 5.3.5.

In contrast, an interesting question is whether EXTRA emerges rather naturally from (5.24), for someone inspecting (5.24) without "knowledge" of EXTRA. This section addresses this problem. Suppose that $\{\lambda_1, \dots, \lambda_{dN}\}$ are the eigenvalues of L and recall that $0 \leq \lambda_i < 2$. We wish to find $\eta > 0$ and $\beta > 0$ such that $|y + 1| < 1$ for every

$$y \in \bigcup_{i=1,\dots,dN, \lambda_i \neq 0} \Big\{x \in \mathbb{C} : x^2 + \eta\lambda_i x + \lambda_i \beta^2 = 0\Big\}. \tag{5.25}$$

Redefine $2a = \eta$ and $b = \beta^2$, and consider the equation

$$y^2 + 2a\lambda_i y + \lambda_i b = 0$$

which can be equivalently written as

$$(y + a\lambda_i)^2 + \lambda_i(b - \lambda_i a^2) = 0.$$

Since $\lambda_i > 0$, the roots of this equation are real if $b \leq \lambda_i a^2$ and have non-negative imaginary parts if $b > \lambda_i a^2$. Assume that $L = I - (W \otimes I_d)$, where W is the Metropolis Weight Matrix (see Chapter 2). From Chapter 2, the only information that agent n needs to "learn" are the values of row n of W, the degrees of its neighbors; no other topological information about W is available at agent n. For this reason, to choose a and b (equivalently, η and β), agent n should only rely on upper and lower bounds of the non-zero eigenvalues of $I - W$, one such bound being $0 < \lambda_i < 2$. However, a natural question is whether this bound can be improved by leveraging only on the topological information required for building the Metropolis Weight Matrix; we show that there is no "universal" lower bound better than $0 < \lambda_i$.

Lemma 5.5.1. *For every $\epsilon > 0$ there exists a Metropolis Weight Matrix W such that $I - W$ has a positive eigenvalue smaller than ϵ.*

Proof. Let N be a natural number and define the vector $x \in \mathbb{R}^N$ by $x_m = 0$ if $m \neq 2, N$ and $x_2 = x_N = 1$. Consider the matrix A_N with the first row equal to x and row i obtained by a cyclic shift to the right of $i - 1$ positions of x (*e.g.* the second row of A_N is $[1, 0, 1, 0, \ldots, 0]$). Clearly, A_N is the adjacency matrix[4] of a graph $\mathcal{G}$ that can be drawn with N nodes lying in a circle with every node having degree 2. Moreover, the Metropolis Weight Matrix of $\mathcal{G}$ is

$$W_N = \frac{1}{3}(I + A_N),$$

since each node has degree 2. Consequently,

$$I - W_N = \frac{1}{3}\big(2I - A_N\big).$$

We will now show that, by choosing N sufficiently large, $I - W_N$ has an arbitrarily small positive eigenvalue. It can be proved (see [81]) that A_N has an eigenvalue of the form $2\cos(2\pi/N)$, and, hence $I - W_N$ has an eigenvalue of the form

$$\frac{1}{3}\big(2 - 2\cos(2\pi/N)\big).$$

By choosing N sufficiently large, we can have $0 < \frac{1}{3}\big(2 - 2\cos(2\pi/N)\big) < \epsilon$. □

[4]Recall that the adjacency matrix of a graph is of the form $A_{ij} = 1$, if i and j are neighbors, and $A_{ij} = 0$ otherwise.

The relevance of this result is that if an agent has degree two and, to build his row of the Metropolis Weight Matrix, "learns" that his two neighbors have degree two as well, then, from his perspective, he might as well be in a complete graph with three nodes. The takeaway is that, without any further topological information, no node can distinguish whether he is part of a complete graph with three nodes or a two-regular circular graph on N nodes. As a consequence, no lower bound better than $0 < \lambda_i$ can be used.

Suppose, for the sake of argument, that there was (we now know that there isn't) a better lower bound than $0 < \lambda_i$, *i.e.,* suppose that $\lambda_i > \epsilon$, then we could look at $b \leq \epsilon a^2$, which would imply $b \leq \lambda_i a^2$, workout the real roots of $(y + a\lambda_i)^2 + \lambda_i(b - \lambda_i a^2) = 0$, and proceed to choose a and b to ensure that $(y+1)^2 < 1$. In the absence of such a lower bound, we work with the upper bound and restrict $b \geq 2a^2$, which implies that $b > \lambda_i a^2$ for all λ_i. The roots are, in this case, given by

$$-a\lambda_i \pm i\sqrt{\lambda_i b - \lambda_i^2 a^2},$$

and, hence,

$$|1 + y|^2 = (1 - a\lambda_i)^2 + \lambda_i b - \lambda_i^2 a^2 = 1 - \lambda_i(2a - b).$$

Consequently, we should have $2a > b$, in order to ensure that $|1 + y|^2 < 1$.

Now comes the final "trick": we want $|1 + y|^2 < 1$, for all y in the set (5.25), because this, from Lemma 5.3.3, will ensure that $\rho\big(I + \mathcal{A}(\eta, \beta)\big) < 1$. This last matrix is unitarily similar to $\mathcal{B}(\eta, \beta)$ (see Section 5.5.3) and this matrix appears to have the role of driving the system towards consensus. Therefore, the lower its spectral radius, the faster we should expect consensus to be achieved. Consequently, to choose a and b, a natural heuristic is the choice

$$\begin{array}{ll} \underset{a>0, b>0}{\text{maximize}} & 2a - b \\ \text{subject to} & b \geq 2a^2 \end{array}. \tag{5.26}$$

To find the solution, observe that it must live in the compact region $2a \geq b$, $b, a \geq 0$ and $b \geq 2a^2$ (by drawing a picture, this region is between a parabola and a line). It is clear that the solution cannot be attained at a point (a, b) outside the parabola, since, otherwise, we could slightly decrease b and obtain a larger value. Consequently, we should solve

$$\underset{0 \leq a \leq 1}{\text{maximize}} \quad 2a - 2a^2 \quad , \tag{5.27}$$

and we easily see the solution to be $a = 1/2$, which implies $b = 1/2$. Since $\eta = 2a$ and $b = \beta^2$, we obtain $\eta = 1$ and $\beta^2 = 1/2$, the choices leading to EXTRA.

Chapter 6

Distributed PCA

6.1 Introduction

In this Chapter, we show how the results from Chapter 5 lead to an algorithm for distributed *principal component analysis* (PCA). To this end, we consider a map H that can be implicitly written as an average of local maps and having as a fixed point the solution to the PCA problem. The Banach-Picard iteration of H was proposed in [21] and appears to have been inspired by [16], hence, in [21], $x^{k+1} = H(x^k)$ is termed a "mini-batch variant" of Sanger's algorithm (SA). From SA, the authors of [21] arrive at a distributed algorithm they term *accelerated distributed Sanger's algorithm* (ADSA), which is nothing but the EXTRA-distributed Banach-Picard iteration (5.19) for the maps $H_1, \ldots, H_N$ for which $H = 1/N \sum_{n=1}^{N} H_n$. In [21], there is no proof of convergence for ADSA; we fill this gap by appealing to the results of Chapter 5. This Chapter is mainly based on our work [14], which is currently under review in the *IEEE Transactions on Signal Processing*.

6.2 Problem Statement: Distributed PCA

Consider a network of N agents, where the interconnection structure is represented by an undirected and connected graph. Each agent n holds a finite set $\mathbf{Y}_n \subseteq \mathbb{R}^d$ and the agents seek to collectively find the m top eigenvectors (*i.e.,* the m eigenvectors associated to the largest m eigenvalues) of the matrix

$$C = \frac{1}{M} \sum_{n=1}^{N} C_n,$$

where $M = \sum_{n=1}^{N} |\mathbf{Y}_n|$, *i.e.,* the sum of the cardinalities of each $\mathbf{Y}_n$, and

$$C_n = \sum_{y \in \mathbf{Y}_n} yy^T.$$

Observe that each C_n is positive semi-definite and, hence, the same holds for C. We assume that C is positive definite, *i.e.,* $C \succ 0$, and that the eigenvalues of C are $\lambda_1 > \lambda_2 > \ldots > \lambda_m > \lambda_{m+1} \geq \ldots \geq \lambda_d > 0$. By a solution to the PCA problem we mean a $d \times m$ matrix $X^\star$ such that $(X^\star)^T X^\star = I_m$ and $CX^\star = X^\star \mathrm{diag}(\lambda_1, \ldots, \lambda_m)$. Note that, given such a solution $X^\star$, we can multiply any of its columns by minus one and we obtain another solution. Since $\lambda_1 > \lambda_2 > \ldots > \lambda_m$, *i.e.,* the top m eigenvectors are all distinct (one-dimensional eigenspaces), all the solutions are of this form and there are 2^m of them. In many situations we will use the slightly imprecise term "the solution".

6.3 Sanger's Algorithm

Let $X^\star \in \mathbb{R}^{d \times m}$ be the solution to the PCA problem described in the previous section, that is, $X^\star$ is a $d \times m$ matrix with unit-norm, orthogonal columns such that $Cx_i^\star = \lambda_i x_i^\star$, where $x_i^\star$ denotes the ith column of $X^\star$ and λ_i is the ith largest eigenvalue of C. To arrive at a distributed algorithm for PCA from the results of Chapter 5, we need to find a map H that can be written as $H(X) = 1/N \sum_{n=1}^{N} H_n(C_n, X)$ and such that $H(X^\star) = X^\star$. Moreover, to have at least local linear convergence towards $X^\star$, we need to have as well $\rho\big(\mathbf{J}_H(X^\star)\big) < 1$. The maps H_n were written as $H_n(C_n, \cdot)$, rather than $H_n(\cdot)$, to stress out the dependence on agent n's local data; however, to simplify the notation the dependence on C_n will now be dropped.

Consider the map $H : \mathbb{R}^{d \times m} \to \mathbb{R}^{d \times m}$ defined, for $\gamma > 0$, by

$$H(X) = X + \gamma\Big(CX - X\mathcal{U}\big(X^T CX\big)\Big), \tag{6.1}$$

where $\mathcal{U}(\cdot)$ maps a square matrix M to an upper triangular matrix with the same dimension and upper triangular part of M. The Banach-Picard iteration of H will be called *Sanger's Algorithm* (see [21] and [16]). We will show that $H(X^\star) = X^\star$ and that we can choose $\gamma > 0$ such that $\rho\big(\mathbf{J}_H(X^\star)\big) < 1$. By observing that $H = 1/N \sum_{n=1}^{N} H_n$, with H_n defined by

$$H_n(X) = X + \gamma\Big(\frac{N}{M} C_n X - X\mathcal{U}\big(X^T \frac{N}{M} C_n X\big)\Big), \tag{6.2}$$

we see that we have all the necessary ingredients to appeal to the results developed in

Chapter 5. As a consequence, we conclude that the EXTRA-distributed Banach-Picard iteration (5.19) with the maps H_n enjoys at least local linear convergence with respect to $X^\star$. The EXTRA-distributed Banach-Picard iteration (5.19) in this case is nothing but the *accelerated distributed Sanger's algorithm* (ADSA) proposed in [21], thus, it will be referred as ADSA from now on.

6.3.1 The Case $m = 1$

In this initial Section we focus on the case $m = 1$, *i.e.,* the problem of estimating only the top eigenvector $x^\star$ of C. This is already an interesting case in its own and it serves as a motivation for the general case. Its analysis is more straightforward because the map $\mathcal{U}$ reduces to the identity. Given that X is, in this case, a vector, we will denote it by x.

Let V be an orthogonal matrix that satisfies $V^TCV = D$, where $D = \operatorname{diag}(\lambda_1, \ldots, \lambda_d)$, with $\lambda_1, \ldots, \lambda_d$ being the eigenvalues of C in decreasing order. Instead of H, consider the map G that corresponds to H up to a change in coordinates $x \mapsto Vx$, that is, $G(x) = V^TH(Vx)$. Observe that to understand H it is enough to understand G. In fact, x is a fixed point of H if and only if V^Tx is a fixed point of G. Moreover, from the chain rule we have that $\mathbf{J}_G(x) = V^T\mathbf{J}_H(Vx)V$, and, hence, the eigenvalues of $\mathbf{J}_H(x)$ are those of $\mathbf{J}_G(V^Tx)$. For these reasons, we will focus on G rather than H.

The first relevant observation is that, although G appears to be a gradient step, it really is not. To see this consider the ith component function of G, *i.e.,*

$$\big(G(x)\big)_i = x_i + \gamma\big(\lambda_i x_i - x_i \sum_{l=1}^{d} \lambda_l x_l^2\big).$$

and, hence,

$$\frac{\partial\big(G(x)\big)_i}{\partial x_j} = \begin{cases} 1 + \gamma\big(\lambda_i - 2\lambda_i x_i^2 - \sum_{l=1}^{d} \lambda_l x_l^2\big), & i = j \\ -2\gamma\lambda_j x_i x_j, & i \neq j \end{cases}. \tag{6.3}$$

Therefore, for $i \neq j$ and $\gamma \neq 0$, we obtain

$$\frac{\partial\big(G(x)\big)_i}{\partial x_j} = \frac{\partial\big(G(x)\big)_j}{\partial x_i}$$

if and only if $\lambda_j x_i x_j = \lambda_i x_i x_j$, showing that if C has at least two distinct eigenvalues λ_i and λ_j, then the Jacobian of G at a point x with $x_i \neq 0$ and $x_j \neq 0$ is not symmetric. As a consequence of Theorem 1.3.1 in [82], even though $G(x) = x + \gamma(Dx - xx^TDx)$ looks like

a gradient step, it is not, *i.e.*, we cannot define a function f such that $G(x) = x+\gamma\nabla f(x)$.

To identify the fixed points of G, consider the equation $G(x) = x$ and, since $\gamma \neq 0$, we obtain $Dx = (x^T Dx)x$. If $x \neq 0$, this equation shows that x is an eigenvalue of D associated to the eigenvector $x^T Dx$. Multiplying both sides of $Dx = (x^T Dx)x$ by x^T, we obtain $x^T Dx(1 - \|x\|^2) = 0$, and, since we are assuming that $D \succ 0$ (which follows from $C \succ 0$), we obtain that $\|x\| = 1$. We conclude that the fixed points of G are either zero or unit norm eigenvectors of D. Now assume that the eigenvalues of D (those of C) are all distinct, that is, all elements in the diagonal of D appear only once. In this case, G has a $2d + 1$ fixed points, the set of fixed points being $\{0, \pm e_1, \ldots, \pm e_d\}$, where e_i is the ith canonical vector of $\mathbb{R}^d$, *i.e.,* $(e_i)_j = 0$, if $i \neq j$ and $(e_i)_i = 1$.

Consider now the local behavior of G near a fixed point $x^\star$, that is, consider $\mathbf{J}_G(x^\star)$. From (6.3), we see that $\mathbf{J}_G(0)$ is a diagonal matrix, with the diagonal entries being $1+\gamma\lambda_i$ and, hence, for any $\gamma > 0$, we obtain that $\rho\big(\mathbf{J}_G(0)\big) = 1 + \gamma\lambda_1 > 1$. If $x^\star = \pm e_i$, then $\mathbf{J}_G(\pm e_i)$ is a diagonal matrix with elements being

$$\big(\mathbf{J}_G(\pm e_i)\big)_{jj} = \begin{cases} 1 - 2\gamma\lambda_j, & j = i \\ 1 + \gamma(\lambda_j - \lambda_i), & i \neq j. \end{cases}$$

Recall that we assumed that the eigenvalues are distinct and in decreasing order: $\lambda_1 > \lambda_2 > \ldots > \lambda_d > 0$. From this, we obtain that, for $\gamma > 0$ sufficiently small, $\rho\big(\mathbf{J}_G(\pm e_1)\big) < 1$ and, for $i \geq 2$, $\rho\big(\mathbf{J}_G(\pm e_i)\big) > 1$, with $\mathbf{J}_G(\pm e_i)$ having a real eigenvalue larger than 1. Informally, $\pm e_1$ are stable fixed points and the others (a finite number of them) are unstable. This is already a nice result, because we know that the map $\tilde{F}$ (see Chapter 5) from where the distributed algorithm emerges preserves this feature.

To finish this section, we show that we can further reduce γ in such a way that the conditions of Theorem 3.2.4 of Chapter 3 hold. Consequently, we obtain that the set of initial conditions x^0 that satisfy $\lim_k H^k(x^0) \in \{0, \pm e_2, \ldots, \pm e_d\}$ has Lebesgue measure zero. Informally, if $H^k(x^0)$ converges, then it "almost surely converges" to $\pm e_1$ at least linearly. We show this in two steps: first, we show that we can trap the iteration in a compact set, that is, $H(\bar{B}) \subseteq \bar{B}$ for a compact set $\bar{B}$; second, we use the compactness of $\bar{B}$ to ensure that $\det(\mathbf{J}_H(x)\big) \neq 0$ for all x, thus establishing the conditions of Theorem 3.2.4 of Chapter 3.

Let $\bar{B}$ be the Euclidean ball of radius $\sqrt{2}$ centered at the origin, that is, the set of x such that $\|x\|^2 = x^T x \leq 2$. Observe that

$$\big\|G(x)\big\|^2 = \|x\|^2 + 2\gamma x^T Dx(1 - \|x\|^2) + \gamma^2\big(x^T D^2 x - (2 - \|x\|^2)(x^T Dx)^2\big)$$

and, since we are assuming $\|x\|^2 \leq 2$, we obtain

$$\begin{aligned}\|G(x)\|^2 &\leq \|x\|^2 + 2\gamma x^T Dx(1 - \|x\|^2) + \gamma^2 x^T D^2 x \\ &\leq \|x\|^2 + 2\gamma x^T Dx(1 - \|x\|^2) + 2\gamma^2 \lambda_1^2.\end{aligned}$$

If $3/2 \leq \|x\|^2 \leq 2$, then

$$\begin{aligned}\|G(x)\|^2 &\leq 2 - \gamma x^T Dx + 2\gamma^2 \lambda_1^2 \\ &\leq 2 - \gamma \lambda_d \|x\|^2 + 2\gamma^2 \lambda_1^2 \\ &\leq 2 - \frac{3}{2}\gamma \lambda_d + 2\gamma^2 \lambda_1^2.\end{aligned}$$

If $\|x\|^2 \leq \frac{3}{2}$, then

$$\begin{aligned}\|G(x)\|^2 &\leq \frac{3}{2} + 2\gamma x^T Dx + 2\gamma^2 \lambda_1^2 \\ &\leq \frac{3}{2} + 3\gamma \lambda_1 + 2\gamma^2 \lambda_1^2.\end{aligned}$$

These inequalities are enough to define $\gamma^\star$ such that, for $0 \leq \gamma \leq \gamma^\star$, $\|G(x)\|^2 \leq 2$, whenever $\|x\|^2 \leq 2$ We conclude that $G(\bar{B}) \subseteq \bar{B}$.

To finish, consider $\det(\mathbf{J}_G(x))$. Given that the entries of $\mathbf{J}_G(x)$ only involve products and sums, there are polynomials in x, denoted by $p_0(x), \ldots, p_d(x)$, such that

$$\det(\mathbf{J}_G(x)) = \sum_{j=0}^{d} \gamma^j p_j(x).$$

For $\gamma = 0$, $\mathbf{J}_G(x)$ reduces to the identity for all x, *i.e,* $p_0(x) = \det(I) = 1$. Let $m_1, \ldots, m_d$ be the minima of $p_1(x), \ldots, p_d(x)$ in $\bar{B}$; since $\gamma \geq 0$, we obtain, for all $x \in \bar{B}$,

$$1 + \sum_{j=1}^{d} \gamma^j m_j \leq \det(\mathbf{J}_G(x)).$$

This inequality is enough to choose $\gamma^\star$ such that, for $0 \leq \gamma \leq \gamma^\star$, $\det(\mathbf{J}_G(x)) \neq 0$, for all $x \in \bar{B}$.

We summarize the results of this section in the following lemma.

Lemma 6.3.1. *Consider $C \succ 0$, with distinct eigenvalues $\lambda_1 > \lambda_2 > \ldots > \lambda_d > 0$, and let*

$$H(x) = x + \gamma(Cx - xx^T Cx),$$

and $\bar{B} = \{x \in \mathbb{R}^d : \|x\|^2 \leq 2\}$. *Then, for* $\gamma \neq 0$, *the set of fixed points of* H, *Fix*(H), *is the set* $\{0, \pm v_1, \ldots, \pm v_d\}$, *where* $\|v_i\|^2 = 1$ *and* $Cv_i = \lambda_i$, *for all* i. *Moreover, there exists* $\gamma^\star$ *such that, for* $0 < \gamma \leq \gamma^\star$,

1) $\rho\big(\mathbf{J}_H(\pm v_1)\big) < 1$ *and, for* $x^\star \in$ *Fix*$(H) \setminus \{\pm v_1\}$, $\mathbf{J}_H(x^\star)$ *has a real eigenvalue strictly larger than* 1 *(consequently,* $\rho\big(\mathbf{J}_H(x^\star)\big) > 1$*).*

2) *The set of* $x^0 \in \bar{B}$ *such that* $\lim_k H^k(x^0) \in$ *Fix*$(H) \setminus \{\pm v_1\}$ *has Lebesgue measure zero.*

6.3.2 The General Case $m \geq 1$

In the general case H maps matrices to matrices, and, hence, it is slightly more complicated to analyze. The contents of this section are essentially those of the first part of our work [14].

6.3.2.1 Fixed Points of H

The following lemma characterizes the fixed points of H.

Lemma 6.3.2. *Let* $C \succ 0$. *If* $X^\star \in \mathbb{R}^{d \times m}$ *satisfies*

$$CX^\star = X^\star \mathcal{U}((X^\star)^T C X^\star), \tag{6.4}$$

then, each column of $X^\star$ *is either* 0 *or a unit-norm eigenvector of* C. *Moreover, the columns are orthogonal,* i.e.*,* $(X^\star)^T X^\star$ *is diagonal with the diagonal elements being either one or zero.*

Proof. Suppose $X^\star$ satisfies (6.4). Throughout this proof, $x_i^\star$ denotes the ith column of $X^\star$. Consider the equation imposed by the first column, $x_1^\star$, *i.e.*,

$$Cx_1^\star = \big((x_1^\star)^T C x_1^\star\big) x_1^\star,$$

and multiply both sides by $(x_1^\star)^T$, which yields

$$\big((x_1^\star)^T C x_1^\star\big)\big(1 - \|x_1^\star\|^2\big) = 0.$$

From the two equalities

$$\begin{aligned} \big((x_1^\star)^T C x_1^\star\big) x_1^\star &= Cx_1^\star, \\ \big((x_1^\star)^T C x_1^\star\big)\big(1 - \|x_1^\star\|^2\big) &= 0, \end{aligned}$$

we conclude that either $x_1^\star = 0$ or $x_1^\star$ is a unit-norm eigenvector of C.

Considering the second column, we prove that $x_2^\star$ is either zero or a unit-norm eigenvector of C that is orthogonal to $x_1^\star$. Observe that

$$Cx_2^\star = \big((x_1^\star)^T C x_2^\star\big)x_1^\star + \big((x_2^\star)^T C x_2^\star\big)x_2^\star. \tag{6.5}$$

Now recall that $x_1^\star = 0$ or $x_1^\star$ is a unit-norm eigenvector of C. If $x_1^\star = 0$, then (6.5) reduces to

$$Cx_2^\star = \big((x_2^\star)^T C x_2^\star\big)x_2^\star$$

and the result follows as in the case of $x_1^\star$. If $x_1^\star \neq 0$, then it is a unit-norm eigenvector of C and, hence, there exists β such that $(x_1^\star)^T C = \beta (x_1^\star)^T$ and (6.5) reduces to

$$Cx_2^\star = \beta\big((x_1^\star)^T x_2^\star\big)x_1^\star + \big((x_2^\star)^T C x_2^\star\big)x_2^\star. \tag{6.6}$$

Multiply on the left by $(x_1^\star)^T$ and use $\|x_1^\star\|^2 = 1$ to obtain

$$\big((x_2^\star)^T C x_2^\star\big)(x_1^\star)^T x_2^\star = 0.$$

If $x_2^\star = 0$, we are done. If not, then $0 = (x_1^\star)^T x_2^\star$ and, returning to (6.6), it holds that

$$Cx_2^\star = \big((x_2^\star)^T C x_2^\star\big)x_2^\star.$$

This establishes the claim for $x_1^\star$ and $x_2^\star$. Proceeding as we did for the second column, it is possible to construct a proof by induction establishing the result. □

6.3.2.2 Stability Properties of the Fixed Points of H

We start by looking at the stability properties of the fixed points of H that have a zero column.

Lemma 6.3.3. *Let $X^\star$ be a fixed point of H and suppose that $x_j^\star = 0$, where $x_j^\star$ denotes the jth column of $X^\star$. Then, for $\gamma > 0$, $\mathbf{J}_H(X^\star)$ as a real eigenvalue larger than one.*

Proof. Let $X^\star$ be a fixed point of H such that $x_j^\star = 0$ and consider the matrix curve $X^\star(t)$, where $x_i^\star(t) = x_i^\star$ for $i \neq j$ and $x_j^\star(t) = tx^\star$, where $x^\star$ is the (unique up to sign change) unit norm eigenvector associated to λ_1, that is, $Cx^\star = \lambda_1 x^\star$. Observe that, at zero this curve coincides with $X^\star$, and that $CX^\star(t) = X^\star(t)D$, where $D = \mathrm{diag}(a_1, \ldots, a_{j-1}, \lambda_1, a_{j+1}, \ldots, a_m)$, where each a_i is an eigenvalue of C if $x_i^\star \neq 0$ and can be

assumed to be one if $x_i^\star = 0$. Moreover $\big(X^\star(t)\big)^T X^\star(t) = \mathrm{diag}(b_1, \ldots, b_{j-1}, t^2, b_j, \ldots, b_m) := B(t)$, where each b_i is either zero or one. Consider now the curve $H(X^\star(t))$,

$$\begin{aligned} H\big(X^\star(t)\big) &= X^\star(t) + \gamma\Big(X^\star(t)D - X^\star(t)\mathcal{U}\big(B(t)D\big)\Big) \\ &= X^\star(t) + \gamma\Big(X^\star(t)D - X^\star(t)B(t)D\Big) \\ &= X^\star(t)\Big(I + \gamma\big(D - B(t)D\big)\Big) \end{aligned}$$

and, thus, the jth column of $H(X^\star(t))$, denoted by $hx_j^\star(t)$, is given by $hx_j^\star(t) = tx^\star\big(1 + \gamma\lambda_1(1-t^2)\big)$. Considering the derivative at zero, we see that

$$\frac{dhx_j^\star(t)}{dt}\Big|_{t=0} = x^\star(1+\gamma\lambda_1). \tag{6.7}$$

For readers familiar with smooth manifolds, this is already enough to see that, for $\gamma > 0$, $\mathbf{J}_H(X^\star)$ has a real eigenvalue larger than one. In fact, $\mathbf{J}_H(X^\star)$ can be seen as a linear map from derivatives of curves at $X^\star$ to derivative of curves at $H(X^\star)$ and, the previous calculation shows that this map has an expanding direction. However, we do not assume any knowledge of smooth manifold theory, hence, we spell out the details: let Z be the $d \times m$ matrix whose jth column is $x^\star$ and the remaining columns are zero, thus, $X^\star(t) = X^\star + tZ$. The linear map $\mathbf{J}_H(X^\star)$ must, by definition, satisfy

$$\lim_{t\to 0} \frac{\big\|H(X^\star + tZ) - H(X^\star) - t\mathbf{J}_H(X^\star)Z\big\|}{|t|\|Z\|} = 0. \tag{6.8}$$

What we have shown in (6.7) is that

$$\lim_{t\to 0} \frac{H(X^\star + tZ) - H(X^\star)}{t} = (1+\gamma\lambda_1)Z$$

and, from (6.8), we conclude straightforwardly that $\mathbf{J}_H(X^\star)Z = (1+\gamma\lambda_1)Z$. □

From now on we only focus on fixed points $X^\star$ of H that have non-zero columns, *i.e.*, the columns of $X^\star$ are unit norm eigenvectors of C. In the previous lemma we essentially appealed to the fact that $\mathbf{J}_H(X^\star)$ is the linear map that maps a matrix Z to the matrix

$$\frac{dH(X^\star + tZ)}{dt}\Big|_{t=0}.$$

We now identify, skipping straightforward details, the form of this map. Consider a curve

of the form $X^\star(t) = X^\star + tZ$, and observe that

$$X^\star(t)^T C X^\star(t) = D + tD(X^\star)^T Z + tZ^T X^\star D + t^2 Z^T C Z,$$

where D is a diagonal matrix with eigenvalues of C, *i.e.*, the diagonal matrix that satisfies $CX^\star = X^\star D$. A straightforward calculation reveals that

$$\frac{dH(X^\star(t))}{dt}\Big|_{t=0} = Z + \gamma\Big(CZ - X^\star\mathcal{U}\big(D(X^\star)^T Z + Z^T X^\star D\big) - ZD\Big)$$

and, hence, $\mathbf{J}_H(X^\star)$ is the linear map given by

$$Z \to Z + \gamma\Big(CZ - X^\star\mathcal{U}\big(D(X^\star)^T Z + Z^T X^\star D\big) - ZD\Big).$$

Consider now the matrix $\hat{X}^\star$ that extends $X^\star$ to an orthonormal basis of eigenvectors of C, that, is $(\hat{X}^\star)^T\hat{X}^\star = I_d$ and $C\hat{X}^\star = \hat{X}^\star\hat{D}$, where

$$\hat{D} = \mathrm{diag}(\lambda_{P(1)}, \ldots, \lambda_{P(m)}, \lambda_{P(m+1)}, \ldots, \lambda_{P(d)}),$$

P is a permutation of the set $\{1, \ldots, d\}$, and $D = \mathrm{diag}(\lambda_{P(1)}, \ldots, \lambda_{P(m)})$. Finally, consider the unitary transformation $(\hat{X}^\star)^T\mathbf{J}_H(X^\star)\hat{X}^\star$, which corresponds to the linear map

$$W \to W + \gamma\Big(\hat{D}W - (\hat{X}^\star)^T X^\star\mathcal{U}\big(D(X^\star)^T\hat{X}^\star W + W^T(\hat{X}^\star)^T X^\star D\big) - WD\Big).$$

Now observe that $(\hat{X}^\star)^T X^\star$ is a $d \times m$ matrix that, since $\hat{X}^\star$ is an extension of $X^\star$ to an orthonormal basis, has the form

$$A := (\hat{X}^\star)^T X^\star = \begin{bmatrix} I_m \\ \mathbf{0}_{d-m,m} \end{bmatrix}.$$

Let W have a block partition compatible with that of A, *i.e.,*

$$W = \begin{bmatrix} \tilde{W} \\ \bar{W} \end{bmatrix},$$

where $\tilde{W}$ and $\bar{W}$ are, respectively, $m \times m$ and $(d-m) \times m$ matrices. The linear map $(\hat{X}^\star)^T\mathbf{J}_H(X^\star)\hat{X}^\star$ can thus be written in block form as

$$\begin{bmatrix} \tilde{W} \\ \bar{W} \end{bmatrix} \to \begin{bmatrix} \tilde{W} + \gamma\Big(D\tilde{W} - \tilde{W}D - \mathcal{U}\big(D\tilde{W} + \tilde{W}^T D\big)\Big) \\ \bar{D}\bar{W} - \bar{W}D \end{bmatrix},$$

where $\bar{D} = \text{diag}(\lambda_{P(m+1)}, \ldots, \lambda_{P(d)})$. We summarize these results in the following Lemma.

Lemma 6.3.4. *Let $X^\star$ be a fixed point of H with no zero columns. From Lemma 6.3.3, there exists a permutation P on the set $\{1, \ldots, d\}$ such that $CX^\star = X^\star D$ with $D = diag(\lambda_{P(1)}, \ldots, \lambda_{P(m)})$. Given P, let $\bar{D} = diag(\lambda_{P(m+1)}, \ldots, \lambda_{P(d)})$. Then, $\mathbf{J}_H(X^\star)$ is unitarily similar to*

$$\begin{bmatrix} \tilde{W} \\ \bar{W} \end{bmatrix} \to \begin{bmatrix} \tilde{W} + \gamma\Big(D\tilde{W} - \tilde{W}D - \mathcal{U}\big(D\tilde{W} + \tilde{W}^T D\big)\Big) \\ \bar{W} + \gamma\big(\bar{D}\bar{W} - \bar{W}D\big) \end{bmatrix}.$$

This lemma is enough to characterize the eigenvalues of $\mathbf{J}_H(X^\star)$, since eigenvalues are preserved by similarity. Let β be an eigenvalue of $\mathbf{J}_H(X^\star)$; then, there exist matrices $\tilde{W}$, and $\bar{W}$ not both equal to zero, such that

$$\begin{aligned} \tilde{W} + \gamma\Big(D\tilde{W} - \tilde{W}D - \mathcal{U}\big(D\tilde{W} + \tilde{W}^T D\big)\Big) &= \beta\tilde{W} \\ \bar{W} + \gamma\big(\bar{D}\bar{W} - \bar{W}D\big) &= \beta\bar{W}. \end{aligned} \tag{6.9}$$

Conversely, given a non-zero matrix that satisfies $\bar{W} + \gamma\big(\bar{D}\bar{W} - \bar{W}D\big) = \beta\bar{W}$, then β is an eigenvalue of $\mathbf{J}_H(X^\star)$ (take $\tilde{W} = 0$). Similarly, given a non-zero matrix that satisfies $W + \gamma\big(\bar{D}\bar{W} - \bar{W}D\big) = \beta\bar{W}$, then β is an eigenvalue of $\mathbf{J}_H(X^\star)$ (take $\bar{W} = 0$).

Lemma 6.3.5. *Let $X^\star$ be a fixed point of C other than the solution to the PCA problem. Then, for $\gamma > 0$, there exists an eigenvalue of $\mathbf{J}_H(X^\star)$ that is real and larger than one.*

Proof. From Lemma 6.3.3, we can just focus on $X^\star$ with non-zero columns. First, suppose that there exists $i < j$ such that $\lambda_i < \lambda_j$, where $Cx_i^\star = \lambda_i x_i^\star$ and $Cx_j^\star = \lambda_j x_j^\star$. Let $\tilde{W}$ be the $m \times m$ matrix such that $\tilde{W}_{ji} = 1$ and $\tilde{W}_{st} = 0$ if $(s,t) \neq (i,j)$. Then,

$$\begin{aligned} D\tilde{W} - \tilde{W}D - \mathcal{U}\big(D\tilde{W} + \tilde{W}^T D\big) &= (\lambda_j - \lambda_i)\tilde{W} - \lambda_j\tilde{W}^T \\ D\tilde{W}^T - \tilde{W}^T D - \mathcal{U}\big(D\tilde{W}^T + \tilde{W}D\big) &= -2\lambda_j\tilde{W}^T. \end{aligned}$$

Let $(a,b) \neq (0,0)$ be such that

$$\begin{bmatrix} \lambda_j - \lambda_i & 0 \\ -\lambda_j & -2\lambda_j \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = (\lambda_j - \lambda_i) \begin{bmatrix} a \\ b \end{bmatrix},$$

which must exist since the matrix is lower triangular. Taking $\hat{W} = a\tilde{W} + b\tilde{W}^T$, we must

have that

$$\begin{aligned} D\hat{W} - \hat{W} - \mathcal{U}\big(D\hat{W} + \hat{W}^T D\big) &= a(\lambda_j - \lambda_i)\tilde{W} - a\lambda_j \tilde{W}^T - 2\lambda_j b\tilde{W}^T \\ &= (\lambda_j - \lambda_i)\hat{W}. \end{aligned}$$

We conclude that, for $\gamma > 0$, the vector (it is in fact a matrix)

$$\begin{bmatrix} \hat{W} \\ \mathbf{0} \end{bmatrix}$$

is an eigenvector of (6.9) associated to the eigenvalue $1 + \gamma(\lambda_j - \lambda_i) > 1$. Moreover, by Lemma 6.3.4, $1 + \gamma(\lambda_i - \lambda_j) > 1$ is an eigenvalue of $\mathbf{J}_H(X^\star)$.

To finish the result, suppose now that, for $1 \leq t \leq m - d$, $Cx_i^\star = \lambda_{m+t} x_i^\star$ and, for this case, we look at the second equation in (6.9). This case is easier than the previous one since it implies that one of the top m eigenvalues, let's say λ_s with $1 \leq s \leq m$, appears in the diagonal of $\bar{D}$, let's say in position $1 \leq j \leq d - m$. Moreover $\lambda_s > \lambda_{m+t}$, and thus, let $\bar{W}$ be the $(d - m) \times m$ matrix such that $\bar{W}_{ji} = 1$ and all other entries equal to zero. Then

$$\bar{W} + \gamma\big(\bar{D}\bar{W} - \bar{W}D\big) = \big(1 + \gamma(\lambda_s - \lambda_{m+t})\big)\bar{W}.$$

Again, by Lemma 6.3.4, $1 + \gamma(\lambda_s - \lambda_{m+t}) > 1$ is an eigenvalue of $\mathbf{J}_H(X^\star)$. □

We now complete the characterization by looking at the eigenvalues of $\mathbf{J}_H(X^\star)$, when $X^\star$ is the solution to PCA. In this case, by Lemma 6.3.4, $\mathbf{J}_H(X^\star)$ is unitarily similar to

$$\begin{bmatrix} \tilde{W} \\ \bar{W} \end{bmatrix} \to \begin{bmatrix} \tilde{W} + \gamma\Big(D\tilde{W} - \tilde{W}D - \mathcal{U}\big(D\tilde{W} + \tilde{W}^T D\big)\Big) \\ \bar{W} + \gamma\big(\bar{D}\bar{W} - \bar{W}D\big) \end{bmatrix},$$

with $D = \operatorname{diag}(\lambda_1, \ldots, \lambda_d)$ and $\bar{D}$ can be taken to be $\bar{D} = \operatorname{diag}(\lambda_{m+1}, \ldots, \lambda_{m-d})$.

Lemma 6.3.6. *Let $X^\star$ be the solution to the PCA problem (there is exactly one solution up to a sign switch in each column). Then, there exists $\gamma^\star$ such that, for $0 < \gamma \leq \gamma^\star$, $\rho\big(\mathbf{J}_H(X^\star)\big) < 1$.*

Proof. Let Z be an eigenvector of (6.9) associated to an eigenvalue β. Consider a block partition of Z of the form

$$Z = \begin{bmatrix} \tilde{Z} \\ \bar{Z} \end{bmatrix},$$

where $\tilde{Z}$ and $\bar{Z}$ are, respectively, $m \times m$ and $(d-m)\times m$ matrices. There are two non-mutually-exclusive cases to consider: $\tilde{Z} \neq 0$ or $\bar{Z} \neq 0$ ($Z \neq 0$, by virtue of being an eigenvector).

Case 1 Suppose that $\bar{Z}_{st} \neq 0$. Then, the second eigenvalue equation in (6.9) implies that

$$\bar{Z}_{st} + \gamma\big(\lambda_{m+s}\bar{Z}_{st} - \lambda_t \bar{Z}_{st}\big) = \beta \bar{Z}_{st},$$

and, hence, $\beta < 1$, for sufficiently small γ.

Case 2 Suppose that $\tilde{Z}_{st} \neq 0$. This case splits in two: either $s > t$ or $s \leq t$. If $s > t$, then the first eigenvalue equation in (6.9) and the "upper triangularization" operation yields

$$\tilde{Z}_{st} + \gamma(\lambda_s \tilde{Z}_{st} - \lambda_t \tilde{Z}_{st}) = \beta \tilde{Z}_{st}, \tag{6.10}$$

which, after dividing by $\tilde{Z}_{st}$, yields, for γ sufficiently small, $\beta < 1$. If $s \leq t$, then,

$$\begin{aligned}
\beta \tilde{Z}_{st} &= \tilde{Z}_{st} + \gamma\Big(\lambda_s \tilde{Z}_{st} - \lambda_t \tilde{Z}_{st} - \mathcal{U}(D\tilde{Z} + \tilde{Z}^T D)_{st}\Big) \\
&= \tilde{Z}_{st} + \gamma\Big(\lambda_s \tilde{Z}_{st} - \lambda_t \tilde{Z}_{st} - \lambda_s \tilde{Z}_{st} - \lambda_t \tilde{Z}_{ts}\Big) \\
&= \tilde{Z}_{st} + \gamma\Big(- \lambda_t (\tilde{Z}_{st} + \tilde{Z}_{ts})\Big).
\end{aligned}$$

Next, notice that if $s < t$, then $\tilde{Z}_{ts}$ can be assumed to be 0, since, otherwise, we could deal with it as in (6.10) with the roles of s and t reversed to conclude $\beta < 1$. Hence, assuming $\tilde{Z}_{ts} = 0$, we obtain, after division by $\tilde{Z}_{st}$, that, for γ sufficiently small, $\beta < 1$. Finally, if $s = t$, then, again for γ sufficiently small, $\beta < 1$. □

To finish, suppose that the eigenvalues of C are all distinct, that is, $\lambda_1 > \ldots > \lambda_m > \lambda_{m+1} > \lambda_{m+2} > \ldots > \lambda_d > 0$, then, the eigenspace of each eigenvalue is one-dimensional and, thus, there are only two unit-norm eigenvectors that span it. Consequently, H has only a finite number of fixed points; in fact, the columns of a fixed point $X^\star$ are either one, or unit-norm eigenvectors of C and, hence, there are only finitely many choices for $X^\star$. There are 2^m stable fixed points and if $X^\star$ is one of the remaining fixed points, then $\mathbf{J}_H(X^\star)$ as a real eigenvalue larger than one. According to Remark 5.3.2 of Chapter 5, the map $\tilde{F}$ preserves this feature for α sufficiently small, that is, for all α sufficiently small, $\tilde{F}$ has 2^m stable fixed points corresponding to the stable fixed points of H and the remaining (a finite number of them) fixed points of $\tilde{F}$ are unstable.

6.4 Comments and References

The distributed PCA problem constitutes a relevant area of research – see, *e.g.*, [83, 84, 85, 86, 87, 88, 89, 90, 91] (master-slave communication architecture), and [92, 93, 94, 95, 96, 97, 98, 99] (arbitrarily meshed network communication architecture). For a recent and comprehensive review on these works, see, *e.g.*, [3]; for a very recent work see [100].

6.4.1 Simulations

We have not performed computer simulations of the distributed PCA algorithm that emerges from Sanger's algorithm, since, as previously noted, the EXTRA-distributed Banach-Picard iteration (5.19) with the maps H_n defined as in (6.2) reduces to the algorithm termed ADSA proposed in [21]. In [21], the authors show the results of the computer simulations of ADSA, hence we refer to that work for the comparison with other algorithms for distributed PCA.

6.4.2 Extensions of Our Previous Work

This chapter differs from our work [14] in two main aspects. First, in [14], we only addressed the stable case, that is, we only showed that $X^\star$, where $X^\star$ is the solution to the PCA problem, satisfies $\mathbf{J}_H(X^\star) < 1$ for sufficiently small γ. In this chapter, however, we have also addressed the unstable case and this constitutes a relevant extension of the results in [14]. Finally, in [14] the proofs relied on *matrix differential calculus* (see [101]) and in this chapter we avoided that route. In fact, given that the Sanger map defined in (6.1) is so "well-behaved", we didn't find a good enough reason not to treat $\mathbf{J}_H(X^\star)$ simply as the linear map that satisfies

$$\mathbf{J}_H(X^\star)Z = \lim_{t\to 0} \frac{H(X^\star + tZ) - H(X^\star)}{t}.$$

Of course, one could argue that this is "how the rules of matrix differential calculus are derived". However, we believe that avoiding these rules, allows for a more self-contained approach.

6.4.3 ADSA Almost Surely Escapes the Unstable Fixed Points

There is a piece missing in this picture. In the case $m = 1$, we proved that if Sanger's algorithm converges, then it "almost surely" converges, at least linearly, to a solution to the PCA problem. Even if the map $\tilde{F}$ from Chapter 5 preserves the stable and the unstable fixed points, the local diffeomorphism condition of Theorem 3.2.4 from Chapter

3 should be verified, if one is to conclude that the Banach-Picard iteration of $\tilde{F}$ "almost surely" escapes the unstable fixed points. The road to prove this, should, in principal, pass through an argument similar to that used for the case $m = 1$, *i.e.,* to prove that there exists a compact set K such that $\tilde{F}(K) \subseteq K$. Observe that

$$\mathbf{J}_{\tilde{F}}(z, \tilde{w}) = A(\eta, \beta) \begin{bmatrix} z \\ \tilde{w} \end{bmatrix} + \alpha \begin{bmatrix} \mathbf{J}_R(z) \\ \mathbf{0} \end{bmatrix}$$

is a polynomial in $z, \tilde{w}$, and α (we assume that η and β are chosen to lead to the EXTRA-distributed Banach-Picard iteration (5.19)). It is easy to show that $A(\eta, \beta)$ is invertible and, hence, its determinant is non-zero. Therefore, if $\tilde{F}(K) \subseteq K$ we have, in K, a uniform lower bound

$$\det\big(\mathbf{J}_{\tilde{F}}(z, \tilde{w})\big) \geq \det\big(A(\eta, \beta)\big) + p(\alpha),$$

where p is a polynomial that satisfies $p(0) = 0$. This would lead to $\det\big(\mathbf{J}_{\tilde{F}}(z, \tilde{w})\big) \neq 0$, for sufficiently small α, thus verifying, for $\tilde{F}$ seen as a map from K to K, the conditions of Theorem 3.2.4 in Chapter 3.

Chapter 7

Distributed Parameter Estimation with Noisy and Faulty Measurements

7.1 Introduction

In this chapter, as in the previous one, we consider an application of the results from Chapter 5, this time to distributed estimation. Informally, throughout this chapter, we consider a collection of spatially distributed sensors monitoring a possibly harsh environment. The sensors communicate wirelessly and the environment harsh conditions may result in faulty communications or sensor malfunctions. Ultimately, the goal is that of estimating a fixed and unknown parameter $\mu^\star$ of which each agent has, with probability p, a noisy linear measurement, and a faulty measurement with probability $1-p$.

A natural way to model this scenario is to describe it with a parametric mixture model depending on μ and to let the estimate of $\mu^\star$ be the "best" μ that "explains" the measurements, or, more formally, to be the *maximum likelihood estimate* (MLE). The standard approach for finding the MLE of a mixture model, assuming the measurements are at a single location, is the *expectation maximization* (EM) algorithm, from now onward referred to as centralized EM. In [12], the authors propose an algorithm (DA-DEM) for the distributed estimation of $\mu^\star$ that corresponds to an extension to distributed settings of the centralized EM algorithm; in the light of Chapter 4, the general idea behind DA-DEM can be described as follows: 1) the centralized EM algorithm for the underlying mixture model has the form $z^{k+1} = G(1/N \sum_n H_n(z^k))$, where the map G does not depend on the measurements and each map H_n depends on what agent n measured; 2) even though the centralized EM map, $z \to G(1/N \sum_n H_n(z))$, is not an average of local maps, the

map $w \to \sum_n H_n(G(w))$ is; 3) to obtain the MLE in a distributed fashion, consider the distributed algorithm from Chapter 4 with local maps $H_n \circ G$, and, at iteration k, agent n's estimate of the MLE is given by $G(w_n^k)$.

As observed several times, the downside of opting for a distributed algorithm such as the one suggested in Chapter 4 is the sacrifice in convergence rate. However, by accepting the memory-convergence rate trade-off, we can have an algorithm that converges at least at a linear rate, provided the sensors pay the cost of having to store at each iteration a $2d$-dimensional, rather than a d-dimensional vector. Crucially, to use the results of Chapter 5 to obtain guarantees of local linear convergence for the distributed algorithm, the corresponding property for the centralized counterpart has to be verified; this constitutes the goal of this chapter. In fact, we show the local linear convergence property, not for the centralized EM map, but for a slightly modified map that emerges from the fixed points equations satisfied by the MLE. The key challenge is that, like the EM map, the "modified" EM map depends on the agent's measurements which are, in turn, samples from a probability distribution and, hence, any statement regarding the map (*e.g.,* that it has a fixed point) is of probabilistic nature.

This chapter, mainly based on our work [14], is organized as follows: Section 2 reviews the basics of maximum likelihood estimation, mixture models, and the EM algorithm; Section 3 presents the mathematical description of the problem statement; Section 4 gives a high-level view of the rest of the chapter; Section 5 presents explicit expressions for the MLE and the modified EM map from where the distributed algorithm emerges; Section 6 provides the convergence analysis of the modified EM algorithm modulo some technicalities that can be found in [14]; Section 7, the final one, shows the result of Monte Carlo simulations comparing DA-DEM with our algorithm, which confirm the sub-linear convergence of DA-DEM and the linear convergence of our algorithm.

7.2 Preliminaries

7.2.1 MLE

Consider N real numbers $y_1, \dots, y_N$ known to have been independently sampled from a probability density function $f_Y(\cdot|\theta^\star)$, where $\theta^\star \in \mathbb{R}$ is termed the *ground truth* parameter. The goal is to estimate $\theta^\star$ and an estimator is a function $\theta(y_1, \dots, y_N)$ that provides, desirably, a "good estimate" thereof. The several formal ways that the phrasing "good estimate" can take are beyond scope the of this work; as an example, we mention *consistency.* The probability distribution over $y_1, \dots, y_N$ and the estimator $\theta(\cdot)$ induce a

natural probability distribution on $\mathbb{R}$: let $\mathcal{A} = \theta^{-1}((-\infty, a])$ and define

$$\mathbb{P}(-\infty < x \leq a) = \int_{(y_1,\ldots,y_N)\in\mathcal{A}} \prod_{n=1}^{N} f_Y(y_n|\theta^\star)dy;$$

that is, the probability that $\theta(y_1, \ldots, y_N)$ is less than a is that of sampling $y_1, \ldots, y_N$ from the set $\mathcal{A}$ that satisfies $\theta(\mathcal{A}) \leq a$. Consistency corresponds to the requirement of having increasingly more mass on small intervals centered at $\theta^\star$, as the number of samples goes to infinity; formally, for every $\epsilon > 0$, the sequence of real numbers $\mathbb{P}(\theta^\star - \epsilon \leq \theta(y_1, \ldots, y_N) \leq \theta^\star + \epsilon)$ should tend to one, as N tends to infinity. Equivalently, for every $\epsilon > 0$ and $\delta > 0$, there should exist N_0 for which the probability that $N \geq N_0$ independent samples $y_1, \ldots, y_N$ map, via the estimator function $\theta(\cdot)$, to the interval $[-\epsilon + \theta^\star, \theta^\star + \epsilon]$ is at least $1 - \delta$.

An estimator that has, for "sufficiently well behaved probability densities", many (*e.g.,* consistency) desired properties is the *maximum likelihood estimator* (see *e.g.* [102]), formally defined as

$$\theta(y_1, \ldots, y_N) = \arg\max_{\eta} \sum_{n=1}^{N} \log\big(f_Y(y_n, \eta)\big). \tag{7.1}$$

To finish, we mention that everything said so far is easily generalizable to d-dimensional spaces, *i.e.,* with $y_1, \ldots, y_N \in \mathbb{R}^d$.

7.2.2 Mixture Models

Mixture models are essentially probability density functions that can be written as a convex combination of (usually simpler) probability density functions. A common way to think of a mixture model is in terms of "missing class labels". Imagine that there are two probability density functions $f_Y(\cdot, \theta^\star)$ and $\hat{f}_Y(\cdot, \theta^\star)$ and that there is an "entity", call it E, that acts as follows: 1) flips a biased coin (probability $p^\star$ of heads and probability $1 - p^\star$ of tails); 2) if the result is heads, it samples y from $f_Y(\cdot, \theta^\star)$, otherwise, it samples y from $\hat{f}_Y(\cdot, \theta^\star)$; 3) hands us y, omitting whether it was sampled from $f_Y(\cdot, \theta^\star)$ or $\hat{f}_Y(\cdot, \theta^\star)$; 4) repeats the process N times. The idea is that E obtains $(y_1, a_1), \ldots, (y_N, a_N)$, where $a_i \in \{\text{heads, tails}\}$, but only reveals $y_1, \ldots, y_N$, hiding the class (heads or tails) from which y_n was sampled. Moreover, to add difficulty, we are also ignorant of $p^\star$, the probability of obtaining heads. Nevertheless, the goal is to estimate $\theta^\star$.

The formal way to describe the process above is as follows. E tosses the coin before

sampling y and, thus, this can be modeled by letting let $z \in \{0, 1\}$ and defining

$$\mathbb{P}(a = z) = (1 - p^\star)^z (p^\star)^{1-z},$$

where heads an tails are identified, respectively, with $a = 0$ and $a = 1$. What happens after the coin flip is naturally modeled via condition probability: define the density $f_{Y|a}(\cdot|\theta^\star, a) = f_Y^{1-a}(\cdot, \theta^\star)\hat{f}_Y^a(\cdot, \theta^\star)$ and note that E samples according to the probability density function

$$f_{Y,a}(\cdot, a = z|\theta^\star) = f_{Y|a}(\cdot|\theta^\star, z)\mathbb{P}(a = z) = f_Y^{1-z}(\cdot, \theta^\star)\hat{f}_Y^z(\cdot, \theta^\star)(1 - p^\star)^z (p^\star)^{1-z}.$$

We, however, only observe y, thus, to us, the samples come from a probability density function $f_Y(\cdot, \theta^\star)$ obtained by marginalizing a, *i.e.,*

$$f_Y(y, \theta^\star) = \sum_{z=0}^{1} f_{Y,a}(y, a = z|\theta^\star) = p^\star f_Y(y|\theta^\star) + (1 - p^\star)\hat{f}_Y(y|\theta^\star);$$

as anticipated, this is a convex combination of f_Y and $\hat{f}_Y$. Because we are also ignorant about $p^\star$, this can be treated as a parametric model in θ and p, which leads to the MLE:

$$\arg\max_{\theta,p} \sum_{n=1}^{N} \log\Big(p f_Y(y_n|\theta) + (1 - p)\hat{f}_Y(y_n|\theta)\Big). \tag{7.2}$$

7.2.3 The EM Algorithm

The EM algorithm is most useful when the MLE is easy to find given the class labels, that is, when E is "kind enough" to hand us $(y_1, a_1), \ldots, (y_N, a_N)$ instead of hiding the a_i's. In that scenario we face the problem

$$\begin{aligned}
&\arg\max_{\theta,p} \sum_{n=1}^{N} \log\Big(f_Y^{1-a_n}(y_n, \theta)\hat{f}_Y^{a_n}(y_n, \theta)(1 - p)^{a_n} p^{1-a_n}\Big) \\
&= \arg\max_{\theta,p} \Big(\sum_{n:a_n=1} \log(1 - p) + \log\big(\hat{f}_Y(y_n, \theta)\big)\Big) + \Big(\sum_{n:a_n=0} \log(p) + \log\big(f_Y(y_n, \theta)\big)\Big).
\end{aligned} \tag{7.3}$$

This has a nice feature: the estimation of $\theta^\star$ can be carried out independently of $p^\star$. In fact, this optimization problem does not "couple" p and θ, and, thus, it is separable, a property absent in (7.2). A mixture of Gaussian densities is an example of when (7.3) is "easy" to solve.

As previously mentioned, the EM algorithm is most useful when (7.2) is hard to solve, but (7.3) is easy to solve. The idea is: 1) start with a guess (θ^0, p^0); 2) from

$\theta^0, p^0, y_1, \ldots, y_N$, estimate $\mathbb{P}(a_n = 1|\theta^0, p^0, y_1, \ldots, y_N)$; 3) let (θ^1, p^1) be the solution of (7.3) with the objective function replaced by its "average" with respect to $\mathbb{P}(a_1^0 = 1), \ldots, \mathbb{P}(a_N^0 = 1)$; 4) repeat the process with (θ^0, p^0) replaced by (θ^1, p^1). The probabilities $\mathbb{P}(a_n = 1|\theta^0, p^0, y_1, \ldots, y_N)$ are easily obtained by noting that, by Bayes law,

$$\begin{aligned}\mathbb{P}(a_n = 1|\theta^0, p^0, y_1, \ldots, y_N) &= f_{Y,a}(y_n, a_n = 1|\theta^0, p^0)\big(f_Y(y_n|\theta^0, p^0)\big)^{-1} \\ &= \frac{(1-p^0)\hat{f}_Y(y_n|\theta^0)}{p^0 f_Y(y_n|\theta^0) + (1-p^0)\hat{f}_Y(y_n|\theta^0)}.\end{aligned}$$

The EM algorithm is summarized as: given (θ^t, p^t)

1) Compute the probabilities for the class labels for $n = 1, \ldots, N$ according to

$$\mathbb{P}(a_n = 1|\theta^t, p^t, y_n) = \frac{(1-p^t)\hat{f}_Y(y_n|\theta^t)}{p^t f_Y(y_n|\theta^t) + (1-p^t)\hat{f}_Y(y_n|\theta^t)} := w_n^t; \tag{7.4}$$

2) Compute (θ^{t+1}, p^{t+1}) by solving (7.3) with the objective function replaced by its average with respect to the class label's probabilities, *i.e.*,

$$\begin{aligned}(\theta^{t+1}, p^{t+1}) &= \arg\max_{\theta,p} \sum_{n=1}^{N}\sum_{z=0}^{1} \mathbb{P}(a_n = z|\theta^t, p^t, y_n) \log\Big(f_Y^{1-z}(y_n,\theta)\hat{f}_Y^z(y_n,\theta)(1-p)^z p^{1-z}\Big) \\ &= \arg\max_{\theta,p} \Big[\sum_{n=1}^{N}\Big(\big(1-w_n^t\big)\big(\log(p) \\ &\quad + \log(f_Y(y_n|\theta)\big) + w_n^t\big(\log(1-p) + \log(\hat{f}_Y(y_n|\theta))\big)\Big)\Big].\end{aligned}$$

Observe that the optimization problem of step 2) is separable in p and θ.

7.3 Problem Statement

Consider a network of N agents, where the interconnection structure is represented by an undirected and connected graph. Each agent n holds an observation y_n sampled according to

$$y_n = \begin{cases} h_n^T\mu^\star + w_n, & \text{with probability } p^\star, \\ w_n, & \text{with probability } 1-p^\star, \end{cases}$$

where: $\mu^\star \in \mathbb{R}^d$ is a fixed and unknown parameter; each $h_n \in \mathbb{R}^d$ is assumed to be known only at agent n; $\{w_n\}_{n=1}^N$ are samples of independent and identically distributed

(i.i.d.) zero-mean Gaussian random variables with variance $(\sigma^\star)^2$. The agents seek to collectively estimate $\mu^\star$, treating $p^\star$ and $(\sigma^\star)^2$, which are also fixed and unknown, as nuisance parameters.

7.3.1 Mixture Model Formulation

Agent n observes y_n, with y_n sampled according to the mixture density

$$f_{Y_n}\big(y|p^\star,\mu^\star,(\sigma^\star)^2\big) = p^\star\mathcal{N}\big(y|h_n^T\mu^\star,(\sigma^\star)^2\big) + (1-p^\star)\mathcal{N}\big(y|0,(\sigma^\star)^2\big) \tag{7.5}$$

and, thus, the random variables Y_n and Y_m associated, respectively, to distinct agents n and m, are not identically distributed. For reasons that will be clear later, it is important to have identically distributed samples, and, to this end, we assume that the value h_n that agent n holds is also a sample of a random variable. Let $Z \in \{0,1\}$, $H \in \mathbb{R}^d$, and $Y \in \mathbb{R}$ be, respectively a binary random variable, a random vector, and a real random variable. Assuming Z and H are independent, the joint density on (Y,H,Z) factors as

$$f_{Y,H,Z}(y,h,z|\theta^\star) = f_H(h)f_Z(z|p^\star)f_{Y|H,Z}\big(y|h,z,\mu^\star,(\sigma^\star)^2\big), \tag{7.6}$$

where $\theta^\star = (\mu^\star,p^\star,(\sigma^\star)^2) \in \Omega = \mathbb{R}^d \times (0,1) \times (0,+\infty)$ is a fixed and unknown vector, which we term the ground truth. Finally, let

$$\begin{aligned} f_H(h) &= \mathcal{N}(h|0,I_d) \\ f_Z(z|p^\star) &= (p^\star)^z(1-p^\star)^{1-z} \\ f_{Y|H,Z}\big(y|h,z,\mu^\star,(\sigma^\star)^2\big) &= \mathcal{N}\big(y|h^T\mu^\star,(\sigma^\star)^2\big)^z\mathcal{N}\big(y|0,(\sigma^\star)^2\big)^{1-z}. \end{aligned}$$

Instead of assuming that agent n has a measurement y_n, we assume that it has a measurement (y_n,h_n), where (y_n,h_n,z_n) was sampled according to (7.6), but agent n does not observe z_n (compare this with the "entity" that hides the result of the coin flip, or, in this case, hides the knowledge about the measurement being faulty or not). Since z_n is not observed, we consider the joint density of (Y,H), which can be computed from (7.6) by marginalization,

$$f_{Y,H}(y,h|\theta^\star) = f_H(h)\Big(p^\star\mathcal{N}\big(y|h_n^T\mu^\star,(\sigma^\star)^2\big) + (1-p^\star)\mathcal{N}\big(y|0,(\sigma^\star)^2\big)\Big).$$

Note that $f_{Y,H}$ is itself a mixture model, equal to the density in (7.5) multiplied by $f_H(h)$. Moreover, $f_H(\cdot)$ does not depend on $\theta^\star$.

7.3.2 Problem Statement in Terms of the MLE

The agents' goal is formulated as that of finding a stationary point of the log-likelihood of (Y, H), *i.e.,* to estimate $\mu^\star$, the agents seek $\theta \in \Omega$ such that

$$\frac{1}{N}\sum_n \nabla_\theta \log\big(f_{Y,H}(y_n, h_n|\theta)\big) = 0.$$

Since $f_H(h)$ does not depend on θ, the agents seek θ such that

$$\frac{1}{N}\sum_{n=1}^N \nabla_\theta \log\big(f_{Y|H}(y_n|h_n, \theta)\big) = 0. \tag{7.7}$$

Let f_{Y_n} be defined as in (7.5) and observe that (7.7) is nothing but

$$\frac{1}{N}\sum_{n=1}^N \nabla_\theta \log(f_{Y_n}(y_n|\theta^\star)) = 0,$$

indicating that the introduction of $f_H(h)$ does not affect our considerations. The convenience of introducing $f_H(h)$, as it will shortly be seen, is that many instrumental results assume i.i.d. observations. To finish, note as well that the MLE defined in (7.1) satisfies (7.7) but the converse does not necessarily hold, *i.e,* a point $\theta \in \Omega$ that satisfies (7.7) is not necessarily a MLE, since it may correspond to a minimum or a saddle point.

7.4 Roadmap

This is a rather long chapter, hence the need for a road map, itself rather long. To apply the results of Chapter 5, we first need to rewrite the solution of (7.7) as a fixed point of a map H that can be written as an average of local maps. A first naive attempt is to let $\gamma \neq 0$ and observe that θ satisfies (7.7) if and only if

$$\theta = \theta + \gamma\big(\frac{1}{N}\sum_{n=1}^N \nabla_\theta \log\big(f_{Y,H}(y_n, h_n|\theta)\big).$$

We could then define

$$H_n = \theta + \gamma\nabla_\theta \log\big(f_{Y,H}(y_n, h_n|\theta)\big),$$

which is a map known at agent n, and check whether there exists $\gamma^\star$ such that, for $0 < \gamma \leq \gamma^\star$, $\rho\big(\mathbf{J}_H(\theta)\big) < 1$, with $H = 1/N\sum_n H_n$. This approach presents some difficulties: the map $G(\theta) = 1/N\sum_{n=1}^N \nabla_\theta \log\big(f_{Y,H}(y_n, h_n|\theta)\big)$ is an average of gradients which implies

that $\mathbf{J}_G(\theta)$ is an Hessian and, as a result, it is a symmetric matrix. Consequently, there exists $\gamma^\star$ such that, for $0 < \gamma \leq \gamma^\star$, $\rho\big(\mathbf{J}_H(\theta)\big) < 1$ if and only if $\mathbf{J}_G(\theta)$ is negative definite. Now, the map $G(\theta)$ depends on observations coming from a probability distribution; thus, the statement "$\mathbf{J}_G(\theta)$ is negative definite" is a probabilistic one. Even if we proved that, for N sufficiently large, the probability that $\mathbf{J}_G(\theta)$ is negative definite approaches one, we would still need to prove that its smallest eigenvalue is lower bounded by some $\beta < 0$ with probability tending to 1. In fact, this would be instrumental to define $\gamma^\star$. Summarizing, we would need to prove a statement stronger than "$\mathbf{J}_G(\theta)$ is negative definite, with probability tending to 1".

The approach we follow is to write (7.7) explicitly and to try to "put" all the terms that can be written as an average of local maps on one side and on the other side "put a map in θ", *i.e.,* to rearrange (7.7) in a form like $g(\theta, y_1, h_1, \ldots, y_N, h_N) = 1/N \sum_n f(\theta, y_n, h_n)$, for maps g and f. Crucially, this should not include a tunable parameter γ. Of course, for this to work, g cannot be just any map and what we will show is that it has the form $1/N \sum_{n=1}^N A(y_n, h_n, \theta)\theta$, where each $A(y_n, h_n, \theta)$ is a matrix.

Suppose for a moment that $1/N \sum_{n=1}^N A(y_n, h_n, \theta)$ is invertible with probability tending to one as N goes to infinity. In this case we obtain (7.7) as a fixed point equation

$$\theta = \Big(\frac{1}{N}\sum_{n=1}^{N} A(y_n, h_n, \theta)\Big)^{-1}\Big(\frac{1}{N}\sum_{n=1}^{N} f(\theta, y_n, h_n)\Big). \tag{7.8}$$

Given that the matrix inversion destroys the property of being an average of local maps, (7.8) is not in the form required for the distributed extension described in Chapter 5. Nevertheless, (7.8) can be broken into two steps,

$$\begin{aligned}\theta &\mapsto \Big(\frac{1}{N}\sum_{n=1}^{N} A(y_n, h_n, \theta), \frac{1}{N}\sum_{n=1}^{N} f(\theta, y_n, h_n)\Big)\\ &\mapsto \Big(\frac{1}{N}\sum_{n=1}^{N} A(y_n, h_n, \theta)\Big)^{-1}\Big(\frac{1}{N}\sum_{n=1}^{N} f(\theta, y_n, h_n)\Big),\end{aligned}$$

where the second map, $(\Gamma, v) \mapsto \Gamma^{-1}v$, does not depend on the observations. More importantly, the first map,

$$\theta \mapsto \Big(\frac{1}{N}\sum_{n=1}^{N} A(y_n, h_n, \theta), \frac{1}{N}\sum_{n=1}^{N} f(\theta, y_n, h_n)\Big)$$

is an average of local maps.

To proceed, let

$$\theta^{k+1} = \big(\frac{1}{N}\sum_n A(y_n, h_n, \theta^k)\big)^{-1}\big(\frac{1}{N}\sum_{n=1}^N f(\theta^k, y_n, h_n)\big)$$

and, with a slight abuse of notation, let

$$\psi^{k+1/2} = \big(\frac{1}{N}\sum_n A(y_n, h_n, \theta^k), \frac{1}{N}\sum_{n=1}^N f(\theta^k, y_n, h_n)\big).$$

Note that $\theta^0 \to \theta^1 \to \theta^2 \ldots \to \theta^k \to \theta^{k+1} \to \ldots$ can be seen as $\theta^0 \to \psi^{1/2} \to \theta^1 \to \psi^{1+1/2} \to \theta^2 \to \ldots \theta^k \to \psi^{k+1/2} \to \theta^{k+1} \to \ldots$, where $\psi^{k+1/2}$ is obtained from θ^{k+1} by applying the map $(\Gamma, v) \mapsto \Gamma^{-1}v$, which, as previously observed, does not depend on the observations. In essence, we can focus on the sequence $\psi^{k+1/2}$ with initialization

$$\psi^{1/2} = \big(\frac{1}{N}\sum_n A(y_n, h_n, \theta^0), \frac{1}{N}\sum_{n=1}^N f(\theta^0, y_n, h_n)\big).$$

What is relevant is that the sequence $\psi^{k+1/2}$ is produced by the Banach-Picard iteration of a map that can be written as an average of local maps.

To summarize: θ satisfies (7.7) if and only if θ is a fixed point of a map

$$\tilde{H}(\theta) = Q\big(\frac{1}{N}\sum_n T_n(y_n, h_n, \theta)\big),$$

where Q does not depend on the observations (Q is the map $(\Gamma, v) \mapsto \Gamma^{-1}v$). To arrive at a distributed algorithm, instead of the Banach-Picard iteration of $\tilde{H}$, consider the Banach-Picard iteration of $H(\psi) = 1/N\sum_n T_n(y_n, h_n, Q(\psi))$, because H is an average of local maps.

7.4.1 Why this Makes Sense

This section explains why switching the roles of two maps, *i.e.,* why considering H rather than $\tilde{H}$, is not a "big deal". Let $g_1 : \mathbb{R}^d \to \mathbb{R}^q, g_2 : \mathbb{R}^q \to \mathbb{R}^d$ be two maps and suppose that $x^\star \in \mathbb{R}^q$ is a fixed point of $g_1 \circ g_2$. Clearly, $g_2(x^\star)$ is a fixed point of $g_2 \circ g_1$, and conversely. Suppose further that $x^\star$ is a fixed point of $g_1 \circ g_2$ such that

$$\rho\big(\mathbf{J}_{g_1 \circ g_2}(x^\star)\big) < 1. \tag{7.9}$$

The chain rule of differentiation implies that

$$\mathbf{J}_{g_1 \circ g_2}(x^\star) = \mathbf{J}_{g_1}\big(g_2(x^\star)\big)\mathbf{J}_{g_2}(x^\star). \tag{7.10}$$

We want to conclude that if $x^\star$ is a fixed point satisfying (7.9), then $g_2(x^\star)$ is a fixed point of $g_2 \circ g_1$ such that $\rho\big(\mathbf{J}_{g_2\circ g_1}\big(g_2(x^\star)\big)\big) < 1$; this reduces to proving that $\rho(AB) = \rho(BA)$. In fact, again by the chain rule and the fact that $x^\star$ is assumed to be a fixed point of $g_1 \circ g_2$, we obtain

$$\mathbf{J}_{g_2\circ g_1}\big(g_2(x^\star)\big) = \mathbf{J}_{g_2}\big(g_1(g_2(x^\star))\big)\mathbf{J}_{g_1}\big(g_2(x^\star)\big) = \mathbf{J}_{g_2}(x^\star)\mathbf{J}_{g_1}\big(g_2(x^\star)\big),$$

which is (7.10) with the matrix product reversed.

To see that $\rho(AB) = \rho(BA)$, let A be an $n \times m$ matrix, B be an $m \times n$ matrix and suppose that $0 \neq v \in \mathbb{R}^n$ satisfies $ABv = \lambda v$ with $\lambda \neq 0$. Then, $\mathbb{R}^m \ni Bv \neq 0$, since, otherwise, we could not have $ABv = \lambda v$ with both λ and v non-zero. Moreover, $BA(Bv) = B(ABv) = B(\lambda v) = \lambda Bv$. From $Bv \neq 0$, we conclude that Bv is an eigenvector of BA associated to the eigenvalue λ. The converse follows from reversing the roles of A and B. We conclude that the non-zero eigenvalues of AB are the non-zero eigenvalues of BA and this is enough to conclude that $\rho(AB) = \rho(BA)$.

7.4.2 Probability of Having a Fixed Point

From the previous section, we can focus on the stable fixed points of

$$\tilde{H}_N(\theta) = \Big(\frac{1}{N}\sum_{n=1}^{N} A(y_n, h_n, \theta)\Big)^{-1}\Big(\frac{1}{N}\sum_{n=1}^{N} f(\theta, y_n, h_n)\Big),$$

rather than those of $H(\psi)$, the map from which the distributed algorithm emerges. The inclusion of a subscript in $\tilde{H}_N$ is to indicate that it depends on samples (not to be confused with the local map belonging to agent N). Given that $\tilde{H}_N$ is "built" after observing $y_1, h_1, \ldots, y_N, h_N$, it is clear that the statement *"the map $\tilde{H}_N$ has a fixed point"* is a probabilistic one. In fact, let $\mathcal{A}_N = \big\{(y_1, h_1, \ldots, y_N, h_N) | \exists \theta \text{ s.t. } H_N(\theta) = \theta\big\}$ and, provided we can integrate over $\mathcal{A}_N$, the map $\tilde{H}_N$ has a fixed point with probability

$$\int_{\mathcal{A}_N} \prod_{n=1}^{N} f_H(h_n)\big(p^\star \mathcal{N}\big(y_n | h_n^T \mu^\star, (\sigma^\star)^2\big) + (1-p^\star)\mathcal{N}\big(y_n | 0, (\sigma^\star)^2\big) dy_1 dh_1 \ldots dy_N dh_N. \tag{7.11}$$

The set $\mathcal{A}_N$, being defined by an existence condition, is not easy to characterize, and we are ultimately interested in proving that, as $N \to \infty$, (7.11) tends to one. To circumvent this issue, we use *Brouwer's fixed point theorem* (see Chapter 3): consider the set

$$\mathcal{B}_N = \big\{(y_1, h_1, \ldots, y_N, h_N) \in \mathbb{R}^N | \tilde{H}_N\big(\bar{B}(\theta^\star, \delta)\big) \subseteq \bar{B}(\theta^\star, \delta)\big\},$$

where $\bar{B}(\theta^\star, \delta)$ is the closed ball with respect to a certain norm, centered at the ground truth $\theta^\star$, and with radius δ. Brouwer's fixed point theorem can be restated as $\mathcal{B}_N \subseteq \mathcal{A}_N$, because $\bar{B}(\theta^\star, \delta)$ is a closed, bounded, and convex set; what we will show is that

$$\lim_{n\to\infty} \mathbb{P}\big((y_1, h_1 \ldots, y_N, h_N) \in \mathcal{B}_N\big) = 1.$$

7.4.3 The Missing Piece

So far, all of this may look very complicated and, in this section, we reveal the missing piece. Let θ be fixed for now and observe that, if the probability density over $y_1, h_1, \ldots, y_N, h_N$ is "sufficiently well behaved", then, by the *weak law of large numbers*, the random vector $Z_N = \tilde{H}_N(\theta)$ converges, in probability, to the expected value, *i.e.,* to

$$\Big(\int A(y, h, \theta) f_{Y,H}(y, h|\theta^\star) dydh\Big)^{-1} \int f(\theta, y, h) f_{Y,H}(y, h|\theta^\star) dydh.$$

This suggests that it is natural to look at the map

$$\tilde{H}(\theta) := \Big(\int A(y, h, \theta) f_{Y,H}(y, h|\theta^\star) dydh\Big)^{-1} \int f(\theta, y, h) f_{Y,H}(y, h|\theta^\star) dydh, \tag{7.12}$$

which has the following relevant properties: $\tilde{H}(\theta^\star) = \theta^\star$ and, under mild conditions, $\rho\big(\mathbf{J}_{\tilde{H}}(\theta^\star)\big) < 1$. Given that $\tilde{H}$ is a type of "infinite sample" version of $\tilde{H}_N$ and that, for each θ, $\tilde{H}_N(\theta)$ converges in probability to $\tilde{H}(\theta)$, it is tempting to infer the probabilistic properties of $\tilde{H}_N$ from the non-probabilistic properties of $\tilde{H}$; this is the path we take. It turns out that "pointwise convergence in probability" is not enough, and, hence, a stronger version of the weak law of large numbers will be required, something like $\tilde{H}_N$ "converges uniformly in probability" to $\tilde{H}$.

7.4.4 Summary

The key points are summarized as:

1) θ satisfies (7.7) if and only if θ satisfies (7.8), which can be written as a fixed point

of $\tilde{H}_N(\theta) = Q\big(1/N \sum_n T_n(y_n, h_n, \theta)\big)$;

2) To arrive at a distributed algorithm, switch the roles of Q and the average map, that is, consider $H_N(\psi) = 1/N \sum_n T_n(y_n, h_n, Q(\psi))$;

3) The existence of a stable fixed point of H_N follows from the existence of a stable fixed point of $\tilde{H}_N$;

4) $\theta^\star$ is, under mild conditions, a stable fixed point of the infinite sample version of $\tilde{H}_N$, *i.e.,* (7.12);

5) The probabilistic properties of $\tilde{H}_N$ follow from "uniform convergence in probability" of $\tilde{H}_N$ to $\tilde{H}$. For example, for N sufficiently large, the probability that a closed ball centered at $\theta^\star$ is invariant under $\tilde{H}_N$ is close to one. Consequently, via Brouwer's fixed point theorem, the probability that $\tilde{H}_N$ has a fixed point is close to one, for N sufficiently large.

7.5 Gradient of the Log-Likelihood

The stability equations (7.7) can be explicitly written by differentiating with respect to μ, p, and σ^2. Let $\phi(y, h, \theta) = \log\big(f_{Y|H}(y|h, \theta)\big) = \log\big(p\mathcal{N}(y|h^T\mu, \sigma^2) + (1-p)\mathcal{N}(y|0, \sigma^2)\big)$; then,

$$\begin{aligned}
\nabla_\mu \phi(y, h, \theta) &= \frac{1}{\sigma^2} \frac{p\mathcal{N}(y|h^T\mu, \sigma^2)}{p\mathcal{N}(y|h^T\mu, \sigma^2) + (1-p)\mathcal{N}(y|0, \sigma^2)} (y - h^T\mu) h, \\
\nabla_p \phi(y, h, \theta) &= \frac{\mathcal{N}(y|h^T\mu, \sigma^2)}{p\mathcal{N}(y|h^T\mu, \sigma^2) + (1-p)\mathcal{N}(y|0, \sigma^2)} - \frac{\mathcal{N}(y|0, \sigma^2)}{p\mathcal{N}(y|h^T\mu, \sigma^2) + (1-p)\mathcal{N}(y|0, \sigma^2)}, \\
\nabla_{\sigma^2} \phi(y, h, \theta) &= -\frac{1}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \Big(\frac{p\mathcal{N}(y|h^T\mu, \sigma^2)}{p\mathcal{N}(y|h^T\mu, \sigma^2) + (1-p)\mathcal{N}(y|0, \sigma^2)} (y - h^T\mu)^2 \\
&\quad + \frac{(1-p)\mathcal{N}(y|0, \sigma^2)}{p\mathcal{N}(y|h^T\mu, \sigma^2) + (1-p)\mathcal{N}(y|0, \sigma^2)} y^2 \Big).
\end{aligned}$$

Observe that the function

$$r(y, h, \theta) = \frac{p\mathcal{N}(y|h^T\mu, \sigma^2)}{p\mathcal{N}(y|h^T\mu, \sigma^2) + (1-p)\mathcal{N}(y|0, \sigma^2)} \tag{7.13}$$

is a "building block" that appears in all partial derivatives. Furthermore, note the similarity between $r(y, h, \theta)$ and the probabilities for the missing class labels computed in the first step of the EM algorithm, *i.e.,* (7.4) ($r(y, h, \theta)$ is the probability that the measurement was not faulty, conditioned on y, h, θ); in the EM literature (see [103]), these

functions are often called the *responsibility functions.* Writing the partial derivatives in terms of the responsibility functions yields

$$\nabla_\mu \phi(y,h,\theta) = \frac{1}{\sigma^2} r(y,h,\theta)(y - h^T\mu)h, \tag{7.14}$$

$$\nabla_p \phi(y,h,\theta) = \frac{1}{p} r(y,h,\theta) - \frac{1}{1-p}(1 - r(y,h,\theta)), \tag{7.15}$$

$$\nabla_{\sigma^2} \phi(y,h,\theta) = -\frac{1}{\sigma^2} + \frac{1}{2(\sigma^2)^2}\Big(r(y,h,\theta)(y - h^T\mu)^2 + (1 - r(y,h,\theta)y^2\Big). \tag{7.16}$$

7.5.1 Modified EM

This section presents the centralized Banach-Picard iteration that will be used to coordinate the agents towards $\mu^\star$. The reason to call it a "modified EM" is explained in Section 7.7.1.

Since we are ultimately interested in obtaining a fixed point equation, it is natural to try to isolate μ, p, and σ^2. To this end, note that θ satisfies $\nabla\phi = 0$ if an only if

$$\begin{aligned} r(y,h,\theta)hh^T\mu &= r(y,h,\theta)yh \\ (1-p)r(y,h,\theta) - p(1 - r(y,h,\theta)) &= 0 \\ \sigma^2 &= r(y,h,\theta)(y - h^T\mu)^2 + (1 - r(y,h,\theta))y^2. \end{aligned}$$

The second equation simplifies to $r(y,h,\theta) = p$, and, hence, (7.7) can be rewritten as

$$\begin{aligned} \Big(\frac{1}{N}\sum_{n=1}^N r(y_n,h_n,\theta)h_n h_n^T\Big)\mu &= \frac{1}{N}\sum_{n=1}^N r(y_n,h_n,\theta)y_n h_n \\ p &= \frac{1}{N}\sum_{n=1}^N r(y_n,h_n,\theta) \\ \sigma^2 &= \frac{1}{N}\sum_{n=1}^N r(y_n,h_n,\theta)(y_n - h_n^T\mu)^2 + (1 - r(y_n,h_n,\theta))y_n^2. \end{aligned}$$

This is "almost" a fixed point equation; in fact, this leads to a fixed point equation, provided the matrix $1/N\sum_n r(y_n,h_n,\theta)h_n h_n^T$ can be inverted. Since each h_n is independently sampled from a Gaussian with zero mean and variance I, the invertibility holds with probability one for N sufficiently large (at least larger than d) [104]. As a

consequence, we consider

$$\mu^{k+1} = \Big(\frac{1}{N}\sum_{n=1}^{N}\Gamma(y_n, h_n, \theta^k)\Big)^{-1}\frac{1}{N}\sum_{n=1}^{N}\psi(y_n, h_n, \theta^k) \tag{7.17}$$

$$p^{k+1} = \frac{1}{N}\sum_{n=1}^{N} r(y_n, h_n, \theta^k) \tag{7.18}$$

$$(\sigma^2)^{k+1} = \frac{1}{N}\sum_{n=1}^{N}\gamma(y_n, h_n, \theta^k), \tag{7.19}$$

where

$$\begin{aligned}
\Gamma(y, h, \theta) &= r(y, h, \theta)hh^T \\
\psi(y, h, \theta) &= r(y, h, \theta)yh \\
\gamma(y, h, \theta) &= r(y, h, \theta)(y - h^T\mu)^2 + (1 - r(y, h, \theta))y^2.
\end{aligned}$$

Comparing this with (7.8) and, from the observations made therein, the map underlying (7.17)-(7.19) is of the form $\tilde{H}(\theta) = Q(1/N\sum_n T_n(y_n, h_n, \theta))$. Consequently, by interchanging the roles of Q and $1/N\sum_n T_n$, we obtain a map that is an average of local maps. Moreover, it is enough to focus on the properties of (7.17)-(7.19).

7.6 Convergence Analysis

This section provides the convergence analysis, that is, it addresses the existence of a stable fixed point of (7.17)-(7.19), specifically, the probability that the map T_N underlying (7.17)-(7.19) has a fixed point θ_N satisfying $\rho\big(\mathbf{J}_{T_N}(\theta_N)\big) < 1$. The convergence proof is merely a skeleton proof with several unenlightening technicalities (which can be found in [14]) omitted.

7.6.1 Infinite Sample Map

Let T_N denote[1] the map underlying the Banach-Picard iteration (7.17)-(7.19). Straightforward manipulation, using (7.14)-(7.16), shows that

$$T_N(\theta) = \theta + \big(A_N(\theta)\big)^{-1}\frac{1}{N}\sum_{n=1}^{N}\nabla_\theta\phi(y_n, h_n, \theta), \tag{7.20}$$

[1]The subscript N emphasizes that T_N depends on N observations, not that T_N is the map of agent N.

where

$$A_N(\theta) = \begin{bmatrix} \frac{1}{N}\sum_{n=1}^{N} \frac{1}{\sigma^2}\Gamma(y_n, h_n, \theta) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \frac{1}{p(1-p)} & 0 \\ \mathbf{0} & 0 & \frac{1}{2(\sigma^2)^2} \end{bmatrix}.$$

As mentioned in Section 7.4.3, from the weak law of large numbers, it is natural to look at the "infinite sample" version of (7.20), *i.e.,* to consider the map T that corresponds to replacing the finite averages by expected values. This infinite sample map is given by

$$T(\theta) = \theta + \big(A(\theta)\big)^{-1}\mathcal{L}(\theta), \tag{7.21}$$

where

$$A(\theta) = \begin{bmatrix} \mathbb{E}_{\theta^\star}\left[\frac{1}{\sigma^2}\Gamma(y, h, \theta)\right] & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \frac{1}{p(1-p)} & 0 \\ \mathbf{0} & 0 & \frac{1}{2(\sigma^2)^2} \end{bmatrix},$$

and

$$\mathcal{L}(\theta) = \mathbb{E}_{\theta^\star}\big[\nabla_\theta \phi(y, h, \theta)\big].$$

A straightforward verification reveals that $\mathcal{L}(\theta^\star) = 0$ and, hence, $T(\theta^\star) = \theta^\star$. It is, thus, natural to search for a condition that guarantees that $\rho\big(\mathbf{J}_T(\theta^\star)\big) < 1$; this is the goal of the next section.

7.6.2 Assumption on the Model

In order to consider $\rho\big(\mathbf{J}_T(\theta^\star)\big) < 1$, we first need to check whether T is differentiable at $\theta^\star$. Suppose for a moment that both $A(\theta)$ and $\mathcal{L}(\theta)$ are differentiable at $\theta^\star$. Then, T is also differentiable at $\theta^\star$ and, since $\mathcal{L}(\theta^\star) = 0$,

$$\mathbf{J}_T(\theta^\star) = I + A(\theta^\star)^{-1}\mathbf{J}_{\mathcal{L}}(\theta^\star).$$

A sufficient condition to have differentiablity of $A(\theta)$ and $\mathcal{L}(\theta)$ is to be able to differentiate under the integral sign. Conditions that ensure this are addressed in detail in our article [14] and we will refer to them as *regularity conditions.* Such regularity conditions are also enough to guarantee that

$$-\mathbf{J}_{\mathcal{L}}(\theta^\star) = -\mathbb{E}_{\theta^\star}\big[\nabla^2\phi(y, h, \theta^\star)\big] = -\mathbb{E}_{\theta^\star}\Big[\big(\nabla_\theta\phi(y, h, \theta^\star)\big)\big(\nabla_\theta\phi(y, h, \theta^\star)\big)^T\Big] =: I(\theta^\star),$$

where $\nabla^2\phi(y,h,\theta^\star) = \mathbf{J}_{\nabla_\theta\phi}(y,h,\theta^\star)$. The matrix $I(\theta^\star)$ is called, in the literature (see *e.g.* [105]), the *Fisher Information*. Suppose that $I(\theta^\star)$ is singular and let v be a non-zero vector such that $I(\theta^\star)v = 0$; then,

$$\mathbf{J}_T(\theta^\star)v = v,$$

that is, v is an eigenvector associated to the unit eigenvalue. This shows that a necessary condition to have $\rho\big(\mathbf{J}_T(\theta^\star)\big) < 1$ is non-singularity of $I(\theta^\star)$.

Assumption S: The ground truth $\theta^\star$ satisfies $\det\big(I(\theta^\star)\big) \neq 0$.

It turns out that Assumption S is not only a necessary condition for $\rho\big(\mathbf{J}_T(\theta^\star)\big) < 1$, but it is also sufficient. To see this, consider the Fisher information of the *complete data model*, that is,

$$I_c(\theta^\star) = \mathbb{E}_{\theta^\star}\Big[\big(\nabla_\theta\phi_c(y,h,z,\theta^\star)\big)\big(\nabla_\theta\phi_c(y,h,z,\theta^\star)\big)^T\Big],$$

where $\phi_c(y,h,z,\theta) = \log\big(f_{Y,H,Z}(y,h,z|\theta)\big)$; this is called the complete data model because it is the Fisher information for the density with no missing class labels. A simple calculation reveals that $I_c(\theta^\star) = A(\theta^\star)$, which leads to

$$\mathbf{J}_T(\theta^\star) = I - \big(I_c(\theta^\star)\big)^{-1} I(\theta^\star).$$

The following result, which implicitly uses Assumption S, is well known and its proof can be found in [15].

Lemma 7.6.1 (Principle of Missing Information)**.** *The matrices $I(\theta^\star)$ and $I_c(\theta^\star)$ satisfy*

$$0 \prec I(\theta^\star) \preceq I_c(\theta^\star).$$

The principle of missing information (Lemma 7.6.1) together with Theorem 7.7.3 of [57] imply that

$$\rho\big(\mathbf{J}_T(\theta^\star)\big) < 1. \tag{7.22}$$

7.6.3 Probability of Having a Fixed Point

Recall that for a given θ, the weak law of large numbers implies that the random vector $T_N(\theta)$ converges, in probability, to the non-random vector $T(\theta)$. As noted in the roadmap presented in Section 7.4, "pointwise convergence" in probability, *i.e.,* convergence for

each θ, is not enough to infer probabilistic properties of T_N from the non-probabilistic properties of T. A stronger version of the weak law of large numbers is required; the next result can be seen as a "uniform law of large numbers". It is stated in terms of the Frobenius norm, denoted by $\|\cdot\|_F$, but, from the equivalence between all norms, it applies to any other norm.

Theorem 7.6.1 ([106]). *Let $a(z,\theta)$ be a matrix of functions of an observation z and the parameter $\theta \in \Omega$. If $z_1, \ldots, z_N$ are i.i.d., Ω is compact, $a(z,\theta)$ is continuous at each θ, and there is $d(z)$ with $\|a(z,\theta)\|_F \leq d(z)$ for all $\theta \in \Omega$, where $\mathbb{E}[d(z)]$ exists and is finite, then $\mathbb{E}[a(z,\theta)]$ is continuous and*

$$\sup_{\theta\in\Omega} \Big\| \frac{1}{N}\sum_{j=1}^{N} a(z_j,\theta) - \mathbb{E}[a(z,\theta)] \Big\|_F \to 0,$$

in probability.

Remark 7.6.1. *The conditions on $a(z,\theta)$ appearing in the statement of Theorem 7.6.1 are again regularity conditions that are similar to the ones required for differentiation under the integral sign. We again refer the reader to our article [14] where these are addressed in detail.*

We will now give a sketch of the proof that establishes that, for sufficiently large N, the probability that T_N has a fixed point is very close to one. From Ostrowski's theorem (see Chapter 3) and $\rho\big(\mathbf{J}_T(\theta^\star)\big) < 1$, there exists a norm $\|\cdot\|$ and $\lambda < 1$ such that

$$\|T(\theta) - \theta^\star\| \leq \lambda \|\theta - \theta^\star\|,$$

for all $\theta \in \bar{B}(\theta^\star, \delta)$, where $\bar{B}(\theta^\star, \delta)$ denotes the closed ball of center $\theta^\star$ and radius δ with respect to $\|\cdot\|$. For any $\theta \in \bar{B}(\theta^\star, \delta)$, the triangular inequality implies that

$$\|T_N(\theta) - \theta^\star\| \leq \|T_N(\theta) - T(\theta)\| + \|T(\theta) - \theta^\star\| \leq \|T_N(\theta) - T(\theta)\| + \lambda\delta.$$

As a consequence, we obtain that

$$\sup_{\theta\in\bar{B}(\theta^\star,\delta)} \|T_N(\theta) - \theta^\star\| \leq \sup_{\theta\in\bar{B}(\theta^\star,\delta)} \|T_N(\theta) - T(\theta)\| + \lambda\delta. \tag{7.23}$$

Without going into details (see [14]), one can show that Theorem 7.6.1 implies that

$$\sup_{\theta\in\bar{B}(\theta^\star,\delta)} \|T_N(\theta) - T(\theta)\| \to 0,$$

in probability. Consequently, the non-random sequence

$$\mathbb{P}_{\theta^\star}\Big(\sup_{\theta \in \bar{B}(\theta^\star, \delta)} \|T_N(\theta) - T(\theta)\| \leq (1-\lambda)\delta \Big)$$

converges to one. From (7.23), we obtain that

$$\mathbb{P}_{\theta^\star}\Big(\sup_{\theta \in \bar{B}(\theta^\star, \delta)} \|T_N(\theta) - T(\theta)\| \leq (1-\lambda)\delta \Big) \leq \mathbb{P}_{\theta^\star}\Big(\sup_{\theta \in \bar{B}(\theta^\star, \delta)} \|T_N(\theta) - \theta^\star\| \leq \delta \Big),$$

and, thus, the non-random sequence

$$\mathbb{P}_{\theta^\star}\Big(\sup_{\theta \in \bar{B}(\theta^\star, \delta)} \|T_N(\theta) - \theta^\star\| \leq \delta \Big)$$

converges to one as well. We conclude that

$$\mathbb{P}_{\theta^\star}\Big(T_N\big(\bar{B}(\theta^\star, \delta)\big) \subseteq \bar{B}(\theta^\star, \delta) \Big)$$

converges to one. By Brouwer's fixed point theorem (see Chapter 3), the probability that T_N has a fixed point is, for N sufficiently large, as close to one as desired.

7.6.4 Probability of Having a Stable Fixed Point

Many technicalities need to be checked to prove that the probability of having a stable fixed point approaches one as N goes to infinity. Similarly to the previous section, we only sketch the proof and refer the reader to [14] for the technical details. Start by noting that, if θ_N is a fixed point of T_N, then

$$T_N'(\theta_N) := \mathbf{J}_{T_N}(\theta_N) = I + \big(A_N(\theta_N)\big)^{-1} \frac{1}{N} \sum_{n=1}^{N} \nabla_\theta^2 \phi(y_n, h_n, \theta_N). \tag{7.24}$$

The idea is to consider an infinite sample version of $\mathbf{J}_{T_N}(\theta_N)$, in the same spirit as when we considered the infinite sample version of T_N, *i.e.,* T. Let

$$T'(\theta) = I + A(\theta)^{-1} \mathbb{E}_{\theta^\star}\Big[\nabla_\theta^2 \phi(y, h, \theta) \Big]. \tag{7.25}$$

Observe that $T'(\theta)$ only coincides with $\mathbf{J}_T(\theta)$ if θ is a fixed point of T. Because $\theta^\star$ is a fixed point of T, at $\theta^\star$,

$$T'(\theta^\star) = \mathbf{J}_T(\theta^\star) = I - I_c(\theta^\star)^{-1} I(\theta^\star).$$

If necessary, by reducing δ (the radius of the ball $\bar{B}(\theta^\star, \delta)$ from the previous section), we may, from (7.22) and the continuity of T', assume that

$$\|T'(\theta)\| < 1,$$

for all $\theta \in \bar{B}(\theta^\star, \delta)$. Omitting the details (see [14]) as in the previous section, one can appeal to Theorem 7.6.1 to show that

$$\sup_{\theta \in \bar{B}(\theta^\star, \delta)} \|T'_N(\theta) - T'(\theta)\| \to 0,$$

in probability. This is is enough to conclude, from $\|T'(\theta)\| < 1$, that

$$\mathbb{P}_{\theta^\star}\Big(\sup_{\theta \in \bar{B}(\theta^\star, \delta)} \|T'_N(\theta)\| < 1 \Big) \to 1.$$

7.6.5 Putting Everything Together

Modulo some technicalities that the reader can find in [14], the summary of the proof skeleton is as follows. Let: 1) T_N be the map underlying (7.17)-(7.19), *i.e.,* (7.20); 2) T be the infinite sample version of T_N, *i.e.,* (7.21); 3) T'_N be the form of the Jacobian of T_N at fixed points of T_N, *i.e.,* (7.24); 4) T' be the infinite sample version of T'_N, *i.e.,* (7.25). Suppose that Assumption S holds. From the principle of missing information (Lemma 7.6.1) and the "uniform law of large numbers" theorem (Theorem 7.6.1), the next theorem follows.

Theorem 7.6.2. *There exists $\delta > 0$ and a norm $\|\cdot\|$ such that*

$$\mathbb{P}_{\theta^\star}\Big(\sup_{\theta \in \bar{B}(\theta^\star, \delta)} \big\|T_N(\theta) - \theta^\star\| \leq \delta \Big) \to 1,$$
$$\mathbb{P}_{\theta^\star}\Big(\sup_{\theta \in \bar{B}(\theta^\star, \delta)} \big\|T'_N(\theta)\| < 1 \Big) \to 1,$$

where $\|T'_N(\theta)\|$ is the induced matrix norm.[2]

We now show why Theorem 7.6.2 encapsulates the notion that, with probability approaching 1, the map T_N has a fixed point θ_N satisfying $\rho\big(\mathbf{J}_{T_N}(\theta_N)\big) < 1$. Let

$$\mathcal{A}_N = \big\{(\mathbf{y}, \mathbf{h}) \in \mathbb{R}^N \times \mathbb{R}^{dN} : \sup_{\theta \in \bar{B}(\theta^\star, \delta)} \big\|T_N(\theta) - \theta^\star\| \leq \delta \big\},$$
$$\mathcal{B}_N = \big\{(\mathbf{y}, \mathbf{h}) \in \mathbb{R}^N \times \mathbb{R}^{dN} : \sup_{\theta \in \bar{B}(\theta^\star, \delta)} \big\|T'_N(\theta)\| < 1 \big\}.$$

[2]The measurability of the maps in this theorem are a consequence of Proposition 7.32 in [107].

Remark 7.6.2. *Informally, observe that the set $\mathcal{A}_N$ is the set of "samples" where the ball $\bar{B}(\theta^\star, \delta)$ is invariant under T_N, i.e,*

$$T_N\big(\bar{B}(\theta^\star, \delta)\big) \subseteq \bar{B}(\theta^\star, \delta),$$

and that the set $\mathcal{B}_N$ is the set of "samples" where the Jacobian of T_N satisfies $\|\mathbf{J}_{T_N}(\theta_N)\| < 1$ at a fixed point θ_N. By noting that a continuous map from a convex compact space into itself has a fixed point (Brouwer's fixed point theorem), it follows that if $(\mathbf{y}, \mathbf{h})$ is in $\mathcal{A}_N$, then T_N has a fixed point. Moreover, if $(\mathbf{y}, \mathbf{h})$ is in $\mathcal{A}_N \cap \mathcal{B}_N$ then T_N has a fixed point θ_N satisfying $\|\mathbf{J}_{T_N}(\theta_N)\| < 1$. All of this is made precise below.

The statement of Theorem 7.6.2 is that the (non-random) sequences $\mathbb{P}_{\theta^\star}(\mathcal{A}_N)$ and $\mathbb{P}_{\theta^\star}(\mathcal{B}_N)$ both tend to 1, as $N \to \infty$. The inequalities

$$\mathbb{P}_{\theta^\star}(\mathcal{A}_N) + \mathbb{P}_{\theta^\star}(\mathcal{B}_N) - 1 \leq \mathbb{P}_{\theta^\star}(\mathcal{A}_N \cap \mathcal{B}_N) \leq \mathbb{P}_{\theta^\star}(\mathcal{A}_N)$$

imply that

$$\mathbb{P}_{\theta^\star}(\mathcal{A}_N \cap \mathcal{B}_N) \to 1.$$

If both inequalities hold, namely

$$\sup_{\theta \in \bar{B}(\theta^\star, \delta)} \big\|T_N(\theta) - \theta^\star\| \leq \delta, \tag{7.26}$$

$$\sup_{\theta \in \bar{B}(\theta^\star, \delta)} \big\|T_N'(\theta)\| < 1, \tag{7.27}$$

then (7.26), together with Brouwer's fixed point theorem (see Chapter 3) implies that T_N has a fixed point θ_N in $\bar{B}(\theta^\star, \delta)$ (this idea is loosely inspired by [108]). Moreover, at a fixed point θ_N, it holds that $T_N'(\theta_N) = \mathbf{J}_{T_N}(\theta_N)$, so, (7.27) implies that

$$\rho\big(\mathbf{J}_{T_N}(\theta_N)\big) \leq \big\|\mathbf{J}_{T_N}(\theta_N)\| \leq \sup_{\theta \in \bar{B}(\theta^\star, \delta)} \big\|T_N'(\theta)\| < 1.$$

This explains why Theorem 7.6.2 expresses the notion that we can "expect" T_N to have a stable fixed point. In fact, from the above, the event

$$\mathcal{C}_N = \big\{(\mathbf{y}, \mathbf{h}) : T_N \text{ has a fixed point } \theta_N \text{ satisfying } \rho\big(\mathbf{J}_{T_N}(\theta_N)\big) < 1\big\}$$

contains the event $\mathcal{A}_N \cap \mathcal{B}_N$, and the probability of this last event approaches 1.

7.7 Simulations

7.7.1 EM Algorithm for (7.7)

Consider the updates of Modified EM, that is, (7.17)-(7.19). The σ^2-update is given by

$$\begin{aligned}(\sigma^2)^{k+1} &= \frac{1}{N}\sum_{n=1}^{N} r(y_n, h_n, \theta^k)(y_n - h_n^T\mu^k)^2 + \big(1 - r(y_n, h_n, \theta^k)\big)y_n^2 \\ &= \frac{1}{N}\sum_{n=1}^{N} y_n^2 - 2r(y_n, h_n, \theta^k)y_n h_n^T\mu^k + r(y_n, h_n, \theta^k)(h_n^T\mu^k)^2.\end{aligned}$$

Now suppose that all instances of μ^k above are replaced by μ^{k+1}, *i.e.,*

$$(\sigma^2)^{k+1} = \frac{1}{N}\sum_{n=1}^{N} y_n^2 - 2r(y_n, h_n, \theta^k)y_n h_n^T\mu^{k+1} + r(y_n, h_n, \theta^k)(h_n^T\mu^{k+1})^2.$$

A simple calculation using (7.17) shows that, in this case,

$$(\sigma^2)^{k+1} = \frac{1}{N}\sum_{n=1}^{N} y_n^2 - r(y_n, h_n, \theta^k)y_n h_n^T\mu^{k+1}. \tag{7.28}$$

The EM algorithm for (7.7) is derived in [12] and it is given by (7.17)-(7.18) and (7.28), instead of (7.19). This is why we called (7.17)-(7.19) a modified EM algorithm.

Remark 7.7.1. *The EM algorithm has the virtue of never decreasing the log-likelihood, that is, if θ^k is the sequence produced by the EM algorithm, then $\phi(\theta^{k+1}) \geq \phi(\theta^k)$ (see [15]). This property may not hold for the Modified EM algorithm.*

7.7.2 Two Distributed Algorithms

Let

$$\begin{aligned}g_2(\theta) &= \frac{1}{N}\Big(\sum_{n=1}^{N}\Gamma(y_n, h_n, \theta), \sum_{n=1}^{N}\psi(y_n, h_n, \theta), \sum_{n=1}^{N} r(y_n, h_n, \theta), \sum_{n=1}^{N}\gamma(y_n, h_n, \theta)\Big) \\ g_1(\Gamma, \psi, p, \sigma^2) &= \big(\Gamma^{-1}\psi, p, \sigma^2\big) \\ \hat{g}_2(\theta) &= \frac{1}{N}\Big(\sum_{n=1}^{N}\Gamma(y_n, h_n, \theta), \sum_{n=1}^{N}\psi(y_n, h_n, \theta), \sum_{n=1}^{N} r(y_n, h_n, \theta), \sum_{n=1}^{N} y_n^2\Big) \\ \hat{g}_1(\Gamma, \psi, p, a) &= \big(\Gamma^{-1}\psi, p, a - \psi^T\Gamma^{-1}\psi\big).\end{aligned}$$

The modified EM algorithm and the standard EM algorithm can be, respectively, written as $\theta^{k+1} = g_1 \circ g_2(\theta^k)$ and $\theta^{k+1} = \hat{g}_1 \circ \hat{g}_2(\theta^k)$. Neither algorithm can be written as an average of local maps, but switching the roles of the composition leads to algorithms that can. Throughout the rest of the section, when we say "our algorithm" we are referring to the EXTRA-distributed Banach-Picard iteration (5.19) stemming from the Banach-Picard iteration of $g_1 \circ g_1$. When we refer to DA-DEM, we mean the algorithm proposed in [12], which coincides with the distributed algorithm from Chapter 4 for the Banach-Picard iteration of $\hat{g}_2 \circ \hat{g}_1$.

7.7.3 Simulation Results

In this section, we compare our algorithm with DA-DEM through Monte Carlo simulations. The parameters generated once and fixed throughout all Monte Carlo runs were: $d = 3$, $N = 100$, a unit-norm vector $\mu^\star \in \mathbb{R}^d$, $p^\star = 0.7$, and an undirected connected graph on N nodes with connectivity radius[3] $r_c = 0.18$. In what follows $z_n^k(\alpha)$ denotes the k-th iterate of agent n of our algorithm with parameter α, and $z_n^k(\rho)$ denotes the k-th iterate of agent n of DA-DEM with parameter ρ. In the language of Chapter 4, parameter ρ corresponds to the shrinking step-size sequence given, for $k = 0, \ldots,$ by

$$\alpha_k = \frac{\rho}{k + \rho}.$$

Each Monte Carlo run consisted in

1) Generating a data set: each h_n was independently sampled from a Gaussian with zero mean and covariance I_3; the variance of the noise $(\sigma^\star)^2$ was set to

$$(\sigma^\star)^2 = \frac{\|\mathbf{H}\|_F^2}{N \times \mathrm{SNR}},$$

with $\mathbf{H}^T = [h_1 \ldots h_N]$ and where SNR is the desired signal to noise ratio (we experimented with $\mathrm{SNR} \in \{10\mathrm{dB}, 20\mathrm{dB}\}$). Finally, each y_n was sampled according to $f_{Y|H}$ (see (7.5)), with h_n, $\mu^\star$, $p^\star$, and $(\sigma^\star)^2$.

2) Computing 10000 iterations of DA-DEM, with $\rho \in \{2, 3, 4\}$, and of our algorithm, with $\alpha \in \{0.001, 0.005, 0.01\}$. Both algorithms were initialized according to the initialization suggested in [12] for DA-DEM.

The performance metrics consisted in finding a fixed point using the centralized maps as

[3] N points were randomly deployed on the unit square; two points were then connected by an edge if their distance was less than r_c.

follows. We first compute

$$\theta^0(\alpha) = \frac{1}{N}\sum_{n=1}^{N} g_1\big(z_n^{10000}(\alpha)\big) \tag{7.29}$$

$$\theta^0(\rho) = \frac{1}{N}\sum_{n=1}^{N} \hat{g}_1\big(z_n^{10000}(\rho)\big), \tag{7.30}$$

where: $\alpha \in \{0.001, 0.005, 0.01\}$; $\rho \in \{2, 3, 4\}$.

We ran the centralized modified EM and the standard EM algorithms, with initialization as in (7.29) and (7.30), that is, we computed

$$\theta^{k+1}(\alpha) = g_1 \circ g_2\big(\theta^k(\alpha)\big)$$
$$\theta^{k+1}(\rho) = \hat{g}_1 \circ \hat{g}_2\big(\theta^k(\rho)\big),$$

until we found $\theta^\star(\alpha)$ and $\theta^\star(\rho)$ satisfying

$$\Big\|\theta^\star(\alpha) - g_1 \circ g_2\big(\theta^\star(\alpha)\big)\Big\| \leq 10^{-10}$$
$$\Big\|\theta^\star(\rho) - \hat{g}_1 \circ \hat{g}_2\big(\theta^\star(\rho)\big)\Big\| \leq 10^{-10}.$$

The error at iteration k of the distributed algorithms was then computed as

$$\frac{1}{N}\sum_{n=1}^{N} \Big\|\pi_1 \circ g_1\big((z_n^k(\alpha)\big) - \theta^\star(\alpha)\Big\|$$
$$\frac{1}{N}\sum_{n=1}^{N} \Big\|\pi_1 \circ \hat{g}_1\big((z_n^k(\rho)\big) - \theta^\star(\rho)\Big\|,$$

where π_1 is the projection onto the average, *i.e.*, $\pi_1(\mu, p, \sigma^2) = \mu$ (as mentioned before, p and σ^2 were treated as nuisance parameters).

The number of Monte Carlo tests was 100 and the errors at iteration k are averages for each α and ρ. The results for two different SNR values are shown in logarithmic scale in Figures 1 and 2.

The simulations show, as expected from the theory, that our algorithm converges linearly and clearly outperforms the algorithm from [12], which, given its diminishing step-size, is bound to converge only sub-linearly. Moreover, both algorithms require just one round of communications per iteration.

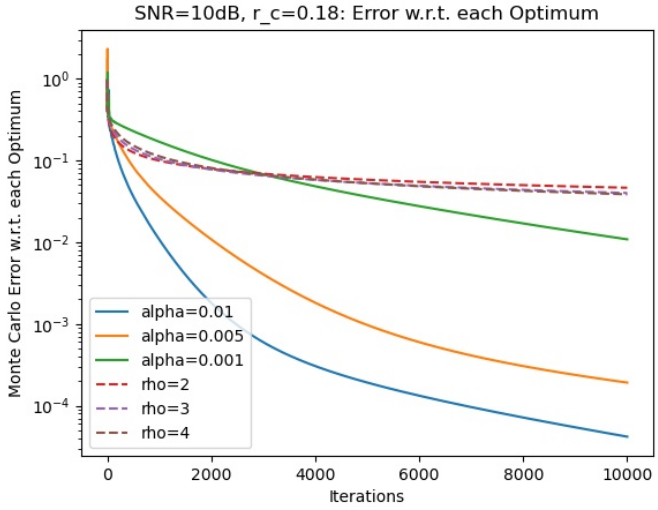


Figure 7.1: Result of the Monte Carlo simulation of the error with respect to each optimum for an SNR = 10dB and a connectivity radius of 0.18. The dashed curves correspond to DA-DEM with parameter $\rho \in \{2, 3, 4\}$; the solid curves correspond to our algorithm with parameter $\alpha \in \{0.001, 0.005, 0.01\}$.

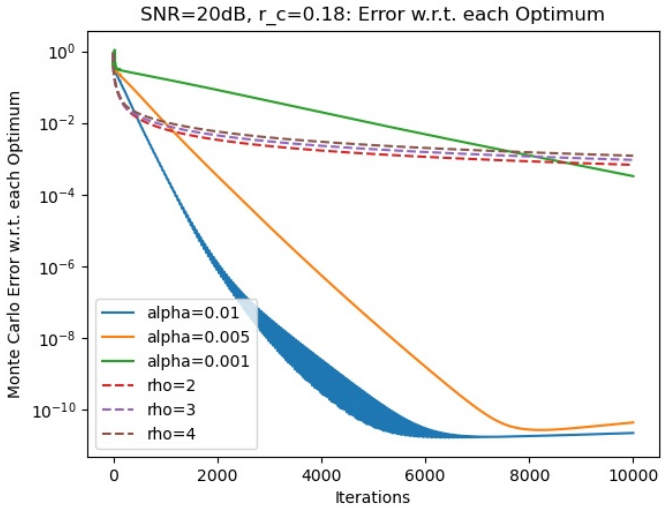


Figure 7.2: Result of the Monte Carlo simulation of the error with respect to each optimum for an SNR = 20dB and a connectivity radius of 0.18. The dashed curves correspond to DA-DEM with parameter $\rho \in \{2, 3, 4\}$; the solid curves correspond to our algorithm with parameter $\alpha \in \{0.001, 0.005, 0.01\}$.

7.8 Comments and References

There is considerable work on the "probabilistic linear convergence" of EM [50], [51], [52]. However, neither the results in [50], nor those in [51] encompass the mixture model underlying (7.5). The mixture of regressions presented in [52] bears some similarity

with the model underlying (7.5), but it is not the same: in [52], p is fixed at $1/2$ and $Z_n \in \{-1, 1\}$ (rather than $\{0, 1\}$), thus there are no measurements that are just noise. Furthermore, [52] is primarily concerned with statistical guarantees for the error with respect to the ground truth, while we address the goal of establishing the existence of a stable fixed point.

As mentioned in [12], there are two other relevant works on distributed EM, namely, [53] and [54]. However (see [12]), both those works address a different problem of Gaussian mixture density estimation. Moreover, in the case of [53], the algorithm demands a cyclic network topology, and, in [54] the algorithm requires higher computational load on each node, since it is based on *alternating direction method of multipliers* (see [55] for a reference on ADMM).

7.8.1 Convergence Towards the Ground Truth

It was shown that with probability tending to one, the map T_N underlying (7.17)-(7.19) and defined in (7.24) has a fixed point θ_N in $\bar{B}(\theta^\star, \delta)$. This was established by proving that

$$\mathbb{P}_{\theta^\star}\Big(\sup_{\theta \in \bar{B}(\theta^\star, \delta)} \|T_N(\theta) - \theta^\star\| \leq \delta \Big) \to 1$$

and appealing to Brouwer's fixed point theorem. However, the agents' goal is to estimate $\theta^\star$ ($\mu^\star$ in reality, since $p^\star$ and $(\sigma^\star)^2$ are treated as nuisance parameters). Therefore, it seems that we cheated since θ_N is just a point in $\bar{B}(\theta^\star, \delta)$ and no connection with $\theta^\star$ was made. We didn't even quantify δ, hence, without any further investigation, the "best" we can say is that we have local linear convergence towards a point at a distance of at most δ from $\theta^\star$; to make things worse, that distance arises from a norm that we didn't even bother to identify. In this section, we "fill" this gap by appealing to an argument in the non-random version of the problem. To this end, suppose that the maps T_N are non-random (just think that, for each θ, $T_N(\theta)$ is a just a vector, *i.e.,* not a random vector) and suppose that the sequence converges uniformly to T in $\bar{B}(\theta^\star, \delta)$, i.e.,

$$\lim_{N \to \infty} \sup_{\theta \in \bar{B}(\theta^\star, \delta)} \|T_N(\theta) - T(\theta)\| = 0.$$

In this case, the existence of a fixed point θ_N of T_N is easy to establish by noting that

$$\|T_N(\theta) - \theta^\star\| \leq \|T_N(\theta) - T(\theta)\| + \|T(\theta) - \theta^\star\| \leq \|T_N(\theta) - T(\theta)\| + \lambda\delta \leq \delta$$

for N sufficiently large. Similar to the random case, by Brouwer's fixed point theorem, there exists θ_N such that $T_N(\theta_N) = \theta_N$. What we now show is that $\theta_N \to \theta^\star$. Without loss of generality, assume that, for $N \geq 1$, T_N has a fixed point in $\bar{B}(\theta^\star, \delta)$ (if it does not, consider the sequence T_N as starting in the first N_0 for which T_N has such a fixed point for $N \geq N_0$). For every $\epsilon' > 0$, there must exist $N_0(\epsilon')$ such that

$$\|\theta_N - \theta^\star\| = \|T_N(\theta_N) - \theta^\star\| \leq \|T_N(\theta_N) - T(\theta_N)\| + \lambda\|\theta_N - \theta^\star\| \leq \epsilon' + \lambda\|\theta_N - \theta^\star\|,$$

for $N \geq N_0(\epsilon')$. Choose $\epsilon' = \epsilon(1-\lambda)$ and, for $N \geq N_0(\epsilon(1-\lambda))$, it holds that

$$\|\theta_N - \theta^\star\| \leq \epsilon(1-\lambda) + \lambda\|\theta_N - \theta^\star\|;$$

this implies that $\|\theta_N - \theta^\star\| \leq \epsilon$, for $N \geq N_0(\epsilon(1-\lambda))$. We conclude that $\theta_N \to \theta^\star$.

By the reasoning above, for N sufficiently large, the fixed points of T_N are at a distance of at most ϵ from the "ground truth". However, without characterizing the "rate" at which T_N converges to T, we cannot characterize the rate at which θ_N converges to $\theta^\star$. To "complete" the picture of this chapter, we should obtain a probabilistic version of the argument above; this route, however, was not pursued and is left as a direction for future research.

Chapter 8

Conclusion and Future Work

8.1 Conclusion

The thesis proposed an algorithmic framework that extends to distributed settings the iteration of a map H that is an average of local maps, *i.e.,* maps held by agents in a communication network. Specifically, given a map H defined on $\mathbb{R}^d$ that can be written as

$$H = \frac{1}{N}\sum_{n=1}^{N} H_n, \tag{8.1}$$

where each H_n is known at an agent (*e.g.* a sensor) in a communication network (*e.g.,* a wireless network); we showed how to build a parametric[1] family of maps F_H defined on $\mathbb{R}^{2dN}$ with the following properties:

1) The Banach-Picard iteration of F_H, termed *distributed Banach-Picard iteration* (DBPI), *i.e.,* the iteration

$$(z^{k+1}, w^{k+1}) = F_H(z^k, w^k), \tag{8.2}$$

has distributed implementation respecting the links of the communication network, that is, by letting agent n maintain the variables (z_n^k, w_n^k), the nth component of (8.2) is of the form

$$(z_n^{k+1}, w_n^{k+1}) = \tilde{F}_{H_n}\Big((z_n^k, w_n^k), \big\{(z_m^k, w_m^k)\big\}_{m\in\mathcal{N}_n}\Big),$$

where $\tilde{F}_{H_n}$ is a map depending only on the variables of agent n, the variables of its

[1]The parameters are omitted for ease of notation.

neighbors and its private map H_n;

2) The DBPI lifts the fixed points of H, that is, if $H(x^\star) = x^\star$, then, for some $w^\star \in \mathbb{R}^{dN}$,

$$F_H\big((x^\star, \ldots, x^\star), w^\star\big) = \big((x^\star, \ldots, x^\star), w^\star\big);$$

3) The contractive properties, either local or global of H, are inherited by F_H. Consequently, if the Banach-Picard iteration of H, *i.e.,* the iteration

$$x^{k+1} = H(x^k) \tag{8.3}$$

has local (global) linear convergence, then the DBPI, *i.e.,* (8.2), has local (global) linear convergence.

The reason to view the DBPI as an algorithmic framework, rather than simply an algorithm, is its dependence on H. In fact, the DBPI can be seen as a recipe for distributed inference for which the convergence properties follow from those of centralized inference, *i.e,* the iteration (8.3). The skeleton of the recipe is as follows: suppose $x^\star$ is to be inferred from $\theta_1, \ldots, \theta_N$, where θ_n is known at an agent n in a communication network, then, inferring $x^\star$ can be carried out in a distributed fashion by

1) Finding a map H for which $x^\star = H(x^\star)$;

2) Showing that H can be written as an average of the form (8.1), where H_n is a map that can be computed by agent n, *i.e.,* a map possibly depending on θ_n but not on θ_m, for $m \neq n$;

3) Proving that H is either locally or globally contractive towards $x^\star$.

The end result is that distributed inference will be carried out with the same "qualitative speed" as centralized inference, that is, all agents will be increasingly in agreement over $x^\star$, at a rate of the same order as if the inference was performed at an agent knowing the full data set. Crucially, this certificate follows only from the properties of H.

To demonstrate the virtues of the DBPI, the second part of the thesis obtained, by following the recipe abovementioned, extensions to distributed scenarios of two inference problems – *principal component analysis* (PCA) and *estimation from noisy and faulty measurements*. As dictated by the recipe, to arrive at a map for distributed inference, two centralized inference maps H_P (PCA map) and H_E (estimation map) were studied and their fixed points analyzed. Specifically, for desired fixed points $x_P^\star$ (PCA solution)

and $x_E^\star$ (an estimate), it was shown that

$$\rho\big(\mathbf{J}_{H_Q}(x_Q^\star)\big) < 1, \tag{8.4}$$

where $Q \in \{P, E\}$, thus completing step 3) of the recipe. Condition (8.4), a sufficient condition for local contractiveness towards $x_Q^\star$ if H_Q is differentiable, is inherited by F_{H_Q} as a consequence of the theoretical results established in Chapter 5. The iteration of H_P corresponds to a well known algorithm termed *Sanger's* algorithm (SA) and the iteration of H_E to a slightly modified *expectation-maximization* (EM) algorithm. Both the SA and the EM algorithm are relevant algorithms that do not benefit from global properties such as strong convexity, therefore contrasting with several algorithms stemming from optimization problems. Consequently, the DBPI encapsulates relevant algorithms for which only rather weak guarantees (local linear convergence) can be provided. On the other side, the extension, via DBPI, to a distributed scenario of a globally contractive map (*e.g.* a gradient map of a strongly convex and Lipschitz function) also inherits the globally contractive property. In fact, if H is a gradient map, the DBPI reduces to a distributed gradient descent algorithm and, upon elimination of the second variable and a wise choice of parameters, it recovers well-known distributed descent algorithms such as EXTRA and DIGing.

8.2 Remarks on the Drawbacks of the DBPI

In addition to preserving a differential local contraction condition such as (8.4), this thesis also shows that a type of unstable (a particular case of the inequality in (8.4) reversed) fixed points of H are lifted to unstable fixed points of F_H. Formally, if $x^\star$ is a fixed point for which $\mathbf{J}_H(x^\star)$ has an eigenvalue with real part larger than one, then,

$$\rho\big(\mathbf{J}_F\big((x^\star, \ldots, x^\star), w^\star\big) > 1.$$

As a consequence, we argued in Chapter 5 that if H has only a finite number of fixed points, some of which satisfying a condition such as (8.4) and the remaining being unstable fixed points of the type described above, then the parameters of F_H (recall that F_H is has tunable parameters) can be tuned to preserve the qualitative character of each fixed point. Under technical conditions on H, the iteration (8.3) escapes unstable fixed points and, hence, if it converges, it almost surely converges to a fixed point satisfying a condition such as (8.4). The relevance of preserving instability can thus be seen as an almost certainty that (8.2) will not converge to an unstable fixed point, and, consequently, the agents will

not agree on an unstable fixed point (think, for example, of reaching a maximum when they are seeking a minimum).

Suppose that H has only two fixed points $x_S^\star$ and $x_U^\star$, with $x_S^\star$ satisfying

$$\mathbf{J}_H(x_S^\star) < 1,$$

and $\mathbf{J}_H(x_U^\star)$ having an eigenvalue equal to -2, while the remaining eigenvalues have magnitude less than one. A careful inspection of the proofs in Appendix A reveals that, in this case, for a sufficiently small step-size (one of the parameters of F_H), both $x_U^\star$ and $x_S^\star$ are lifted to fixed points satisfying

$$\rho\big(\mathbf{J}_F\big((x_Q^\star, \ldots, x_Q^\star), w^\star\big) < 1,$$

where $Q \in \{U, S\}$. If we additionally assume that H satisfies the technical conditions that ensure that H almost surely escapes unstable fixed points, we see that the distributed extension of H via the DBPI destroys this property, that is, with a random initialization, the DBPI has a non-zero probability of coordinating the agents towards $x_U^\star$. This, from our point of view, constitutes a drawback of the DBPI and suggests that the DBPI might not be the "right" general extension of an arbitrary map H to a distributed configuration.

8.3 Future Work

We conclude with four questions that were highlighted along the thesis and that were left unanswered.

8.3.1 Asymptotically Stable but not Exponentially Stable

A fixed point $x^\star$ may be a stable atractor[2] (asymptotically stable fixed point), while satisfying

$$\mathbf{J}_H(x^\star) = 1, \tag{8.5}$$

the frontier case between (8.4) and the reverse strict inequality. In this scenario, there might be a neighborhood U of $x^\star$ whose points are attracted (by iterating H) to $x^\star$ at a rate slower than linear; as an example, see Chapter 3, where we consider the fixed point 0 of the map $g(x) = x - x^3$. In the dynamical systems literature this is termed a

[2]For the purposes of this discussion think of a fixed point $x^\star$ for which there exists a neighborhood U such that $H(U) \subseteq U$ and, if $x^0 \in U$, then $\lim_k H^k(x^0) = x^\star$.

locally asymptotically stable but not locally exponentially stable fixed point. Whereas local exponential stability, *i.e.,* (8.3), is "detected" by looking at first derivatives (the Jacobian), the same is not true for (8.5). Consequently, the analysis in Chapter 5 is not enough to determine whether the lift of an asymptotically stable but not exponentially stable fixed point preserves its qualitative character. A positive answer to this question could take the form: Let $x^\star$ be a fixed point of H for which there exists a Lyapunov function V_H certifying its asymptotic stability (the validity of this formulation follows from the *converse of Lyapunov Theorem*) with respect to H, then, for a sufficiently small step-size, there exists a Lyapunov function V_{F_H} (possibly depending on the step-size) certifying the asymptotic stability of $\big((x^\star, \dots, x^\star, w^\star\big)$ with respect to F_H. The challenge is, thus, to construct V_{F_H} from V_H.

8.3.2 Non-Differential Local Contraction

The continuous analog of

$$\rho\big(\mathbf{J}_H(x^\star)\big) < 1 \tag{8.6}$$

is, via *Ostrowski's theorem*, the existence of a norm $\|\cdot\|$, a number $0 \leq \mu < 1$, and a positive number $\delta > 0$ such that, for all x satisfying $\|x - x^\star\| \leq \delta$,

$$\|H(x) - x^\star\| \leq \mu \|x - x^\star\|;$$

in this scenario, H is said to be a local contraction towards $x^\star$. Chapter 5 proves that if H is a continuous and not-necessarily differential local contraction with respect to $x^\star$, then F_H preserves this feature with respect to $\big((x^\star, \dots, x^\star), w^\star\big)$. This, however, was proved by assuming that each H_n is globally Lipschitz, an absent condition in the proof of the differential case. An interesting question is whether this can be proved without assuming global Lipschitzianity.

8.3.3 The Local Diffeomorphism Condition for Distributed PCA

A sufficient condition that ensures that a map H almost surely escapes unstable fixed points is being a local diffeomorphism. Chapter 6 proves that the Sanger's map inducing the Sanger's algorithm has, under mild conditions a finite number of fixed points, the solutions to PCA satisfying (8.6) and the remaining fixed points being unstable. Moreover, we showed that the Sanger's Map is a local diffeomorphism. Consequently, the Sanger's algorithm almost surely escapes the non-desired fixed points. Even though, the results of

Chapter 6 imply that the distributed extension preserves both the number of fixed points and their qualitative character, we left unanswered whether F_H can be tuned to be a local diffeomorphism and, as a consequence, almost surely escape the undesired fixed points as well.

8.3.4 Ground Truth of Variant of EM

Chapter 7 looks at a variant of the EM algorithm for the estimation of a parameter $\mu^\star$ which corresponds to the Banach-Picard iteration of a map H_E. Being dependent on samples from a probability distributed, the results therein are of probabilistic nature. Specifically, we show that, as the number of agents N tends to infinity, the probability that H_E has a fixed point $\tilde{\mu}_N^\star$ satisfying

$$\rho\big(\mathbf{J}_{H_E}(\tilde{\mu}_N^\star)\big) < 1$$

tends to one. The agents are ultimately, interested in estimating $\mu^\star$ and not some unrelated value $\tilde{\mu}^\star$. Therefore, a question that should be answered is whether $\tilde{\mu}_N^\star$ converges in probability to $\mu^\star$, as N tends to infinity. In the comments section of Chapter 7, it is argued that this should be the case, by appealing to the non-random version of this problem; nevertheless, this argument should be "turned into" a probabilistic one.

Appendix A

Proof of Theorem 5.3.2

Theorem A.0.1 (Theorem 2.4.7.2, [57]). *Let M be a square matrix. For every $\epsilon > 0$, there exists a non-singular matrix S_ϵ such that*

1) *$T_\epsilon := S_\epsilon^{-1} M S_\epsilon$ is upper triangular;*

2) *The magnitude of the elements above the diagonal does not exceed ϵ,* i.e., *for $i < j$, $|(T_\epsilon)_{ij}| \leq \epsilon$.*

Theorem A.0.2 (Geršgorin Theorem). *Let A be an $n \times n$ square matrix, let*

$$R_i'(A) = \sum_{i \neq j} |A_{ij}|, \quad i = 1, \ldots, n$$

denote the deleted absolute row sums *of A, and consider the* Geršgorin discs

$$\big\{ z \in \mathbb{C} : |z - A_{ii}| \leq R_i'(A) \big\}, \quad i = 1, \ldots, n.$$

The eigenvalues of A are in the union of Geršgorin discs

$$G(A) = \bigcup_{i=1}^{n} \big\{ z \in \mathbb{C} : |z - A_{ii}| \leq R_i'(A) \big\}.$$

Furthermore, if the union of k of the n discs that comprise $G(A)$ forms a set $G_k(A)$ that is disjoint from the remaining $n - k$ discs, then $G_k(A)$ contains exactly k eigenvalues of A, counted according to their multiplicities.

To prove Theorem 5.3.2, we begin with the following lemma. In the statement, $\bar{B}(x, \delta)$ denotes the closed ball in $\mathbb{C}$ of center x and radius δ.

Lemma A.0.1. *For every $\epsilon > 0$, there exists $\alpha(\epsilon)$ such that, for $|\alpha| \leq \alpha(\epsilon)$, the matrix $A(\alpha)$ from Theorem 5.3.2 is similar to a matrix with Geršgorin discs given by*

$$\bar{B}(1 + \mu\alpha, |\alpha|\epsilon)$$
$$\bar{B}(\tilde{\mu} + \alpha(\mathcal{A}_\epsilon)_{ii}, \epsilon),$$

with $\mu \in \mathcal{U} := \{$eigenvalues of $B_{11}\}$, $\tilde{\mu} \in \tilde{\mathcal{U}} := \{$eigenvalues of $A_{22}\}$, and where the numbers $(\mathcal{A}_\epsilon)_{ii}$ depend on ϵ but not on α.

Proof. Let $\epsilon > 0$ and, from Theorem A.0.1, let S_ϵ be a $d \times d$ non-singular matrix such that $T_\epsilon = S_\epsilon^{-1} B_{11} S_\epsilon$ is upper triangular and with absolute deleted row sums not exceeding $\epsilon/2$. Similarly, let $\hat{S}_\epsilon$ be a non-singular $k \times k$ matrix such that $\hat{T}_\epsilon = \hat{S}_\epsilon^{-1} A_{22} \hat{S}_\epsilon$ is upper triangular and with absolute deleted row sums not exceeding $\epsilon/3$. For any α we have that $A(\alpha)$ is similar to

$$\begin{bmatrix} S_\epsilon^{-1} & \mathbf{0} \\ \mathbf{0} & \hat{S}_\epsilon^{-1} \end{bmatrix} \begin{bmatrix} I + \alpha B_{11} & \alpha B_{12} \\ \alpha B_{21} & A_{22} + \alpha B_{22} \end{bmatrix} \begin{bmatrix} S_\epsilon & \mathbf{0} \\ \mathbf{0} & \hat{S}_\epsilon \end{bmatrix} = \begin{bmatrix} I + \alpha T_\epsilon & \alpha S_\epsilon^{-1} B_{12} \hat{S}_\epsilon \\ \alpha \hat{S}_\epsilon^{-1} B_{21} S_\epsilon & \hat{T}_\epsilon + \alpha \hat{S}_\epsilon^{-1} B_{22} \hat{S}_\epsilon \end{bmatrix}.$$

For simplicity of notation let $\mathcal{B}_\epsilon = S_\epsilon^{-1} B_{12} \hat{S}_\epsilon$, $\mathcal{C}_\epsilon = \hat{S}_\epsilon^{-1} B_{21} S_\epsilon$, and $\mathcal{A}_\epsilon = \hat{S}_\epsilon^{-1} B_{22} \hat{S}_\epsilon$. Let $r > 0$ and consider the further similarity

$$\begin{bmatrix} I & \mathbf{0} \\ \mathbf{0} & r^{-1} I \end{bmatrix} \begin{bmatrix} S_\epsilon^{-1} & \mathbf{0} \\ \mathbf{0} & \hat{S}_\epsilon^{-1} \end{bmatrix} A(\alpha) \begin{bmatrix} S_\epsilon & \mathbf{0} \\ \mathbf{0} & \hat{S}_\epsilon \end{bmatrix} \begin{bmatrix} I & \mathbf{0} \\ \mathbf{0} & rI \end{bmatrix} = \begin{bmatrix} I + \alpha T_\epsilon & \alpha r \mathcal{B}_\epsilon \\ \alpha r^{-1} \mathcal{C}_\epsilon & \hat{T}_\epsilon + \alpha \mathcal{A}_\epsilon \end{bmatrix} =: \tilde{A}(\alpha).$$

The following remarks are due:

a) The matrix $I + \alpha T_\epsilon$ is an upper triangular matrix with elements in the diagonal of the form $1 + \alpha\mu$, where μ is an eigenvalue of B_{11}. Moreover, the absolute deleted row sums of $I + \alpha T_\epsilon$ do not exceed $|\alpha|\epsilon/2$.

b) The ith diagonal entry of $\hat{T}_\epsilon + \alpha\mathcal{A}_\epsilon$ is of the form $\tilde{\mu} + \alpha(\mathcal{A}_\epsilon)_{ii}$, where $\tilde{\mu}$ is an eigenvalue of A_{22}, and $(\mathcal{A}_\epsilon)_{ii}$ is the ith diagonal entry of $\mathcal{A}_\epsilon$.

Choose $r(\epsilon) > 0$ such that $r(\epsilon)\mathcal{B}_\epsilon$ has absolute row sums not exceeding $\epsilon/2$, and, with this choice, the absolute deleted row sums of the upper part of $\tilde{A}(t)$, *i.e.*, $[I + \alpha T_\epsilon \quad \alpha r \mathcal{B}_\epsilon]$, do not exceed $\epsilon|\alpha|$. Finally, choose $\alpha\big(r(\epsilon), \epsilon\big)$ such that, for $|\alpha| < \alpha\big(r(\epsilon), \epsilon\big)$, the absolute row sums of both $\alpha r^{-1}\mathcal{C}_\epsilon$ and $\alpha\mathcal{A}_\epsilon$ do not exceed $\epsilon/3$. As a consequence, the absolute deleted row sums of the lower part of $\tilde{A}(t)$, that is, $[\alpha r^{-1}\mathcal{C}_\epsilon \quad \hat{T}_\epsilon + \alpha\mathcal{A}_\epsilon]$, do not exceed ϵ. This completes the proof with $\tilde{A}(\alpha)$ being the matrix similar to $A(\alpha)$. □

A.1 Proof of Part 1) of Theorem 5.3.2

Throughout the rest of the appendix, the *weak Geršgorin Theorem* is the first part of Theorem A.0.2, *i.e.,* the statement that the eigenvalues of A are in $G(A)$; the *strong Geršgorin Theorem* is the second part of Theorem A.0.2, *i.e,* the part that begins with "Furthermore,...".[1]

Lemma A.0.1 and the weak Geršgorin Theorem are already enough to prove 1) of Theorem 5.3.2. In fact, the proof follows from the following simple lemmas.

Lemma A.1.1. *Let $\mu \in \mathbb{C}$ such that $Re(\mu) < 0$. Then, there exists ϵ_μ and α_μ such that, for $0 < \epsilon < \epsilon_\mu$ and $0 < \alpha \leq \alpha_\mu$,*

$$\bar{B}(1 + \mu\alpha, \alpha\epsilon) \subseteq B(0, 1).$$

Proof. The proof is easily understood by making a drawing. Nevertheless, we spell out the details. Let $S(0,1) := \{z \in \mathbb{C} : |z| = 1\}$ and $d(z, w) := |z - w|$ be the distance in $\mathbb{C}$. The first thing we need to observe is that if $x \in B(0,1)$ and $\delta < d(x, S(0,1))$, then $\bar{B}(x, \delta) \subseteq B(0,1)$ (draw a picture). Consider $|1 + \alpha\mu| = 1$, or equivalently, $2\alpha\mathrm{Re}(\mu) + \alpha^2|\mu|^2 = 0$. From $\mathrm{Re}(\mu) < 0$, we conclude that, for $0 < \alpha < -2\mathrm{Re}(\mu)/|\mu|^2$, $1 + \mu\alpha \in B(0,1)$.

Let $g(\alpha) = |1 + \alpha\mu|$ and observe that, for $0 \leq \alpha \leq -2\mathrm{Re}(\mu)/|\mu|^2$,

$$d(1 + \alpha\mu, S(0,1)) = g(0) - g(\alpha).$$

The idea is to use the Mean Value Theorem which gives $d(1+\alpha\mu, S(0,1)) = -\alpha g'(\xi)$, for some $\xi \in (0, \alpha)$. Note that

$$-g'(\alpha) = -\frac{\alpha|\mu|^2 + \mathrm{Re}(\mu)}{|1 + \alpha\mu|},$$

the sign of which coincides with that of $-\alpha|\mu|^2 - \mathrm{Re}(\mu)$. Observe that, for $0 \leq \alpha \leq -\mathrm{Re}(\mu)/(2|\mu|^2)$, $-g'(\alpha) > 0$ and, so, restrict α to this interval which contains the interval $0 \leq \alpha \leq -2\mathrm{Re}(\mu)/|\mu|^2$. In this interval we have that $|1 + \alpha\mu| \leq 1$, and, hence, for $0 \leq \alpha \leq -\mathrm{Re}(\mu)/(2|\mu|^2)$,

$$-g'(\alpha) \geq -\alpha|\mu|^2 - \mathrm{Re}(\mu) \geq -\frac{\mathrm{Re}(\mu)}{2}.$$

[1]The reason for these names is that the proof of the weak Geršgorin Theorem is a straightforward and elementary one, whereas the proof of the strong Geršgorin Theorem is not exactly trivial.

We conclude, via the Mean Value Theorem, that, for $0 \leq \alpha \leq -\mathrm{Re}(\mu)/(2|\mu|^2)$,

$$d(1+\alpha\mu, S(0,1)) \geq -\frac{\mathrm{Re}(\mu)}{2}\alpha.$$

This finishes the proof: let $\epsilon_\mu = -\mathrm{Re}(\mu)/2$ and restrict $0 < \alpha \leq -\mathrm{Re}(\mu)/(2|\mu|^2)$, then

$$\bar{B}(1+\mu\alpha, \epsilon\alpha) \subseteq B(0,1),$$

provided $0 < \epsilon < \epsilon_\mu$. □

proof of part 1) of Theorem 5.3.2. Given $\mathcal{U} = \{\text{eigenvalues of } B_{11}\}$ let $\epsilon_1^\star = \min_{\mu\in\mathcal{U}}\{\epsilon_\mu\}$ and let $\alpha_1^\star = \min_{\mu\in\mathcal{U}}\{\alpha_\mu\}$, where ϵ_μ and α_μ are defined in Lemma A.1.1. Given $\tilde{\mathcal{U}} = \{\text{eigenvalues of } A_{22}\}$, let

$$\epsilon_2^\star = \min_{\tilde{\mu}\in\tilde{\mathcal{U}}}\left\{\frac{d(\tilde{\mu}, S(0,1))}{2}\right\}$$

and let

$$\epsilon^\star = \min\{\epsilon_1^\star, \epsilon_2^\star\}.$$

Let $(\mathcal{A}_{\epsilon^\star})_{ii}$ be the numbers from Lemma A.0.1 and let $\alpha_2^\star$ be sufficiently small such that, for $0 < \alpha < \alpha_2^\star$, $\alpha(\mathcal{A}_{\epsilon^\star})_{ii} < \epsilon^\star$, for all i. Since $\epsilon^\star \leq \epsilon_2^\star \leq d(\tilde{\mu}, S(0,1))/2$, we obtain that, for $0 < \alpha < \alpha_2^\star$,

$$\bar{B}(\tilde{\mu} + \alpha(A_{\epsilon^\star})_{ii}, \epsilon^\star) \subseteq B(0,1).$$

Finally, let

$$\alpha^\star = \min\left\{\alpha_1^\star, \alpha_2^\star, \alpha(\epsilon^\star)\right\},$$

where $\alpha(\epsilon^\star)$ is defined in Lemma A.0.1 and the result is proved for $0 < \alpha < \alpha^\star$. □

A.2 Proof of Part 2) of Theorem 5.3.2

For the second part we really need the strong Geršgorin Theorem. The idea is that if B_{11} has at least one eigenvalue $\mu^\star$ with $\mathrm{Re}(\mu^\star) > 0$ then, for sufficiently small and positive α, the ball $\bar{B}(1+\alpha\mu^\star, \alpha\epsilon)$ can be forced to be outside $\bar{B}(0,1)$ and to be disjoint from all other balls of the form $\bar{B}(1+\alpha\mu, \alpha\epsilon)$, where μ is an eigenvalue of B_{11} other than $\mu^\star$. Finally, we argue, as in the proof of part 1), that the balls $\bar{B}(\tilde{\mu} + \alpha(A_\epsilon)_{ii}, \epsilon)$ can be trapped inside $B(0,1)$ and are, consequently, also disjoint from $\bar{B}(1+\alpha\mu^\star, \alpha\epsilon)$. The

strong Geršgorin Theorem then implies that there are eigenvalues outside $\bar{B}(0,1)$, the number of them being at least the multiplicity of $\mu^\star$ (the number could be larger than the multiplicity of $\mu^\star$ because B_{11} can have eigenvalues other than $\mu^\star$ with positive real part). We begin with an analog of Lemma A.1.1.

Lemma A.2.1. *Let $\mu \in \mathbb{C}$ such that $Re(\mu) > 0$. Then, there exist ϵ_μ and α_μ such that, for $0 < \epsilon < \epsilon_\mu$ and $0 < \alpha < \alpha_\mu$,*

$$\bar{B}(1 + \mu\alpha, \alpha\epsilon) \subseteq \mathbb{C} \setminus \bar{B}(0,1).$$

Proof. The proof is very similar to that of Lemma A.1.1. The first observation is that if $x \in \mathbb{C} \setminus \bar{B}(0,1)$ and $\delta < d(x, S(0,1))$, then $\bar{B}(x,\delta) \subseteq \mathbb{C} \setminus \bar{B}(0,1)$ (draw a picture). Define $g(\alpha) = |1 + \alpha\mu|$ and observe that for $\alpha \geq 0$,

$$d(1 + \alpha\mu, S(0,1)) = g(\alpha) - g(0).$$

Similar to Lemma A.1.1,

$$g'(\alpha) = \frac{\alpha|\mu|^2 + \mathrm{Re}(\mu)}{|1 + \alpha\mu|}$$

and the Mean Value Theorem implies that

$$d(1 + \alpha\mu, S(0,1)) = g'(\xi)\alpha,$$

where $\xi \in (0, \alpha)$. Restrict $\alpha \in [0,1]$ and we obtain

$$g'(\alpha) \geq \frac{\mathrm{Re}(\mu)}{|1 + \mu|}.$$

The proof is now finished as in Lemma A.1.1, *i.e.,* let $\epsilon_\mu = \mathrm{Re}(\mu)/(2|1+\mu|)$ and $\alpha_\mu = 1$. □

proof of part 2) of Theorem 5.3.2. Let $\mu^\star \in \mathcal{U} := \{\text{eigenvalues of } B_{11}\}$ be such that $\mathrm{Re}(\mu^\star) > 0$, and let $\epsilon_{\mu^\star}$ and $\alpha_{\mu^\star}$ be defined as in Lemma A.2.1. Observe that a sufficient condition for $\bar{B}(x,\delta) \cap \bar{B}(y,\delta) = \emptyset$, where $x \neq y$, is that

$$\delta < \frac{|x - y|}{2}.$$

Let

$$\epsilon_1^\star = \min_{\mu\in\mathcal{U}:\mu\neq\mu^\star} \Big\{ \frac{|\mu-\mu^\star|}{2} \Big\}.$$

For $\alpha \neq 0$ and $\mu \neq \mu^\star$, $1+\alpha\mu \neq 1+\alpha\mu^\star$. Consequently, for $0 < \epsilon < \epsilon_1^\star$, $\alpha \neq 0$, and $\mu \neq \mu^\star$, we have that

$$\bar{B}(1+\alpha\mu, \alpha\epsilon) \cap \bar{B}(1+\alpha\mu^\star, \alpha\epsilon) = \emptyset.$$

Similar to Lemma A.1.1, let

$$\epsilon_2^\star = \min_{\tilde{\mu}\in\tilde{\mathcal{U}}} \Big\{ \frac{d(\tilde{\mu}, S(0,1))}{2} \Big\}.$$

Define

$$\epsilon^\star = \min\{\epsilon_{\mu^\star}, \epsilon_1^\star, \epsilon_2^\star\}.$$

Let $(\mathcal{A}_{\epsilon^\star})_{ii}$ be the numbers from Lemma A.0.1 and let $\alpha_2^\star$ be sufficiently small such that, for $0 < \alpha < \alpha_2^\star$, $\alpha(\mathcal{A}_{\epsilon^\star})_{ii} < \epsilon^\star$, for all i. Since $\epsilon^\star \leq \epsilon_2^\star \leq d(\tilde{\mu}, S(0,1))/2$, we obtain that, for $0 < \alpha < \alpha_2^\star$,

$$\bar{B}(\tilde{\mu} + \alpha(A_{\epsilon^\star})_{ii}, \epsilon^\star) \subseteq B(0,1).$$

Finally, let

$$\alpha^\star = \min\big\{\alpha_1^\star, \alpha_{\mu^\star}, \alpha(\epsilon^\star)\big\},$$

where $\alpha(\epsilon^\star)$ is defined in Lemma A.0.1. Observe that this choice ensures that for $0 < \alpha < \alpha^\star$,

1) $\bar{B}(1+\alpha\mu, \alpha\epsilon^\star) \cap \bar{B}(1+\alpha\mu^\star, \alpha\epsilon^\star) = \emptyset$, for $\mu \in \mathcal{U}$ with $\mu \neq \mu^\star$;

2) $\bar{B}(1+\mu^\star\alpha, \alpha\epsilon^\star) \subseteq \mathbb{C} \setminus \bar{B}(0,1)$;

3) $\bar{B}(\tilde{\mu} + \alpha(A_{\epsilon^\star})_{ii}, \epsilon^\star) \subseteq B(0,1)$, for all $\tilde{\mu} \in \tilde{\mathcal{U}}$.

Consequently, the ball $\bar{B}(1+\mu^\star\alpha, \alpha\epsilon^\star) \subseteq \mathbb{C} \setminus \bar{B}(0,1)$ is disjoint from all the others which implies the result from the strong Geršgorin Theorem and Lemma A.0.1. □

Appendix B

Proof of Remark 3.2.7 (Sublinear Convergence)

For any $x \in (0, \frac{1}{\sqrt{2}})$, a straightforward manipulation shows that $g(x) > \frac{1}{2}x$ and that $g(x) \in (0, \frac{1}{\sqrt{2}})$, implying that, for any $x^0 \in (0, \frac{1}{\sqrt{2}})$, the orbit generated by x^0 satisfies $x^{k+1} = g(x^k) > \frac{1}{2}x^k$. To conclude, note that

$$\begin{aligned}
K+1 &= \sum_{k=0}^{K} \frac{x^k - x^k + (x^k)^3}{(x^k)^3} = \sum_{k=0}^{K} \frac{x^k - x^{k+1}}{(x^k)^3} = \sum_{k=0}^{K} \frac{x^k - x^{k+1}}{(x^k)^3 - (x^{k+1})^3} \frac{(x^k)^3 - (x^{k+1})^3}{(x^k)^3} \\
&= \sum_{k=0}^{K} \frac{1}{(x^k)^2 + x^k x^{k+1} + (x^{k+1})^2} \Big(1 - \frac{(x^{k+1})^3}{(x^k)^3}\Big) \\
&\geq \sum_{k=0}^{K} \frac{1}{\frac{1}{4}(x^{k+1})^2 + \frac{1}{2}(x^{k+1})^2 + (x^{k+1})^2} \Big(1 - \frac{(x^{k+1})^3}{(x^k)^3}\Big) = \frac{4}{7} \sum_{k=0}^{K} \frac{1}{(x^{k+1})^2} \Big(1 - \frac{(x^{k+1})^3}{(x^k)^3}\Big) \\
&= \frac{4}{7}\Big(\sum_{k=0}^{K} \frac{1}{(x^{k+1})^2} - \frac{x^{k+1}}{(x^k)^3}\Big) = \frac{4}{7}\Big(\sum_{k=0}^{K} \frac{1}{(x^{k+1})^2} - \frac{x^k - (x^k)^3}{(x^k)^3}\Big) \\
&= \frac{4}{7}\Big(\sum_{k=0}^{K} \frac{1}{(x^{k+1})^2} - \frac{1}{(x^k)^2} + 1\Big) = \frac{4}{7}\Big(\frac{1}{(x^{K+1})^2} - \frac{1}{(x^0)^2} + K + 1\Big),
\end{aligned}$$

where the fourth equality is found by polynomial division, *i.e.*, dividing $y^3 - x^3$ by $y - x$, and the inequality follows from $x^{k+1} > \frac{1}{2}x^k$. After rearrangement, we obtain

$$(x^{k+1})^2 \geq \frac{1}{\frac{3}{4}(k+1) + \frac{1}{(x^0)^2}},$$

showing that the orbit generated by $x^0 \in (0, \frac{1}{\sqrt{2}})$ converges slower than linearly to zero.

Bibliography

[1] P. Combettes and J.-C. Pesquet. Fixed point strategies in data science. *IEEE Transactions on Signal Processing*, 69:3878–3905, 2021.

[2] Karl Pearson. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

[3] Sissi Xiaoxiao Wu, Hoi-To Wai, Lin Li, and Anna Scaglione. A review of distributed algorithms for principal component analysis. *Proceedings of the IEEE*, 106(8):1321–1340, 2018.

[4] Alexandros G Dimakis, Soummya Kar, José MF Moura, Michael G Rabbat, and Anna Scaglione. Gossip algorithms for distributed signal processing. *Proceedings of the IEEE*, 98(11):1847–1864, 2010.

[5] Jin-Jun Xiao, Alejandro Ribeiro, Zhi-Quan Luo, and Georgios B Giannakis. Distributed compression-estimation using wireless sensor networks. *IEEE Signal Processing Magazine*, 23(4):27–41, 2006.

[6] Stephen Boyd, Arpita Ghosh, Balaji Prabhakar, and Devavrat Shah. Randomized gossip algorithms. *IEEE transactions on information theory*, 52(6):2508–2530, 2006.

[7] Sergio Barbarossa and Gesualdo Scutari. Decentralized maximum-likelihood estimation for sensor networks composed of nonlinearly coupled dynamical systems. *IEEE Transactions on Signal Processing*, 55(7):3456–3470, 2007.

[8] Tong Zhao and Arye Nehorai. Information-driven distributed maximum likelihood estimation based on gauss-newton method in wireless sensor networks. *IEEE Transactions on Signal Processing*, 55(9):4669–4682, 2007.

[9] Ioannis D Schizas, Alejandro Ribeiro, and Georgios B Giannakis. Consensus in ad hoc wsns with noisy links—part i: Distributed estimation of deterministic signals. *IEEE Transactions on Signal Processing*, 56(1):350–364, 2007.

[10] Srdjan S Stanković, Miloš S Stankovic, and Dušan M Stipanovic. Decentralized parameter estimation by consensus based stochastic approximation. *IEEE Transactions on Automatic Control*, 56(3):531–543, 2010.

[11] Ali H Sayed. Diffusion adaptation over networks. In *Academic Press Library in Signal Processing*, volume 3, pages 323–453. Elsevier, 2014.

[12] Silvana Silva Pereira, Roberto López-Valcarce, and Alba Pages-Zamora. Parameter estimation in wireless sensor networks with faulty transducers: A distributed EM approach. *Signal Processing*, 144:226–237, 2018.

[13] Francisco de Lima Andrade, Mario Figueiredo, and Joao Xavier. Distributed Banach-Picard iteration for locally contractive maps. *IEEE Transactions on Automatic Control*, 2022.

[14] Francisco de Lima Andrade, Mario Figueiredo, and Joao Xavier. Distributed Picard iteration: Application to distributed EM and distributed PCA. *arXiv preprint arXiv:2106.10665*, 2021.

[15] Geoffrey J McLachlan and Thriyambakam Krishnan. *The EM algorithm and Extensions*, volume 382. John Wiley & Sons, 2007.

[16] Terence D Sanger. Optimal unsupervised learning in a single-layer linear feedforward neural network. *Neural networks*, 2(6):459–473, 1989.

[17] Tosio Kato. *Perturbation Theory for Linear Operators*. Springer, 2013.

[18] Doron Blatt, Alfred O Hero, and Hillel Gauchman. A convergent incremental gradient method with a constant step size. *SIAM Journal on Optimization*, 18(1):29–51, 2007.

[19] Huanyu Zhao, Ju H Park, Yulin Zhang, and Hao Shen. Distributed output feedback consensus of discrete-time multi-agent systems. *Neurocomputing*, 138:86–91, 2014.

[20] Francisco Facchinei and Jong-Shi Pang. *Finite-dimensional Variational Inequalities and Complementarity Problems*. Springer, 2007.

[21] Arpita Gang, Haroon Raja, and Waheed Bajwa. Fast and communication-efficient distributed PCA. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7450–7454, 2019.

[22] Daniel Fullmer and A Stephen Morse. A distributed algorithm for computing a common fixed point of a finite family of paracontractions. *IEEE Transactions on Automatic Control*, 63(9):2833–2843, 2018.

[23] Xiuxian Li, Min Meng, and Lihua Xie. A linearly convergent algorithm for multi-agent quasi-nonexpansive operators in real Hilbert spaces. In *59th IEEE Conference on Decision and Control (CDC)*, pages 4903–4908, 2020.

[24] Xiuxian Li and Lihua Xie. Distributed algorithms for computing a fixed point of multi-agent nonexpansive operators. *Automatica*, 122:109286, 2020.

[25] Angelia Nedic and Asuman Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54:48–61, 2009.

[26] Dušan Jakovetić, Joao Xavier, and J. Moura. Fast distributed gradient methods. *IEEE Transactions on Automatic Control*, 59:1131–1146, 2014.

[27] Dušan Jakovetić, Joao Xavier, and J. Moura. Linear convergence rate of a class of distributed augmented Lagrangian algorithms. *IEEE Transactions on Automatic Control*, 60:922–936, 2015.

[28] Wei Shi, Qing Ling, Gang Wu, and Wotao Yin. Extra: An exact first-order algorithm for decentralized consensus optimization. *SIAM Journal on Optimization*, 25(2):944–966, 2015.

[29] Jinming Xu, Shanying Zhu, Yeng Chai Soh, and Lihua Xie. Augmented distributed gradient methods for multi-agent optimization under uncoordinated constant stepsizes. In *54th IEEE Conference on Decision and Control (CDC)*, pages 2055–2060, 2015.

[30] Georgios B Giannakis, Qing Ling, Gonzalo Mateos, Ioannis D Schizas, and Hao Zhu. Decentralized learning for wireless communications and networking. In *Splitting Methods in Communication, Imaging, Science, and Engineering*, pages 461–497. Springer, 2016.

[31] Guannan Qu and Na Li. Harnessing smoothness to accelerate distributed optimization. *IEEE Transactions on Control of Network Systems*, 5(3):1245–1260, 2017.

[32] Angelia Nedic, Alex Olshevsky, and Wei Shi. Achieving geometric convergence for distributed optimization over time-varying graphs. *SIAM Journal on Optimization*, 27(4):2597–2633, 2017.

[33] Zheng Xu, Gavin Taylor, Hao Li, Mário Figueiredo, Xiaoming Yuan, and Tom Goldstein. Adaptive consensus ADMM for distributed optimization. In *International Conference on Machine Learning*, pages 3841–3850, 2017.

[34] Dušan Jakovetić. A unification and generalization of exact distributed first-order methods. *IEEE Transactions on Signal and Information Processing over Networks*, 5:31–46, 2018.

[35] Sulaiman A Alghunaim and Ali H Sayed. Linear convergence of primal–dual gradient methods and their performance in distributed optimization. *Automatica*, 117:109003, 2020.

[36] César A Uribe, Soomin Lee, Alexander Gasnikov, and Angelia Nedić. A dual approach for optimal algorithms in distributed optimization over networks. In *Information Theory and Applications Workshop*, pages 1–37, 2020.

[37] Guannan Qu and Na Li. Accelerated distributed nesterov gradient descent. *IEEE Transactions on Automatic Control*, 65(6):2566–2581, 2019.

[38] Fatemeh Mansoori and Ermin Wei. A general framework of exact primal-dual first-order algorithms for distributed optimization. In *58th IEEE Conference on Decision and Control (CDC)*, pages 6386–6391, 2019.

[39] Dušan Jakovetić, Dragana Bajović, Joao Xavier, and Jose Moura. Primal–dual methods for large-scale and distributed convex optimization and data analytics. *Proceedings of the IEEE*, 108(11):1923–1938, 2020.

[40] Puya Latafat, Nikolaos M Freris, and Panagiotis Patrinos. A new randomized block-coordinate primal-dual proximal algorithm for distributed optimization. *IEEE Transactions on Automatic Control*, 64(10):4050–4065, 2019.

[41] Puya Latafat, Lorenzo Stella, and Panagiotis Patrinos. New primal-dual proximal algorithm for distributed optimization. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pages 1959–1964. IEEE, 2016.

[42] Alireza Fallah, Mert Gurbuzbalaban, Asuman Ozdaglar, Umut Simsekli, and Lingjiong Zhu. Robust distributed accelerated stochastic gradient methods for multi-agent networks. *arXiv:1910.08701*, 2019.

[43] Paolo Di Lorenzo and Gesualdo Scutari. Next: In-network nonconvex optimization. *IEEE Transactions on Signal and Information Processing over Networks*, 2(2):120–136, 2016.

[44] Tatiana Tatarenko and Behrouz Touri. Non-convex distributed optimization. *IEEE Transactions on Automatic Control*, 62(8):3744–3757, 2017.

[45] Stefan Vlaski and Ali H Sayed. Distributed learning in non-convex environments—part I: Agreement at a linear rate. *IEEE Transactions on Signal Processing*, 69:1242–1256, 2021.

[46] Susana Faria and Gilda Soromenho. Fitting mixtures of linear regressions. *Journal of Statistical Computation and Simulation*, 80(2):201–225, 2010.

[47] A. Gang and W. Bajwa. A linearly convergent algorithm for distributed principal component analysis. available at arXiv:2101.01300, 2021.

[48] G. McLachlan and D. Peel. *Finite Mixture Models*. John Wiley & Sons, 2004.

[49] Soummya Kar and José MF Moura. Distributed consensus algorithms in sensor networks with imperfect communication: Link failures and channel noise. *IEEE Transactions on Signal Processing*, 57(1):355–369, 2008.

[50] Richard A Redner and Homer F Walker. Mixture densities, maximum likelihood and the EM algorithm. *SIAM review*, 26(2):195–239, 1984.

[51] Rolf Sundberg. Maximum likelihood theory for incomplete data from an exponential family. *Scandinavian Journal of Statistics*, 1(2):49–58, 1974.

[52] S. Balakrishnan, M. Wainwright, and B. Yu. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *Annals of Statistics*, 45(1):77–120, 2017.

[53] Robert D Nowak. Distributed em algorithms for density estimation and clustering in sensor networks. *IEEE transactions on signal processing*, 51(8):2245–2253, 2003.

[54] Pedro A Forero, Alfonso Cano, and Georgios B Giannakis. Distributed clustering using wireless sensor networks. *IEEE Journal of Selected Topics in Signal Processing*, 5(4):707–724, 2011.

[55] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

[56] John Adrian Bondy, Uppaluri Siva Ramachandra Murty, et al. *Graph theory with applications*, volume 290. Macmillan London, 1976.

[57] Roger A Horn and Charles R Johnson. *Matrix Analysis*. Cambridge university press, 2012.

[58] Christoforos N Hadjicostis, Alejandro D Domínguez-García, and Themistokis Charalambous. *Distributed averaging and balancing in network systems*. Now Foundations and Trends, 2018.

[59] Lin Xiao and Stephen Boyd. Fast linear iterations for distributed averaging. *Systems & Control Letters*, 53(1):65–78, 2004.

[60] Kazimierz Goebel and William A Kirk. *Topics in metric fixed point theory*. Number 28. Cambridge university press, 1990.

[61] James M Ortega and Werner C Rheinboldt. *Iterative Solution of Nonlinear Equations in Several Variables*. SIAM, 2000.

[62] James M Ortega and Werner C Rheinboldt. On a class of approximate iterative processes. Technical report, 1966.

[63] Andreas Hefti. A differentiable characterization of local contractions on banach spaces. *Fixed Point Theory and Applications*, 2015(1):1–5, 2015.

[64] Joseph P LaSalle. *The stability and control of discrete processes*, volume 62. Springer Science & Business Media, 2012.

[65] Jason D Lee, Ioannis Panageas, Georgios Piliouras, Max Simchowitz, Michael I Jordan, and Benjamin Recht. First-order methods almost always avoid strict saddle points. *Mathematical programming*, 176(1):311–337, 2019.

[66] Walter Rudin et al. *Principles of mathematical analysis*, volume 3. McGraw-hill New York, 1976.

[67] James M Ortega. Stability of difference equations and convergence of iterative processes. *SIAM Journal on Numerical Analysis*, 10(2):268–282, 1973.

[68] Vasile Berinde and F Takens. *Iterative Approximation of Fixed Points*, volume 1912. Springer, 2007.

[69] Morris W Hirsch, Stephen Smale, and Robert L Devaney. *Differential equations, dynamical systems, and an introduction to chaos*. Academic press, 2012.

[70] John Milnor. On the concept of attractor. In *The theory of chaotic attractors*, pages 243–264. Springer, 1985.

[71] Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in optimization*, 1(3):127–239, 2014.

[72] Marian Muresan and Marian Muresan. *A concrete approach to classical analysis*, volume 14. Springer, 2009.

[73] Kai Lai Chung. On a stochastic approximation method. *The Annals of Mathematical Statistics*, pages 463–483, 1954.

[74] Xinlong Weng. Fixed point iteration for local strictly pseudo-contractive mapping. *Proceedings of the American Mathematical Society*, 113(3):727–731, 1991.

[75] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

[76] Kun Yuan, Qing Ling, and Wotao Yin. On the convergence of decentralized gradient descent. *SIAM Journal on Optimization*, 26(3):1835–1854, 2016.

[77] Akhil Sundararajan, Bryan Van Scoy, and Laurent Lessard. A canonical form for first-order distributed optimization algorithms. In *2019 American Control Conference (ACC)*, pages 4075–4080. IEEE, 2019.

[78] Peter Lancaster. On eigenvalues of matrices dependent on a parameter. *Numerische Mathematik*, 6(1):377–387, 1964.

[79] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[80] Dimitri P Bertsekas. *Constrained Optimization and Lagrange Multiplier Methods*. Academic press, 2014.

[81] Norman Biggs, Norman Linstead Biggs, and Biggs Norman. *Algebraic graph theory*. Number 67. Cambridge university press, 1993.

[82] Francisco Facchinei and Jong-Shi Pang. *Finite-dimensional variational inequalities and complementarity problems*. Springer, 2003.

[83] Yongming Qu, George Ostrouchov, Nagiza Samatova, and Al Geist. Principal component analysis for dimension reduction in massive distributed data sets. In *Proceedings of IEEE International Conference on Data Mining (ICDM)*, volume 1318, page 1788, 2002.

[84] Yingyu Liang, Maria-Florina F Balcan, Vandana Kanchanapally, and David Woodruff. Improved distributed principal component analysis. *Advances in Neural Information Processing Systems*, 27:3113–3121, 2014.

[85] Ravi Kannan, Santosh Vempala, and David Woodruff. Principal component analysis and higher correlations for distributed data. In *Conference on Learning Theory*, pages 1040–1057. PMLR, 2014.

[86] Christos Boutsidis, David P Woodruff, and Peilin Zhong. Optimal principal component analysis in distributed and streaming models. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 236–249, 2016.

[87] Dan Garber, Ohad Shamir, and Nathan Srebro. Communication-efficient algorithms for distributed stochastic principal component analysis. In *International Conference on Machine Learning*, pages 1203–1212. PMLR, 2017.

[88] Zheng-Jian Bai, Raymond H Chan, and Franklin T Luk. Principal component analysis for distributed data sets with updating. In *International Workshop on Advanced Parallel Processing Technologies*, pages 471–483. Springer, 2005.

[89] Hillol Kargupta, Weiyun Huang, Krishnamoorthy Sivakumar, and Erik Johnson. Distributed clustering using collective principal component analysis. *Knowledge and Information Systems*, 3(4):422–448, 2001.

[90] Hairong Qi, Tsei-Wei Wang, and J Douglas Birdwell. Global principal component analysis for dimensionality reduction in distributed data mining. *Statistical data mining and knowledge discovery*, pages 327–342, 2004.

[91] Faisal N Abu-Khzam, Nagiza F Samatova, George Ostrouchov, Michael A Langston, and Al Geist. Distributed dimension reduction algorithms for widely dispersed data. In *IASTED PDCS*, pages 167–174, 2002.

[92] Anna Scaglione, Roberto Pagliari, and Hamid Krim. The decentralized estimation of the sample covariance. In *2008 42nd Asilomar Conference on Signals, Systems and Computers*, pages 1722–1726. IEEE, 2008.

[93] Yann-Aël Le Borgne, Sylvain Raybaud, and Gianluca Bontempi. Distributed principal component analysis for wireless sensor networks. *Sensors*, 8(8):4821–4850, 2008.

[94] Mehmet E Yildiz, Frank Ciaramello, and Anna Scaglione. Distributed distance estimation for manifold learning and dimensionality reduction. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3353–3356. IEEE, 2009.

[95] Wassim Suleiman, Marius Pesavento, and Abdelhak M Zoubir. Performance analysis of the decentralized eigendecomposition and esprit algorithm. *IEEE Transactions on Signal Processing*, 64(9):2375–2386, 2016.

[96] Satish Babu Korada, Andrea Montanari, and Sewoong Oh. Gossip pca. *ACM SIGMETRICS Performance Evaluation Review*, 39(1):169–180, 2011.

[97] Lin Li, Anna Scaglione, and Jonathan H Manton. Distributed principal subspace estimation in wireless sensor networks. *IEEE Journal of Selected Topics in Signal Processing*, 5(4):725–738, 2011.

[98] Ioannis D Schizas and Abiodun Aduroja. A distributed framework for dimensionality reduction and denoising. *IEEE Transactions on Signal Processing*, 63(23):6379–6394, 2015.

[99] Sissi Xiaoxiao Wu, Hoi-To Wai, Anna Scaglione, and Neil A Jacklin. The power-oja method for decentralized subspace estimation/tracking. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3524–3528. IEEE, 2017.

[100] A. Gang, B. Xiang, and W. Bajwa. Distributed principal subspace analysis for partitioned big data: Algorithms, analysis, and implementation. *IEEE Transactions on Signal and Information Processing over Networks*, 7:699–715, 2021.

[101] Jan R Magnus and Heinz Neudecker. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. John Wiley & Sons, 2019.

[102] Larry Wasserman. All of statistics: a concise course in statistical inference. 2013. *Cited on*, page 22.

[103] C. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[104] Madan Lal Mehta. *Random matrices*. Elsevier, 2004.

[105] Erich L Lehmann and George Casella. *Theory of point estimation*. Springer Science & Business Media, 2006.

[106] Whitney K Newey and Daniel McFadden. Large sample estimation and hypothesis testing. *Handbook of Econometrics*, 4:2111–2245, 1994.

[107] D. Bertsekas and S. Shreve. *Stochastic Optimal Control: the Discrete-time Case*. Athena Scientific, 1996.

[108] A. Van der Vaart. *Asymptotic Statistics*, volume 3. Cambridge University Press, 2000.