Contraction Maps and Applications to the Analysis of Iterative Algorithms

by

Emmanouil Zampetakis

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2017

Signature redacted

Author .....
Department of Electrical Engineering and Computer Science
January 31, 2017

Signature redacted

Certified by... .....
Constantinos Daskalakis
Associate Professor EECS
Thesis Supervisor

Signature redacted

Accepted by ... .
Professor Leslie A. Kolodziejski
Chair, Department Committee on Graduate Theses

Contraction Maps and Applications to the Analysis of Iterative Algorithms

by

Emmanouil Zampetakis

Submitted to the Department of Electrical Engineering and Computer Science
on January 31, 2017, in partial fulfillment of the
requirements for the degree of
Master of Science

Abstract

The increasing interest of the scientific community, and especially machine learning, on non-convex problems, has made non-convex optimization one of the most important and challenging areas of our days. Despite of this increasing interest too little is known from a theoretical point of view. The main reason for this is that the existing and well understood techniques used for the analysis of convex optimization problem are not applicable or meaningful in the non-convex case. The purpose of this thesis is to make a step in the direction of investigating a rich enough toolbox, to be able to analyze non-convex optimization.

Contraction maps and Banach's Fixed Point Theorem are very important tools for bounding the running time of a big class of iterative algorithms used to solve non-convex problems. But when we use the natural distance metric, of the spaces that we are working on, the applicability of Banach's Fixed Point Theorem becomes limited. The reason is that only few functions have the contraction property with the natural metrics. We explore how generally we can apply Banach's fixed point theorem to establish the convergence of iterative methods when pairing it with carefully designed metrics. Our first result is a strong converse of Banach's theorem, showing that it is a universal analysis tool for establishing uniqueness of fixed points and convergence of iterative maps to a unique solution.

We next consider the computational complexity of Banach's fixed point theorem. Making the proof of our converse theorem constructive, we show that computing Banach's fixed point theorem is CLS-complete, answering a question left open in the work of Daskalakis and Papadimitriou [23].

Finally, we turn to applications proving global convergence guarantees for one of the most celebrated inference algorithms in Statistics, the EM algorithm. Proposed in the 70's [26], the EM algorithm is an iterative method for maximum likelihood estimation whose behavior has vastly remained elusive. We show that it converges to the true optimum for balanced mixtures of two Gaussians.

Thesis Supervisor: Constantinos Daskalakis
Title: Associate Professor EECS

Acknowledgments

During the preparation of my thesis I was fortunate to work with many highly remarkable people. In this few lines I want to deeply thank them for all the knowledge that they gave me and all the nice moments that we spent together.

First I want to thank Prof. Constantinos Daskalakis who is supervising this work! Thanks to his valuable advices we could together do a great job and successfully completed most of the initial objectives that we had for this work. His guidance is invaluable and I feel very fortunate to have him as my advisor!

I owe special thanks to my colleague Christos Tzamos, for his constant support and help during my academic (and non-academic) life at MIT. It's his excitement and dedication in solving elegant and important problems, that made, and are still making, the research at MIT an enjoyable experience.

I also want to thank all of the members of the Theory of Computer Science lab at MIT for being very polite and supportive every time I need them. From the undergrad students to the Professors and secretaries, all create an academic enviroment very pleasant for work and research. Thank you all!

When it is time to thank families, I am very fortunate to have to thank two of them! The first one from Athens consisting of a huge number of people, is always supporting me during any step that I do in my life. It is so important to always feel that there is a place in the world with a great number of people that really care about you and will provide you a welcoming corner whenever you want. Very very special thanks to my parents Vassilis and Maria for everything that they have done for me. I will always be thanking you for the rest of my life!

The second one is the so called "Marney" family that we managed to create here in Boston with my roommates Ilias, Konstantinos and Vassilis. Master and Ph.D. would be a much more difficult experience without having created this very warm and hospitable environment, where everyone is there for everyone else! Living all together feels like a very nice and supportive family environment and surely not just sharing house with friends!

I can never forget the importance of a lot of close friends, siblings, cousins and Professors

in my life! I will try to name all of them and I hope I don't forget anyone! From my family I want to thank everybody: Vassilis, Maria, Tasos, Marina, Fotis, Eleni, Dimitris, Costas, Spiros, Vasso, Stelios, Fotis, Charalampos, Olympia, Eleni, Alexandra, Lambros, Maria, Costas, Olympia, Giorgos, Aggelos, Despoina, Katerina, Sofia, Costas, Constantinos. I want also to thank friends from Greece: Thanasis, Natalia, Markos, Thodoris, Manolis, Lydia, Irini, Sotiris, Fotis, Makis, Giannis, Giorgos, Alkisti, Vassia. I could not of course forget the friends that I made here in Boston and make my everyday life enjoyable: Artemis, Christos, Dimitris, Evgenia, G, Giorgos, Katerina, Konstantina, Kyriakos, Marieta, Madalina, Maryam, Nishanth, Sofi, Themis, Theodora, Vassilis, Vasso, Afrodite, Chara, Dimitris. Except from my parents, I want to especially thanks those who I'm spending more time with, talking, thinking and making dreams for the future: Eleni, Alexandra, Costas, Katerina, Artemis, Natalia, Thanasis, Marney Family, Markos, Madalina, Giorgakis Group, Alkisti, Thodoris.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The field of optimization is concerned with solving the problem:

$$\max_{x \in \mathcal{D}} \psi(x)$$

In this general version of the problem, without any restriction on the domain $\mathcal{D}$ and on ψ it is very easy to show that this problem is very difficult to solve [7]. This suggests that it is unavoidable to make some assumptions. A most helpful assumption that we can make is $\mathcal{D}$ to be a convex set and ψ to be a convex function. In this case the problem is called *convex optimization problem.* There is a very long line of work investigating and analyzing algorithms that solve *convex optimization problems* [9]. But in a lot of areas of science, the optimization problems that arise are not convex. This creates the area of *non-convex optimization.* In contrast with convex optimization, there is no general theory and tools for analyzing and finding theoretically proven algorithms for solving the non-convex problems. Actually, most of the techniques that we know are heuristic and sometimes we cannot even prove that they converge to local optima.

Non-convex optimization lies at the heart of some exciting recent developments in machine learning, optimization, statistics and signal processing. Deep networks, Bayesian inference, matrix and tensor factorization and dynamical systems are some representative examples where non-convex methods constitute efficient – and, in many cases, even more accurate – alternatives to convex ones. However, unlike convex optimization, these non-convex

approaches often lack theoretical justification.

The above facts have triggered attempts in providing answers to when and why non-convex methods perform well in practice in the hope that it might provide a new algorithmic paradigm for designing faster and better algorithms.

A diverse set of approaches have been devised to solve non-convex problems in a variety of approaches. They range from simple local search approaches such as gradient descent and alternating minimization to more involved frameworks such as simulated annealing, continuation method, convex hierarchies, Bayesian optimization, branch and bound, and so on. Moreover, for solving special classes of non-convex problems there are efficient methods such as quasi convex optimization, star convex optimization, submodular optimization.

The goal of this master thesis is to prove the generality of a technique in analyzing the performance of iterative heuristic algorithms. This technique is based on the notion of *contraction maps* and Banach's Fixed Point Theorem. We approach this proof of generality both from a pure mathematical point of view and from a complexity theoretic point of view.

In the second part of the thesis, inspired by the results of the previous part, we analyze a very well known heuristic algorithm for a non-convex problem, the EM algorithm. EM algorithm, is defined to solve a very important and generally non-convex problem, namely the *maximum-likelihood* maximization problem. We prove a positive result, analyzing EM for a paradigmatic case of finding the centers of a mixture of two Gaussians. This performance of EM even in this restricted case was an open problem since the definition of EM at 1977 [26].

1.1 Solution of a Non-Convex Optimization as a Fixed Point and the Basic Iterative Method

Working on a domain $\mathcal{D}$, the abstract goal of a huge class of algorithms is to find a point $x^* \in \mathcal{D}$ with some desired properties. In many cases these properties might be difficult to express. Sometimes even given a solution x^* there is no obvious way to verify that this is actually a solution. A common way to overcome these difficulties, is to express the solutions as *fixed points* of an easily described function. More formally we can define a function

$f : \mathcal{D} \mapsto \mathcal{D}$ such that the solution point $x^* \in \mathcal{D}$ satisfies $f(x^*) = x^*$. This way of expressing solutions is very common in a lot of scientific areas, e.g. equilibria in games [46], solutions of differential equations [14], a huge class of numerical methods [42].

Because of the importance of such a representation, a lot of interesting and important questions arise. Given a function $f : \mathcal{D} \mapsto \mathcal{D}$: is there any fixed point? is there a procedure that converges to this fixed point?

The first question, can be handled from important theorems in the field of topology called *Fixed Point Theorems.* Some of the most known once are : Brouwer's Fixed Point Theorem, Tikhonov's Fixed Point Theorem, Kakutani's Fixed Point Theorem and others [30].

The second question, while seemingly more difficult, has a very simple and intuitive candidate solution. If f has a fixed point and under some regularity conditions, like continuity, we can define the following sequence of point

$$x_{n+1} = f(x_n)$$

where the starting point x_0 can be picked arbitrarily. If (x_n) converges to a point $\bar{x}$ then

$$\lim_{n\to\infty} x_{n+1} = \lim_{n\to\infty} f(x_n) \Rightarrow \lim_{n\to\infty} x_{n+1} = f\left(\lim_{n\to\infty} x_n\right) \Rightarrow \bar{x} = f(\bar{x})$$

This observation means that a candidate procedure for computing such a fixed point is to *iteratively* apply the function f starting from an arbitrary point x_0. If this procedure converges we have the wanted fixed point x^*.

So one last question that we have to answer is whether this sequence (x_n) actually converges. One of the most known techniques to prove that (x_n) converges is to find a *potential function* ϕ. Usually a potential function, or Lyapunov function, is a lower bounded, real valued function that decreases with every application of f. More formally $\phi : \mathcal{D} \mapsto \mathbb{R}^+$ and $\phi(f(x)) < \phi(x)$. If we provide such a function and under some regularity conditions we can make sure that the sequence (x_n) converges and provides us with an algorithm that finds a fixed point of f. We will refer to this method of computing fixed points as the *Basic Iterative Method.*

1.1.1 An Algorithmic Point of View

The main performance guarantee of an algorithm is its running time. Therefore the first question that arises from a computer science perspective is: what is the running time of computing a fixed point and more specifically of the Basic Iterative Method?

We have seen that the potential function gives a general way of proving that the Basic Iterative Method converges. Moreover in the theory of dynamical systems there is a celebrated result called Conley's Decomposition Theorem [15]. One consequence of this work is that if a continuous analog of the Basic Iterative Method converges then there exist a potential function that can prove it. Therefore looking for potential functions in order to prove convergence does not restrict our power for proving convergence, because if there exists any argument to prove so there also exists a potential function argument.

But potential functions cannot tell us anything about the running time of the Basic Iterative Method. So one the main question that we would like to answer in this thesis is: is there a general way to *upper* and *lower* bound the running time of the Basic Iterative Method?

In the next sections we present our proposed direction for answering both the problems of *lower* and *upper* bounding the running time. Also in the last session we explain some important instantiations of the Basic Iterative Method that we don't know how to analyse and we hope we can get an answer after developing these general techniques.

1.2 Contraction Maps – Banach's Fixed Point Theorem

One of the main Fixed Point Theorems that we haven't mentioned yet is Banach's Fixed Point Theorem [5].

Informal Theorem (Banach's Fixed Point Theorem). *If there is a distance metric function d, such that $(\mathcal{D}, d)$ is a complete metric space and f is a contraction map with contraction constant $c \in (0,1)$ with respect to d, then f has a unique fixed point x^* and the convergence rate of the Basic Iterative Method with respect to d is c^n.*

The last sentence in the statement of the theorem implies that after n iterations of the Basic Iterative Method the distance of x_n from x^* decreases by factor c^n, i.e. $d(x_n, x^*) \leq$

$c^n d(x_0, x^*)$. There is a very nice algorithmic implication of this result. If we are only interested to find a point that is only ε-close with respect to d to x^* then the Basic Iterative Method will finish after $\log_c \varepsilon$ steps.

This theorem therefore provides a way to prove all of: existence of fixed point, uniqueness of fixed point, convergence of Basic Iterative Method and most importantly bounds the running time of the Basic Iterative Method.

The applications of this theorem are very important and distributed in a lot different subjects. One of the most celebrated ones is to prove the existence of a unique solution to differential equations through Picard's Theorem and also for bounding the running time of numerical methods that solve these differential equations [14], [42].

But how general can this contraction map argument be? Is there a sort of converse theorem like there is for the potential function argument?

The answer for most of the implications of Banach's Fixed Point Theorem is yes! There are converses of Banach's Fixed Point Theorems which prove the following [8].

Informal Theorem (Bessaga's Converse Fixed Point Theorem). *If f has a unique fixed point then for every constant $c \in (0,1)$ there exists a distance metric function d_c such that $(\mathcal{D}, d_c)$ is a complete metric space and f is a contraction map with contraction constant c with respect to d_c.*

The implication of this converse theorem is that if we want to prove existence and uniqueness of fixed points of f and convergence of the Basic Iterative Method then Banach's Fixed Point Theorem is the most general way to do it. This also proves that when we have only one fixed point x^*, there exists a potential function of the form $\phi(x) = d_c(x, x^*)$ where d_c is a distance metric function that makes $(\mathcal{D}, d_c)$ complete metric space.

But what are the actual implications for the complexity of computing a fixed point? The problem with the running time implications of the theorem, is that after $\log_c \varepsilon$ steps of the Basic Iterative Method we just have $d_c(x_n, x^*) \leq \varepsilon$. But it is not clear what is the natural meaning of d_c. It might not have any relation to the metric that we are interested in, i.e. the metric d for which we want a point x such that $d(x, x^*) \leq \varepsilon$. So this Converse Theorem cannot prove us the generality of the contraction mapping theorem for the analysis of the

running time of Basic Iterative method.

So our starting point instead of just the function f and the domain $\mathcal{D}$ should also be the complete distance metric function d that we are interested in. One step in this direction has been done by Meyers [44].

Informal Theorem (Meyers's Converse Fixed Point Theorem). *If $(\mathcal{D}, d)$ is a complete metric space, $\mathcal{D}$ is compact, f has a unique fixed point and the Basic Iterative Method converges then for any $c \in (0,1)$ there exists a distance metric function d_c equivalent with d such that $(\mathcal{D}, d_c)$ is a complete metric space and f is a contraction map according to d_c.*

The basic improvement in this theorem is that, instead of an arbitrary metric, it provides a metric equivalent with the metric that we started with. In order to do so Meyers's Converse Fixed Point Theorem has to assume that the Basic Iterative Method converges therefore we have to already have a proof of that in order to use it. Although this is a good step in the direction it is not enough in order to bound the number of steps needed by the Basic Iterative Method in order to get $d(x_n, x^*) \leq \varepsilon$.

Our basic goal for this section is to close this gap under the assumptions of Meyers's Converse Fixed Points Theorem. The basic observation that we have is that in order for a fixed point x^* to be stable there exists, usually, an open neighborhood of x^* of radius $\delta > 0$ where f is a contraction according to d. This allows us to define $d_c = d$ around this neighborhood. For the rest of the space we extend d_c in a way such that the contraction mapping condition is satisfied. This extension is inspired by the techniques that have been used for proving the mentioned Converse Fixed Point Theorems. After succeeding that we will have the guarantee that for any $\varepsilon < \delta$ the condition $d_c(x_n, x^*) \leq \varepsilon$ implies $d(x_n, x^*) \leq \varepsilon$. This will let us prove the generality of Banach's Fixed Point Theorem for upper bounding the running time of the Basic Iterative Method in a statement like the following

Informal Theorem. *Let $(\mathcal{D}, d)$ be a complete metric space and $f : \mathcal{D} \to \mathcal{D}$ be a self-made that has a unique fixed point x^* and every x converges to it. Then for any $c \in (0,1)$ there exists a complete distance metric d_c, such that f is a contraction with constant c with respect to d_c. Additionally, closeness of an arbitrary point x to x^* with respect to d_c implies closeness of x to x^* with respect to d.*

1.3 Computational Complexity of Fixed Points

Thus far we have discussed fixed point theorems, iterative methods, and the role of potential functions in establishing their convergence. We have also seen their interplay in Banach's fixed point theorem. In this section, we explore how the computational complexity of fixed points, potential function arguments and Banach's theorem are related. This question was one of the main motivations for a long line of research work starting with the papers of Johnson, Papadimitriou, Yannakakis [37] and Papadimitriou [48]. These papers define respectively the complexity class PLS, capturing the complexity of computing local optima of a given potential function ϕ, and the class PPAD, capturing the complexity of finding a fixed point of a continuous function f. Daskalakis and Papadimitriou [23] define the class CLS which relates to these classes as follows:

1. PLS [37]: when ϕ satisfies some continuity condition
2. PPAD [48]: when f satisfies some continuity condition
3. CLS [23]: when both ϕ and f satisfy some continuity condition

A main problem that has been left open in the work of [23] is which of these classes captures the complexity of computing a fixed point whose existence is guaranteed by Banach's Fixed Point Theorem. We will refer to this problem as BANACH. More precisely in [23] they have shown that BANACH $\in$ CLS but it was left open whether BANACH is complete for CLS.

Our main idea for solving this problem is to adjust the proofs of the Converse Fixed Point Theorems appropriately so that the procedure that constructs d_c becomes computationally efficient. This would be a reduction that shows the completeness of BANACH for CLS. Obviously one bottleneck here is that Banach's Fixed Point Theorem guarantees the existence of a unique fixed point whereas general CLS allows multiple fixed points to exist. We were able to overcome this difficulty by accepting as solutions points that don't satisfy the contraction property. This way we can also have a nice correspondence between the promise class that promises the existence of only one fixed point and the promise class that BANACH defines and guarantees the validity of the contraction map condition.

1.4 Runtime Analysis of EM Algorithm

In the field of *algorithms for inference* one of the most important problems, is finding the hypothesis that *maximizes the likelihood* from a predefined set of hypotheses. The problem of directly maximizing such an objective might be too difficult. Even to verify the optimality of the solution is not a trivial problem in this case. The only obvious way to solve it might be the brute force algorithm.

In [26], the authors follow the direction that we described in the Introduction defining a function f whose set fixed points includes the solutions to the maximum likelihood problem. More precisely if x^* is the hypothesis, from the set of hypothesis $\mathcal{D}$, that maximizes the likelihood function then $f(x^*) = x^*$. This function f is called the *EM Iteration* and the Basic Iterative Method of the EM Iteration is called *EM algorithm* [1] [2]. Since then, the EM algorithm became the main tool for computing the solution to the maximum likelihood problem with very good practical guarantees and running times.

Despite its practical importance, too little is known theoretically about the convergence of the algorithm. For example it is known that when the likelihood is a unimodular function then the EM algorithm converges [57]. Under some other conditions the convergence of EM algorithm is also understood [59]. Although a little has been known about the convergence of the EM algorithm, only recently did some results appear about the running time of the EM algorithm. These results apply only to some restricted cases which are nevertheless the paradigmatic applications of the EM algorithm.

In this part of our work, we provide general unconditional guarantees about the performance of the EM algorithm based on the techniques and the intuition that we developed in the work described in the previous sections. This achievement would is a great step in analyzing algorithms that use the Basic Iterative Method. This kind of algorithms are very common in the area of Machine Learning and EM is a very important and paradigmatic

[1] The name *EM* comes from the fact that f is a back to back application of an *Expectation* and a *Maximization* operators.

[2] Notice that there are fixed points which do not correspond to maximum likelihood solutions. Nevertheless in a lot of important instantiations, the region of attraction of these dummy fixed points is limited, as it has been observed in practice. Therefore choosing only a few random starting points we can be almost sure that we have found the appropriate region of attraction.

one. This analysis of this case of the EM algorithm, we hope that will help in the theoretical analysis of other important algorithms in machine learning.

1.4.1 Mixture of Two Gaussians with Known Covariance Matrices

Let us assume that we are getting samples from a mixture of two Gaussians in n dimensions and we know their covariance matrices. We would like to recover the means of the two Gaussians, that is to find the hypothesis about the means that maximizes the likelihood function. This is a well studied problem and although there is a growing number of theoretical guarantees to solve this problem [17], [38], [21], [51], in practice the most useful technique is to run the EM algorithm. Although the model is very restricted and the EM iteration has a nice form in this case, still too little is known about the theoretical performance of EM algorithm in this case. Very recently Balakrishnan, Wainwright and Yu [3] and Yang, Balakrishnan and Wainwright [60] were the first to find a way to prove that there is an ball B around the correct hypothesis x^* such that if $x_0 \in B$ then the EM algorithm converges to the correct solution and also with very good convergence rate.

Based on the intuition we have from what we explained at Section 2, we find a way to analyze EM algorithm when applied to the mixture of two Gaussians case. This way we provide the first unconditional theoretical guarantees for EM applied to the mixture of two Gaussians.

Chapter 2

Notation and Preliminaries

We start this chapter by defining the notation that we will use for the rest of the thesis. Then we also give and explain the basic definitions that span the basic concepts of topological spaces, metric spaces, continuity and computational complexity theory. While giving these basic definitions we also state and prove some basic results of the literature that we are going to use later.

2.0.2 Basic Notation and Definitions

– $\mathbb{R}$	set of real numbers.
– $\mathbb{R}_+$	set of non-negative real numbers.
– $\mathbb{R}^n$	n-dimensional Euclidean space.
– $\mathbb{R}^{n\times n}$	the space of real valued $n \times n$ matrices.
– A^T	the transpose of the real matrix A.
– $\mathbb{N}$	set of natural numbers.
– $\mathbb{N}_1$	set of natural numbers except 0.
– $A \setminus B$	all the set A apart from the members that belong to B too.
– A^c	the complement of the set A.
– $f^{[n]}$	n times composition f with it self, i.e. $\underbrace{f(f(\dots f(\cdot)))}_{n \text{ times}}$.
– $\|\cdot\|_p$	ℓ_p norm of a vector in $\mathbb{R}^n$.
– $\|\cdot\|_\Sigma$	Mahalanobis distance norm.
– $\mathcal{D}/\sim$	set of equivalence classes of the equivalence relation $\sim$ on a set $\mathcal{D}$.
– $\mathrm{Int}(S)$	interior of the set S.
– $\mathrm{Clos}(S)$	closure of the set S.
– $\mathrm{diam}_d[W]$	diameter of the set W with respect to the distance metric d.
– $B(x,r)$	open ball around x of diameter r.
– $\bar{B}(x,r)$	closed ball around x of diameter r.
– S^*	Kleene star of a set S.

Table 2.1: Basic Notation

A real valued function $g : \mathcal{D}^2 \to \mathbb{R}$ is called *symmetric* if $g(x,y) = g(y,x)$ and *anti-symmetric* if $g(x,y) = -g(y,x)$.

2.1 Set Theoretic Definitions

For this section we assume the reader is familiar with the basic set theoretic definitions. Based on those we define the notion equivalence relation and equivalence classes and we

define and explain the Axiom of Choice.

2.1.1 Equivalence Relations

A *relation* R on a set $\mathcal{X}$ is a subset of $\mathcal{X} \times \mathcal{X}$. An *equivalence relation* is a relation R that satisfies the following three properties

Reflexivity for all $x \in \mathcal{X}$, $(x, x) \in R$.
Symmetry for all $x, y \in \mathcal{X}$, $(x, y) \in R \Leftrightarrow (y, x) \in R$.
Transitivity for all $x, y, z \in \mathcal{X}$, $(x, y) \in R$ and $(y, z) \in R \implies (x, z) \in R$.

If $\sim$ is an equivalence relation on $\mathcal{X}$ and $x \in \mathcal{X}$, the set $E_x = \{y \mid y \in S \text{ and } x \sim y\}$ is called the *equivalence class* of x with respect to $\sim$. The set of all the equivalence classes of $\mathcal{X}$ for $\sim$ is

$$\mathcal{X}/\sim \;=\; \{E_x \mid x \in \mathcal{X}\}$$

2.1.2 Axiom of Choice

The importance of the *Axiom of Choice* to a huge range of pure mathematics can be indicated by the following sentence of an introduction to the Axiom of Choice from the Stanford Encyclopedia of Philosophy.

> The principle of set theory known as the *Axiom of Choice* has been hailed as "probably the most interesting and, in spite of its late appearance, the most discussed axiom of mathematics, second only to Euclid's axiom of parallels which was introduced more than two thousand years ago".

Axiom of Choice. *If A is a family of nonempty sets, then there is a function f with domain A such that $f(a) \in a$ for every $a \in A$. Such a function f is called a choice function for A.*

Axiom of Choice is known to be equivalent with a lot of very well known theorems, lemmas or principles. Some of them are: Zorn's Lemma, Well-ordering principle, Tychonoff's theorem and more. One such equivalent theorem is the Bessaga's theorem that we present, prove and explain in Chapter 3.

2.2 Topological and Metric Spaces

In this section we give the basic definitions and properties of topological and metric spaces that we are going to use in Chapter 3, when discussing about the converses of Banach's Fixed Point Theorem. Throughout this section we are working on a domain set $\mathcal{D}$ which is an arbitrary set. The material of this chapter is based on the notes by [41].

2.2.1 Topological Spaces

We first give the definition of a *topology* and then based on this we define *topological spaces.*

Definition 1. *Let $\mathcal{D}$ be a set and τ a collection of subsets of $\mathcal{D}$ with the following properties.*

(a) The empty set $\emptyset \in \tau$ and the space $\mathcal{D} \in \tau$.

(b) If $U_a \in \tau$ for all $a \in A$ then $\bigcup_{a \in A} Ua \in \tau$.

(c) If $U_j \in \tau$ for all $1 \leq j \leq n \in \mathbb{N}$, then $\bigcap_{j=1}^{n} U_j \in \tau$.

Then we say that τ is a topology on $\mathcal{D}$ and that (X, τ) is a topological space.

Example. If $\mathcal{D}$ is a set and $\tau = \{\emptyset, \mathcal{D}\}$, then τ is a topology. We call $\{\emptyset, \mathcal{D}\}$ the *indiscrete* topology on $\mathcal{D}$. The reason of the name will become clear when we will define the discrete metric and its induced topology.

If $(\mathcal{D}, \tau)$ is a topological space we define the notion of *open set* by calling the members of τ open sets. Now a subset C of $\mathcal{D}$ is called *closed* if $\mathcal{D} \setminus C$ is an open set, i.e. belongs to τ.

Definition 2. *Let $(\mathcal{D}, \tau)$ and $(\mathcal{X}, \tau)$ be topological spaces. A function $f : \mathcal{D} \to \mathcal{X}$ is said to be continuous if and only if $f^{-1}(U)$ is open in $\mathcal{D}$ whenever U is open in Y.*

Remark. We could start with closed sets the basic notion and then define topology, topological spaces and continuity with respect to collections of closed sets, instead of open sets. The two types of definitions are completely equivalent and for historical reasons the definition based on open sets is the normal one.

2.2.2 Interior and Closure

Definition 3. *Let $(\mathcal{D}, \tau)$ be a topological space and A a subset of $\mathcal{D}$. We write*

$$\text{Int}(A) = \bigcup\{U \in \tau \mid U \subseteq A\} \tag{2.1}$$

$$\text{Clos}(a) = \bigcap\{U closed \mid A \subseteq U\} \tag{2.2}$$

and we call $\text{Clos}(A)$ *the closure of A and* $\text{Int}(A)$ *the interior of A.*

We now give some useful, basic lemmas without proof. A proof of those can be found in [41].

Lemma 1. *(a)* $\text{Int}(A) = \{x \in A \mid \exists U \in \tau \text{ with } x \in U \subseteq A\}$.

(b) $\text{Int}(A)$ *is the unique $V \in \tau$ such that $V \subseteq A$ and if $W \in \tau$ and $V \subseteq W \subseteq A$, then $V = W$. In other words,* $\text{Int}(A)$ *is the largest open set contained in A.*

Lemma 2. *(a)* $\text{Clos}(A) = \{x \in \mathcal{D} \mid \forall U \in \tau \text{ with } x \in U, \text{ we have } U \cap A \neq \emptyset\}$.

(b) $\text{Clos}(A)$ *is the unique closed set V such that $A \subseteq V$ and if W is closed and $A \subseteq W \subseteq V$, then $V = W$. In other words,* $\text{Clos}(A)$ *is the smallest closed set containing in A.*

2.2.3 Metric Spaces

For the definition of metric spaces the most important is the definition of a *distance metric function.*

Definition 4. *Let $\mathcal{D}$ be a set and $d : \mathcal{D}^2 \to \mathbb{R}$ a function with the following properties:*

(i) $d(x, y) \geq 0$ *for all $x, y \in$ Domain.*

(ii) $d(x, y) = 0$ *if and only if $x = y$.*

(iii) $d(x, y) = d(y, x)$ *for all $x, y \in$ Domain.*

(iv) $d(x, y) \leq d(x, z) + d(z, x)$ *for all $x, y, z \in$ Domain. This is called triangle inequality.*

Then we say that d is a metric on $\mathcal{D}$ and $(\mathcal{D}, d)$ is a metric space.

Definition 5. *The diameter of a set $W \subseteq \mathcal{D}$ according to the metric d is defined as*

$$\text{diam}_d\,[W] = \max x, y \in W d(x, y)$$

A metric space $(\mathcal{D}, d)$ is called *bounded* if $\text{diam}_d[\mathcal{D}]$ is finite.

Definition 6. *If $\mathcal{D}$ is a set and we define $d_S : \mathcal{D}^2 \to \mathbb{R}$ by*

$$d_S(x,y) = \begin{cases} 0 & x = y \\ 1 & x \neq y \end{cases}$$

then d_S is called the discrete metric on $\mathcal{D}$.

Remark. It is very easy to see that discrete metric is indeed a metric, that is satisfies the conditions of Definition 4.

Now we can define the notion of continuity.

Definition 7. *Let $(\mathcal{D}, d)$ and $(\mathcal{X}, d')$ be metric spaces. A function $f : \mathcal{D} \to \mathcal{X}$ is called continuous if, given $x \in \mathcal{D}$ and $\varepsilon > 0$, we can find a $\delta(x, \varepsilon)$ such that*

$$d'(f(x), f(y)) < \varepsilon \text{ whenever } d(x,y) < \delta(x,\varepsilon)$$

Definition 8. *Let $(\mathcal{D}, d)$ be a metric space. We say that a subset $E \subseteq \mathcal{D}$ is open in $\mathcal{D}$ if, whenever $e \in E$, we can find a $\delta > 0$ (depending on e) such that*

$$x \in E \text{ whenever } d(x,e) < \delta$$

The next lemma connects the definition of open sets according to some metric with the definition of open sets in a topological space.

Lemma 3. *If $(\mathcal{D}, d)$ is a metric space, then the collection of open sets forms a topology.*

Example. If $(\mathcal{D}, d)$ is a metric space with the discrete metric, show that the induced topology consists of all the subsets of $\mathcal{D}$.

We define the *open ball* of radius r around x to be $B(x,r) = \{y \in \mathcal{D} | d(x,y) < r\}$.

2.2.4 Closed Sets for Metric Spaces

Definition 9. *Consider a sequence (x_n) in a metric space $(\mathcal{D}, d)$. If $x \in \mathcal{D}$ and, given $\varepsilon > 0$, we can find an integer $N \in \mathbb{N}_1$ (depending maybe on ε such that*

$$d(x_n, x) < \varepsilon \text{ for all } n \geq N$$

then we say that $x_n \to x$ as $n \to \infty$ and that x is the limit of the sequence (x_n).

Remark. It is easy to see that if a sequence has a limit then this limit is unique.

Definition 10. *Let $(\mathcal{D}, d)$ be a metric space. A set $G \subseteq \mathcal{D}$ is said to be closed if, whenever $x_n \in G$ and $x_n \to x$ then $x \in G$.*

Now that we have the notion of convergence of a sequence we can give the definition of compactness.

Lemma 4. *Let $(\mathcal{D}, d)$ be a metric space and A a subset of $\mathcal{D}$. Then* Clos(A) *consists of all those $x \in \mathcal{D}$ such that we can find (x_n) with $x_n \in A$ with $d(x_n, x) \to 0$.*

Definition 11. *A subset G of a metric space $(\mathcal{D}, d)$ is called compact if G is closed and every sequence in G has a convergent subsequence.*

A metric space $(\mathcal{D}, d)$ is called compact if $\mathcal{D}$ is compact and locally compact if for any $x \in \mathcal{D}$, x has a neighborhood that is compact.

We define the *closed ball* of radius r around x to be $\bar{B}(x, r) = \{y \in \mathcal{D} | d(x, y) \leq r\}$.

One of the most important and exiting applications of the definitions of metric spaces and continuity is the following fixed point theorem by Brouwer.

Theorem 1. *Let $S \subseteq \mathbb{R}^n$ be convex and compact. If $T : S \to S$ is continuous, then there exists a fixed point, i.e., there exists $x^* \in S$ such that*

$$T(x^*) = x^*$$

2.2.5 Complete Metric Spaces

We present the notion of complete metric spaces starting from the definition Cauchy sequences.

Definition 12. *If $(\mathcal{D}, d)$ is a metric space, we say that a sequence (x_n) in $\mathcal{D}$ is Cauchy sequence (or d-Cauchy sequence if the distance metric is not clear from the context) if, given $\varepsilon > 0$, we can find $N(\varepsilon) \in \mathbb{N}_1$ with*

$$d(x_n, x_m) < \varepsilon \text{ whenever } n, m \geq N(\varepsilon)$$

Definition 13. *A metric space $(\mathcal{D}, d)$ is complete if every Cauchy sequence converges.*

Definition 14. *Two metrics d, d' of the same set $\mathcal{D}$ are called topologically equivalent (or just equivalent) if for every sequence (x_n) in $\mathcal{D}$, (x_n) is d-Cauchy sequence if and only if it is d'-Cauchy sequence.*

2.2.6 Lipschitz Continuity

Definition 15. *Let $(\mathcal{D}, d)$ and $(\mathcal{X}, d')$ be metric spaces. A function $f : \mathcal{D} \to \mathcal{X}$ is Lipschitz continuous (or d-Lipschitz continuous if the distance metric is not clear from the context) if there exists a positive constant $\lambda \in \mathbb{R}_+$ such that for all $x, y \in \mathcal{D}$*

$$d'(f(x), f(y)) \leq \lambda d(x, y)$$

Lemma 5. *If a function $f : \mathcal{D} \to \mathcal{X}$ is Lipschitz continuous then it is continuous.*

Definition 16. *Let $(\mathcal{D}, d)$ and $(\mathcal{X}, d')$ be metric spaces. A function $f : \mathcal{D} \to \mathcal{X}$ is contraction (or d-contraction if the distance metric is not clear from the context) if there exists a positive constant $1 > c \in \mathbb{R}_+$ such that for all $x, y \in \mathcal{D}$*

$$d'(f(x), f(y)) \leq cd(x, y)$$

Definition 17. *Let $(\mathcal{D}, d)$ and $(\mathcal{X}, d')$ be metric spaces. A function $f : \mathcal{D} \to \mathcal{X}$ is non-expansion (or d-non-expansion if the distance metric is not clear from the context) if for all*

$x, y \in \mathcal{D}$

$$d'(f(x), f(y)) \leq d(x, y)$$

Definition 18. *Let $(\mathcal{D}, d)$ and $(\mathcal{X}, d')$ be metric spaces. A function $f : \mathcal{D} \to \mathcal{X}$ is a similarity (or a d-similarity if the distance metric is not clear from the context) if there exists a positive constant $1 > c \in \mathbb{R}_+$ such that for all $x, y \in \mathcal{D}$*

$$d'(f(x), f(y)) = cd(x, y)$$

Chapter 3

Converse Banach Fixed Point Theorems

In this chapter we state prove and analyze the existing inverses of the Banach's Fixed Point Theorems. We focus on the applications of these theorems on the analysis of the running time of the Basic Iterative Method. Finally we prove an inverse theorem that has richer meaning from an algorithmic point of view.

For this chapter we will assume that the domain $\mathcal{D}$ admits a topology τ according to which the function $f : \mathcal{D} \to \mathcal{D}$ is continuous.

3.1 Banach's Fixed Point Theorem

Before presenting the inverses we state and prove the original Banach Fixed Point Theorem. The following presentation and proofs follow [16].

Theorem 2 (Banach's Fixed Point Theorem). *If there is a distance metric function d, such that $(\mathcal{D}, d)$ is a complete metric space and f is a contraction map according to d, i.e.*

$$d(f(x), f(y)) \leq c \cdot d(x, y) \text{ with } c < 1 \tag{3.1}$$

then f has a unique fixed point x^* and the convergence rate of the Basic Iterative Method with respect to d is c^n.

Proof. First lets assume that there exist two different fixed points, x_1, x_2. Then by (3.1) we have

$$d(f(x_1), f(x_2)) \leq cd(x_1, x_2) \Rightarrow d(x_1, x_2) < d(x_1, x_2)$$

which gives a contradiction.

Now we let x_n be the sequence produced by the Basic Iterative Method on f starting from x_0. We have that

$$d(x_n, x_{n-1}) \leq c \cdot d(x_{n-1}, x_{n-2}) \leq c^2 \cdot d(x_{n-2}, x_{n-3}) \leq \cdots \leq c^n \cdot d(x_1, x_0)$$

Therefore the distance between x_n and x_{n-1} decreases an n increases. Using this property we can prove that (x_n) is a Cauchy sequence.

Let $N > n$ we get by triangle inequality

$$\begin{aligned} d(x_N, x_n) &\leq d(x_N, x_{N-1}) + \cdots + d(x_{n+1}, x_n) && (3.2) \\ &\leq c^N d(x_1, x_0) + c^{N-1} d(x_1, x_0) + \cdots + c^n d(x_1, x_0) && (3.3) \\ &\leq \frac{c^n}{1-c} d(x_1, x_0) && (3.4) \end{aligned}$$

Therefore for any $\varepsilon > 0$ we can pick M such that $c^M/(1-c) \leq \varepsilon$ and then the Cauchy property holds for any $n, N \geq M$. Since (x_n) is a Cauchy sequence we and the $\mathcal{D}$ is complete we have that (x_n) converges. Let x^* be the limit of (x_n), that is $x^* = \lim_{n\to\infty} x_n$.

Now from the previous chapter we know that 5 every Lipschitz continuous function is also continuous. Therefore f is continuous by (3.1). So

$$x_{n+1} = f(x_n) \Rightarrow \lim_{n\to\infty} x_{n+1} = f(\lim_{n\to\infty} x_n) \Rightarrow x^* = f(x^*)$$

Therefore x^* is the unique fixed point of f and the Basic Iterative Method converges to this fixed point. □

One interesting generalization of Banach's Fixed Point theorem is one given by Edelstein [28]. This generalization will become useful to get some counter examples later when we present the Converse Theorems. Also the techniques used in the proof are very useful for the rest of this thesis.

Theorem 3. *Let $(\mathcal{D}, d)$ be a compact metric space. If $f : \mathcal{D} \to \mathcal{D}$ satisfies*

$$d(f(x), f(y)) < d(x, y) \text{ for } x \neq y \in \mathcal{D}$$

then f has a unique fixed point in $\mathcal{D}$ and the fixed point can be found as the limit of $f^{[n]}(x_0)$ as $n \to \infty$ for any $x_0 \in \mathcal{D}$.

Proof. To show f has at most one fixed point in $\mathcal{D}$, suppose f has two fixed points $a \neq b$. Then $d(a, b) = d(f(a), f(b)) < d(a, b)$. This is impossible, so $a = b$.

To prove that f has actually a fixed point, we will look at the function $\mathcal{D} \to [0, \infty)$ given by $x \mapsto d(x, f(x))$. This measures the distance between each point and its f-value. A fixed point of f is where this function takes the value zero.

Since $\mathcal{D}$ is compact, the function $d(x, f(x))$ takes in $\mathcal{D}$ its minimum value: there is an $a \in \mathcal{D}$ such that $d(a, f(a)) \leq d(x, f(x))$ for all $x \in \mathcal{D}$. We will show by contradiction that a is a fixed point for f. If $f(a) \neq a$ then the hypothesis about f in the theorem says

$$d(f(a), f(f(a))) < d(a, f(a))$$

, which contradicts the minimality of $d(a, f(a))$ among all numbers $d(x, f(x))$. So $f(a) = a$.

Finally, we show for any $x_0 \in \mathcal{D}$ that the sequence $x_n = f^{[n]}(x_0)$ converges to a as $n \to \infty$. This can't be done as in the proof of the Banach's Fixed Point Theorem since we don't have the contraction constant to help us out. Instead we will exploit compactness.

If for some $k \geq 0$ we have $x_k = 1$ then $x_{k+1} = f(x_k) = f(a) = a$, and more generally $x_n = a$ for all $n \geq k$, so $x_n \to a$ since the terms of the sequence equal a for all large n. Now we may assume instead that $x_n \neq a$ for all n. Then

$$0 < d(x_{n+1}, a) = d(f(x_n), f(a)) < d(x_n, a),$$

so the sequence of numbers $d(x_n, a)$ is decreasing and positive. Thus it has a limit $\ell = \lim_{n\to\infty} d(x_n, a) \geq 0$. We will show $\ell = 0$. By compactness of X, the sequence $\{x_n\}$ has a convergent subsequence x_{n_i}, say $x_{n_i} \to y \in \mathcal{D}$. The function f is continuous, so $f(x_{n_i}) \to f(y)$, which says $x_{n_i+1} \to f(y)$ as $i \to \infty$. Since $d(x_n, a) \to \ell$ as $n \to \infty$,

$d(x_{n_i}, a) \to \ell$ and $d(x_{n_1+1}, a) \to \ell$ as $i \to \infty$. Since the metric, $d(x_{n_i}, a) \to d(y, a)$ and $d(x_{n_1+1}, a) = d(f(x_{n_i}), a) \to d(f(y), a)$. Having already shown these limits are ℓ,

$$d(y, a) = \ell = d(f(y), a) = d(f(y), f(a))$$

If $y \neq a$ the $d(f(y), f(a)) < d(y, a)$, but this contradicts the previous equation. So $y = a$, which means $\ell = d(y, a) = 0$. That shows $d(x_n, a) \to 0$ as $n \to \infty$. □

3.2 Bessaga's Converse Fixed Point Theorem

In this section we give the first result that tries to capture the generality of the Banach's Fixed Point Theorem. This result is due to Bessaga [8] and was first published in 1959. There are at least four different proofs of this result. The first one, due to Bessaga [8], uses a special form of the Axiom of Choice. This original version is not long, however, some statements are left to the reader for verifying.

The second proof, from Deimling's book [25] is a special case of that given by Wong [56], and it uses the Kuratowski–Zorn Lemma. In fact, Wong extended Bessaga's theorem to a finite family of commuting maps.

The third proof, due to Janos [36], is based on combinatorial techniques with a use of Ramsey's theorem. Actually, the existence of a separable metric is shown here (under the assumption that $\mathcal{D}$ has at most continuum many elements), though this metric need not be complete.

The last one, to the best of our knowledge, is due to Jachymski [35] and provides a nice direct connection with the existence of a potential function that decreases sufficiently at every step. Also, this proof enables us to get rid of the use of the Axiom of Choice when the metric space $\mathcal{D}$ is bounded and provides conditions under which f is a similarity.

We first state and prove a simple version of the theorem and we provide the first proof given by Bessaga in a simplified way. Then, we present the proof given by Jachymski and comment on the connection with the existence of a potential function. Finally we focus on the bounded case and the proof that is independent of the Axiom of Choice.

3.2.1 First Statement of Bassaga's Theorem

Formally, in the paper [8] Bessaga proved the following.

Theorem 4 (Bessaga's Converse Fixed Point Theorem). *Let $\mathcal{D} \neq \emptyset$ be an arbitrary set, $f : \mathcal{D} \to \mathcal{D}$ and $c \in (0,1)$. Then, if for all $n \in \mathbb{N}$, $f^{[n]}$ has a unique fixed point $x^* \in \mathcal{D}$, then there is a complete metric d for $\mathcal{D}$ such that $d(f(x), f(y)) \leq c \cdot d(x,y)$ for all $x, y \in \mathcal{D}$.*

This proof of Bessaga's theorem is the one found on the blog "Bubbles Bad; Ripples Good", in an article with title "Bessaga's converse to the contraction mapping theorem" by Willie Wong.

Proof. First we define an equivalence relation on $\mathcal{D}$. We say $x \sim y$ if there exists positive integers $p, q \in \mathbb{N}$ such that $f^{[p]}(x) = f^{[q]}(y)$. If $x \sim y$ we define

$$\rho(x,y) = \min\{p + q \mid f^{[p]}(x) = f^{[q]}(y)\} \text{ and}$$

$$\xi(x,y) = f^{(p)}(x) \text{ where } p \text{ is the value that attains } \rho(x,y)$$

We also define

$$\sigma(x,y) = \rho(x, \xi(x,y)) - \rho(y, \xi(x,y))$$

It is easy to prove that ρ is symmetric and σ antisymmetric.

Now, by the axiom of choice, there exists a choice function that chooses for each equivalence class of $\mathcal{D}/\sim$ a representative, this extends to a function $h : \mathcal{D} \to \mathcal{D}$ by setting the same value to all the members of the same equivalence class. We can now define a function $\lambda : X \to \mathbb{Z}$ by

$$\lambda(x) = \sigma(h(x), x)$$

We are now in a position to define our distance function d. Let $K = 1/c$.

If $x \sim y$, we define

$$d(x,y) = K^{-\lambda(x)} + K^{-\lambda(y)} - K \cdot K^{-\lambda(\xi(x,y))}$$

If $x \not\sim x^*$, we define

$$d(x, x^*) = K^{-\lambda(x)}$$

If $x \not\sim y$ and neither x, y is x^*, we define

$$d(x, y) = d(x, x^*) + d(y, x^*)$$

By the definition of d we can see that it is symmetric and non-negative. It is also easy to check that $d(x, y) = 0 \implies x \sim y$ and $x = y = \xi(x, y)$.

Triangle inequality involves a little bit more work, but most cases are immediately obvious except when $x \sim y \sim z$. Here we need to check that

$$K \cdot K^{-\lambda(y)} - K \cdot K^{-\lambda(\xi(x,y))} - K \cdot K^{-\lambda(\xi(y,z))} \geq -K \cdot K^{-\lambda(\xi(x,z))}$$

Suppose $f^{[p]}(x) = f^{[q_1]}(y)$ and $f^{[q_2]}(y) = f^{[r]}(z)$. Without loss of generality we can take $q_1 \geq q_2$. Then we have that $f^{[p]}(x) = f^{[r+q_1-q_2]}(z)$. This shows that $\lambda(\xi(x, z)) \leq \min(\lambda(\xi(x, y)), \lambda(\xi(y, z)))$. And this proves the inequality above.

Finally it is easy to see that any Cauchy sequence either is eventually constant, or must converge to x^* : if $x \neq y$ we have that

$$d(x, y) \geq 2^{-\min(\lambda(x),\lambda(y))-1} \geq \frac{1}{4}(d(x, \xi) + d(y, \xi))$$

and this shows that d is a complete metric. Now, it remains to verify that f is a contraction. Noting that $\lambda(f(x)) = \lambda(x) + 1$ and $\xi(f(x), f(y)) = \xi(x, y)$ we see easily that f is a contraction with Lipschitz constant $1/K = c$. □

3.2.2 Correspondence with Potential Function and Applications

In this section we present the work of Jachymski on a more simple proof of the Bessaga's theorem that also provides a connection with the existence of a potential function that decreases by a constant factor after every iteration of f. We start with the definition of Schröder functional inequality, that captures this behaviour of the potential function.

Definition 19. *We say that the function (potential)* $\phi : \mathcal{D} \to \mathbb{R}_+$ *satisfies the Schröder*

functional inequality for $f : \mathcal{D} \to \mathcal{D}$ *with constant* $c \in (0, 1)$ *if,*

$$\phi(f(x)) \leq c \cdot \phi(x) \tag{3.5}$$

The following lemma proves that the existence of a potential function ϕ that satisfies the Schröder functional inequality implies the existence of a complete metric d for which f is a contraction map.

Lemma 6. *Let* $f : \mathcal{D} \to \mathcal{D}$ *and* $c \in (0, 1)$. *The following statements are equivalent:*

i. there exists a complete metric d *for* $\mathcal{D}$ *such that*

$$d(f(x), f(y)) \leq c \cdot d(x, y) \text{ for all } x, y \in \mathcal{D}$$

ii. there exists a function $\phi : \mathcal{D} \to \mathbb{R}_+$ *such that* $\phi^{-1}(\{0\})$ *is a singleton and the Schröder functional inequality (3.5) holds.*

The forward direction of the lemma is easy to see and explain while gives a proof that a potential of the form $\phi(x) = d(x, x^*)$ always exists. The inverse direction is not that simple and is surprising because the intuition suggests that a complete metric has much more structure than a simple potential function.

Proof. i. $\implies$ ii. By the Banach's Fixed Point Theorem f has a fixed point x^* and we can easily see that the potential function $\phi(x) = d(x, x^*)$ satisfies the Schröder functional inequality (3.5).

ii. $\implies$ i. Define d by $d(x, y) = \phi(x) + \phi(y)$ if $x \neq y$ and $d(x, x) = 0$. It is easily seen that d is a metric for $\mathcal{D}$ and by (3.5) f is a contraction with respect to d. To see this we verify that the conditions of Definition 4 are satisfied.

(i). Since $\phi(x) \geq 0$ obviously $d(x, y) \geq 0$.

(ii). If $x \neq y$ then at least one of them is not equal to $x^* = \phi^{-1}(\{0\})$, let $x \neq x^*$. Obviously then $\phi(x) > 0$ and therefore $d(x, y) \neq 0$. The case $x = y$ is captured by the definition of d.

(iii). Obvious by the definition.

(iv). Since $\phi(z) \geq 0$ we have

$$d(x,y) = \phi(x) + \phi(y) \leq \phi(x) + \phi(y) + 2 \cdot \phi(z) = d(x,z) + d(z,y)$$

Finally we have to prove that the metric space $(\mathcal{D}, d)$ is complete. Let (x_n) be a Cauchy sequence. We may assume that the set $\{x_n : n \in \mathbb{N}\}$ is infinite, otherwise (x_n) contains a constant subsequence and then (x_n). Then there is a subsequence (x_{k_n}) of distinct elements so that

$$d(x_{k_n}, x_{k_m}) = \phi(x_{k_n}) + \phi(x_{k_m}) \quad \text{for } n \neq m$$

Now since (x_n) is a Cauchy sequence we know that $d(x_{k_n}, x_{k_m}) \to 0$ as n goes to $+\infty$ and m remains to the same distance from n. Therefore since $\phi(\cdot) \geq 0$ we also get that $\phi(x_{k_n}) \to 0$ as n goes to infinity. But by the assumption of ii. we have that $\phi(z) = 0$ only for a unique $z \in \mathcal{D}$. Therefore $d(x_{k_n}, z) = \phi(x_{k_n})$ which means that $d(x_{k_n}, z) \to 0$ and so (x_n) converges to z. □

The above lemma can be used in order to prove the Bessaga's theorem as stated before.

3.2.3 Necessary and Sufficient Conditions for Similarity

Given the conditions of the Bessaga's Theorem and that f is injective we can prove not only that f is a contraction with respect to some complete metric of $\mathcal{D}$ but also that it is a similarity.

Theorem 5. *Let $f : \mathcal{D} \to \mathcal{D}$ and $c \in (0,1)$. The following statements are equivalent.*

(a) f is injective and f has a unique periodic point,

(b) f is an c-similarity with respect to some complete metric d.

One interesting question is what happens when the function f does not have any fixed points. Surprisingly enough there is a version of Lemma 6 that captures the case where f has no fixed point which we present next. Obviously if a function has more that one fixed point then any contraction or similarity conditions are impossible. To see this assume that x_1^*, x_2^* are fixed points of f. Then $d(f(x_1^*), f(x_2^*)) = d(x_1^*, x_2^*)$ and so the Lipschitz constant of f cannot be anything less than one.

Lemma 7. *Let $f : \mathcal{D} \to \mathcal{D}$ and $c \in (0,1)$. The following statements are equivalent:*

i. f has no periotic points

ii. the Schröder equation $\phi(f(x)) = c \cdot \phi(x)$ has solution $\phi : \mathcal{D} \to (0, \infty)$

This lemma proves the following two theorems on the existence of metrics that make f contraction or similarity.

Theorem 6. *Let $\mathcal{D} \neq \emptyset$ be an arbitrary set, $f : \mathcal{D} \to \mathcal{D}$ and $c \in (0,1)$. Then if $f^{[n]}$ has at most one fixed point for every $n \in \mathbb{N}$, then there exists a metric d such that $d(f(x), f(y)) \leq c \cdot d(x,y)$ for all $x, y \in \mathcal{D}$.*

Observe, that this theorem does not guarantee that the topological space $(\mathcal{D}, d)$ is complete. Therefore the natural meaning of d is very limited. But without the hypothesis that there is actually a fixed point we couldn't hope for something stronger since then the application of Banach's fixed point theorem would actually prove the existence of a fixed point. This cannot be true since there are functions without any fixed point. For example for $\mathcal{D} = \mathbb{R}$ the function $f(x) = x + 1$. This theorem proves that even those function can be viewed as contraction maps!

Similarly we take a result on similarity

Theorem 7. *Let $f : \mathcal{D} \to \mathcal{D}$ and $c \in (0,1)$. The following statements are equivalent.*

(a) f is injective and f has no periodic points,

(b) f is an c-similarity with respect to some metric d for $\mathcal{D}$.

To prove this last theorem we use exactly the same proof we used for Theorem 6 and we replace the inequality with equality.

3.2.4 Avoiding Axiom of Choice

The proofs of Bessaga's theorem are all based on the Axiom of Choice. In his initial work Bessaga proved that this is unavoidable, since Bessaga's Theorem is equivalent to some version of the Axiom of Choice. In this section we show how we can escape the use of Axiom of Choice by putting some restrictions on the domain. More specifically we are going to assume that the domain is *bounded.* This analysis is based on the paper of Jachymski [35].

Theorem 8. *Let $f : \mathcal{D} \to \mathcal{D}$ and $c \in (0,1)$. The following statements are equivalent.*

(a) the intersection $\bigcap_{n\in\mathbb{N}} f^{[n]}(\mathcal{D})$ is a singleton,

(b) the Schröder inequality (3.5) has a bounded solution $\phi : \mathcal{D} \to \mathbb{R}_+$ such that $\phi^{-1}(\{0\})$ is a singleton.

(c) there exists a complete and bounded metric d for $\mathcal{D}$ such that f is contraction with constant c with respect to d.

Proof. (a) $\implies$ (b). Let $\bigcap_{n\in\mathbb{N}} f^{[n]}(\mathcal{D}) = \{x^*\}$. For $x \neq x^*$ define

$$n(x) = \sup_{n\in\mathbb{N}_1} \{n \mid x \in f^{[n]}(\mathcal{D})\}$$

Since the sequence $(f^{[n]}(\mathcal{D}))$ is decreasing, condition (a) implies that $n(x)$ is finite. Define the function ϕ as follows

$$\phi(x^*) = 0 \quad , \quad \phi(x) = c^{n(x)} \text{ for } x \neq z$$

Clearly ϕ is bounded and $\phi^{-1}(\{0\}) = \{z\}$. Fix an $x \in \mathcal{D}$. If $f(x) = x^*$ then (3.5) holds. So let $f(x) \neq x^*$. Then $n(f(x)) \geq n(x) + 1$ and hence

$$\phi(f(x)) = c^{n(f(x)} \leq c^{n(x)+1} = a\phi(x)$$

Thus (b) holds.

(b) $\implies$ (c). This can be proved by repeating the proof of Lemma 6.

(c) $\implies$ (a). Direct application of the Banach Fixed Point Theorem. □

3.2.5 An non-intuitive application of Bessaga's Theorem

The initial version of the Bessaga's Theorem does not have any conditions on the convergence of f to the unique fixed point. This means that even if the fixed point is unstable the theorem guarantees the existence of a complete metric that makes the function f contraction.

For example lets take $f(x) = 2x$. This function is defined on the domain $\mathcal{D} = \mathbb{R}$ and has only one fixed point at $x = 0$. It is obvious that this fixed point is an unstable one,

since the Basic Iterative Method for f goes exponentially fast to infinity, i.e. $f^{[n]} = 2^n x$. But Bessaga's theorem suggests that $\mathbb{R}$ admit a complete metric that makes f contraction! Then by applying the Banach's Fixed Point Theorem we get that $(f^{[n]})$ converges to this fixed point according to this metric! In order to understand how this can happen, in this section we construct a metric d that has this properties. Bessaga's theorem only guarantees the existence of such a metric d.

One very good starting point when looking for a metric that satisfies some properties is to compute a natural metric of the space that we are working on after applying some arbitrary function to the arguments. Formally let d be a natural metric for a domain $\mathcal{D}$, then we define the new metric $d'(x, y) = d(h(x), h(y))$ where $h : \mathcal{D} \to \mathcal{D}$ is an injective function. Once h is injective is easy to see that d' defines a metric on $\mathcal{D}$.

For our case with $\mathcal{D} = \mathbb{R}$ and $f(x) = 2x$ we can see that we want h to be a decreasing function. Therefore we pick $h(x) = 1/x$ for $x \neq 0$ and $h(0) = 0$. Then we get that

$$d'(f(x), f(y)) = |h(2x) - h(2y)| = \left|\frac{1}{2x} - \frac{1}{2y}\right| = \frac{1}{2}\left|\frac{1}{x} - \frac{1}{y}\right| = \frac{1}{2}d'(x, y)$$

Therefore f is a contraction with constant $1/2$ with respect to the metric d'.

3.3 Meyers's Converse Fixed Point Theorems

As we have seen in the last part of the previous section, the guaranteed by Bessaga's theorem distance metric, that makes a function f contraction might not have clear natural meaning. It might not have any relation to the topology of the domain $\mathcal{D}$. It is this natural metric d, that produces the correct topology, that we would be interested in. In this metric d finding a point x such that $d(x, x^*) \leq \varepsilon$ has some meaning. So Bessaga's Theorem cannot prove us the generality of the contraction mapping theorem for the analysis of the running time of Basic Iterative method, as described in the introductory chapter.

So our starting point instead of just the function f and the domain $\mathcal{D}$ should also be the complete distance metric function d that we are interested in. One step towards this direction has been done by Meyers [44, 43] and this work we are presenting in this section.

Theorem 9 (First Meyers's Converse Fixed Point Theorem). *If $(\mathcal{D}, d)$ is a complete metric*

space, $\mathcal{D}$ is locally compact and the following hold :

1. *f has a unique fixed point x^**
2. *for every $x \in \mathcal{D}$, the sequence $(f^{[n]}(x))$ converges to x^* with respect to d*

then for any $c \in (0,1)$ there exists a distance metric function d_c topologically equivalent with d such that $(\mathcal{D}, d_c)$ is a complete metric space and

$$d_c(f(x), f(y)) \leq c \cdot d_c(x, y) \qquad \text{for all } \; x, y \in \mathcal{D}$$

This theorem has a very important improvement compared to Bessaga's one. The distance metric d_c that it produces is topologically equivalent with the initial given metric d. This means that d_c and d produce the same topology and therefore they capture the same basic structural properties of the space $\mathcal{D}$.

The second interesting result of Meyers's describes one different aspect, which is the *local to global contraction.* More precisely suppose that we have a distance metric that makes f contraction only when x, y are close enough. The question how can we get a metric that makes f globally a contraction map? This question is what this second theorem by Meyer tries to capture. We first give the definition of local contraction and then we state the theorem by Meyers's.

Definition 20. *Let $(\mathcal{D}, d)$ be a complete metric space. A function $f : \mathcal{D} \to \mathcal{D}$ is a local contraction if there exist real-valued functions $\mu(x), c(x)$, with $\mu(x) > 0$ and $0 < c(x) < 1$, such that whenever y, z are in the ball*

$$B(x, \mu(x)) = \{u \mid d(x) \leq \mu(x)\}$$

it follows that

$$d(f(y), f(z)) \leq c(x) d(y, z)$$

Theorem 10 (Second Meyers's Converse Fixed Point Theorem). *If $(\mathcal{D}, d)$ is a complete metric space, $\mathcal{D}$ is locally compact and f is a local contraction then there exists a distance metric function d' topologically equivalent with d such that $(\mathcal{D}, d')$ is a complete metric space*

and

$$d'(f(x), f(y)) \leq c \cdot d'(x, y) \qquad \text{for some } \; c \in (0, 1)$$

The proofs of the above results are very interesting and innovative. We choose not to present them here because we follow the path of these proofs in the next chapter where we present one more precise Converse Fixed Point Theorem.

3.4 A New Converse to Banach's Fixed Point Theorem

In this section we present our basic result and improvement on the converses of Banach's Fixed Point Theorem. To proceed we see that the basic improvement of Meyers's theorem is that, instead of an arbitrary metric, it provides a metric equivalent with the metric that we started with. Although this is a good step, it is not enough in order to bound the number of steps needed by the Basic Iterative Method in order to get $d(x_n, x^*) \leq \varepsilon$.

Our result in this section closes this gap under the assumptions of Meyers's Converse Fixed Points Theorem. The main technical idea is that there is a way to change the proof of Meyers's Theorem such that we can get a distance metric d_c with the property $d_c(x, y) \geq d(x, y)$ everywhere except maybe from the region $d(x, x^*) \leq \varepsilon$. This implies that if we guarantee that $d_c(x_n, x^*) \leq \varepsilon$ then $d(x_n, x^*) \leq \varepsilon$.

We start by proving the result and then we discuss on the implication and the corollaries that we can get based on this.

Theorem 11. *If $(\mathcal{D}, d)$ is a complete metric space, $\mathcal{D}$ is locally compact and the following hold :*

1. *f has a unique fixed point x^**
2. *for every $x \in \mathcal{D}$, the sequence $(f^{[n]}(x))$ converges to x^* with respect to d*

then for any $c \in (0, 1)$ and any $\varepsilon > 0$ there exists a distance metric function d_c topologically equivalent with d such that $(\mathcal{D}, d_c)$ is a complete metric space and

$$d_c(f(x), f(y)) \leq c \cdot d_c(x, y) \qquad \textit{for all} \;\; x, y \in \mathcal{D} \tag{3.6a}$$

$$d_c(x, y) \leq \varepsilon \implies x \in \bar{B}(x^*, 2\varepsilon) \text{ or } y \in \bar{B}(x^*, 2\varepsilon) \tag{3.6b}$$

Proof. The construction of d_c starts with an open neighborhood of x^* with some desired properties. In order to satisfy (3.6b), this open neighborhood W must have $\mathrm{diam}_d[W] \leq \varepsilon$.

Lemma 8. *There exists an open neighborhood W of x^* such that We first prove that there exists an open neighborhood W of x^* such that*

$$f^{[n]}(W) \to \{x^*\} \tag{3.7a}$$

$$f(W) \subset W \tag{3.7b}$$

$$\mathrm{diam}_d[W] \leq \varepsilon \tag{3.7c}$$

Proof. We start by showing (3.7a) for an open neighborhood U of x^* with $\mathrm{diam}_d[U] \leq \varepsilon$. Let $C = \{x : d(x, x^*) \leq \varepsilon\}$. Observe that since $\mathcal{D}$ is locally compact and complete the set C is a compact neighborhood of x^*. We define $U = \mathrm{Int}(C)$, an open neighborhood of x^*. Consider any other open neighborhood V of x^*. For each $x \in C$, there exists by the hypothesis 2. of the theorem, a smallest $n(x)$ such that $f^{[n]}(x) \in V$ for all $n \geq n(x)$. We need only to show that

$$n(V) = \sup_{x \in C} n(x)$$

is finite. For the sake of contradiction we assume its not, then C contains a sequence (x_i) such that $n(x_i) \geq i$, and since C is compact, we may assume $x_i \to y$ for some $y \in C$. If this is not the case then we can take any converging subsequence of (x_i) and will satisfy the same properties. The desired contradiction follows by observing that $n(y) < \infty$ and that by continuity of f it holds that $n(x) \leq n(y) + 1$ for all x in some neighborhood of y. Therefore there exist an open neighborhood U such that $f^{[n]}(U) \to \{x^*\}$.

Now starting from U we prove the existence of W. For this, we will prove that there exists an open neighborhood W of x^* such that $f(W) \subset W$ and $W \subset U$. The latter implies $f^{[n]}(W) \to \{x^*\}$ and $\mathrm{diam}_d[W] \leq \varepsilon$.

Since $f^{[n]}(U) \to \{x^*\}$, there is an integer k such that $f^{[k]}(U) \subset U$. Let

$$W = \bigcap_{j=0}^{k-1} f^{[-j]}(U) \subset U$$

Then for $x \in W$ we have, for $1 \leq j \leq k-1$, $x \in f^{[-j]}(U)$ and thus $f(x) \in f^{[-(j-1)]}(U)$. Moreover $x \in U$, so that $f^{[k]}(x) \in f^{[k]}(U) \subset U$ and thus $f(x) \in f^{[-(k-1)]}(U)$. Hence $x \in W$ implies $f(x) \in W$, which was to be shown. ■

We now proceed to the main line of the proof. The construction follows three steps:

I. construction of a metric d_M, topologically equivalent to d, with respect to which f is non-expanding.

II. given d_M we proceed with the construction of a function ρ_c with all the desired properties except maybe from the triangle inequality.

III. given ρ_c we construct the final wanted metric d_c by defining the ρ_c-geodesic distance.

I. Construction of d_M

We set

$$d_M(x,y) = \max_{n\in\mathbb{N}}\{d(f^{[n]}(x), f^{[n]}(y))\}$$

The fact that this maximum is finite can be proved using the condition 2. of the theorem. Indeed, since $d(f^{[n]}(x), x^*) \to 0$ and $d(f^{[n]}(y), x^*) \to 0$, for any $\delta > 0$ there is a number $N \in \mathbb{N}$ such that $d(f^{[n]}(x), x^*) \leq \delta$ and $d(f^{[n]}(y), x^*) \leq \delta$ for all $n > N$. Now if let $\delta = d(x,y)$ we get that $\max_{n\geq N}\{d(f^{[n]}(x), f^{[n]}(y)) \leq d(x,y)$ and therefore $\max_{n\in\mathbb{N}}\{d(f^{[n]}(x), f^{[n]}(y)) = \max_{0\leq n\leq N}\{d(f^{[n]}(x), f^{[n]}(y))$. Hence the maximum has a finite value.

Observe now that by definition the following is obvious

$$d_M(f(x), f(y)) \leq d_M(x,y)$$

Therefore it only remains to prove that this function satisfies the properties of a distance metric function. The positive definiteness and symmetry of d_M follow from the corresponding properties of d. The fact that $d_M(x,y) \neq 0$ for $x \neq y$ follows from the fact that $d(x,y) \leq d_M(x,y)$, which follows directly from the definition of d_M since $f^{[0]}(x) = x$. It remains to prove the triangle inequality.

For the triangle inequality we observe by the definition of d_M that there exists $n \in \mathbb{N}$

such that

$$d_M(x,y) = d(f^{[n]}(x), f^{[n]}(\dot{z})) \leq \tag{3.8}$$

$$\leq d(f^{[n]}(x), f^{[n]}(y)) + d(f^{[n]}(y), f^{[n]}(x)) \leq \tag{3.9}$$

$$\leq d_M(x,y) + d_M(y,z) \tag{3.10}$$

Thus d_M is indeed a metric, which must be shown to be topologically equivalent to d.

From the inequality $d(x,y) \leq d_M(x,y)$ it follows that any d_M-convergent sequence is also d-convergent, with the same limit point. To prove the implication in the opposite direction, note that (3.7a) implies the existence for each $\eta > 0$ of an N such that

$$\mathrm{diam}_d \left[f^{[n]}(W)\right] < \varepsilon \quad \text{for } n > N$$

For each $x \in \mathcal{D}$, it follows from 2. that

$$\nu(x) = \min_{n \in \mathbb{N}, f^{[n]}(x) \in W} \{n\} \tag{3.11}$$

is finite. Since f is continuous, there is an $\delta > 0$ so small that $d(x,y) < \delta$ implies

$$f^{[\nu(x)]}(y) \in W \text{ and } d(f^{[j]}(x), f^{[j]}(y)) < \delta \text{ for } 0 \leq j \leq N + \nu(x) \tag{3.12}$$

By (b) $f^{[n+N+\nu(x)]}(x) \in f^{[n+N]}(W)$ and $f^{[n+N+\nu(x)]}(y) \in f^{[n+N]}(W)$ for all $n > 0$, so that the (3.12) implies

$$d(f^{[j]}(x), f^{[j]}(y)) < \delta \quad \text{for } j > N + \nu(x)$$

Thus $d(x,y) \leq \delta$ implies $d_M(x,y) \leq \eta$. This shows that a sequence which is d-convergent to x is also d_M-convergent to x, completing the proof of topological equivalence. Finally since d and d_M are topologically equivalent and d is complete for $\mathcal{D}$ it follows that d_M is also complete for $\mathcal{D}$.

II. Construction of ρ_c

We begin by defining K_n to be the closure of $f^n(W)$ for $n \geq 0$, and $K_{(-n)} = f^{[-n]}(K_0)$, so

that (3.7a) implies

$$K_n \to \{x^*\} \qquad \text{as} \quad n \to \infty \tag{3.13}$$

For $x \in K_0 \setminus \{x^*\}$, set

$$n(x) = \max_{x \in K_n}\{n\} \geq 0$$

finiteness is assured by (3.13). Let also $n(x^*) = \infty$, and for $x \in \mathcal{D} \setminus K_0$ set

$$n(x) = -\min_{f^{[m]}(x) \in K_0}\{m\} = \max_{x \in K_n}\{n\} < 0$$

which must exists by condition 2. Letting also $\kappa(x, y) = \min\{n(x), n(y)\}$, we define ρ_c to be

$$\rho_c(x, y) = c^{\kappa(x,y)} d_M(x, y)$$

We can now prove that ρ_c satisfies all the distance metric requirements except maybe triangle inequality. Positive definiteness and symmetry is obvious. Also since d_M is a distance metric and $\kappa(x, y) \geq 0$ is finite at every point except from $\kappa(x^*, x^*)$, we get that $\rho_c(x, y) = 0 \Leftrightarrow x = y$. Now from the non-expansion property of f with respect to d_M and from the fact that $n(f(x)) \geq n(x) + 1$ we get that

$$\rho_c(f(x), f(y)) \leq c \cdot \rho_c(x, y) \tag{3.14}$$

and this concludes the proof of this step.

III. Construction of $\mathbf{d_c}$

In this last step what we do is that we assign the distance between two points to be the length of the shortest path that connects these two points, with the lengths computed according to ρ_c. Then the distance satisfies the triangle inequality because of the shortest path property.

Formally, denote by S_{xy} the set of chains $s_{xy} = (x = x_0, x_1, \ldots, x_m = y)$ from x to y with associated lengths $L_c(s_{xy}) = \sum_{i=1}^{m} \rho_c(x_i, x_{i-1})$. We define

$$d_c(x, y) = \inf\{L_c(s_{xy}) \mid s_{xy} \in S_{xy}\} \tag{3.15}$$

We will prove that d_c is the desired metric.

That f is a contraction with constant c with respect to d_c follows by applying (3.14) to the links $[x_{i-1}, x_i]$ of any chain s_{xy}. Clearly d_c is symmetric and $d_c(x,x) = 0$. The triangle law holds since following a s_{xy} with a s_{yz} yields a s_{xz}. It remains to show positive definite.

Consider any $x \neq x^*$ and $y \neq x$ and assume $n(x) \leq n(y)$ without loss of generality. If $y \neq x^*$, any chain s_{xy} either lies in $\mathcal{D} \setminus K_{n(y)+1}$, or has a last link which leaves $K_{n(y)+1}$, so that

$$d_c(x,y) \geq c^{n(y)} \min\{d_M(x,y), d_M(x, K_{n(y)+1})\} > 0 \tag{3.16}$$

The remaining case, $y = x^*$ is covered by

$$d_c(x,y) \geq c^{n(x)} d_M(x, K_{n(x)+1}) > 0 \tag{3.17}$$

Thus d_c is a distance metric. We now have to prove that d_c is equivalent to d_M.

Let $B_\nu = \mathcal{D} \setminus f^{[-\nu]}(U)$ for $\nu > 0$, so that the definition of $\nu(x)$ (3.11) implies $d_M(x, B_{\nu(x)}) > 0$ and $n(x) \geq \nu(x)$. For any $x \neq x^*$, if y obeys

$$d_M(x,y) < \delta(x) = \min\{d_M(x, K_{n(x)+1}), d_M(x, B_{\nu(x)})\} \tag{3.18}$$

then $n(x) \geq -\nu(x)$, so that (3.15) and (3.16), the latter with x and y interchanged, imply

$$c^{n(x)} d_M(x,y) \leq d_c(x,y) \leq \rho_c(x,y) \leq c^{-\nu(x)} d_M(x,y) \tag{3.19}$$

Choose $k(x) > \max\{0, n(x)\}$ such that $z \in K_{k(x)}$ implies $d_M(z, x^*) < d_c(x, x^*)/2$. Then $d_c(x, K_{k(x)}) \geq d_c(x, x^*)/2$, so that if y obeys

$$d_c(x,y) < d_c(x,x^*)/2 \tag{3.20}$$

then only chains disjoint from $K_{k(x)}$ need enter (3.15), implying

$$d_c(x,y) \geq c^{k(x)} d_M(x,y) \tag{3.21}$$

In particular, if

$$d_c(x,y) < \min\{d_c(x,x^*)/2, c^{k(x)}\delta(x)\}$$

then with (3.20) and (3.21) this implies (3.18) and hence (3.19) applies. Thus $d_c(x_n,x) \to 0$ whenever $d_M(x_n,x) \to 0$.

Now if $x = x^*$, note first that if $d_M(x^*,y) < d_M(x^*,B_0)$, then

$$d_c(x^*,y) \le \rho_c(x^*,y) \le d_M(x^*,y) \tag{3.22}$$

Second, for any $\eta > 0$, (3.7a) guarantees an $N(\eta) > 0$ such that $d_M(x^*,z) < \eta/2$ for all $z \in K_{N(\eta)}$. Then $d_M(x^*,y) > \eta$ implies that $d_M(y,K_{N(\eta)}) \ge \eta/2$ and thus that

$$d_c(x^*,y) \ge d_c(K_{N(\eta)},y) \ge c^{N(\eta)}\eta/2$$

Hence $d_c(x_n,x^*) \to 0$ if and only if $d_M(x_n,x^*) \to 0$.

To show that d_M-completeness is preserved, assume that (x_n) is a d_c-Cauchy sequence and that (X,d_M) is complete. If (x_n) does not converge to x^* then since d_c and d_M are equivalent, for some $N \in nats$ and all sufficiently large n, $n(x_n) < N$.

Now exactly as above choose $k((x_n)) = P > \max 0, N$ such that $z \in K_{k((x_n))}$ implies

$$d_M(x^*,z) < \inf_{i\in\mathbb{N}}\left\{\frac{d_c(x_i,x^*)}{2}\right\} = \frac{R}{2}$$

then since (x_n) is a Cauchy sequence there is an $i \in \mathbb{N}_1$ such that

$$d_c(x_p,x_{p+j}) < \frac{R}{2}$$

for all $p > i$, and using (3.21) with $k(x) = P$, we have

$$c^{-P}d_c(x_p,x_{p+j}) \ge d_M(x_p,x_{p+j})$$

so that (x_n) is a d_M-Cauchy sequence. Therefore since $(\mathcal{D},d_M)$ is complete, the topological space $(\mathcal{D},d_c)$ is complete too.

The final step is to prove (3.6b). Let $A = \text{diam}_d\left[\bar{B}(x^*, 2\varepsilon)\right]$ and without loss of generality $d(x, x^*) \geq d(y, x^*)$.

If either $d_M(x, x^*) \leq \varepsilon$ or $d_M(y, x^*) \leq \varepsilon$ then we are done since as we have seen in the construction of d_M, $d_M(x, y) \geq d(x, y)$, thus either $d(x, x^*) \leq \varepsilon$ or $d(y, x^*) \leq \varepsilon$ and (3.6b) is satisfied. So we may assume that $d(x, x^*) \geq \varepsilon$ and $d(y, x^*) \geq \varepsilon$. Therefore $x, y \in \mathcal{D} \setminus K_0$ and which translates to $n(x), n(y) < 0$. So using the same argument as when we derived (3.16) but with K_0 instead of $K_{n(y)+1}$ we get

$$d_c(x, y) \geq \min\{d_M(x, y), d_M(x, K_0)\} \tag{3.23}$$

Now we consider two cases according to the value of $d_M(x, K_0)$. If $d_M(x, K_0) \geq \varepsilon$ then

$$d_c(x, y) \leq \varepsilon \implies d(x, y) \leq \varepsilon \leq A$$

Otherwise if $d_M(x, K_0) \leq \varepsilon$ then $d(x, K_0) \leq \varepsilon$ and by triangle inequality $d(x, x^*) \leq 2\varepsilon$. By our assumption for the relative position of x and y we also get $d(y, x^*) \leq 2\varepsilon$ and therefore $x, y \in \bar{B}(x^*, 2\varepsilon)$. Thus, $d(x, y) \leq \text{diam}_d\left[\bar{B}(x^*, 2\varepsilon)\right]$. □

3.4.1 Corollaries of the Converse Fixed Point Theorem

The main disadvantage of the result presented in the previous section is that the metric d_c is not the same for all $\varepsilon > 0$ but it depends on ε. Besides this disadvantage this new Converse Theorem has some very interesting corollaries. The first one, it that we can now express with respect to d_c the number of f iterations in order to get close to the fixed point x^*.

Corollary 1. *Under the assumptions of Theorem 11, starting from a point $x_0 \in \mathcal{D}$, the Basic Iterative Method finds the fixed point with additive error ε after*

$$\frac{\log(d_c(x_0, f(x_0))) + \log(2/\varepsilon)}{\log(1/c)}$$

iterations.

Proof. We choose d_c that satisfies the Theorem 11 with parameters $c, \varepsilon/2$. Let also (x_n), the sequence produced by the Basic Iterative Method. Then we have that since f is a contraction

with respect to d_c

$$d_c(x_n, x_{n+1}) \leq c^n d_c(x_0, x_1)$$

If we ensure that $d_c(x_n, x_{n+1}) \leq \varepsilon/2$ then according to Theorem 11 $d(x_n, x^*) \leq \varepsilon$ or $d(x_{n+1}, x^*) \leq \varepsilon$. So we need

$$c^n d_c(x_0, x_1) \leq \frac{\varepsilon}{2} \Leftrightarrow n \geq \frac{\log(d_c(x_0, x_1)) + \log(2/\varepsilon)}{\log(1/c)}$$

□

The above corollary only describes for a fixed ε how many steps we need to get into a ball of radius ε from the fixed point. If we want to have the same kind of argument for any $\delta > 0$ we have to make additional assumptions on f. If for example f is a contraction with respect to d locally for $x, y \in \bar{B}(x^*, \varepsilon)$ in an then we get the following result.

Corollary 2. *Under the assumptions of Theorem 11, and the assumption that there exists* $0 < c < 1$, $\varepsilon > 0$ *such that*

$$d(f(x), f(y)) \leq cd(x, y) \text{ for all } x, y \in \bar{B}(x^*, \varepsilon)$$

then starting from a point $x_0 \in \mathcal{D}$, *the Basic Iterative Method finds the fixed point with additive error* $\delta > 0$ *after*

$$\frac{\log(d_c(x_0, f(x_0))) + \log(1/\delta) + 1}{\log(1/c)}$$

iterations.

Proof. Using the same idea as in the previous Corollary we get that after

$$\frac{\log(d_c(x_0, f(x_0))) + \log(2/\varepsilon)}{\log(1/c)}$$

iterations we will have $d(x_n, x^*) \leq \varepsilon$. Now since from now on f is a contraction with respect to d we will have

$$d_c(x_{n+m}, x^*) \leq c^m d_c(x_n, x^*)$$

Therefore for $d(x_{n+m}, x^*) \leq \delta$ we have

$$m \geq \frac{-\log(1/\varepsilon) + \log(1/\delta)}{\log(1/c)}$$

So in total we need $n + m$ iterations and we have the number of iterations that the corollary says. □

3.5 Application to Computation of Eigenvectors

In this section we show how we can apply the ideas of this chapter to the analysis of very famous iterative algorithm for computing eigenvalues and eigenvectors, namely the power method. We start with the definition of the power method and we proceed on the investigation of a metric $d(\cdot, \cdot)$ that makes the power method contraction. For the introductory part we follow the survey of [47].

3.5.1 Introduction to Power Method

Let $A \in \mathbb{R}^{n \times n}$. Recall that if q is an eigenvector for A with eigenvalue λ, then $Aq = \lambda q$, and in general, $A^k = \lambda^k q$ for all $k \in \mathbb{N}$. This observation is the foundation of the *power iteration method.*

Suppose that the set $\{q_i\}$ of unit eigenvectors of A forms a basis of $\mathbb{R}^n$, and has corresponding set of real eigenvalues $\{\lambda_i\}$ such that $|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n|$. Let v_0 be an arbitrary initial vector, not perpendicular to q_1, with $\|v_0\| = 1$. We can write v_0 as a linear combination of the eigenvectors of A for some $c_1 \ldots, c_n \in \mathbb{R}$ we have that

$$v_0 = c_1 q_1 + c_2 q_2 + \cdots + c_n q_n$$

and since we assumed that v_0 is not perpendicular to q_1 we have that $c_1 \neq 0$.

Now

$$Av_0 = c_1 \lambda_1 q_1 + c_2 \lambda_2 q_2 + \cdots + c_n \lambda_n q_n$$

and therefore

$$\begin{aligned} Av_k &= c_1\lambda_1^k q_1 + c_2\lambda_2^k q_2 + \cdots + c_n\lambda_n^k q_n \\ &= \lambda_1^k \left(c_1 q_1 + c_2 \left(\frac{\lambda_2}{\lambda_1}\right)^k q_2 + \cdots + c_n \left(\frac{\lambda_n}{\lambda_1}\right)^k q_n \right) \end{aligned}$$

Since the eigenvalues are assumed to be real, distinct, and ordered by decreasing magnitude, it follows that

$$\lim_{k\to\infty} \left(\frac{\lambda_i}{\lambda_1}\right)^k = 0$$

So, as k increases, $A^k v_0$ approaches $c_1\lambda_1^k q_1$, and thus for large values,

$$\frac{A^k v_0}{\|A^k v_0\|} \to q_1 \text{ as } k \to \infty$$

The power iteration method is simple and elegant, but suffers some major drawbacks. The method only returns a single eigenvector estimate, and it is always the one corresponding to the eigenvalue of largest magnitude. In addition, convergence is only guaranteed if the eigenvalues are distinct—in particular, the two eigenvalues of largest absolute value must have distinct magnitudes. The rate of convergence primarily depends upon the ratio of these magnitudes, so if the two largest eigenvalues have similar sizes, then the convergence will be slow.

In spite of its drawbacks, the power method is still used in some applications, since it works well on large, sparse matrices when only a single eigenvector is needed. However, there are other methods that overcome the difficulties of the power iteration method.

The main problem that we answer in the next section is: what is the exact tradeoff between the number of iterations and the error that we get?

3.5.2 Power Method as Contraction Map

We define the following metric with respect to two vectors $v = (v_1, \ldots, v_n)^T$ and $u = (u_1, \ldots, u_n)^T$

$$d_1(v, u) = \sum_{j=2}^{n} \left| \frac{v_j}{v_1} - \frac{u_j}{u_1} \right|$$

it is then easy to see the following

Lemma 9. *The power method is a contraction with respect to d_1, with contraction constant $c = \lambda_2/\lambda_1$.*

Proof. Let f be the iteration of the power method, we have that

$$d_1(f(v), f(u)) = \sum_{j=2}^{n} \left| \frac{\lambda_j v_j}{\lambda_1 v_1} - \frac{\lambda_j u_j}{\lambda_1 u_1} \right| = \sum_{j=2}^{n} \frac{\lambda_j}{\lambda_1} \left| \frac{v_j}{v_1} - \frac{u_j}{u_1} \right| \leq \frac{\lambda_2}{\lambda_1} \sum_{j=2}^{n} \left| \frac{v_j}{v_1} - \frac{u_j}{u_1} \right| \Rightarrow$$

$$d_1(f(v), f(u)) \leq \frac{\lambda_2}{\lambda_1} d_1(v, u)$$

Therefore the lemma holds. □

But as in the discussions that we had in the previous chapters, d_1 is not a metric that we care for. Such a metric is the norm of the space $\|\cdot\|_1$. We also observe that we can without loss of generality assume that $\|v_0\|_1 = 1$ since we can do a normalization at each step. Among the vector with unit ℓ_1 norm the only fixed point obviously is the $(1, 0, \ldots, 0)^T$. Now let's con cider the case where after k iterations $d_1(u = v_k, e_1) \leq \varepsilon$, then we have

$$\sum_{j=2}^{n} \left| \frac{u_j}{u_1} \right| \leq \varepsilon \Rightarrow \sum_{j=2}^{n} |u_j| \leq \varepsilon \, |u_1| \leq \varepsilon$$

But since $\|u\|_1 = 1$ and since $u_1 > 0$ without loss of generality we also get

$$|1 - u_1| \leq \varepsilon$$

Therefore

$$\|u - e_1\| = |1 - u_1| + \sum_{j=2}^{n} |u_j| \leq 2\varepsilon$$

This is what we wanted that proves the following theorem.

Theorem 12. *The power method is a contraction with respect to d_1, with contraction constant $c = \lambda_2/\lambda_1$. Furthermore for any $\varepsilon > 0$ and any $u \in \mathbb{R}^n$ such that $\|u\|_1$ we have that*

$$d_1(u, e_1) \leq \varepsilon \implies \|u - e_1\| \leq 2\varepsilon$$

The same way we got the Corollaries in the previous section we can get the following Corollary

Corollary 3. *Starting from a vector v_0 not perpendicular to q_1, the power method finds a vector u, such that $\|u - e_1\| \varepsilon$ after*

$$\frac{\log(d_1(v_0, e_1)) + \log(2/\varepsilon)}{\log(\lambda_1/\lambda_2)}$$

iterations.

Chapter 4

Computational Complexity of Computing Fixed Point of Contraction Maps

The purpose of this chapter is to capture the computational complexity of computing fixed points, the existence of which is guaranteed by Banach's Fixed Point Theorem (Theorem 2).

We first present an introduction to the tools that have been developed in the area of computational complexity in order to capture the computational complexity of computing fixed points starting with the works of Johnson Yannakakis, Papadimitriou and Daskalakis. We define, explain and prove the basic properties of the following complexity classes:

- PLS [37]: when ϕ satisfies some continuity condition
- PPAD [48]: when f satisfies some continuity condition
- CLS [23]: when both ϕ and f satisfy some continuity condition

The main result of this chapter is :

$$\text{BANACH is CLS} - \text{complete}$$

We show how the ideas from the previous chapter can be used in order to prove this result. We finish the chapter with a discussion about future directions and problems in this

area.

4.1 The PLS Complexity Class

The complexity class PLS first appeared in the seminal paper of Johnson, Papadimitriou and Yannakakis at 1988 [37]. The motivation is to capture the complexity of computing a local optimum according to some objective or potential function $\phi : \mathcal{D} \to \mathbb{R}_+$. According to the authors:

> One of the few general approaches to difficult combinatorial optimization problems that has met with empirical success is local (or neighborhood) search. In a typical combinatorial optimization problem, each instance is associated with a finite set of feasible solutions, each feasible solution has a cost, and the goal is to find a solution of minimum (or maximum) cost. In order to derive a local search algorithm for such a problem, one superimposes on it a neighborhood structure that specifies a "neighborhood" for each solution, that is, a set of solutions that are, in some sense "close" to that solution. For example, in the traveling salesman problem (TSP), a classical neighborhood is to assign to each tour the set of tours that differ from it in just two edges (this is called the 2-change neighborhood). In the graph partitioning problem (given a graph with $2n$ vertices and weights on the edges, partition the vertices into two sets of n vertices such that the sum of the weights of the edges going from one set to the other is minimized) a reasonable neighborhood would be the so-called "swap" neighborhood: Two partitions are neighbors if one can be obtained from the other by swapping two vertices.
>
> Given a combinatorial optimization problem with a superimposed neighborhood structure, the local search heuristic operates as follows. Starting from an independently obtained initial solution, we repeatedly replace the current solution by a neighboring solution of better value, until no such neighboring solution exists, at which point we have identified a solution that is "locally optimal." Typically, we repeat this procedure for as many randomly chosen initial solutions as is computationally feasible and adopt the best local optimum found. Variants of this methodology have been applied to dozens of problems, often with impressive

success.

The importance and the success of this complexity class is based on the fact that a lot of important and useful in practice local search techniques can be proved to be *complete* for this class. The first one, that already appeared in [37] is the one based on the Kernighan-Lin neighborhood structure for the graph partitioning problem [40].

In this section we will give the formal definition of PLS and then prove that the problem of computing a fixed point given a potential function is complete for this class. This provides a new formulation of PLS, first appeared in the work of Daskalakis, Papadimitriou [23]. This formulation will become very useful latter when we define the CLS class and argue about its relation with the NONMETRICBANACH problem.

4.1.1 Formal Definition and Basic Properties

General NP search problems and TFNP

In general a *search problem* L consists of the following ingredients:

(I) a set D_L of instances, which can be taken to be a polynomial-time recognizable subset of $\{0,1\}^*$. That is, there exists a polynomial time computable characteristic function $R : \{0,1\}^* \to \{0,1\}$ such that $R(x) = 1 \Leftrightarrow x \in D_L$.

(II) for each instance x, we have a finite set $F_L(x)$ of solutions, which are considered also as strings in $\{0,1\}^*$ that have without loss of generality all with the same polynomially bounded length $p(|x|)$.

(III) a polynomial time relation $R_L : \{0,1\}^* \times \{0,1\}^* \to \{0,1\}$. That could take value 1 only when computed to $R_L(x,y)$ with $x \in D_L$, $y \in F_L(x)$.

The answer to an instance x of a search problem L is given (I), (II) and (III) to provide a $y \in F_L(x)$ such that $R_L(x,y) = 1$.

The above description captures all the search problems in NP and defines that class FNP. The class of *total search problems* TFNP can be defined given the promise that there exists at least one $y \in F_L(x)$ such that $R_L(x,y) = 1$.

The PLS Complexity Class

If additionally to these a search problem L also satisfies the properties

(A1) for each solution $s \in F_L(x)$ we have a non-negative integer *cost* $c_L(s,x) \in \mathbb{N}$ that is computed in polynomial time.

(A2) for each solution $s \in F_L(x)$ we have a subset $N(s,x) \subseteq F_L(x)$ called the neighborhood of s.

(A3) there exists a polynomial-time algorithm A_L such that given $x \in D_L$ produces a particular standard solution $A_L(x) \in F_L(x)$.

(A4) there exists a polynomial-time algorithm C_L such that given $x \in D_L$ and a solution $s \in F_L(x)$, has two possible types of output, depending on s. If there is any solution $s' \in N(s,x)$ with better cost than that for s (i.e., such that $c_L(s',x) < c(s,x)$) C_L produces such a solution. Otherwise it reports that no such solution exists and hence that s is *locally optimal.*

(A5) the result of $R_L(x,s)$ is 1 if and only if s is locally optimal.

The first observation is that the properties (A) guarantee the existence of a polynomial local search algorithm to find a solution to the search problem L. We call this algorithm *standard local search algorithm.* The steps of this algorithm are the following

1. Given x, use A_L, to produce a starting solution s.
2. Repeat until locally optimal:
   - Apply algorithm C_L to x and s.
   - If C_L, yields a better cost neighbor s' of s, set $s \leftarrow s'$.

Note that this algorithm seems to be the only thing one can do in order to find a locally optimal solution. If this was the case then the problem would be NP-hard. That is, given L, if instead of an arbitrary locally optimal solution one wants to compute the exact result of the above algorithm then the problem is NP-hard as shown in the next result proven in [37].

Theorem 13. *There is a* PLS *problem L whose computing the final state of the standard local search in* NP*-hard.*

So at this point one could ask what's the difference between NP and PLS?

The difference, comes from the fact that PLS is asking for a local optimal and not for the specific one. This local optima could be found not by using the standard local search

algorithm, but by structurally trying to find a local optima using the white box access to the algorithms A_L and C_L. Is the difficulty of that problem that the PLS class tries to capture.

So one interesting question is if NP-hardness can still capture the complexity of this easier problem. Here comes another result, also appeared in the paper by Johnson [37].

Theorem 14. *If any search problem L in* TFNP *is* NP*-hard, then* NP = coNP

Proof. If L is NP-hard, then by definition there is an algorithm A for an NP-complete problem N such that calls an algorithm for L as a subroutine and takes polynomial time (if the time spent executing the subroutine is ignored). But the existence of such an algorithm implies that we can verify that x is a no-instance for N in nondeterministic polynomial time: simply guess a computation of A on input x, including the inputs and outputs of the calls to the subroutine for L. The validity of the computation of A outside of the subroutines can be checked in polynomial time because A is deterministic; the validity of the subroutine outputs can be verified using the polynomial-time algorithm R_L, (whose existence is implied by the fact that L is in TFNP) to check whether the output is really an answer for this input. Thus the set of no-instances of N is in NP, i.e., $N \in$ coNP. Since N was NP-complete, this implies that NP = coNP. □

This result suggests that for any problem in TFNP is unlike to be NP-hard. At least based on the strong belief of the computer scientists that NP $\neq$ coNP.

One interesting result in the work of [37], in that the following problem is PLS-complete.

Definition 21. *For the* LocalOpt *problem we are given the following:*

(PI1) a boolean circuit that computes a function f.

(PI2) a boolean circuit that computes a function - potential p.

and we ask for one of the following:

(PO1) a binary string $x \in \{0,1\}^*$ such that $p(f(x)) \geq p(x)$.

Theorem 15. LocalOpt *is* PLS*-complete.*

4.1.2 Reductions Among Search Problems

We now explain what a reduction means for search problems. We define the notion of PLS-reducibility which naturally extends to all the search problems and to other search classes.

We say that a problem L in PLS is PLS-reducible to another, K, if there are polynomial-time computable functions f and g such that: f maps instances x of L to instances $f(x)$ of K, g maps (solution of $f(x)$, x) pairs to solutions of x. Finally for all instances x of L, if s is a local optimum for instance $f(x)$ of K, then $g(s, x)$ is a local optimum for x. Note that this notion of reduction has the standard desirable properties.

4.1.3 Characterization of PLS in terms of Fixed Point Computing

In this section we define a fixed point computing problem and prove that is PLS-complete. This gives a new characterization of PLS that we are going to use latter on. The material of this section comes from the work of Daskalakis and Papadimitriou [23].

Our discussion from now on focuses on functions from continuous domains to continuous domains, and we shall represent these functions in terms of arithmetic circuits with operations $+, -, \cdot$, max, min, and $>$, the latter defined as $> (x, y) = 1$ if $x > y$ and 0 otherwise; rational constants are also allowed. The result of the arithmetic circuits has to be able to be computed in polynomial time by a Turing Machine. The outputs of arithmetic circuits can be restricted in $[0, 1]$ by redefining the arithmetic gates to output 0 or 1 when the true output is negative or greater than one, respectively.

We now define the REAL LOCALOPT problem that we will prove it is PLS complete.

Definition 22. *For the* REAL LOCALOPT *problem we are given the following:*

(PI1) an arithmetic circuit that computes a function $f : [0, 1]^3 \to [0, 1]^3$.

(PI2) an arithmetic circuit that computes a function - potential $p : [0, 1]^3 \to [0, 1]$.

(PI3) a rational number $\varepsilon > 0$.

(PI4) a rational number $\lambda > 0$.

and we ask for one of the following:

(PO1) a point $x \in [0, 1]^3$ *such that* $p(f(x)) \geq p(x) - \varepsilon$.

(PO2) two points x, x' *violating the* λ*-Lipschitz continuity of* p*, i.e.* $|p(x) - p(x')| > \lambda |x - x'|$.

The first thing to notice is that REAL LOCALOPT is a total search problem, i.e. belongs to TFNP. Indeed, starting at an arbitrary point $x \in [0, 1]^3$, we can just follow the chain $x, f(x), f(f(x)), \ldots$ for $p(x)/\varepsilon$ steps, as long as a step result in a more than ε decrease of the

value of p. Then the existence of a point satisfying (PO1) is guaranteed because $p(\cdot) \geq 0$. Also notice the running time $p(x)/\varepsilon$ is not polynomial in the input size but pseudo-polynomial instead.

Theorem 16. REAL LOCALOPT *is* PLS*-complete.*

Proof. We will see that any problem in PLS can be reduced to REAL LOCALOPT, by embedding the solution space in small cubelets of $[0,1]^3$. At the centers of the cubelets the values of f and p are defined so that they capture the neighborhood function and cost function of the original problem. Then p is extended to the rest of the cube continuously by interpolation. We can see that f need not be continuous and we extended carefully so that no new solutions are introduced.

More precisely, for a given instance (f,p) of LOCALOPT, our instance $(f',p',\lambda,\varepsilon)$ of REAL LOCALOPT satisfies the property that, for all $x,y,z,y',z' \in [0,1]$, $p'(x,y,z) = p'(x,y',z')$ and $f'(x,y,z) = f'(x,y',z')$ in other words only the value of x is important in determining the values of p' and f'. Now, for every n-bit string s, if $x(s) \in \{0,\ldots,2^n-1\}$ is the number corresponding to s, we define $p'(x(s)\cdot 2^{-n},y,z) = p(s)\cdot 2^{-n}$, for all $y,z \in [0,1]$ and $f'(x(s)\cdot 2^{-n},y,z) = (x(f(s))\cdot 2^{-n},y,z)$, for all $y,z \in [0,1]$. To extend f' and p' to the rest of the cube we do the following: p' is extended simply by linear interpolation. We have to be a bit more careful in how we extend f' to the rest of the cube, so that we do not introduce spurious solutions. For all $i \in \{0,\ldots,2^n-2\}$, if $p'(i\cdot 2^{-n},y,z) < p'((i+1)\cdot 2^{-n},y,z)$, we set $f'(i\cdot 2^{-n}+t\cdot(i+1)2^{-n},y,z)$ to be equal to $f'(i2^{-n},y,z)$, for all $t,y,z \in [0,1]$, while if $p'(i\cdot 2^{-n},y,z) \geq p'((i+1)\cdot 2^{-n},y,z)$, we set we set $f'(i\cdot 2^{-n}+t\cdot(i+1)2^{-n},y,z)$ to be equal to $f'((i+1)2^{-n},y,z)$ for all $t,y,z \in [0,1]$. Exceptionally, for $x > 1-2^{-n}$, we define $f'(x,y,z) = f'(1-2^{-n},y,z)$, for all $y,z \in [0,1]$. Finally, we choose $\varepsilon = 0$ and $\lambda = 2^n$, so that p is guaranteed to be λ-Lipschitz continuous. It is easy to verify that any solution to the REAL LOCALOPT instance that we just created can be mapped to a solution of the PLS instance (f,p) that we departed from.

We point out next that REAL LOCALOPT is in PLS, by showing a reduction in the opposite direction, from real-valued to discrete. Suppose we are given an instance of REAL LOCALOPT defined by a four-tuple $(f,p,\varepsilon,\lambda)$. We describe how to reduce this instance

to an instance (f', p') of LOCALOPT. For a choice of n that makes $2^{-n/3}$ sufficiently small with respect to ε and λ, we identify the n-bit strings that are inputs to f' and p' with the points of the 3-dimensional unit cube whose coordinates are integer multiples of $2^{-n/3}$. The value of p' on an n-bit string x is then defined to be equal to the value of p on x (viewed as a point in the cube). Similarly, the value of f' on an n-bit string x is defined by rounding each coordinate of $f(x)$ down to the closest multiple of $2^{-n/3}$. Suppose now that we have found a solution to LOCALOPT, that is an n-bit string x such that $p'(f'(x)) \geq p'(x)$. Notice that $f(x)$ and $f'(x)$ are within $2^{-n/3}$ to each other in the ℓ_∞ norm. Hence, it should be that $\|p(f(x)) - p(f'(x))\|_\infty < \lambda \cdot 2^{-n/3}$ (assuming that λ is the purported Lipschitz constant of p in the ℓ_∞ norm). If this is not the case, we have found a violation of the Lipschitzness of p. Otherwise, we obtain from the above that $p(f(x)) \geq p(x) - \lambda 2^{-n/3}$, where we used that $p(x) = p'(x)$ and $p(f(x)) = p'(f(x))$. If n is chosen large enough so that $\varepsilon > \lambda \cdot 2^{-n/3}$, x is a solution to REAL LOCALOPT. In the above argument we assumed that λ is the Lipschitz constant of f in the ℓ_∞ norm, but this is not important for the argument to go through. □

4.2 The PPAD Complexity Class

The complexity class PPAD first appeared in the seminal paper of Papadimitriou at 1994 [48]. The motivation is to capture the complexity of computing a fixed point whose existence is guaranteed by the Brouwers Fixed Point Theorem. This was shown to be hard if we only have black box access to the circuit that computes the values of f, by [32]. According to the author:

> In [37] the authors defined a broad and natural subclass of TFNP, namely PLS (for polynomial local search). For a problem L in PLS we wish to find a solution such that no neighbor has better cost. Thus, totality for functions in PLS is established by invoking the following lemma:
>
> *Every finite directed acyctic graph has a sink.*
>
> The dag for invoking the lemma is the graph whose adjacency lists are the $N_L(x, s)$, with arcs leading to nodes with no better c_L omitted. In other words, we

can view N_L and c_L as an implicit syntactic way for specifying an exponentially large dag. Class PLS contains a host of problems that are not known to be in FP (the difficulty is, of course, that the dag may have exponential depth). Several important problems are now known to be PLS-complete, including computing a local optimum in the Lin-Kernighan heuristic for the TSP and computing a stable configuration in Hopfield neural nets [37].

Like PLS, each of our new complexity classes can be seen as based on a graph-theoretic lemma. Perhaps the most basic one is the parity argument:

Any finite graph has an even number of odd − degree nodes.

This last graph theoretic lemma is the basis for the definition of a class called PPA for *polynomial parity argument.* An interesting whose complexity seems to be captured by PPA is the Smith's theorem [53]. Which describes a procedure to find an alternative Hamilton path given an initial Hamilton path. Although the relation of Smith's theorem with PPA is clear, is not yet known if the corresponding computational problem is PPA-complete or not.

We will concentrate to the directed version of PPA which based on the following graph-theoretic lemma.

Any directed graph with total degree at most two,
has an even number of degree one nodes.

We call this class PPAD which stands for *polynomial parity argument directed.* In contrast with PPA, for PPAD there are a lot of interesting and important computational problems that are known to be PPAD-complete. The first two, SPERNER, BROUWER came from the first publication of PPAD at 1992 and are closely related. A third very important one is the NASH problem, which describes the complexity of computing a Nash Equilibrium in polymatrix games. In the seminal work of Daskalakis, Goldberg and Papadimitriou [20] they proved that NASH is PPAD-complete.

Our interest in this class comes from the fact that BROUWER is complete for this class. As we will see later BROUWER is very closely related with the REAL LOCALOPT problem that characterizes the PLS class. We proceed with a formal definition of PPAD and BROUWER and we state some important properties of them.

4.2.1 Formal Definition and Basic Properties

As in the case of PLS a problem L in PPAD satisfies (I), (II) and (III) and additionally satisfies the following

(B1) the string $0 \ldots 0$ belongs to $F_L(x)$.

(B2) there exists polynomial time algorithm P_L, called *predecessor*, that for any solution s returns another solution s'.

(B3) there exists polynomial time algorithm S_L, called *successor*, that for any solution s returns another solution $s' \neq 0 \ldots 0$.

(B4) the result of $R_L(x, s)$ is equal to 1 if and only if

- $s \neq 0 \ldots 0$ and
- $P_L(S_L(s)) \neq s$ or $S_L(P_L(s)) \neq s$.

From a first reading its not clear why this definition captures the graph-theoretic lemma that we started from. The situation becomes more clear if we define the underline graph $G_L(x)$ as follows:

- the vertex set of $G_L(x)$ is $F_L(x)$,
- an edge (s, s') belongs to $G_L(x)$ if and only if $P_L(S_L(s)) = s$ and $S_L(P_L(s)) = s$.

The reason for this way of definition of the class is that this definition, as well as the definition of PLS, is a *syntactic* definition and not a *semantic* one. The existence of a syntactic definition is very important since it enables the possibility of finding complete problems for the class. We remind the reader that a semantic class unlikely has a complete problem because of the Rice's Theorem [50]. This theorem states that any non-trivial property of a Turing machine is undecidable and therefore there is no hope that we can check if a given machine has the property or not.

As in the case of PLS, if instead of an arbitrary node of degree one, we ask for the node of degree one that appears in the other end of the path that starts from $0\ldots0$ then this problem is much harder. Namely in this case is PSPACE-hard.

Theorem 17. *There exists a problem $L \in$ PPAD such that computing be the end of the directed path in G_L starting from $0\ldots0$ is* PSPACE*-hard.*

4.2.2 Characterization of PPAD in terms of Fixed Point Computing

In this section we define a fixed point computing problem and prove that is PPAD-complete. This gives a new characterization of PPAD that we are going to use latter on. The material of this section comes from the work of Papadimitriou [48]. We are going to use again the notion of arithmetic circuits as defined in the previous section.

Definition 23. *For the* BROUWER *problem we are given the following:*

(BI1) an arithmetic circuit that computes a function $f : [0,1]^3 \to [0,1]^3$.

(BI2) a rational number $\varepsilon > 0$.

(BI3) a rational number $\lambda > 0$.

and we ask for one of the following:

(BO1) a point $x \in [0,1]^3$ such that $|f(x) - x| \leq \varepsilon$.

(BO2) two points $x, x' \in [0,1]^3$ violating the λ-Lipschitz continuity of f, i.e.
$|f(x) - f(x')| > \lambda|x - x'|$.

It was proven in [48] that the BROUWER problem is PPAD-complete. Thus, BROUWER gives an alternative definition of PPAD.

Theorem 18. BROUWER *is* PPAD*-complete.*

For the proof of the above theorem we refer to initial paper by Papadimitriou [48].

Looking at PLS and PPAD this way, as close relatives so to speak, is particularly helpful when one considers the class PLS $\cap$ PPAD.

4.2.3 The Class PLS ∩ PPAD

Unlike the complexity class NP∩coNP, the class PLS∩PPAD can be defined syntactically. Therefore is has a complete problem, namely the EITHER FIXED POINT problem.

Definition 24. *For the* EITHER FIXED POINT *problem we are given the following:*

(EI1) *an arithmetic circuit that computes a function* $f : [0,1]^3 \to [0,1]^3$.

(EI2) *an arithmetic circuit that computes a function* $g : [0,1]^3 \to [0,1]^3$.

(EI3) *an arithmetic circuit that computes a function - potential* $p : [0,1]^3 \to [0,1]$.

(EI4) *a rational number* $\varepsilon > 0$.

(EI5) *a rational number* $\lambda > 0$.

and we ask for one of the following:

(EO1) *a point* $x \in [0,1]^3$ *such that* $|f(x) - x| \leq \varepsilon$.

(EO2) *a point* $x \in [0,1]^3$ *such that* $p(g(x)) \geq p(x) - \varepsilon$.

(EO3) *two points* $x, x' \in [0,1]^3$ *violating the* λ*-Lipschitz continuity of* f*, i.e.*
$|f(x) - f(x')| > \lambda |x - x'|$.

(EO4) *two points* x, x' *violating the* λ*-Lipschitz continuity of* p*, i.e.* $|p(x) - p(x')| > \lambda |x - x'|$.

It has been proved in the work of Daskalakis and Papadimitrou [23] that this problem is PLS ∩ PPAD-complete.

Theorem 19. EITHER FIXED POINT *is* PLS ∩ PPAD*-complete.*

Proof. The problem is clearly in both PLS and PPAD, because it can be reduced to both REAL LOCALOPT and BROUWER. To show completeness, consider any problem C in PPAD ∩ PLS. Since BROUWER is PPAD-complete and C is in PPAD, there is a reduction such that, given an instance x of C, produces an instance $f(x)$ of BROUWER of A, such that from any solution of $f(x)$ we can recover a solution of C. Similarly, there is a reduction g from C to REAL LOCALOPT. Therefore, going from x to $(f(x), g(x))$ is a reduction from C to EITHER FIXED POINT. □

Now we are ready to define a main class that we will consider in this chapter and belong to the intersection of PLS and PPAD, namely CLS.

4.3 The CLS Complexity Class

From the definition of EITHER FIXED POINT, we can see that there should be an interesting subproblem where the functions f and g in the definition of EITHER FIXED POINT problem are the same. In this class the existence of a solution is guaranteed both by the continuity of f and the Brouwer's Fixed Point theorem and by the continuity of p and the guaranteed local optimum. We also observe that because of the continuity of p if we have a point such that $|f(x) - x| \leq \varepsilon$ then we also get that $|p(f(x)) - p(x)| \leq \lambda\varepsilon$. Therefore a fixed point of f satisfies also the local optimum of p requirement. For this reason we don't need to keep the (EO1) possibility for the output but we can only ask for a point x that satisfies (EO2). This gives us the definition of the CONTINUOUS LOCALOPT problem.

Definition 25. *For the* CONTINUOUS LOCALOPT *problem we are given the following:*

(CI1) *an arithmetic circuit that computes a function* $f : [0,1]^3 \to [0,1]^3$.

(CI2) *an arithmetic circuit that computes a function - potential* $p : [0,1]^3 \to [0,1]$.

(CI3) *a rational number* $\varepsilon > 0$.

(CI4) *a rational number* $\lambda > 0$.

and we ask for one of the following:

(CO1) *a point* $x \in [0,1]^3$ *such that* $p(f(x)) \geq p(x) - \varepsilon$.

(CO2) *two points* $x, x' \in [0,1]^3$ *violating the* λ*-Lipschitz continuity of* f*, i.e.*
$|f(x) - f(x')| > \lambda|x - x'|$.

(CO3) *two points* x, x' *violating the* λ*-Lipschitz continuity of* p*, i.e.* $|p(x) - p(x')| > \lambda|x - x'|$.

We are now ready to define the CLS class.

Definition 26. *The complexity class* CLS *is the set of search problems that can be reduced to the* CONTINUOUS LOCALOPT *problem.*

The main observation about CLS is that it is a subset of PLS $\cap$ PPAD. The proof of this just formalizes the argument that we present in the beginning of the section.

Theorem 20. CLS $\subseteq$ PLS $\cap$ PPAD.

Proof. The fact that CONTINUOUS LOCALOPT is in PLS follows from the fact that it is a special case of REAL LOCALOPT.

To show that it is also in PPAD, we provide a reduction to BROUWER. We reduce an instance $(f, p, \lambda, \varepsilon)$ of CONTINUOUS LOCALOPT to an instance $(f, \lambda, \varepsilon/\lambda)$ of BROUWER. If on this instance BROUWER returns a pair of points violating f's Lipschitz continuity, we return this pair of points as a witness of this violation. Otherwise, BROUWER returns a point x such that

$$|f(x) - x| \leq \frac{\varepsilon}{\lambda}$$

In this case, we check whether

$$|p(f(x)) - p(x)| \leq \lambda \frac{\varepsilon}{\lambda} = \varepsilon$$

if this is not the case then we return $x, f(x)$ as witness of violation of p's Lipschitz continuity. Otherwise we have that

$$p(f(x)) \geq p(x) - \varepsilon$$

which again is a solution to CONTINUOUS LOCALOPT. □

The importance of CLS comes from the fact that a lot of interesting problem belong to it. We give here a list of these problems without the proofs.

- ↦ **Approximate Fix point of a Contraction Map** (CONSTRACTION MAP). We are given a function $f : [0,1]^n \to [0,1]^n$ and some constant c with the promise that f is contracting with constant c with respect to $\|\cdot\|_p$ norm. We seek an approximate fix point of this function, or a violation of contraction.
- ↦ **Linear Complementarity Problem for P-matrices**. In this problem we are given an $n \times n$ matrix M and a vector q, and we seek two vectors x, y with positive entries such that

$$y = Mx + q, \ \langle x, y \rangle = 0$$

- ↦ **Finding a stationary point of a polynomial**.

$\mapsto$ **Simple Stochastic Games.**

$\mapsto$ **Nash equilibrium in network coordination games.**

$\mapsto$ **Nash equilibrium in congestion games.**

$\mapsto$ **Nash equilibrium in implicit congestion games.**

Surprisingly, very recently a line of work appeared that bases the hardness of CLS on some cryptographic assumptions [34]. The reason this is a surprising result is that these classes were defined to capture the complexity of computing fixed points and this connection to cryptography looks very strange and interesting!

4.4 Banach's Fixed Point is Complete for CLS

We can see that the CONSTRACTION MAP problem is very closely related with the discussion we had in the previous chapter about the Banach fixed point theorem. The only restriction is the fact that instead of an arbitrary metric function it assume the contraction property with respect to one of the ℓ_p norms.

In this section we consider the general problem of computing a fixed point whose existence is guaranteed by the Banach's fixed point theorem. In order to capture the aspect of an arbitrary metric space we have to somehow provide as input the distance metric with respect to which the input function is a contraction map. We do so by providing the distance metric as an arithmetic circuit.

With these in mind and given the definition of a distance metric (Definition 4) we define the following problem that we call NONMETRICBANACH.

Definition 27. *For the* NONMETRICBANACH *problem we are given the following:*

(Ia) an arithmetic circuit that computes a function

$f : [0,1]^3 \to [0,1]^3$

(Ib) an arithmetic circuit that computes a distance metric function

$d : [0,1]^3 \times [0,1]^3 \to \mathbb{R}$

(Ic) a rational numbers $\varepsilon > 0$, $\lambda > 0$, $1 > c > 0$.

and we ask for one of the following:

(Oa) a point $x \in [0,1]^3$ *such that* $d(x, f(x)) \leq \varepsilon$.

(Ob) *two points* $x, x' \in [0,1]^3$ *violating the contraction property with constant* c*, i.e.* $d(f(x), f(x')) > c \cdot d(x, x')$

(Oc) *two points* $x, x' \in [0,1]^3$ *violating the* λ*-Lipschitz continuity of* f*, i.e.* $|f(x) - f(x')| > \lambda |x - x'|$.

(Od) *two points* $x, x' \in [0,1]^3$ *violating the* λ*-Lipschitz continuity of* $d(x, f(x))$*, i.e.* $|d(x, f(x)) - d(x', f(x'))| > \lambda |x - x'|$.

We are now ready to prove the first statement of this chapter.

Theorem 21. *The* NONMETRICBANACH *problem is* CLS*-complete. In particular,* NONMETRICBANACH *is a total search problem.*

Proof. We will first show that NONMETRICBANACH belongs to CLS. Starting from an instance
$(f, d, \varepsilon, \lambda, c)$ we create the following instance

$$\begin{aligned} f'(x) &= f(x) \\ p(x) &= d(x, f(x)) \\ \varepsilon' &= (1-c) \cdot \varepsilon \\ \lambda' &= \lambda \end{aligned}$$

Now we have to show that any result of the CONTINUOUS LOCALOPT with input $(f, p, \varepsilon', \lambda)$ will give us a result of NONMETRICBANACH with input $(f, d, \varepsilon, \lambda, c)$.
(CO1) $\Longrightarrow$ If $d(f(x), f(f(x))) > c \cdot d(x, f(x))$ then $(x, f(x))$ satisfies (Ob) and therefore is a solution to NONMETRICBANACH. Otherwise

$$\begin{aligned} p(f(x)) \geq p(x) - \varepsilon' \Rightarrow d(f(x), f(f(x))) &\geq d(x, f(x)) - \varepsilon' \Rightarrow \\ c \cdot d(x, f(x)) \geq d(f(x), f(f(x))) &\geq d(x, f(x)) - \varepsilon' \Rightarrow \\ c \cdot d(x, f(x)) &\geq d(x, f(x)) - (1-c) \cdot \varepsilon \Rightarrow \\ (1-c) \cdot d(x, f(x)) &\leq (1-c) \cdot \varepsilon \Rightarrow \\ d(x, f(x)) &\leq \varepsilon \end{aligned}$$

Therefore x satisfies (Oa) and therefore is a solution of NONMETRICBANACH.

(CO2) $\Longrightarrow$ (Oc).

(CO3) $\Longrightarrow$ (Od).

This means that any solution to CONTINUOUS LOCALOPT at the instance $(f', p, \varepsilon', \lambda')$ can produce a solution to the instance $(f, d, \varepsilon, \lambda, c)$ of the NONMETRICBANACH problem. Therefore NONMETRICBANACH $\in$ CLS.

Now we are going to show the opposite direction and reduce CONTINUOUS LOCALOPT to NONMETRICBANACH. Starting from an instance $(f, p, \varepsilon, \lambda)$ of CONTINUOUS LOCALOPT we define for any $x, y \in [0,1]^3$,

$$\kappa(x,y) = \min\left\{-\frac{p(x)}{\varepsilon}, -\frac{p(y)}{\varepsilon}\right\}$$

We also remind the reader the definition of the *discrete metric* (Definition 6)

$$d_S(x,y) = 1 \text{ if } x \neq y \text{ and } d_S(x,x) = 0$$

Based on these definitions we create the following instance of NONMETRICBANACH

$$\begin{aligned} f'(x) &= f(x) \\ d(x,y) &= c^{\kappa(x,y)} d_S(x,y) \\ \varepsilon' &= \frac{1}{c} \\ \lambda' &= \max\left\{\lambda, c^{-1/\varepsilon}\lambda\frac{\ln(1/c)}{\varepsilon}\right\} \\ c &= 1 - 0.1\varepsilon \end{aligned}$$

As in the previous reduction we have to show that any result of the NONMETRICBANACH with input

$(f, d, \varepsilon', \lambda, c)$ will give us a result of CONTINUOUS LOCALOPT with input $(f, p, \varepsilon, \lambda)$.

(Oa) $\Longrightarrow$ If $p(f(x)) \geq p(x)$ then x satisfies (CO1) and therefore gives us a solution of CONTINUOUS LOCALOPT. Otherwise we can see that $\kappa(x, f(x)) = -p(x)/\varepsilon$ and $x \neq f(x)$

so

$$\begin{aligned} d(x, f(x)) \leq \varepsilon' &\Rightarrow c^{-\frac{p(x)}{\varepsilon}} \leq \varepsilon' \\ &\Rightarrow \frac{p(x)}{\varepsilon} \log(1/c) \leq \log(\varepsilon') \\ &\Rightarrow p(x) \leq \varepsilon \frac{\log(\varepsilon')}{\log(1/c)} \\ &\Rightarrow p(x) \leq \varepsilon \end{aligned}$$

now $p(f(x)) \geq 0 \geq p(x) - \varepsilon$ and so x satisfies (CO1) and therefore gives us a solution of CONTINUOUS LOCALOPT.

(Ob) $\Longrightarrow$ As in the previous case we may assume that $p(f(x)) \leq p(x) - \varepsilon$ and that $p(f(y)) \leq p(y) - \varepsilon$. This implies the following

$$\frac{p(f(x))}{\varepsilon} \leq \frac{p(x)}{\varepsilon} - 1 \tag{4.1}$$

$$\frac{p(f(y))}{\varepsilon} \leq \frac{p(y)}{\varepsilon} - 1 \tag{4.2}$$

Also without loss of generality we can assume that $p(x) > p(y)$. If also $p(f(x)) \geq p(f(y))$ then $\kappa(x, y) = -p(x)/\varepsilon$ and $\kappa(f(x), f(y)) = -p(f(x))/\varepsilon$. Therefore

$$d(x, y) = c^{-\frac{p(x)}{\varepsilon}} d(f(x), f(y)) = c^{-\frac{p(f(x))}{\varepsilon}}$$

But because of (4.1) we have that if (Ob) is satisfied then

$$d(f(x), f(y)) = c^{-\frac{p(f(x))}{\varepsilon}} > c \cdot c^{-\frac{p(x)}{\varepsilon}} = cd(x, y)$$

This implies that

$$\frac{p(f(x))}{\varepsilon} > \frac{p(x)}{\varepsilon} - 1 \Rightarrow p(f(x)) > p(x) - \varepsilon$$

Therefore x satisfies (CO1) and therefore gives us a solution of CONTINUOUS LOCALOPT.

Now similarly if $p(f(y)) > p(f(x))$ then $p(f(y)) > p(x) - \varepsilon$. But by our assumption that

$p(x) > p(y)$ we get $p(f(y)) > p(y) - \varepsilon$. Therefore y satisfies (CO1) and therefore gives us a solution of CONTINUOUS LOCALOPT.

(Oc) $\Longrightarrow$ (CO2).

(Od) $\Longrightarrow$ We will analyze the function $h(x) = c^{-x}$ when $x \in [0, 1/\varepsilon]$. By the mean value theorem we have that the Lipschitz constant ℓ_h of h is less that $\max_{x\in[0,1/\varepsilon]} h'(x)$. But

$$h'(x) = \left(e^{-x\ln c}\right)' = \ln(1/c)c^{-x}$$

But because $c < 1$ we have that

$$\max_{x\in[0,1/\varepsilon]} h'(x) = c^{-1/\varepsilon}\ln(1/c)$$

As before if $\kappa(x, f(x)) \neq x$ then $p(f(x)) \geq p(x)$ and therefore x is a solution to CONTINUOUS LOCALOPT and the same for is true for y. Therefore $d(x, f(x)) = c^{-p(x)/\varepsilon}$ and $d(y, f(y)) = c^{-p(y)/\varepsilon}$. We have

$$\begin{aligned} |d(x, f(x)) - d(y, f(y))| = \left|c^{-p(x)/\varepsilon} - c^{-p(y)/\varepsilon}\right| &\leq \left(\max_{x\in[0,1/\varepsilon]} h'(x)\right)\left|\frac{p(x)}{\varepsilon} - \frac{p(y)}{\varepsilon}\right| \\ \Rightarrow \left|c^{-p(x)/\varepsilon} - c^{-p(y)/\varepsilon}\right| &\leq c^{-1/\varepsilon}\frac{\ln(1/c)}{\varepsilon}|p(x) - p(y)| \end{aligned}$$

Now if $|p(x) - p(y)| > \lambda|x - y|$ then x, y satisfy (CO3) and we have a solution for CONTINUOUS LOCALOPT. So $|p(x) - p(y)| \leq \lambda|x - y|$ and from the last inequality we have that

$$|d(x, f(x)) - d(y, f(y))| \leq c^{-1/\varepsilon}\lambda\frac{\ln(1/c)}{\varepsilon}|x - y|$$

But this contradicts with (Od) since

$$\lambda' = \max\left\{\lambda, c^{-1/\varepsilon}\lambda\frac{\ln(1/c)}{\varepsilon}\right\}$$

Finally it is easy to see that the arithmetic circuit that we used for the reduction can be computed in polynomial time by a Turing Machine. The only function that needs for explanation is that of d. The reason is the term $c^{\kappa(x,y)}$. Since $c < 1$ we need to bound the

size of $c^{-\kappa x,y}$. We observe that $c^{-\kappa(x,y)} \leq c^{-1/\varepsilon} = (1-0.1\varepsilon)^{-1/\varepsilon} \leq e^{10}$. Therefore the size of $c^{-1/\varepsilon}$ is bounded and can be computed in polynomial time.

The fact that NONMETRICBANACH is total comes easily from the fact that NONMETRICBANACH $\in$ CLS. □

Notice that the function d in the definition of NONMETRICBANACH does not satisfy the properties of a distance metric function. But in order to be able to describe exactly the application of Banach it's important to also have d to be a distance metric. It would be even more exciting if d was a complete metric for our space which is $[0,1]^3$. In order to satisfy this conditions, surprisingly, we don't need to include these properties of d in the NONMETRICBANACH problem. Instead we can put them in a semantic way and still the problem will be complete for CLS.

We start with the semantic version of the NONMETRICBANACH problem but with also the distance metric and the completeness assumptions.

Definition 28. *The problem* BANACH *has the same input and output with the problem* NONMETRICBANACH, *but we have also the promise that d satisfies the properties of a distance metric function (Definition 4)* [1] *and also* $([0,1]^3, d)$ *is a complete metric space.*

Theorem 22. *The* BANACH *problem is* CLS*-complete.*

Proof. Obviously because of Theorem 21, BANACH belongs to CLS.

For the opposite direction, we use the same reduction as in the proof of Theorem 21. We then prove that d satisfies the desired properties. We remind that we used the following

[1]
(i) for all $x, x' \in [0,1]^3$, $d(x,x') \geq 0$.
(ii) for all $x, x' \in [0,1]^3$, $x \neq x' \Leftrightarrow d(x,x') \neq 0$.
(iii) for all $x, x' \in [0,1]^3$, $d(x,x') = d(x',x)$.
(iv) for all $x, y, z \in [0,1]^3$, $d(x,y) > d(x,z) + d(z,y)$.

instance of BANACH for the reduction

$$\begin{aligned} f'(x) &= f(x) \\ d(x,y) &= c^{\kappa(x,y)} d_S(x,y) \\ \varepsilon' &= \frac{1}{c} \\ \lambda' &= \max\left\{\lambda, c^{-1/\varepsilon}\lambda\frac{\ln(1/c)}{\varepsilon}\right\} \\ c &= 1-\varepsilon \end{aligned}$$

We first prove that d is a distance metric.

(i) Obvious from the definition of d.

(ii) If $x \neq y$ then $d_S(x,y) > 0$. Also always $c^{\kappa(x,y)} > 0$, therefore $d(x,y) > 0$. Now since $d_S(x,x) = 0$ we also have $d(x,x) = 0$.

(iii) It is obvious from the definition of κ that $\kappa(x,y) = \kappa(y,x)$ and since d_S is a distance metric, the same is true for the d_S and thus for d.

(iv) Without loss of generality we assume that $p(x) \geq p(y)$. We consider the following cases

$\mathbf{p(x) \geq p(y) \geq p(z)}$ then we have $d(x,y) = c^{-p(x)/\varepsilon}$, $d(x,z) = c^{-p(x)/\varepsilon}$, $d(z,y) = c^{-p(y)/\varepsilon}$ therefore obviously $d(x,y) \leq d(x,z) + d(z,y)$.

$\mathbf{p(x) \geq p(z) \geq p(y)}$ then we have $d(x,y) = c^{-p(x)/\varepsilon}$, $d(x,z) = c^{-p(x)/\varepsilon}$, $d(z,y) = c^{-p(z)/\varepsilon}$ therefore obviously $d(x,y) \leq d(x,z) + d(z,y)$.

$\mathbf{p(z) \geq p(x) \geq p(y)}$ then we have $d(x,y) = c^{-p(x)/\varepsilon}$, $d(x,z) = c^{-p(z)/\varepsilon}$, $d(z,y) = c^{-p(z)/\varepsilon}$ therefore obviously triangle inequality is equivalent with

$$\frac{p(x)}{\varepsilon}\log(1/c) \leq \frac{p(z)}{\varepsilon}\log(1/c) + \log 2$$

which holds because of the assumption that $p(z) \geq p(x)$.

Finally we will show the completeness of $([0,1]^3, d)$. We first observe that for all $x \neq y$, $d(x,y) > 1$, this comes from the fact that $c < 1$ and so $c^{-p(x)/\varepsilon} > 1$.

Now let (x_n) be a Cauchy sequence then $\forall \delta > 0$, $\exists N \in \mathbb{N}$ such that $\forall n, m > N$, $d(x_n, x_m) \leq \delta$. We set $\delta = 1/2$ then there exists $N \in \mathbb{N}$ such that $\forall n, m > N$, $d(x_n, x_m) < 1/2$. But from the previous observation this implies $d(x_n, x_m) = 0$ and since d defines a metric we get $x_n = x_m$. Therefore (x_n) is constant for all $n > N$ and obviously converges. This means that every Cauchy sequence converges and so $([0,1]^3, d)$ is a complete metric space. $\square$

Chapter 5

Runtime Analysis of EM for Mixtures of Two Gaussians with known Covariances

In this chapter, we provide global convergence guarantees for the expectation-maximization (EM) algorithm applied to mixtures of two Gaussians with known covariance matrices. We show that EM converges geometrically to the correct mean vectors, and provide simple, closed-form expressions for the convergence rate. As a simple illustration, we show that in one dimension ten steps of the EM algorithm initialized at $+\infty$ result in less than 1% error estimation of the means. This chapter is based on the of work of Daskalakis, Tzamos, Zampetakis [24].

5.1 Introduction to EM

The *Expectation-Maximization (EM) algorithm* [27, 57, 49] is one of the most widely used heuristics for maximizing likelihood in statistical models with latent variables. Consider a probability distribution $p_{\boldsymbol{\lambda}}$ sampling $(\boldsymbol{X}, \boldsymbol{Z})$, where $\boldsymbol{X}$ is a vector of observable random variables, $\boldsymbol{Z}$ a vector of non-observable random variables and $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$ a vector of parameters. Given independent samples $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ of the observed random variables, the goal of maximum likelihood estimation is to select $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$ maximizing the log-likelihood of the samples, namely $\sum_i \log p_{\boldsymbol{\lambda}}(\boldsymbol{x}_i)$. Unfortunately, computing $p_{\boldsymbol{\lambda}}(\boldsymbol{x}_i)$ involves summing $p_{\boldsymbol{\lambda}}(\boldsymbol{x}_i, \boldsymbol{z}_i)$ over all possible values of $\boldsymbol{z}_i$, which commonly results in a log-likelihood function that is non-convex with respect to $\boldsymbol{\lambda}$ and therefore hard to optimize. In this context, the EM algorithm proposes

the following heuristic:

- Start with an initial guess $\boldsymbol{\lambda}^{(0)}$ of the parameters.
- For all $t \geq 0$, until convergence:
  - (E-Step) For each sample i, compute the posterior

$$Q_i^{(t)}(\boldsymbol{z}) := p_{\boldsymbol{\lambda}^{(t)}}(\boldsymbol{Z} = \boldsymbol{z} | \boldsymbol{X} = \boldsymbol{x}_i)$$

  - (M-Step) Set

$$\boldsymbol{\lambda}^{(t+1)} := \arg\max_{\boldsymbol{\lambda}} \sum_i \sum_{\boldsymbol{z}} Q_i^{(t)}(\boldsymbol{z}) \log \frac{p_{\boldsymbol{\lambda}}(\boldsymbol{x}_i, \boldsymbol{z})}{Q_i^{(t)}(\boldsymbol{z})}$$

Intuitively, the E-step of the algorithm uses the current guess of the parameters, $\boldsymbol{\lambda}^{(t)}$, to form beliefs, $Q_i^{(t)}$, about the state of the (non-observable) $\boldsymbol{Z}$ variables for each sample i. Then the M-step uses the new beliefs about the state of $\boldsymbol{Z}$ for each sample to maximize with respect to $\boldsymbol{\lambda}$ a lower bound on $\sum_i \log p_{\boldsymbol{\lambda}}(\boldsymbol{x}_i)$. Indeed, by the concavity of the log function, the objective function used in the M-step of the algorithm is a lower bound on the true log-likelihood for all values of $\boldsymbol{\lambda}$, and it equals the true log-likelihood for $\boldsymbol{\lambda} = \boldsymbol{\lambda}^{(t)}$. From these observations, it follows that the above alternating procedure improves the true log-likelihood until convergence.

Indeed we can view EM as an alternating maximization problem of the function

$$D(\boldsymbol{Q}, \boldsymbol{\lambda}) = \sum_i \sum_{\boldsymbol{z}} Q_i^{(t)}(\boldsymbol{z}) \log \frac{p_{\boldsymbol{\lambda}}(\boldsymbol{x}_i, \boldsymbol{z})}{Q_i^{(t)}(\boldsymbol{z})} - \sum_i \sum_{\boldsymbol{z}} Q_i^{(t)}(\boldsymbol{z}) \log Q_i^{(t)}(\boldsymbol{z})$$

For the M-step it is obvious that

$$\boldsymbol{\lambda}^{(t+1)} := \arg\max_{\boldsymbol{\lambda}} D(\boldsymbol{Q}, \boldsymbol{\lambda})$$

It is also not difficult to verify that

$$\boldsymbol{\lambda}^{(t+1)} := \arg\max_{\boldsymbol{Q}} D(\boldsymbol{Q}, \boldsymbol{\lambda})$$

Finally we observe that the log-likelihood is equal to $D(\boldsymbol{Q}^{(t+1)}, \lambda^t)$. Therefore the likelihood at every step increases.

Despite its wide use and practical significance, little is known about whether and under what conditions EM converges to the true maximum likelihood estimator. A few works establish local convergence of the algorithm to stationary points of the log-likelihood function [57, 54, 13], and even fewer local convergence to the MLE [49, 4]. Besides local convergence to the MLE, it is also known that badly initialized EM may settle far from the MLE both in parameter and in likelihood distance [57].

The lack of theoretical understanding of the convergence properties of EM is intimately related to the non-convex nature of the optimization it performs. Our paper aims to illuminate why EM works well in practice and develop techniques for understanding its behavior. We do so by analyzing one of the most basic and natural, yet still challenging, statistical models EM may be applied to, namely balanced mixtures of two multi-dimensional Gaussians with equal and known covariance matrices. In particular, the family of parameterized density functions we will be considering are:

$$p_{\boldsymbol{\mu}_1,\boldsymbol{\mu}_2}(\boldsymbol{x}) = 0.5 \cdot \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_1, \Sigma) + 0.5 \cdot \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_2, \Sigma),$$

where Σ is a known covariance matrix, $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2)$ are unknown parameters, and $\mathcal{N}(\boldsymbol{\mu}, \Sigma; \boldsymbol{x})$ represents the Gaussian density with mean $\boldsymbol{\mu}$ and covariance matrix Σ, i.e.

$$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) = \frac{1}{\sqrt{2\pi \det \Sigma}} \exp\left(-0.5(\boldsymbol{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right).$$

To elucidate the optimization nature of the algorithm and avoid analytical distractions arising in the finite sample regime, it has been standard practice in the literature of theoretical analyses of EM to consider the "population version" of the algorithm, where the EM iterations are performed assuming access to infinitely many samples from a distribution $p_{\boldsymbol{\mu}_1,\boldsymbol{\mu}_2}$ as above. With infinitely many samples, we can identify the mean, $\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2}$, of $p_{\boldsymbol{\mu}_1,\boldsymbol{\mu}_2}$, and re-parametrize

the density around the mean as follows:

$$p_{\boldsymbol{\mu}}(\boldsymbol{x}) = 0.5 \cdot \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) + 0.5 \cdot \mathcal{N}(\boldsymbol{x}; -\boldsymbol{\mu}, \Sigma). \tag{5.1}$$

We will study the convergence of EM when we perform iterations with respect to the parameter $\boldsymbol{\mu}$ of $p_{\boldsymbol{\mu}}(\boldsymbol{x})$ in (5.1). Starting with an initial guess $\boldsymbol{\lambda}^{(0)}$ for the unknown mean vector $\boldsymbol{\mu}$, the t-th iteration of EM amounts to the following update:

$$\boldsymbol{\lambda}^{(t+1)} = M(\boldsymbol{\lambda}^{(t)}, \boldsymbol{\mu}) \triangleq \frac{\mathbb{E}_{\boldsymbol{x} \sim p_{\boldsymbol{\mu}}} \left[\frac{0.5\mathcal{N}(\boldsymbol{x}; \boldsymbol{\lambda}^{(t)}, \Sigma)}{p_{\boldsymbol{\lambda}^{(t)}}(\boldsymbol{x})} \boldsymbol{x} \right]}{\mathbb{E}_{\boldsymbol{x} \sim p_{\boldsymbol{\mu}}} \left[\frac{0.5\mathcal{N}(\boldsymbol{x}; \boldsymbol{\lambda}^{(t)}, \Sigma)}{p_{\boldsymbol{\lambda}^{(t)}}(\boldsymbol{x})} \right]}, \tag{5.2}$$

where we have compacted both the E- and M-step of EM into one update.

To illuminate the EM update formula, we take expectations with respect to $\boldsymbol{x} \sim p_{\boldsymbol{\mu}}$ because we are studying the population version of EM, where we assume access to infinitely many samples from $p_{\boldsymbol{\mu}}$. For each sample $\boldsymbol{x}$, the ratio $\frac{0.5\mathcal{N}(\boldsymbol{x}; \boldsymbol{\lambda}^{(t)}, \Sigma)}{p_{\boldsymbol{\lambda}^{(t)}}(\boldsymbol{x})}$ is our belief, at step t, that $\boldsymbol{x}$ was sampled from the first Gaussian component of $p_{\boldsymbol{\mu}}$, namely the one for which our current estimate of its mean vector is $\boldsymbol{\lambda}^{(t)}$. (The complementary probability is our present belief that $\boldsymbol{x}$ was sampled from the other Gaussian component.) Given these beliefs for all vectors $\boldsymbol{x}$, the update (5.2) is the result of the M-step of EM. Intuitively, our next guess $\boldsymbol{\lambda}^{(t+1)}$ for the mean vector of the first Gaussian component is a weighted combination over all samples $\boldsymbol{x} \sim p_{\boldsymbol{\mu}}$ where the weight of every $\boldsymbol{x}$ is our belief that it came from the first Gaussian component.

Our main result is the following:

Informal Theorem. *Whenever the initial guess $\boldsymbol{\lambda}^{(0)}$ is not equidistant to $\boldsymbol{\mu}$ and $-\boldsymbol{\mu}$, EM converges geometrically to either $\boldsymbol{\mu}$ or $-\boldsymbol{\mu}$, with convergence rate that improves as $t \to \infty$. We provide a simple, closed form expression of the convergence rate as a function of $\boldsymbol{\lambda}^{(t)}$ and $\boldsymbol{\mu}$.*

A formal statement is provided as Theorem 24 in Section 5.4. We start with the proof of the single-dimensional version, presented as Theorem 23 in Section 5.3. As a simple illustration of our result, we show in Section 5.5 that, in one dimension, when our original

guess $\lambda^{(0)} = +\infty$ and the signal-to-noise ratio $\mu/\sigma = 1$, 10 steps of the EM algorithm result in 1% error.

Despite the simplicity of the case we consider, no global convergence results were known prior to our work. Balakrishnan, Wainwright and Yu [4] studied the same setting proving only local convergence, i.e. convergence only when the initial guess is close to the true parameters. In this work, we study the problem under arbitrary starting points and completely characterize the fixed points of EM. We show that other than a measure-zero subset of the space, any initialization of the EM algorithm converges in a few steps to the true parameters of the Gaussians and provide explicit bounds on the convergence rate. To achieve this, we follow an orthogonal approach to [4]: Instead of trying to directly compute the number of steps required to reach convergence *for a specific instance of the problem*, we study the sensitivity of the EM iteration *as the instance varies.* This enables us to relate the behavior of EM on all instances of the Gaussian mixture problem and gain a handle on the convergence rate of EM on all instances at once.

5.1.1 Related Work on Learning Mixtures of Gaussians

We have already outlined the literature on the Expectation-Maximization algorithm. Several results study its local convergence properties and there are known cases where badly initialized EM fails to converge. See above.

There is also a large body of literature on learning mixtures of Gaussians. A long line of work initiated by Dasgupta [18, 2, 55, 1, 39, 19, 12, 10, 11] provides rigorous guarantees on recovering the parameters of Gaussians in a mixture under separability assumptions, while later work [38, 45, 6] has established guarantees under minimal information theoretic assumptions. More recent work [31] provides tight bounds on the number of samples necessary to recover the parameters of the Gaussians as well as improved algorithms, while another strand of the literature studies proper learning with improved running times and sample sizes [52, 22]. Finally, there has been work on methods exploiting general position assumptions or performing smoothed analysis [33, 29].

In practice, the most common algorithm for learning mixtures of Gaussians is the Expectation - Maximization algorithm, with the practical experience that it performs well in a

broad range of scenarios despite the lack of theoretical guarantees. In recent work, Balakrishnan, Wainwright and Yu [4] studied the convergence of EM in the case of an equal-weight mixture of two Gaussians with the same and known covariance matrix, showing local convergence guarantees. In particular, they show that when EM is initialized close enough to the actual parameters, then it converges. In this work, we revisit the same setting considered by [4] but establish *global convergence guarantees.* We show that, for any initialization of the parameters, the EM algorithm converges geometrically to the true parameters. We also provide a simple and explicit formula for the rate of convergence. Concurrent and independent work by Xu, Hsu and Maleki [58] has also provided global and geometric convergence guarantees for the same setting, as well as a slightly more general setting where the mean of the mixture is unknown, but they do not provide explicit convergence rates.

5.2 Preliminary Observations

In this section we illustrate some simple properties of the EM update (5.2) and simplify the formula. First, it is easy to see that plugging in the values $\boldsymbol{\lambda} \in \{-\boldsymbol{\mu}, \mathbf{0}, \boldsymbol{\mu}\}$ into $M(\boldsymbol{\lambda}, \boldsymbol{\mu})$ results into

$$M(-\boldsymbol{\mu}, \boldsymbol{\mu}) = -\boldsymbol{\mu} \quad ; \quad M(\mathbf{0}, \boldsymbol{\mu}) = \mathbf{0} \quad ; \quad M(\boldsymbol{\mu}, \boldsymbol{\mu}) = \boldsymbol{\mu}. \tag{5.3}$$

In particular, for all $\boldsymbol{\mu}$, these values are certainly fixed points of the EM iteration. Next, we rewrite $M(\boldsymbol{\lambda}, \boldsymbol{\mu})$ as follows:

$$M(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \frac{\frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu},\Sigma)}\left[\frac{0.5\mathcal{N}(\boldsymbol{x};\boldsymbol{\lambda},\Sigma)}{p_{\boldsymbol{\lambda}}(\boldsymbol{x})}\boldsymbol{x}\right] + \frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(-\boldsymbol{\mu},\Sigma)}\left[\frac{0.5\mathcal{N}(\boldsymbol{x};\boldsymbol{\lambda},\Sigma)}{p_{\boldsymbol{\lambda}}(\boldsymbol{x})}\boldsymbol{x}\right]}{\frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu},\Sigma)}\left[\frac{0.5\mathcal{N}(\boldsymbol{x};\boldsymbol{\lambda},\Sigma)}{p_{\boldsymbol{\lambda}}(\boldsymbol{x})}\right] + \frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(-\boldsymbol{\mu},\Sigma)}\left[\frac{0.5\mathcal{N}(\boldsymbol{x};\boldsymbol{\lambda},\Sigma)}{p_{\boldsymbol{\lambda}}(\boldsymbol{x})}\right]}.$$

It is easy to observe that by symmetry this simplifies to

$$M(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \frac{\frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu},\Sigma)}\left[\frac{\frac{1}{2}\mathcal{N}(\boldsymbol{x};\boldsymbol{\lambda},\Sigma)-\frac{1}{2}\mathcal{N}(\boldsymbol{x};-\boldsymbol{\lambda},\Sigma)}{\frac{1}{2}\mathcal{N}(\boldsymbol{x};\boldsymbol{\lambda},\Sigma)+\frac{1}{2}\mathcal{N}(\boldsymbol{x};-\boldsymbol{\lambda},\Sigma)}\boldsymbol{x}\right]}{\frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu},\Sigma)}\left[\frac{\frac{1}{2}\mathcal{N}(\boldsymbol{x};\boldsymbol{\lambda},\Sigma)+\frac{1}{2}\mathcal{N}(\boldsymbol{x};-\boldsymbol{\lambda},\Sigma)}{\frac{1}{2}\mathcal{N}(\boldsymbol{x};\boldsymbol{\lambda},\Sigma)+\frac{1}{2}\mathcal{N}(\boldsymbol{x};-\boldsymbol{\lambda},\Sigma)}\right]} = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu},\Sigma)}\left[\frac{\mathcal{N}(\boldsymbol{x};\boldsymbol{\lambda},\Sigma)-\mathcal{N}(\boldsymbol{x};-\boldsymbol{\lambda},\Sigma)}{\mathcal{N}(\boldsymbol{x};\boldsymbol{\lambda},\Sigma)+\mathcal{N}(\boldsymbol{x};-\boldsymbol{\lambda},\Sigma)}\boldsymbol{x}\right].$$

Simplifying common terms in the density functions $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\lambda}, \Sigma)$, we get that

$$M(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \mathbb{E}_{\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)} \left[\frac{\exp\left(\boldsymbol{\lambda}^T \Sigma^{-1} \boldsymbol{x}\right) - \exp\left(-\boldsymbol{\lambda}^T \Sigma^{-1} \boldsymbol{x}\right)}{\exp\left(\boldsymbol{\lambda}^T \Sigma^{-1} \boldsymbol{x}\right) + \exp\left(-\boldsymbol{\lambda}^T \Sigma^{-1} \boldsymbol{x}\right)} \boldsymbol{x} \right].$$

We thus get the following expression for the EM iteration

$$M(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \mathbb{E}_{\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)} \left[\tanh(\boldsymbol{\lambda}^T \Sigma^{-1} \boldsymbol{x}) \boldsymbol{x} \right]. \tag{5.4}$$

5.3 Single-dimensional Convergence

In the single dimensional case the EM algorithm takes the following form according to (5.4).

$$\lambda^{(t+1)} = M(\lambda^{(t)}, \mu) = \mathbb{E}_{x \sim \mathcal{N}(\mu, \sigma^2)} \left[\tanh\left(\frac{\lambda^{(t)} x}{\sigma^2} \right) x \right] \tag{5.5}$$

Observe that the function $M(\lambda, \mu)$ is increasing with respect to λ. Indeed the partial derivative of M with respect to λ is

$$\frac{\partial M(\lambda, \mu)}{\partial \lambda} = \mathbb{E}_{x \sim \mathcal{N}(\mu, \sigma^2)} \left[\tanh'\left(\frac{\lambda^{(t)} x}{\sigma^2} \right) \frac{x^2}{\sigma^2} \right]$$

which is strictly greater than zero since the $\tanh'$ function is strictly positive.

We will show next that the fixed points we identified at (5.3) are the only fixed points of $M(\cdot, \mu)$. When initialized with $\lambda^{(0)} > 0$ (resp. $\lambda^{(0)} < 0$), the EM algorithm converges to $\mu > 0$ (resp. to $-\mu < 0$). The point $\lambda = 0$ is an unstable fixed point.

Theorem 23. *In the single dimensional case, when $\lambda^{(0)}, \mu > 0$, the parameters $\lambda^{(t)}$ satisfy*

$$\left|\lambda^{(t+1)} - \mu\right| \leq \kappa^{(t)} \left|\lambda^{(t)} - \mu\right| \quad \text{where } \kappa^{(t)} = \exp\left(-\frac{\min(\lambda^{(t)}, \mu)^2}{2\sigma^2} \right)$$

Moreover $\kappa^{(t)}$ is a decreasing function of t.

Proof. For simplicity we will use λ for $\lambda^{(t)}$, λ' for $\lambda^{(t+1)}$ and we will assume that $X \sim \mathcal{N}(0, \sigma^2)$.

By a simple change of variables we can see that

$$M(\lambda,\mu)=\mathbb{E}\left[\tanh\left(\frac{\lambda(X+\mu)}{\sigma^2}\right)(X+\mu)\right]$$

The main idea is to use the Mean Value Theorem with respect to the second coordinate of the function M on the interval $[\lambda,\mu]$.

$$\frac{M(\lambda,\mu)-M(\lambda,\lambda)}{\mu-\lambda}=\left.\frac{\partial M(\lambda,y)}{\partial y}\right|_{y=\xi}\text{ with }\xi\in(\lambda,\mu)$$

But we know that $M(\lambda,\lambda)=\lambda$ and $M(\lambda,\mu)=\lambda'$ and therefore we get

$$\lambda'-\lambda\geq\left(\min_{\xi\in[\lambda,\mu]}\left.\frac{\partial M(\lambda,y)}{\partial y}\right|_{y=\xi}\right)(\mu-\lambda)$$

which is equivalent to

$$|\lambda'-\mu|\leq\left(1-\min_{\xi\in[\lambda,\mu]}\left.\frac{\partial M(\lambda,y)}{\partial y}\right|_{y=\xi}\right)|\lambda-\mu|$$

where we have used the fact that $\lambda'<\mu$ which is comes from the fact that $M(\lambda,\mu)$ is increasing with respect to λ and that $M(\mu,\mu)=\mu$.

The only thing that remains to complete our proof is to prove a lower bound of the partial derivative of M with respect to μ.

$$\left.\frac{\partial M(\lambda,y)}{\partial y}\right|_{y=\xi}=\mathbb{E}\left[\frac{\lambda}{\sigma^2}\tanh'\left(\frac{\lambda(X+\xi)}{\sigma^2}\right)(X+\xi)+\tanh\left(\frac{\lambda(X+\xi)}{\sigma^2}\right)\right]$$

The first term is non-negative, Lemma 10. The second term is at least $1-\exp\left[-\frac{\min(\xi,\lambda)\cdot\xi}{2\sigma^2}\right]$, Lemma 11 and the theorem follows. □

Lemma 10. *Let $\alpha,\beta>0$ and $X\sim\mathcal{N}(\alpha,\sigma^2)$ then $\mathbb{E}\left[\tanh'\left(\beta X/\sigma^2\right)X\right]\geq 0$.*

Proof.

$$\mathbb{E}\left[\tanh'\left(\frac{\beta X}{\sigma^2}\right)X\right]=\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty}\tanh'\left(\frac{\beta y}{\sigma^2}\right)y\exp\left(-\frac{(y-\alpha)^2}{2\sigma^2}\right)dy$$

But now we can see that since $\tanh'$ is an even function and since for any $y > 0$ we have $\exp\left(-\frac{(y-\alpha)^2}{2\sigma^2}\right) \geq \exp\left(-\frac{(-y-\alpha)^2}{2\sigma^2}\right)$ then

$$-\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{0}\tanh'\left(\frac{\beta y}{\sigma^2}\right)y\exp\left(-\frac{(y-\alpha)^2}{2\sigma^2}\right)dy \leq \frac{1}{\sqrt{2\pi}\sigma}\int_{0}^{\infty}\tanh'\left(\frac{\beta y}{\sigma^2}\right)y\exp\left(-\frac{(y-\alpha)^2}{2\sigma^2}\right)dy$$

which means that $\mathbb{E}\left[\tanh'\left(\beta X/\sigma^2\right)X\right] \geq 0$. □

Lemma 11. *Let $\alpha, \beta > 0$ and $X \sim \mathcal{N}(\alpha, \sigma^2)$ then $\mathbb{E}\left[\tanh\left(\beta X/\sigma^2\right)\right] \geq 1 - \exp\left[-\frac{\min(\alpha,\beta)\cdot\alpha}{2\sigma^2}\right]$.*

Proof. Note that $\mathbb{E}\left[\tanh\left(\beta X/\sigma^2\right)\right]$ is increasing as a function of β as its derivative with respect to β is positive by Lemma 10. It thus suffices to show that $\mathbb{E}\left[\tanh\left(\beta X/\sigma^2\right)\right] \geq 1 - \exp\left[-\frac{\alpha\beta}{2\sigma^2}\right]$ when $\beta \leq \alpha$. We have that

$$\begin{aligned}
\mathbb{E}\left[1-\tanh\left(\beta X/\sigma^2\right)\right] &= \mathbb{E}\left[\frac{2}{\exp(2\beta X/\sigma^2)+1}\right] \leq \mathbb{E}\left[\frac{1}{\exp(\beta X/\sigma^2)}\right] \\
&= \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty}\frac{\exp\left(-\frac{(x-\alpha)^2}{2\sigma^2}\right)}{\exp(\beta x/\sigma^2)}dx = \frac{\exp\left(\frac{(\alpha-\beta)^2-\alpha^2}{2\sigma^2}\right)}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty}\exp\left(-\frac{(x-\alpha+\beta)^2}{2\sigma^2}\right)dx \\
&= \exp\left(\frac{(\alpha-\beta)^2-\alpha^2}{2\sigma^2}\right) \leq \exp\left(-\frac{\alpha\beta}{2\sigma^2}\right)
\end{aligned}$$

which completes the proof. □

5.4 Multi-dimensional Convergence

In the multidimensional case, the EM algorithm takes the form of (5.4). In this case, we will quantify our approximation guarantees using the *Mahalanobis distance* $\|\cdot\|_\Sigma$ between vectors with respect to matrix Σ, defined as follows:

$$\|\boldsymbol{x}-\boldsymbol{y}\|_\Sigma = \sqrt{(\boldsymbol{x}-\boldsymbol{y})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{y})}.$$

We will show that the fixed points identified in (5.3) are the only fixed points of $M(\cdot, \boldsymbol{\mu})$. When initialized with $\boldsymbol{\lambda}^{(0)}$ such that $\left\|\boldsymbol{\lambda}^{(0)}-\boldsymbol{\mu}\right\|_\Sigma < \left\|\boldsymbol{\lambda}^{(0)}+\boldsymbol{\mu}\right\|_\Sigma$ (resp. $\left\|\boldsymbol{\lambda}^{(0)}-\boldsymbol{\mu}\right\|_\Sigma > \left\|\boldsymbol{\lambda}^{(0)}+\boldsymbol{\mu}\right\|_\Sigma$), the EM algorithm converges to $\boldsymbol{\mu}$ (resp. to $-\boldsymbol{\mu}$). The algorithm converges to $\boldsymbol{\lambda} = \mathbf{0}$ when initialized with $\left\|\boldsymbol{\lambda}^{(0)}-\boldsymbol{\mu}\right\|_\Sigma = \left\|\boldsymbol{\lambda}^{(0)}+\boldsymbol{\mu}\right\|_\Sigma$. In particular,

Theorem 24. *Whenever* $\left\|\boldsymbol{\lambda}^{(0)}-\boldsymbol{\mu}\right\|_{\Sigma}<\left\|\boldsymbol{\lambda}^{(0)}+\boldsymbol{\mu}\right\|_{\Sigma}$, *i.e. the initial guess is closer to* $\boldsymbol{\mu}$ *than* $-\boldsymbol{\mu}$, *the estimates* $\boldsymbol{\lambda}^{(t)}$ *of the EM algorithm satisfy*

$$\left\|\boldsymbol{\lambda}^{(t+1)}-\boldsymbol{\mu}\right\|_{\Sigma}\leq\kappa^{(t)}\left\|\boldsymbol{\lambda}^{(t)}-\boldsymbol{\mu}\right\|_{\Sigma},\qquad \textit{where } \kappa^{(t)}=\exp\left(-\frac{\min\left(\boldsymbol{\lambda}^{(t),T}\Sigma^{-1}\boldsymbol{\lambda}^{(t)},\boldsymbol{\mu}^{T}\Sigma^{-1}\boldsymbol{\lambda}^{(t)}\right)^{2}}{2\boldsymbol{\lambda}^{(t),T}\Sigma^{-1}\boldsymbol{\lambda}^{(t)}}\right).$$

Moreover, $\kappa^{(t)}$ *is a decreasing function of* t. *The symmetric things hold when* $\left\|\boldsymbol{\lambda}^{(0)}-\boldsymbol{\mu}\right\|_{\Sigma}>\left\|\boldsymbol{\lambda}^{(0)}+\boldsymbol{\mu}\right\|_{\Sigma}$. *When the initial guess is equidistant to* $\boldsymbol{\mu}$ *and* $-\boldsymbol{\mu}$, *then* $\boldsymbol{\lambda}^{(t)}=\mathbf{0}$ *for all* $t>0$.

Proof. For simplicity we will use $\boldsymbol{\lambda}$ for $\boldsymbol{\lambda}^{(t)}$, $\boldsymbol{\lambda}'$ for $\boldsymbol{\lambda}^{(t+1)}$.

By applying the following change of variables $\boldsymbol{\lambda}\leftarrow\Sigma^{-1/2}\boldsymbol{\lambda}$ and $\boldsymbol{\mu}\leftarrow\Sigma^{-1/2}\boldsymbol{\mu}$ we may assume that $\Sigma=I$ where I is the identity matrix. Therefore the iteration of EM becomes

$$M(\boldsymbol{\lambda},\boldsymbol{\mu})=\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu},I)}\left[\tanh(\langle\boldsymbol{\lambda},\boldsymbol{x}\rangle)\boldsymbol{x}\right]=\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\mathbf{0},I)}\left[\tanh(\langle\boldsymbol{\lambda},\boldsymbol{x}\rangle+\langle\boldsymbol{\lambda},\boldsymbol{\mu}\rangle)(\boldsymbol{x}+\boldsymbol{\mu})\right]$$

Let $\hat{\boldsymbol{\lambda}}$ be the unit vector in the direction of $\boldsymbol{\lambda}$, $\hat{\boldsymbol{\lambda}}^{\perp}$ be the unit vector that belongs to the plane of $\boldsymbol{\mu},\boldsymbol{\lambda}$ and is perpendicular to $\boldsymbol{\lambda}$, and let $\{\boldsymbol{v}_1=\hat{\boldsymbol{\lambda}},\boldsymbol{v}_2=\hat{\boldsymbol{\lambda}}^{\perp},\boldsymbol{v}_3,...,\boldsymbol{v}_d\}$ be a basis of $\mathbb{R}^d$. We have:

$$\langle\boldsymbol{v}_i,\boldsymbol{\lambda}'\rangle=\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\mathbf{0},I)}\left[\tanh(\langle\boldsymbol{\lambda},\boldsymbol{x}\rangle+\langle\boldsymbol{\lambda},\boldsymbol{\mu}\rangle)(\langle\boldsymbol{v}_i,\boldsymbol{x}\rangle+\langle\boldsymbol{v}_i,\boldsymbol{\mu}\rangle)\right]\tag{5.6}$$

Since the Normal distribution is rotation invariant we can equivalently write:

$$\langle\boldsymbol{v}_i,\boldsymbol{\lambda}'\rangle=\mathbb{E}_{\alpha_1,...,\alpha_d\sim\mathcal{N}(0,1)}\left[\tanh(\langle\boldsymbol{\lambda},\sum_j\alpha_j\boldsymbol{v}_j\rangle+\langle\boldsymbol{\lambda},\boldsymbol{\mu}\rangle)(\langle\boldsymbol{v}_i,\sum_j\alpha_j\boldsymbol{v}_j\rangle+\langle\boldsymbol{v}_i,\boldsymbol{\mu}\rangle)\right]$$

which simplifies to

$$\langle\boldsymbol{v}_i,\boldsymbol{\lambda}'\rangle=\mathbb{E}_{\alpha_1,...,\alpha_d\sim\mathcal{N}(0,1)}\left[\tanh(\alpha_1\left\|\boldsymbol{\lambda}\right\|+\langle\boldsymbol{\lambda},\boldsymbol{\mu}\rangle)(a_i+\langle\boldsymbol{v}_i,\boldsymbol{\mu}\rangle)\right]=$$

$$\mathbb{E}_{\alpha_1\sim\mathcal{N}(0,1)}\left[\tanh(\alpha_1\left\|\boldsymbol{\lambda}\right\|+\langle\boldsymbol{\lambda},\boldsymbol{\mu}\rangle)\cdot(\mathbb{E}_{\alpha_2,...,\alpha_d\sim\mathcal{N}(0,1)}\left[a_i\right]+\langle\boldsymbol{v}_i,\boldsymbol{\mu}\rangle)\right]\tag{5.7}$$

We now consider different cases for i to further simplify Equation (5.7).

- When $i=1$, we have that $\langle\hat{\boldsymbol{\lambda}},\boldsymbol{\lambda}'\rangle=\mathbb{E}_{y\sim\mathcal{N}(0,1)}\left[\tanh(\left\|\boldsymbol{\lambda}\right\|(y+\langle\hat{\boldsymbol{\lambda}},\boldsymbol{\mu}\rangle))\left(y+\langle\hat{\boldsymbol{\lambda}},\boldsymbol{\mu}\rangle\right)\right]$.

This is equivalent with an iteration of EM in one dimension and thus from Theorem 23 we get that

$$|\langle\hat{\boldsymbol{\lambda}}, \boldsymbol{\mu}\rangle - \langle\hat{\boldsymbol{\lambda}}, \boldsymbol{\lambda}'\rangle| \leq \kappa|\langle\hat{\boldsymbol{\lambda}}, \boldsymbol{\mu}\rangle - \langle\hat{\boldsymbol{\lambda}}, \boldsymbol{\lambda}\rangle| \tag{5.8}$$

where

$$\kappa = \exp\left(-\frac{\min(\langle\hat{\boldsymbol{\lambda}}, \boldsymbol{\lambda}\rangle, \langle\hat{\boldsymbol{\lambda}}, \boldsymbol{\mu}\rangle)^2}{2}\right) = \exp\left(-\frac{\min(\langle\boldsymbol{\lambda}, \boldsymbol{\lambda}\rangle, \langle\boldsymbol{\lambda}, \boldsymbol{\mu}\rangle)^2}{2\langle\boldsymbol{\lambda}, \boldsymbol{\lambda}\rangle}\right)$$

- When $i = 2$, $\mathbb{E}_{\alpha_2,\ldots,\alpha_d\sim\mathcal{N}(0,1)}[a_i] + \langle\boldsymbol{v_i}, \boldsymbol{\mu}\rangle = \langle\hat{\boldsymbol{\lambda}}^{\perp}, \boldsymbol{\mu}\rangle$ and thus

$$\langle\hat{\boldsymbol{\lambda}}^{\perp}, \boldsymbol{\lambda}'\rangle = \langle\hat{\boldsymbol{\lambda}}^{\perp}, \boldsymbol{\mu}\rangle\mathbb{E}_{y\sim\mathcal{N}(0,1)}\left[\tanh(\|\boldsymbol{\lambda}\|\,(y + \langle\hat{\boldsymbol{\lambda}}, \boldsymbol{\mu}\rangle))\right]$$

Let κ as defined before and using Lemma 11 we get that

$$\langle\hat{\boldsymbol{\lambda}}^{\perp}, \boldsymbol{\mu}\rangle \geq \langle\hat{\boldsymbol{\lambda}}^{\perp}, \boldsymbol{\lambda}'\rangle \geq (1 - \kappa)\langle\hat{\boldsymbol{\lambda}}^{\perp}, \boldsymbol{\mu}\rangle \tag{5.9}$$

- When $i \geq 3$, $\mathbb{E}_{\alpha_2,\ldots,\alpha_d\sim\mathcal{N}(0,1)}[a_i] + \langle\boldsymbol{v_i}, \boldsymbol{\mu}\rangle = 0$ and thus $\langle\boldsymbol{v_i}, \boldsymbol{\lambda}'\rangle = 0$.

We can now bound the distance of $\boldsymbol{\lambda}'$ from $\boldsymbol{\mu}$:

$$\begin{aligned}\|\boldsymbol{\lambda}' - \boldsymbol{\mu}\| &= \sqrt{\sum_i \langle\boldsymbol{v}_i, \boldsymbol{\lambda}' - \boldsymbol{\mu}\rangle^2} = \sqrt{\langle\hat{\boldsymbol{\lambda}}, \boldsymbol{\lambda}' - \boldsymbol{\mu}\rangle^2 + \langle\hat{\boldsymbol{\lambda}}^{\perp}, \boldsymbol{\lambda}' - \boldsymbol{\mu}\rangle^2} \\ &\overset{(5.8),\,(5.9)}{\leq} \sqrt{\kappa^2\langle\hat{\boldsymbol{\lambda}}, \boldsymbol{\lambda} - \boldsymbol{\mu}\rangle^2 + \kappa^2\langle\hat{\boldsymbol{\lambda}}^{\perp}, \boldsymbol{\lambda} - \boldsymbol{\mu}\rangle^2} \leq \kappa\,\|\boldsymbol{\lambda} - \boldsymbol{\mu}\|\end{aligned}$$

We now have to prove that this convergence rate κ decreases as the iterations increase. This is implied by the following lemmas which show that $\min(\langle\hat{\boldsymbol{\lambda}}, \boldsymbol{\lambda}\rangle, \langle\hat{\boldsymbol{\lambda}}, \boldsymbol{\mu}\rangle) \leq \min(\langle\hat{\boldsymbol{\lambda}}', \boldsymbol{\lambda}'\rangle, \langle\hat{\boldsymbol{\lambda}}', \boldsymbol{\mu}\rangle)$

Lemma 12. *If* $\|\boldsymbol{\lambda}\| \geq \langle\hat{\boldsymbol{\lambda}}, \boldsymbol{\mu}\rangle$ *then* $\langle\hat{\boldsymbol{\lambda}}, \boldsymbol{\mu}\rangle \leq \|\boldsymbol{\lambda}'\|$ *and* $\langle\hat{\boldsymbol{\lambda}}, \boldsymbol{\mu}\rangle \leq \langle\hat{\boldsymbol{\lambda}}', \boldsymbol{\mu}\rangle$.

Proof. The analysis above implies that $\boldsymbol{\lambda}'$ can be written in the form $\boldsymbol{\lambda}' = \alpha \cdot \hat{\boldsymbol{\lambda}} + \beta \cdot \hat{\boldsymbol{\lambda}}^{\perp}$, where $\langle\hat{\boldsymbol{\lambda}}, \boldsymbol{\mu}\rangle \leq \alpha \leq \|\boldsymbol{\lambda}\|$ and $0 \leq \beta \leq \langle\hat{\boldsymbol{\lambda}}^{\perp}, \boldsymbol{\mu}\rangle$. It is easy to see that the first inequality

holds since $\|\boldsymbol{\lambda}'\| \geq \alpha \geq \langle \hat{\boldsymbol{\lambda}}, \boldsymbol{\mu} \rangle$. For the second, we write $\langle \hat{\boldsymbol{\lambda}}', \boldsymbol{\mu} \rangle$ as:

$$\langle \hat{\boldsymbol{\lambda}}', \boldsymbol{\mu} \rangle = \frac{\langle \boldsymbol{\lambda}', \boldsymbol{\mu} \rangle}{\|\boldsymbol{\lambda}'\|} = \frac{\alpha \langle \hat{\boldsymbol{\lambda}}, \boldsymbol{\mu} \rangle + \beta \langle \hat{\boldsymbol{\lambda}}^{\perp}, \boldsymbol{\mu} \rangle}{\sqrt{\alpha^2 + \beta^2}} = \langle \hat{\boldsymbol{\lambda}}, \boldsymbol{\mu} \rangle \frac{1 + \frac{\langle \hat{\boldsymbol{\lambda}}^{\perp}, \boldsymbol{\mu} \rangle}{\langle \hat{\boldsymbol{\lambda}}, \boldsymbol{\mu} \rangle} \frac{\beta}{\alpha}}{\sqrt{1 + \left(\frac{\beta}{\alpha}\right)^2}} \geq \langle \hat{\boldsymbol{\lambda}}, \boldsymbol{\mu} \rangle \frac{1 + \left(\frac{\beta}{\alpha}\right)^2}{\sqrt{1 + \left(\frac{\beta}{\alpha}\right)^2}} \geq \langle \hat{\boldsymbol{\lambda}}, \boldsymbol{\mu} \rangle$$

where we used the fact that $\frac{\langle \hat{\boldsymbol{\lambda}}^{\perp}, \boldsymbol{\mu} \rangle}{\langle \hat{\boldsymbol{\lambda}}, \boldsymbol{\mu} \rangle} \geq \frac{\beta}{\alpha}$ which follows by the bounds on α and β. □

Lemma 13. *If* $\|\boldsymbol{\lambda}\| \leq \langle \hat{\boldsymbol{\lambda}}, \boldsymbol{\mu} \rangle$ *then* $\|\boldsymbol{\lambda}\| \leq \|\boldsymbol{\lambda}'\| \leq \langle \hat{\boldsymbol{\lambda}}', \boldsymbol{\mu} \rangle$.

Proof. We have that $\boldsymbol{\lambda}' = \alpha \cdot \hat{\boldsymbol{\lambda}} + \beta \cdot \hat{\boldsymbol{\lambda}}^{\perp}$, where $\|\boldsymbol{\lambda}\| \leq \alpha \leq \langle \hat{\boldsymbol{\lambda}}, \boldsymbol{\mu} \rangle$ and $0 \leq \beta \leq \langle \hat{\boldsymbol{\lambda}}^{\perp}, \boldsymbol{\mu} \rangle$. We also have $\langle \boldsymbol{\lambda}', \boldsymbol{\mu} \rangle = \alpha \langle \hat{\boldsymbol{\lambda}}, \boldsymbol{\mu} \rangle + \beta \langle \hat{\boldsymbol{\lambda}}^{\perp}, \boldsymbol{\mu} \rangle \geq \alpha^2 + \beta^2 = \|\boldsymbol{\lambda}'\|^2 \geq \alpha^2 \geq \|\boldsymbol{\lambda}\|^2$ so the lemma follows. □

Finally substituting back in the basis that we started before changing coordinates to make the covariance matrix identity we get the result as stated at the theorem. □

5.5 An Illustration of the Speed of Convergence

Using our results in the previous sections we can calculate explicit speeds of convergence of EM to its fixed points. In this section, we present some results with this flavor. For simplicity, we focus on the single dimensional case, but our calculations easily extend to the multidimensional case.

Let us consider a mixture of two single-dimensional Gaussians whose signal-to-noise ratio $\eta = \mu/\sigma$ is equal to 1. There is nothing special about the value of 1, except that it is a difficult case to consider since the Gaussian components are not separated, as shown in Figure 5-1. When the SNR is larger, the numbers presented below still hold and in reality the convergence is even faster. When the SNR is even smaller than one, the numbers change, but gracefully, and they can be calculated in a similar fashion.

We will also assume a completely agnostic initialization of EM, setting $\lambda^{(0)} \to +\infty$.[1] To analyze the speed of convergence of EM to its fixed point μ, we first make the observation that in one step we already get to $\lambda^{(1)} \leq \mu + \sigma$. To see this we can plug $\lambda^{(0)} \to \infty$ into

[1] In the multi-dimensional setting, this would correspond to a very large magnitude $\boldsymbol{\lambda}^{(0)}$ chosen in a random direction.

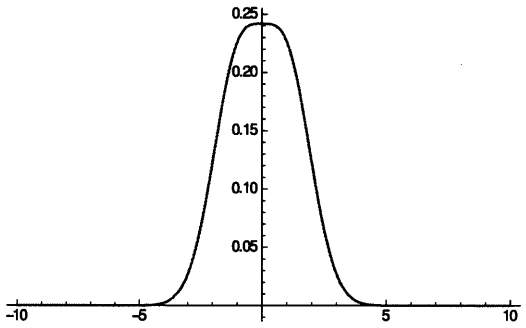

Figure 5-1: The density of $\frac{1}{2}\mathcal{N}(x;1,1) + \frac{1}{2}\mathcal{N}(x;-1,1)$.

equation (5.5) to get:

$$\lambda^{(1)} = \mathbb{E}_{x\sim\mathcal{N}(\mu,\sigma^2)}\left[\text{sign}(x)x\right] = \mathbb{E}_{x\sim\mathcal{N}(\mu,\sigma^2)}\left[|x|\right],$$

which equals the mean of the Folded Normal Distribution. A well-known bound for this mean is $\mu + \sigma$. Therefore the distance from the true mean after one step is $\left|\lambda^{(1)} - \mu\right| \leq \sigma$.

Now, using Theorem 23, we conclude that in all subsequent steps the distance to μ shrinks by a factor of at least $e^{+1/2}$. This means that, if we want to estimate μ to within additive error $1\%\sigma$, then we need to run EM for at most $2 \cdot \ln 100$ steps. That is, 10 iterations of the EM algorithm suffice to get to within error 1% even when our initial guess of the mean is infinitely away from the true value!

In Figure 5-2 we illustrate the speed of convergence of EM as implied by Theorem 24 in multiple dimensions. The plot was generated for a Gaussian mixture with $\boldsymbol{\mu} = (2\ 2)$ and $\Sigma = I$, but the behavior illustrated in this figure is generic (up to a transformation of the space by $\Sigma^{-\frac{1}{2}}$). As implied by Theorem 24, the rate of convergence depends on the distance of $\boldsymbol{\lambda}^{(t)}$ from the origin $\mathbf{0}$ and the angle $\langle\boldsymbol{\lambda}^{(t)}, \boldsymbol{\mu}\rangle$. The figure shows the directions of the EM updates for every point, and the factor by which the distance to the fixed point decays, with deeper colors corresponding to faster decays. There are three fixed points. Any point that is equidistant from $\boldsymbol{\mu}$ and $-\boldsymbol{\mu}$ is updated to $\mathbf{0}$ in one step and stays there thereafter. Points that are closer to $\boldsymbol{\mu}$ are pushed towards $\boldsymbol{\mu}$, while points that are closer to $-\boldsymbol{\mu}$ are pushed towards $-\boldsymbol{\mu}$.

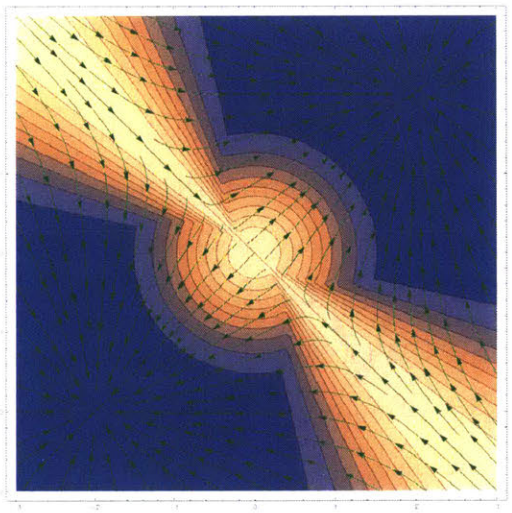

Figure 5-2: Illustration of the Speed of Convergence of EM in Multiple Dimensions as Implied by Theorem 24.

Chapter 6

Conclusions

This thesis is inspired by the increasing recent interest of artificial intelligence and machine learning to the non-convex optimization. Despite this increasing interest, from a theoretical point of view only few thing are known. The reason is that usually the non-convex optimization problems are NP-hard. This prevents the theoretical worst-case analysis of any algorithm that aims to solve these problems without any additional assumption. For this reason the theoretical tools for analyzing the performance of these algorithms is very limited. In this thesis we aim to contribute on creating such a toolbox to attack this kind of problems.

We start by exploiting the generality and power of the well known contraction mapping technique. To this direction we get the following two results presented in Chapter 3 and Chapter 4 respectively:

- we prove a converse to the Banach's Fixed Point Theorem. This converse theorem suggests that any running time analysis of an iterative algorithm, can also be done using the contraction map principle. The basic idea that enables this converse to exists, is that anyone that wants to apply Banach's Fixed Point Theorem has the power of designing its own distance metric for which the iteration is a contraction map. Therefore contraction maps is a much more general tool than one could think. As a proof of this concept we give a contraction mapping argument to bound the running time of the power method for computing eigenvectors of a square matrix.

- we formulate Banach's Fixed Point Theorem in a concrete computational *search prob-*

lem that lives in the TFNP class. We show that the computational complexity of Banach's Theorem is completely captured by the CLS class. This result is important in two ways. It is another evidence of the generality and the power of Banach's Theorem and it also gives a non-trivial complete problem for CLS, a question that it was left open in the work of Papadimitriou and Daskalakis [23].

In the second part of this thesis we consider on the most powerful and widely used algorithm in machine learning and in science in general, the celebrated Expectation - Maximization (EM) algorithm. The theoretical analysis of EM algorithm has been a problem left open for a lot of year since its definition by [26]. We are doing a big step towards this direction by giving the first global analysis of EM when applied to mixtures of two Gaussians. We also believe that the techniques used for this analysis contribute to the toolbox of analyzing algorithms for non-convex optimization problems.

6.1 Future Directions

A lot of open problems and future directions arise when finishing this work. We will refer here to a couple of those.

- it the complexity theoretic area, one interesting and important open problem is if the Banach's Fixed Point Theorem is complete for CLS even in the case where the distance metric function is not an input but it is the natural $\|\cdot\|_p$ of $[0,1]^3$.

- for the analysis of the EM algorithm, a lot of open problems are very interesting. A first obvious one is how general can the techniques become, in order to prove the convergence and the optimality of EM algorithm in other more general settings than the mixture of two Gaussians.

  An other interesting question arises in the cases that we already know that EM gets trapped in local optima. For these cases the theoretical analysis of EM could suggest appropriate initialization such that this bad behaviour disappears. This would be very interesting even from a practical point of view and would indicate the importance of having a theoretical analysis for the machine learning algorithms.

Finally beyond the specific problems that we mentioned the general goal of analyzing heuristic or local search algorithms in machine learning is very important and attracts more and more interest from the scientific community. In this direction the continuing development of a good toolbox to attack these problems has high value and is something that will increase our understanding and abilities to control the recent breakthroughs in the broader area of artificial intelligence.

Bibliography

[1] Dimitris Achlioptas and Frank McSherry. On spectral learning of mixtures of distributions. In *International Conference on Computational Learning Theory*, pages 458–469. Springer, 2005.

[2] Sanjeev Arora and Ravi Kannan. Learning mixtures of arbitrary gaussians. In *Proceedings of the thirty-third annual ACM symposium on Theory of computing*, pages 247–257. ACM, 2001.

[3] Sivaraman Balakrishnan, Martin J. Wainwright, and Bin Yu. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *CoRR*, abs/1408.2156, 2014.

[4] Sivaraman Balakrishnan, Martin J Wainwright, and Bin Yu. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *arXiv preprint arXiv:1408.2156*, 2014.

[5] Stefan Banach. Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales. *Fundamenta Mathematicae*, 3(1):133–181, 1922.

[6] Mikhail Belkin and Kaushik Sinha. Polynomial learning of distribution families. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pages 103–112. IEEE, 2010.

[7] Mihir Bellare and Phillip Rogaway. The complexity of approximating a nonlinear program. *Mathematical Programming*, 69(1):429–441, 1995.

[8] C. Bessaga. On the converse of banach "fixed-point principle". *Colloquium Mathematicae*, 7(1):41–43, 1959.

[9] Stephen Boyd and Lieven Vandenberghe. *Convex optimization.* Cambridge university press, 2004.

[10] S Charles Brubaker and Santosh S Vempala. Isotropic PCA and affine-invariant clustering. In *the 49th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 2008.

[11] Kamalika Chaudhuri, Sanjoy Dasgupta, and Andrea Vattani. Learning mixtures of gaussians using the k-means algorithm. *arXiv preprint arXiv:0912.0086*, 2009.

[12] Kamalika Chaudhuri and Satish Rao. Learning Mixtures of Product Distributions Using Correlations and Independence. In *the 21st International Conference on Computational Learning Theory (COLT)*, 2008.

[13] Stéphane Chrétien and Alfred O Hero. On EM algorithms and their proximal generalizations. *ESAIM: Probability and Statistics*, 12:308–326, 2008.

[14] Earl A. Coddigton and Norman Levinson. *Theory of Ordinary Differential Equations.* McGraw-Hill, 1966.

[15] C.C. Conley. *Isolated Invariant Sets and the Morse Index.* Number $\alpha\rho\iota\theta$. 38 in Conference Board of the Mathematical Sciences Series No. 38. Conference Board of the Mathematical Sciences, 1978.

[16] Keith Conrad. The contraction mapping theorem. *Expository paper. University of Connecticut, College of Liberal Arts and Sciences, Department of Mathematics*, 2014.

[17] S. Dasgupta. Learning mixtures of gaussians. In *40th Annual Symposium on Foundations of Computer Science.* Institute of Electrical & Electronics Engineers (IEEE), 1999.

[18] Sanjoy Dasgupta. Learning mixtures of gaussians. In *Foundations of Computer Science, 1999. 40th Annual Symposium on*, pages 634–644. IEEE, 1999.

[19] Sanjoy Dasgupta and Leonard Schulman. A probabilistic analysis of em for mixtures of separated, spherical gaussians. *Journal of Machine Learning Research*, 8(Feb):203–226, 2007.

[20] Constantinos Daskalakis, Paul W Goldberg, and Christos H Papadimitriou. The complexity of computing a nash equilibrium. *SIAM Journal on Computing*, 39(1):195–259, 2009.

[21] Constantinos Daskalakis and Gautam Kamath. Faster and sample near-optimal algorithms for proper learning mixtures of gaussians. In *Proceedings of The 27th Conference on Learning Theory, COLT 2014, Barcelona, Spain, June 13-15, 2014*, pages 1183–1213, 2014.

[22] Constantinos Daskalakis and Gautam Kamath. Faster and sample near-optimal algorithms for proper learning mixtures of gaussians. In *Proceedings of The 27th Conference on Learning Theory*, pages 1183–1213, 2014.

[23] Constantinos Daskalakis and Christos H. Papadimitriou. Continuous local search. In *Proceedings of the Twenty-Second Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2011, San Francisco, California, USA, January 23-25, 2011*, pages 790–804, 2011.

[24] Constantinos Daskalakis, Christos Tzamos, and Manolis Zampetakis. Ten steps of em suffice for mixtures of two gaussians. *arXiv preprint arXiv:1609.00368*, 2016.

[25] Klaus Deimling. *Nonlinear functional analysis*. Courier Corporation, 2010.

[26] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[27] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

[28] Michael Edelstein. On fixed and periodic points under contractive mappings. *Journal of the London Mathematical Society*, 1(1):74–79, 1962.

[29] Rong Ge, Qingqing Huang, and Sham M Kakade. Learning mixtures of gaussians in high dimensions. In *the 47th Annual ACM on Symposium on Theory of Computing (STOC)*, 2015.

[30] Andrzej Granas and James Dugundji. *Fixed Point Theory.* Springer New York, 2003.

[31] Moritz Hardt and Eric Price. Tight bounds for learning a mixture of two gaussians. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing*, pages 753–760. ACM, 2015.

[32] Michael D Hirsch, Christos H Papadimitriou, and Stephen A Vavasis. Exponential lower bounds for finding brouwer fix points. *Journal of Complexity*, 5(4):379–416, 1989.

[33] Daniel Hsu and Sham M Kakade. Learning mixtures of spherical gaussians: moment methods and spectral decompositions. In *the 4th conference on Innovations in Theoretical Computer Science (ITCS)*, 2013.

[34] Pavel Hubácek and Eylon Yogev. Hardness of continuous local search: Query complexity and cryptographic lower bounds. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1352–1371. SIAM, 2017.

[35] Jacek Jachymski et al. A short proof of the converse to the contraction principle and some related results. *Topol. Methods Nonlinear Anal*, 15:179–186, 2000.

[36] Ludvik Janos. A converse of banach's contraction theorem. *Proceedings of the American Mathematical Society*, 18(2):287–289, 1967.

[37] David S. Johnson, Christos H. Papadimitriou, and Mihalis Yannakakis. How easy is local search? *J. Comput. Syst. Sci.*, 37(1):79–100, 1988.

[38] Adam Tauman Kalai, Ankur Moitra, and Gregory Valiant. Efficiently learning mixtures of two gaussians. In *Proceedings of the Forty-second ACM Symposium on Theory of Computing*, STOC '10, pages 553–562, New York, NY, USA, 2010. ACM.

[39] Ravindran Kannan, Hadi Salmasian, and Santosh Vempala. The spectral method for general mixture models. In *the 18th International Conference on Computational Learning Theory (COLT)*, 2005.

[40] Brian W Kernighan and Shen Lin. An efficient heuristic procedure for partitioning graphs. *The Bell system technical journal*, 49(2):291–307, 1970.

[41] T Körner. Metric and topological spaces, 2010.

[42] J. D. Lambert. *Numerical Methods for Ordinary Differential Systems. The Initial Value Problem.* John Wiley & Sons, 1991.

[43] Philip R Meyers. Some extensions of banach's contraction theorem. *J. Res. Nat. Bur. Standards Sect. B*, 69:179–184, 1965.

[44] Philip R. Meyers. A converse to banach's contraction theorem. *Journal of Research of the National Bureau of Standards Section B Mathematics and Mathematical Physics*, 71B(2 and 3):73, apr 1967.

[45] Ankur Moitra and Gregory Valiant. Settling the polynomial learnability of mixtures of gaussians. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pages 93–102. IEEE, 2010.

[46] John F. Nash. Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences*, 36(1):48–49, 1950.

[47] Maysum Panju. Iterative methods for computing eigenvalues and eigenvectors. *arXiv preprint arXiv:1105.1185*, 2011.

[48] Christos H. Papadimitriou. On the complexity of the parity argument and other inefficient proofs of existence. *J. Comput. Syst. Sci.*, 48(3):498–532, 1994.

[49] Richard A Redner and Homer F Walker. Mixture densities, maximum likelihood and the EM algorithm. *SIAM review*, 26(2):195–239, 1984.

[50] Henry Gordon Rice. Classes of recursively enumerable sets and their decision problems. *Transactions of the American Mathematical Society*, 74(2):358–366, 1953.

[51] Ananda Theertha Suresh, Alon Orlitsky, Jayadev Acharya, and Ashkan Jafarpour. Near-optimal-sample estimators for spherical gaussian mixtures. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1395–1403. Curran Associates, Inc., 2014.

[52] Ananda Theertha Suresh, Alon Orlitsky, Jayadev Acharya, and Ashkan Jafarpour. Near-optimal-sample estimators for spherical gaussian mixtures. In *Advances in Neural Information Processing Systems*, pages 1395–1403, 2014.

[53] Andrew G Thomason. Hamiltonian cycles and uniquely edge colourable graphs. *Annals of Discrete Mathematics*, 3:259–268, 1978.

[54] Paul Tseng. An analysis of the EM algorithm and entropy-like proximal point methods. *Mathematics of Operations Research*, 29(1):27–44, 2004.

[55] Santosh Vempala and Grant Wang. A spectral algorithm for learning mixture models. *Journal of Computer and System Sciences*, 68(4):841–860, 2004.

[56] James Sai-Wing Wong. *Generalizations to the converse of contraction mapping principle.* PhD thesis, California Institute of Technology, 1964.

[57] CF Jeff Wu. On the convergence properties of the EM algorithm. *The Annals of statistics*, pages 95–103, 1983.

[58] Ji Xu, Daniel Hsu, and Arian Maleki. Global analysis of Expectation Maximization for mixtures of two Gaussians. In *the 30th Annual Conference on Neural Information Processing Systems (NIPS)*, 2016.

[59] Lei Xu and Michael I. Jordan. On convergence properties of the em algorithm for gaussian mixtures. *Neural Computation*, 8:129–151, 1995.

[60] Fanny Yang, Sivaraman Balakrishnan, and Martin J. Wainwright. Statistical and computational guarantees for the baum-welch algorithm. In *53rd Annual Allerton Conference on Communication, Control, and Computing, Allerton 2015, Allerton Park &*

Retreat Center, Monticello, IL, USA, September 29 - October 2, 2015, pages 658–665, 2015.