Contributions to Unsupervised and Semi-Supervised Learning

BY
Dávid Pál

A THESIS
PRESENTED TO THE UNIVERSITY OF WATERLOO
IN FULFILMENT OF THE
THESIS REQUIREMENT FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN
COMPUTER SCIENCE

WATERLOO, ONTARIO, CANADA, 2009

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

This thesis studies two problems in theoretical machine learning. The first part of the thesis investigates the statistical stability of clustering algorithms. In the second part, we study the relative advantage of having unlabeled data in classification problems.

Clustering stability was proposed and used as a model selection method in clustering tasks. The main idea of the method is that from a given data set two independent samples are taken. Each sample individually is clustered with the same clustering algorithm, with the same setting of its parameters. If the two resulting clusterings turn out to be close in some metric, it is concluded that the clustering algorithm and the setting of its parameters match the data set, and that clusterings obtained are meaningful. We study asymptotic properties of this method for certain types of cost minimizing clustering algorithms and relate their asymptotic stability to the number of optimal solutions of the underlying optimization problem.

In classification problems, it is often expensive to obtain labeled data, but on the other hand, unlabeled data are often plentiful and cheap. We study how the access to unlabeled data can decrease the amount of *labeled* data needed in the worst-case sense. We propose an extension of the probably approximately correct (PAC) model in which this question can be naturally studied. We show that for certain basic tasks the access to unlabeled data might, at best, halve the amount of labeled data needed.

Acknowledgements

I am indebted to my supervisor Prof. Shai Ben-David for his patience with which he has guided me through my PhD studies. I am very much enjoyed all the discussions we had together about our research, computer science and mathematics in general.

Most of the results in this thesis are based on three conference papers [8, 6, 5] which I co-authored with my advisor, Ulrike von Luxburg, Hans Ulrich Simon, and Tyler Lu. It has been a great pleasure to work with them and I thank them all. I also thank Shalev Ben-David for providing the proof of Lemma 6.5.

Thanks to Michael Spriggs, Steve Bahun, and especially Mustaq Ahmed for being so great office mates. I also thank to all the teachers for the beautiful lectures that I attended. I mention Therese Biedl, Timothy Chan, and Jeff Shallit from the School of Computer Science, Joseph Cheriyan and Nick Wormald from the Department of Combinatorics and Optimization, and Christopher Small from Department of Statistics and Actuarial Science. These people in their lectures explained to me a lot of complicated mathematics in a simple and intuitive, yet rigorous way.

Finally, I would like thank my friends Lenka Kovalčinová, Tomáš Vinař and Broňa Brejová, my father and my brother Martin for their support in so many ways.

Contents

Abstract iii

Acknowledgements iv

I Stability of Clustering Algorithms 1

1 Introduction 2

2 Definitions and Notation 4

3 Optimization Algorithms 7
3.1 Optimization Schemes and Algorithms 7
3.2 Empirical Risk Minimization 9
3.3 Risk Convergence of k-means and k-medians 10

4 Stability of Optimization Algorithms 15
4.1 The Stability Theorem 15
4.2 Stability of k-means and k-medians 16

5 Instability Because of Symmetry 19
5.1 Instability Because of Symmetry for k-means and k-medians 24

6 Instability of k-means 26
6.1 Outline of the Proof 27
6.2 The Technical Lemmas 30

7 Conclusion 39
7.1 Examples 39
7.2 Mismatch between Theory and Practice 41
7.3 Rates of Convergence and Cluster Boundaries 42
7.4 Technical Questions 43

II Comparison of Supervised and Semi-Supervised Learning 44

8 Introduction 45

9 Definitions and Notation 47
9.1 A Folklore Example 49

10 The Hypothesis Class of Thresholds 51
10.1 Kääriäinen's Algorithm 51
10.2 Lower Bound 53
10.3 Semi-supervised Learning Ratio 56

11 Conclusion 58

List of Figures

3.1 Voronoi diagram 8

5.1 Uniform distribution over the vertices of a square. 20
5.2 Symmetric risk minimizers 20
5.3 Symmetry of a probability distribution 21

6.1 Instability of k-means 27
6.2 The decision set Q 29
6.3 The cone T inside the set Q 30

7.1 A two cluster distribution over $\mathbb{R}$ 40
7.2 k-means solutions for $k = 2$ and $k = 3$ 41

"The question of whether a computer can think is no more interesting than the question of whether a submarine can swim."

— Edsger W. Dijkstra

Part I

Stability of Clustering Algorithms

Chapter 1

Introduction

A common machine learning task is to partition a given set of sample points into a collection of groups so that points within each group are similar to each other and points from different groups are dissimilar. These tasks are referred to as clustering problems. As an example of a clustering problem consider a statistician doing market analysis. Her data set consists of a set of (potential) customers of a company and for each customer the data set contains filled in answers of a questionnaire. She might want to identify a few representative groups in the customer base. A more extreme example are Internet websites such as `http://news.google.com/` which aggregate news stories from many other news websites. The goal is to group news stories according to their topic, so that they can be presented coherently to the visitors of the website.

Even small clustering problems are impossible to solve by hand and they are computed automatically on a computer using a clustering algorithm. Various clustering algorithms have been invented and are in use today. Common clustering algorithms, which we discuss in our thesis, are k-means and k-medians; see, for example, the book [15, Chapter 10]. Roughly speaking, these algorithms find a set of k centers by minimizing a certain cost function and then assign each sample point to its closest center.

Practical clusterings tasks are hard to specify a priori in such way so that the clustering algorithm would produce a good and meaningful clustering. Thus, usually, the user has to verify that the produced clustering is as she wanted. This might be possible to do visually for example, but that is prohibitive if the dimensionality of the data is large. Furthermore, sometimes the user trying to cannot even tell whether a given clustering is good or not. This is especially true for problems where the task is to find some unknown structure such as in the market analysis example.

Another issue is that, typically, the clustering algorithm used has several free parameters which are left for the user to choose. A common parameter, for example, is the desired number of clusters. For each setting of the parameters the algorithm produces a clustering and it is the user's task to choose one setting of the parameters. In statistics and machine learning the problem of choosing the setting of parameters is sometimes called a model selection problem.

Ideally, one would like to have an automatic or semi-automatic model selection method that would optimally select the parameters of a clustering algorithm and also somehow verify that the clustering produced is meaningful. The so-called *stability method* was pro-

posed for this purpose by several authors. See for example Ben-Hur et al. [9] and Lange et al. [20, 21].

The basic idea of the stability method is that two (or more) subsamples from the data set are drawn. Each subsample individually is clustered by the same clustering algorithm with the same settings of its parameters. The resulting clusterings are compared and if they are close in a certain metric, it is declared that the clustering algorithm is stable and thus the produced clusterings are good and meaningful. On a high level, we can view the stability method as conducting repetitions of the same "scientific" experiment (i.e. clustering a subsample of the data) with the assumption that if we obtain the same outcome multiple times then, perhaps, the outcome is "verified".

This part of the thesis studies the theoretical properties of the stability method as the sample size approaches infinity. We focus on the stability of clustering algorithms that optimize a cost function, and in particular, we study the stability of the algorithms that minimize the k-means and the k-medians cost. The main conclusion of our investigation is that the stability of such clustering algorithms depends solely on the number of optimal solutions of cost function associated with the data set. Based on our results, we hold the opinion that the stability method is not well justified, which we support by giving several examples.

This part of the thesis is based on conference articles by Ben-David et al. [8, 6].

Chapter 2

Definitions and Notation

We introduce the basic setup that will be used for the rest of this part of the thesis. The idea is that a sample $S = (x_1, x_2, \ldots, x_m)$ is generated i.i.d. according to some probability distribution P over some domain X. The probability distribution P is meant to represent the "whole" data set. A clustering algorithm receives the sample S as an input and outputs a clustering which for our purposes is simply a partition of X. The fundamental notion which we study, is the clustering stability. Roughly speaking, the *in*stability of the clustering algorithm on P is the expected distance between clusterings output by the algorithm on two independent samples of the same size coming from P. Of course, we need to define what is the distance between two clusterings.

In the rest of this chapter we give all the mentioned words their technical mathematical meaning. We assume that the reader is familiar with basics of probability theory and real analysis. See for example the books by Resnick [24] and Shilov [27].

Definition 2.1 (Domain)**.** *A* domain *is a measurable space. That is, it is a pair* $(X, \mathfrak{M})$ *where* X *is a non-empty set and* $\mathfrak{M}$ *is a* σ*-algebra of subsets of* X*.*

As is usual in mathematics, we assume that the σ-algebra $\mathfrak{M}$ is clear from the context and simply talk about the domain X.

Definition 2.2 (Clustering)**.** *Let* $(X, \mathfrak{M})$ *be a domain. A* clustering $\mathcal{C}$ *of the domain is a partition of* X *into finitely many measurable sets. That is,* $\mathcal{C}$ *is a finite collection of pairwise disjoint elements of* $\mathfrak{M}$ *whose union is* X*. An element of a clustering* $\mathcal{C}$ *is called a* cluster*. We denote by* $\sim_{\mathcal{C}}$ *the equivalence relation induced by a clustering* $\mathcal{C}$*. That is,* $x \sim_{\mathcal{C}} y$ *means that points* $x, y \in X$ *belong to the same cluster of* $\mathcal{C}$*. We use* $x \not\sim_{\mathcal{C}} y$ *to denote that* x *and* y *belong to different clusters of* $\mathcal{C}$*. We denote by* $\mathfrak{C}$ *the set of all clusterings of the domain.*[1]

Definition 2.3 (Sample)**.** *Let* $(X, \mathfrak{M})$ *be a domain. A* sample *is a finite sequence* $(x_1, x_2, \ldots, x_m)$ *of the points from* X*. The* size *of a sample is its length.*

Definition 2.4 (Clustering Algorithm)**.** *Let* $(X, \mathfrak{M})$ *be a domain. A* clustering algorithm *is a function* A *that maps a sample of any size to a clustering* $\mathcal{C}$ *of the domain* X*. Formally,* $A : (\bigcup_{m=1}^{\infty} X^m) \to \mathfrak{C}$*.*

[1] Every time we write $\mathfrak{C}$, the domain to which $\mathfrak{C}$ refers to, should be clear from the context.

Note that in practice a clustering computed by a clustering algorithm is only a partition of the sample points. However, clusterings produced by some clustering algorithms have natural extensions to the whole domain.

The next definition formalizes the notion of distance between two clusterings. In practice, people often compare partitions of two different samples and quite naturally the distance measure they define depends on the samples themselves. In our formalism, we are interested in computing distances between two partitions of the domain. It might seem that there is no way how to do this in a data dependent way. However, the scenarios we care about are those where the two samples arise from the same probability distribution P. For such scenarios, it is natural to allow the distance between two partitions of the domain to depend on P. In the next definition, we do exactly that.

Definition 2.5 (Clustering distance)**.** *Let $\mathfrak{P}$ be a family of probability distributions over some domain X. A clustering distance d is a family of pseudo-metrics on the set of all clusterings, $\mathfrak{C}$, with values in $[0,1]$.*[2] *The family is indexed by the elements of $\mathfrak{P}$. That is, formally, d is a function $d : \mathfrak{P} \times \mathfrak{C} \times \mathfrak{C} \to [0,1]$ such that for any probability measure $P \in \mathfrak{P}$ and for any clusterings $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3 \in \mathfrak{C}$ satisfies the following axioms:*

(i) $d_P(\mathcal{C}_1, \mathcal{C}_1) = 0$,[3]

(ii) $d_P(\mathcal{C}_1, \mathcal{C}_2) = d_P(\mathcal{C}_2, \mathcal{C}_1)$ *(symmetry),*

(iii) $d_P(\mathcal{C}_1, \mathcal{C}_3) \leq d_P(\mathcal{C}_1, \mathcal{C}_2) + d_P(\mathcal{C}_2, \mathcal{C}_3)$ *(triangle inequality).*

Note that for a clustering distance it can be the case that $d_P(\mathcal{C}_1, \mathcal{C}_2) = 0$ and yet $\mathcal{C}_1 \neq \mathcal{C}_2$. A prototypic example of a clustering distance is the Hamming clustering distance. For any probability measure P we define the *Hamming clustering distance* between two clusterings $\mathcal{C}_1$ and $\mathcal{C}_2$ as

$$d_P(\mathcal{C}_1, \mathcal{C}_2) = \Pr_{\substack{x \sim P \\ x' \sim P}} [(x \sim_{\mathcal{C}_1} x') \oplus (x \sim_{\mathcal{C}_2} x')]$$

where $\oplus$ denotes the logical XOR operation. In other words, Hamming distance is the probability that the two clusterings $\mathcal{C}_1$ and $\mathcal{C}_2$ disagree on two randomly drawn pair of points x and x'.

It can be easily checked that the Hamming clustering distance is indeed a clustering distance. The first two properties are trivially satisfied and the triangle inequality follows from

$$\begin{aligned}
d_P(\mathcal{C}_1, \mathcal{C}_3) &= \Pr_{\substack{x \sim P \\ x' \sim P}} [(x \sim_{\mathcal{C}_1} x') \oplus (x \sim_{\mathcal{C}_3} x')] \\
&= \Pr_{\substack{x \sim P \\ x' \sim P}} [((x \sim_{\mathcal{C}_1} x') \oplus (x \sim_{\mathcal{C}_2} x')) \oplus ((x \sim_{\mathcal{C}_2} x') \oplus (x \sim_{\mathcal{C}_3} x'))] \\
&\leq \Pr_{\substack{x \sim P \\ x' \sim P}} [(x \sim_{\mathcal{C}_1} x') \oplus (x \sim_{\mathcal{C}_2} x')] + \Pr_{\substack{x \sim P \\ x' \sim P}} [(x \sim_{\mathcal{C}_2} x') \oplus (x \sim_{\mathcal{C}_3} x')] \\
&= d_P(\mathcal{C}_1, \mathcal{C}_2) + d_P(\mathcal{C}_2, \mathcal{C}_3) \,.
\end{aligned}$$

[2] The upper bound 1 on values of d is an arbitrary choice. Any positive finite number suffices.

[3] Note that we write the first parameter of d, the probability measure P, as a subscript.

In this part of the thesis, when we write d or d_P we mean either a generic clustering distance or the Hamming clustering distance.

We finish the chapter with the definition of the fundamental notion of instability.

Definition 2.6 (Instability)**.** *Let* P *be a probability distribution over a domain* X *and* d *be a (generic) clustering distance. The* instability *of a clustering algorithm* A *for sample size* m *on the probability distribution* P *is*

$$\text{instab}(A, P, m) = \mathop{\mathbf{E}}_{\substack{S_1 \sim P^m \\ S_2 \sim P^m}} d_P(A(S_1), A(S_2)) \ .$$

The (asymptotic) instability of a clustering algorithm A *on the probability distribution* P *is*

$$\text{instab}(A, P) = \limsup_{m \to \infty} \ \text{instab}(A, P, m) \ .$$

We say that algorithm A *is* stable *for* P *if* $\text{instab}(A, P) = 0$*. If* $\text{instab}(A, P) > 0$*, we say that* A *is* unstable *on* P*.*

Chapter 3

Optimization Algorithms

3.1 Optimization Schemes and Algorithms

In the statistical setting introduced in the previous chapter, often, the goal of a clustering algorithm is to find the clustering that minimizes some cost function defined by the probability distribution P generating the data. We assume that the distribution P is unknown and the only access to it is through an i.i.d. sample. The task of a clustering algorithm is to minimize the cost function just from the sample. Clearly, for a clustering algorithm there is no hope to minimize a function defined by some unknown probability distribution. What we have in mind, however, is that the algorithm minimizes the cost in the asymptotic sense. That is, the cost of the clustering output by the algorithm converges to the minimum cost as the sample size grows.

In this section we setup the general framework for the setting just described. Instead of the names "cost" or "objective" which are more common in optimization and operations research, we will use the traditional statistical name *risk*. Along the way, we talk about empirical risk minimization (ERM) and we describe two centre based ERM algorithms: k-means and k-medians that naturally fit our model. We close the chapter by proving risk convergence of these two algorithms in Euclidean spaces.

Definition 3.1 (Optimization Scheme)**.** *An* optimization scheme *over a domain X is a quadruple $(Y, \mathfrak{P}, R, \Gamma)$ where Y is a non-empty set of* solutions*, $\mathfrak{P}$ is a family of distributions over X, $R : \mathfrak{P} \times Y \to \mathbb{R}$ is a* risk *function and $\Gamma : Y \to \mathfrak{C}$ is a function that maps solutions to clusterings.*

Given an optimization scheme $(Y, \mathfrak{P}, R, \Gamma)$ an element $P \in \mathfrak{P}$ is called an instance *of the scheme. The* optimal value *of an instance P is*

$$\mathrm{opt}_R(P) = \inf_{y \in Y} R(P, y) \ .$$

A risk minimizer *(or an* optimal solution*) for an instance P is an element $y^* \in Y$ such that $R(P, y^*) = \mathrm{opt}_R(P)$. We denote by $\operatorname{argmin}_{y \in Y} R(P, y)$ the set of risk minimizers of P. This set is formally defined as*

$$\operatorname*{argmin}_{y \in Y} R(P, \mathcal{C}) = \left\{ y^* \in Y \ : \ R(P, y^*) = \mathrm{opt}_R(P) \right\} \ .$$

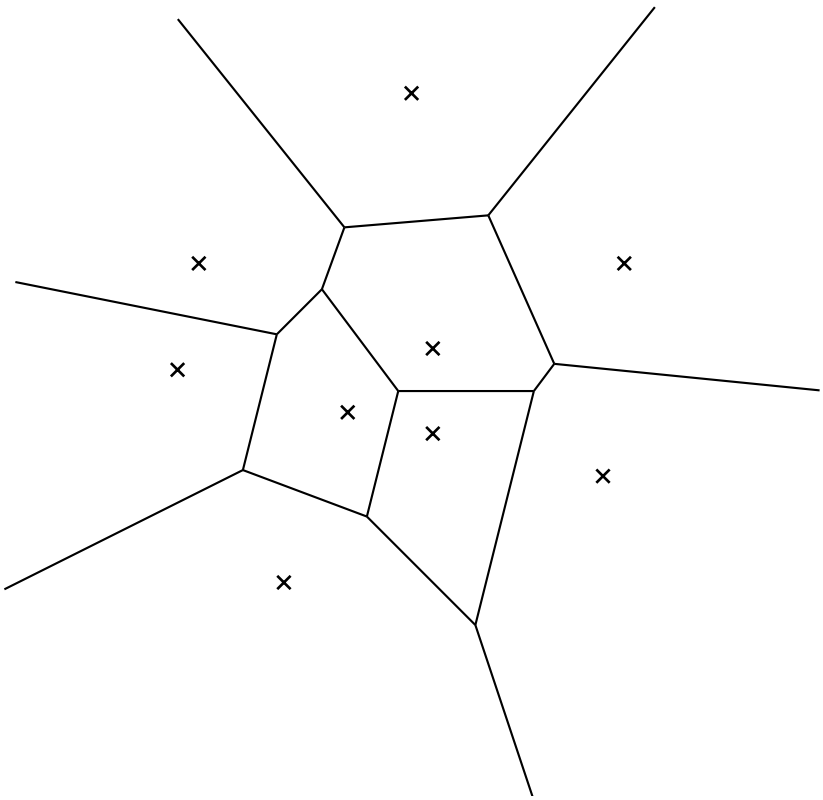

Figure 3.1: The figures shows a Voronoi diagram in the Euclidean plane $\mathbb{R}^2$ with 9 centers. The centers are shown as crosses. The lines represent the boundaries of the clusters. The metric ℓ is the Euclidean metric.

Note that an optimization scheme does not define any clustering algorithm. It barely tells us what risk function we are interested in minimizing. In the next definition we define what meant by an optimization algorithm. Its task is to (asymptotically) minimize the risk. In principle, for an optimization scheme, there might exist many quite different optimization algorithms with that property (or, perhaps, none). For this reason we keep these two notions separate.

Before we proceed with the definition of an optimization algorithm let us take a small detour. We give examples of two generic optimization schemes—the k-*means scheme* and k-*medians scheme*. These two examples are generic in the sense that each of them represents in fact a class of optimization schemes that share a certain common structure.

All the components of the two schemes are the same except for the risk function. For both of them, we require existence of a metric ℓ on the domain X. The second component of the schemes, $\mathfrak{P}$, is an arbitrary family of distributions of interest. The set of solutions Y is the set of k-tuples of elements of the domain, that is, $Y = X^k$.[1] Thus a generic solution can be written as $(c_1, c_2, \ldots, c_k)$ and its components $c_1, c_2, \ldots, c_k$ are called *centers*. The

[1] The reasons why we use k-tuples of centres, as opposed to (multi)sets of centres of size k, is purely technical and it will become obvious later. Namely, we will need to compute the derivative of a function with values in Y^k.

function Γ maps a solution $(c_1, c_2, \ldots, c_k)$ to the *Voronoi diagram with centers* $c_1, c_2, \ldots, c_k$, which we denote Voronoi$(c_1, c_2, \ldots, c_k)$. It is defined as a clustering with k clusters where we assign to the i-th cluster all the domain points that are closer to the center c_i than to any other center $c_j \neq c_i$. (See Figure 3.1.) A domain point to which two or more centers are equally close is assigned to exactly one of them according to some fixed tie breaking rule. As for now, we leave the tie breaking rule unspecified; however in any instantiation of these generic optimizations schemes it needs to be specified. Finally, we define the risk functions. For k-means optimization scheme the risk function is[2]

$$R(P; c_1, c_2, \ldots, c_k) = \mathop{\mathbf{E}}_{x \sim P}\left[\min_{i=1,2,\ldots,k} (\ell(x, c_i))^2\right] . \tag{3.1}$$

For k-medians optimization scheme the risk function is

$$R'(P; c_1, c_2, \ldots, c_k) = \mathop{\mathbf{E}}_{x \sim P}\left[\min_{i=1,2,\ldots,k} \ell(x, c_i)\right] . \tag{3.2}$$

We will need the following technical definition.

Definition 3.2. *Let* $(Y, \mathfrak{P}, R, \Gamma)$ *be an optimization scheme. An* optimization algorithm *is any function* $B : (\bigcup_{m=1}^{\infty} X^m) \to Y$. *The composition* $\Gamma \circ B$ *is called the clustering algorithm* induced *by* B.

Note that despite the name an optimization algorithm does not need to optimize anything. The intention, however, is that an optimization algorithm for any instance P of an optimization scheme converges to the optimal value of P.

Definition 3.3 (Risk Convergence)**.** *Let* $(Y, \mathfrak{P}, R, \Gamma)$ *be an optimization scheme over a domain* X. *Let* B *be an optimization algorithm. We say that* B *is* risk converging *for the optimization scheme whenever for any instance* P *of the optimization scheme if an i.i.d. sample* $S \sim P^m$ *is generated then as* $m \to \infty$

$$R(P, B(S)) \to \mathrm{opt}_R(P) \quad \textit{in probability.}$$

More formally, B *is risk converging if and only if for any* $P \in \mathfrak{P}$*, any* $\epsilon > 0$ *and any* $\delta > 0$ *there exists a positive integer* m_0 *such that for all* $m \geq m_0$

$$\Pr_{S \sim P^m}\left[R(P, B(S)) > \mathrm{opt}(P) + \epsilon\right] \leq \delta .$$

3.2 Empirical Risk Minimization

One natural approach to design an optimization algorithm for some optimization scheme is empirical risk minimization (ERM). The basic idea behind ERM is that the algorithm replaces the data generating distribution P by the empirical distribution defined by the sample. ERM can be applied to the k-means and the k-medians optimization schemes which leads to simple optimization algorithms for the schemes. The sole purpose of this section is to present these two algorithms.

We start with the basic definition of the empirical distribution.

[2]To avoid having too many parentheses we use semicolon to separate the two formal parameters of the risk function.

Definition 3.4 (Empirical Distribution)**.** *Let* $(X, \mathfrak{M})$ *be a domain. For a sample* $S = (x_1, x_2, \ldots, x_m)$ *we denote by* P_S *the* empirical probability distribution *of* S*. We define* $P_S(M)$ *for any measurable set* $M \subseteq X$ *to be the fraction of the sample points of* S *that lies in* M*. Formally, for any* $M \in \mathfrak{M}$

$$P_S(M) = \frac{1}{m}|\{i \,:\, 1 \leq i \leq m,\ x_i \in M\}| \,. \qquad [3]$$

ERM is suitable for optimization schemes with risk function that can be written as $R(P, y) = \mathbf{E}_{x\sim P}[L(x, y)]$ where $L : X \times Y \to \mathbb{R}$ is the so-called *loss function* and it is a function of a domain point x and a solution y. The generic ERM algorithm for such an optimization scheme on an input sample $S = (x_1, x_2, \ldots, x_m)$ minimizes the empirical loss $R(P_S, y) = \frac{1}{m}\sum_{i=1}^{m} L(x_i, y)$.

In both the k-means and the k-medians optimization scheme, introduced in the previous section, the risk function can be written via a loss function. We thus design generic optimization algorithms B and B' for the generic k-means and the k-medians optimization schemes respectively. The algorithms B, which we henceforth call the k-*means ERM algorithm*, on an input $S = (x_1, x_2, \ldots, x_m)$ outputs a solution $(c_1, c_2, \ldots, c_k)$ which minimizes

$$R(P_S; c_1, c_2, \ldots, c_k) = \mathop{\mathbf{E}}_{x\sim P_S}\left[\min_{i=1,2,\ldots,k} (\ell(x, c_i))^2\right] = \frac{1}{m}\sum_{j=1}^{m} \min_{i=1,2,\ldots,k} (\ell(x_j, c_i))^2 \,.$$

Similarly, the algorithm B', which we call k-*medians ERM algorithm*, on the input S outputs a solution that minimizes

$$R'(P_S; c_1, c_2, \ldots, c_k) = \mathop{\mathbf{E}}_{x\sim P_S}\left[\min_{i=1,2,\ldots,k} \ell(x, c_i)\right] = \frac{1}{m}\sum_{j=1}^{m} \min_{i=1,2,\ldots,k} \ell(x_j, c_i) \,.$$

We intentionally leave undefined the behavior of B and B' when $R(P_S, \cdot), R'(P_S, \cdot)$ respectively have multiple minimizers.

A brief comment is necessary here. Computer scientists and statisticians when talking about k-means and k-medians they often refer to a particular local search heuristic algorithms for minimizing the empirical risks R and R'. See for example the book by Duda et al. [15, page 527]. In this thesis, however, when we talk about the k-means and the k-medians algorithms, we mean the algorithms B and B', just defined, that compute exact global minima of the empirical risks R and R' respectively. In particular, we ignore any computational issues of how the minimization is carried out.

3.3 Risk Convergence of k-means and k-medians

In this section, we prove the risk convergence of k-means and k-medians ERM algorithms if the domain is a bounded subset of a Euclidean space. Similar results were obtained by a number of people. For k-means it has been done by Pollard [23] and Bartlett et al. [2]. For

[3] In the notation "P_S" the symbol P bears no meaning. It is only a reminder that P_S is a probability distribution.

both k-means and k-medians it has been done by Ben-David [4]. An independent recent result of Biau et al. [10] show risk convergence of k-means in the real Hilbert space.

Our proof is new and has the advantage of being simple. The only non-trivial fact it relies on is the famous result by Vapnik and Chervonenkis [31]. They prove that for certain "small" classes of events the relative frequencies of the events uniformly converge to their probabilities. For the k-means and k-medians risks R and R' we derive from their result that

$$\sup_{y \in Y} |R(P, y) - R(P_S, y)| \qquad \text{and} \qquad \sup_{y \in Y} |R'(P, y) - R'(P_S, y)|$$

converge in probability to zero. The risk convergence of the algorithms B and B' easily follows from that.

We state the result of Vapnik and Chervonenkis as Lemma 3.6. The proof of the lemma can found in the Vapnik and Chervonenkis' original paper [31] and books [30, 29, 13, 14, 1]. In order to phrase the lemma, we need first to define the crucial notion of Vapnik-Chervonenkis dimension of a class of sets.

Definition 3.5 (VC-dimension)**.** *Let $\mathcal{A}$ be a class of subsets of a domain X.* Vapnik-Chervonenkis dimension *(abbreviated as* VC-dimension*) of $\mathcal{A}$ and denoted by* VC-dim$(\mathcal{A})$ *is the largest integer n such that there exists a subset $S \subseteq X$ of size n such that*

$$|\{M \cap S \ : \ M \in \mathcal{A}\}| = 2^n \, .$$

In other words, VC-dim$(\mathcal{A})$ *is the largest cardinality of subset $S \subseteq X$ such that for any subset $S' \subseteq S$ there exists $M \in \mathcal{A}$ such that $S' = S \cap M$.*

Lemma 3.6 (Vapnik-Chervonenkis)**.** *If $\mathcal{A}$ is a family of measurable subsets of $(X, \mathfrak{M})$ with* VC-dim$(\mathcal{A}) < \infty$*, and P is any probability distribution P over X, then for a sample $S \sim P^m$*

$$\sup_{A \in \mathcal{A}} |P_S(A) - P(A)| \quad \to 0 \quad \textit{in probability.}$$

In the rest of this section we assume that the domain is a bounded subset of a Euclidean space $\mathbb{R}^N$. The metric ℓ we use on the domain is the standard Euclidean metric. It is defined for any $x = (x_1, x_2, \ldots, x_N), x' = (x'_1, x'_2, \ldots, x'_N) \in \mathbb{R}^N$ as

$$\ell(x, x') = \|x - x'\|_2 = \sqrt{\sum_{j=1}^{N} (x_j - x'_j)^2} \, .$$

This metric fully specifies the k-means and k-medians risks R and R' defined at the beginning of the chapter. We make X a measurable space by endowing it with the σ-algebra of the Borel sets of (X, ℓ).

Theorem 3.7 (Uniform Convergence of k-means and k-medians)**.** *Let N be any positive integer, X be any bounded subset of $\mathbb{R}^N$, let P be any probability distribution over (the σ-algebra of Borel sets of) X. Let $S = (x_1, x_2, \ldots, x_m)$ be an i.i.d. sample from P. Then, as $m \to \infty$,*

$$\sup_{c_1, c_2, \ldots, c_k \in X} |R(P; c_1, c_2, \ldots, c_k) - R(P_S; c_1, c_2, \ldots, c_k)| \to 0 \qquad \textit{in probability,}$$

$$\sup_{c_1, c_2, \ldots, c_k \in X} |R'(P; c_1, c_2, \ldots, c_k) - R'(P_S; c_1, c_2, \ldots, c_k)| \to 0 \qquad \textit{in probability.}$$

Proof. Let $B(x, r) \subseteq X$ denote an open ball of radius r centered at $x \in X$. Consider the class of sets

$$\mathcal{A} = \left\{ \bigcup_{i=1}^{k} B(c_i, r) \ : \ r \geq 0, c_1, c_2, \ldots, c_k \in X \right\} .$$

This class has finite VC-dimension. To prove it, we first employ Dudley's result [16] which says that the class of open balls $\mathcal{B} = \{B(x, r) \ : \ x \in X, r \geq 0\}$ has VC-dimension (at most) $N + 1$. Second, by Sauer's Lemma [1, Lemma 3.6], for any finite subset $S \subseteq X$, we bound

$$|\{S \cap B \ : \ B \in \mathcal{B}\}| \leq \sum_{i=0}^{n+1} \binom{|S|}{i} . \tag{3.3}$$

Third, for any finite subset S, we have

$$\begin{aligned} |\{S \cap M \ : \ M \in \mathcal{A}\}| &= \left| \left\{ S \cap \left(\bigcup_{i=1}^{k} B(c_i, r) \right) \ : \ r \geq 0, c_1, c_2, \ldots, c_k \in X \right\} \right| \\ &\leq \left| \left\{ S \cap \left(\bigcup_{i=1}^{k} B_i \right) \ : \ B_1, B_2, \ldots, B_k \in \mathcal{B} \right\} \right| \\ &\leq |\{S \cap B \ : \ B \in \mathcal{B}\}|^k . \end{aligned} \tag{3.4}$$

By combining the inequalities (3.3) and (3.4), we obtain

$$|\{S \cap M \ : \ M \in \mathcal{A}\}| \leq \left(\sum_{i=0}^{N+1} \binom{|S|}{i} \right)^k . \tag{3.5}$$

The last expression is a polynomial in $|S|$ of degree $k(N + 1)$. This polynomial grows "slower" than the exponential $2^{|S|}$. Therefore, there exists an integer n' such that if the subset $S \subseteq X$ has size more than n', then the right hand side of (3.5) is strictly less than $2^{|S|}$. Thus, VC-dim$(\mathcal{A})$ is at most n'.[4]

Now, by the uniform convergence result of Vapnik-Chervonenkis (Lemma 3.6), as $m \to \infty$,

$$\sup_{M \in \mathcal{A}} |P(M) - P_S(M)| \to 0 \qquad \text{in probability.}$$

Let D be the diameter of X. For probability distribution P over X and any centers

[4] One can even compute an explicit upper bound on the VC-dimension. See Exercise 4.2 in the book [14].

$c_1, c_2, \ldots, c_k \in X$,

$$\begin{aligned}
R(P; c_1, c_2, \ldots, c_k) &= \mathop{\mathbf{E}}_{x\sim P}\left[\min_{1\le i\le k} \|c_i - x\|_2^2\right] \\
&= \int_0^{D^2} \Pr_{x\sim P}\left[\min_{1\le i\le k} \|c_i - x\|_2^2 \ge t\right] dt \\
&= \int_0^{D^2} 1 - \Pr_{x\sim P}\left[\min_{1\le i\le k} \|c_i - x\|_2^2 < t\right] dt \\
&= D^2 - \int_0^{D^2} P\left(\bigcup_{i=1}^{k} B(c_i, \sqrt{t})\right) dt
\end{aligned}$$

and similarly

$$\begin{aligned}
R'(P; c_1, c_2, \ldots, c_k) &= \mathop{\mathbf{E}}_{x\sim P}\left[\min_{1\le i\le k} \|c_i - x\|_2\right] \\
&= \int_0^{D} \Pr_{x\sim P}\left[\min_{1\le i\le k} \|c_i - x\|_2 \ge t\right] dt \\
&= \int_0^{D} 1 - \Pr_{x\sim P}\left[\min_{1\le i\le k} \|c_i - x\|_2 < t\right] dt \\
&= D - \int_0^{D} P\left(\bigcup_{i=1}^{k} B(c_i, t)\right) dt\,.
\end{aligned}$$

Therefore, for an i.i.d. sample S from any probability distribution P

$$\begin{aligned}
&\sup_{c_1,c_2,\ldots,c_k\in X} |R(P; c_1, c_2, \ldots, c_k) - R(P_S; c_1, c_2, \ldots, c_k)| \\
&\quad= \sup_{c_1,c_2,\ldots,c_k\in X} \left|\int_0^{D^2} P\left(\bigcup_{i=1}^{k} B(c_i, \sqrt{t})\right) - P_S\left(\bigcup_{i=1}^{k} B(c_i, \sqrt{t})\right)\right| dt \\
&\quad\le \sup_{c_1,c_2,\ldots,c_k\in X} \int_0^{D^2} \left|P\left(\bigcup_{i=1}^{k} B(c_i, \sqrt{t})\right) - P_S\left(\bigcup_{i=1}^{k} B(c_i, \sqrt{t})\right)\right| dt \\
&\quad\le D^2 \sup_{c_1,c_2,\ldots,c_k\in X} \sup_{t\ge 0} \left|P\left(\bigcup_{i=1}^{k} B(c_i, \sqrt{t})\right) - P_S\left(\bigcup_{i=1}^{k} B(c_i, \sqrt{t})\right)\right| \\
&\quad= D^2 \sup_{M\in\mathcal{A}} |P(M) - P_S(M)| \quad \to 0 \quad \text{in probability}
\end{aligned}$$

and similarly

$$\begin{aligned}
&\sup_{c_1,c_2,\ldots,c_k\in X} |R'(P;c_1,c_2,\ldots,c_k) - R'(P_S;c_1,c_2,\ldots,c_k)| \\
&\quad = \sup_{c_1,c_2,\ldots,c_k\in X} \left| \int_0^D P\left(\bigcup_{i=1}^k B(c_i,t)\right) - P_S\left(\bigcup_{i=1}^k B(c_i,t)\right) \right| dt \\
&\quad \le \sup_{c_1,c_2,\ldots,c_k\in X} \int_0^D \left| P\left(\bigcup_{i=1}^k B(c_i,t)\right) - P_S\left(\bigcup_{i=1}^k B(c_i,t)\right) \right| dt \\
&\quad \le D \sup_{c_1,c_2,\ldots,c_k\in X} \sup_{t\ge 0} \left| P\left(\bigcup_{i=1}^k B(c_i,t)\right) - P_S\left(\bigcup_{i=1}^k B(c_i,t)\right) \right| \\
&\quad = D \sup_{M\in\mathcal{A}} |P(M) - P_S(M)| \quad \to 0 \quad \text{in probability.}
\end{aligned}$$

■

Corollary 3.8 (Risk Convergence of k-means and k-medians)**.** *Consider the instantiations of the* k*-means and* k*-medians risks optimization schemes over a bounded subset of a Euclidean space* $\mathbb{R}^N$ *with Euclidean metric. The optimization algorithms* B, B' *are risk converging for the schemes respectively.*

Proof. We prove the corollary only for k-means. The argument for k-medians is the same. Fix any $\epsilon > 0$ and any $\delta > 0$. By the above Theorem, there exists m_0 such that for all $m \ge m_0$ with probability $1-\delta$,

$$\sup_{c_1,c_2,\ldots,c_k\in X} |R(P;c_1,c_2,\ldots,c_k) - R(P_S;c_1,c_2,\ldots c_k)| \le \epsilon/2\ . \tag{3.6}$$

Suppose that the random event (3.6) occurs. Then, if we denote by y^* a minimizer of $R(P,\cdot)$, then since $B(S)$ is a minimizer of $R(P_S,\cdot)$, we have

$$\begin{aligned}
R(P,B(S)) &\le R(P_S,B(S)) + \epsilon/2 \\
&\le R(P_S,y^*) + \epsilon/2 \\
&\le R(P,y^*) + \epsilon/2 + \epsilon/2 \\
&= \mathrm{opt}_R(P) + \epsilon\ .
\end{aligned}$$

Thus with probability $1-\delta$, $|R(P,B(S)) - \mathrm{opt}_R(P)| \le \epsilon$ for any $m \ge m_0$. Since ϵ and δ are arbitrary, risk convergence of B follows. ■

Chapter 4

Stability of Optimization Algorithms

4.1 The Stability Theorem

In this section we investigate the stability of clustering algorithms induced by optimization algorithms that are risk converging. Speaking informally, we show that their stability depends only on the existence of a unique risk minimizer of the risk function. In this section we prove the Stability Theorem (Theorem 4.1) which is essentially one implication of this statement. The theorem states that under a certain condition on the risk function and the clustering distance, which we call *inverse continuity* condition, the existence of a unique risk minimizer guarantees that any clustering algorithm induced by a risk converging optimization algorithm is stable.

Note that the Stability Theorem applies also to algorithms which are not ERM. Its main application, however, are the k-means and k-medians ERM algorithms. We apply the Stability Theorem to these algorithms in Section 4.2 for the case where the domain is a convex compact subset of an Euclidean space $\mathbb{R}^N$.

Theorem 4.1 (Stability Theorem)**.** *Let* $(Y, \mathfrak{P}, R, \Gamma)$ *be a risk optimization scheme over a domain* X*, and let* $d : \mathfrak{P} \times \mathfrak{C} \times \mathfrak{C} \to [0, 1]$ *be clustering distance. Let* $P \in \mathfrak{P}$ *be a probability measure, and let* y^* *be a risk minimizer for* P*. Suppose that the following* inverse continuity *condition holds:*

$$\forall \eta > 0 \; \exists \epsilon > 0 \; \forall y' \in Y \quad \big(R(P, y') \leq \operatorname{opt}_R(P) + \epsilon \implies d_P(\Gamma(y'), \Gamma(y^*)) \leq \eta\big)$$

If A *is any risk converging optimization algorithm for the scheme* $(Y, \mathfrak{P}, R, \Gamma)$*, then the clustering algorithm* $\Gamma \circ A$ *induced by* A *is stable on* P*.*

Roughly speaking, the inverse continuity condition requires that the set of "ϵ-minimizers" of the risk function is contained in a small neighborhood of the unique risk minimizer.

Note that, formally, the inverse continuity condition implies that $d_P(\Gamma(y^*), \Gamma(y'^*)) = 0$ for any (other) risk minimizer $y'^* \in \operatorname{argmin}_{y \in Y} R(P, y)$. Therefore, y^* can be thought of as a *unique* risk minimizer for P.

Proof of the Stability Theorem. Let A be any risk converging optimization algorithm for $(Y, \mathfrak{P}, R, \Gamma)$. Pick any $\delta > 0$ and any $\eta > 0$. The inverse continuity condition implies that there exists $\epsilon > 0$ such that

$$\forall y' \in Y \qquad R(P, y') \leq \operatorname{opt}_R(P) + \epsilon \implies d_P(\Gamma(y'), \Gamma(y^*)) \leq \eta$$

Since A is risk converging there exists m_0 such that for all $m \geq m_0$

$$\Pr_{S \sim P^m} \left[R(P, A(S)) > \mathrm{opt}_R(P) + \epsilon \right] \leq \delta \ .$$

Combining these two facts thus implies that for all $m \geq m_0$

$$\Pr_{S \sim P^m} \left[d_P(\Gamma(A(S)), \Gamma(y^*)) > \eta \right] \leq \Pr_{S \sim P^m} \left[R(P, A(S)) > \mathrm{opt}_R(P) + \epsilon \right] \leq \delta \ .$$

Therefore, for all $m \geq m_0$ we upper bound instability as

$$\begin{aligned}
\mathrm{instab}(\Gamma \circ A, P, m) &= \mathop{\mathbf{E}}_{\substack{S_1 \sim P^m \\ S_2 \sim P^m}} d_P(\Gamma(A(S_1)), \Gamma(A(S_2))) \\
&\leq \mathop{\mathbf{E}}_{\substack{S_1 \sim P^m \\ S_2 \sim P^m}} \left[d_P(\Gamma(A(S_1)), \Gamma(y^*)) + d_P(\Gamma(y^*), \Gamma(A(S_2))) \right] \\
&= 2 \mathop{\mathbf{E}}_{S \sim P^m} d_P(\Gamma(A(S)), \Gamma(y^*)) \\
&\leq 2 \left(\eta \cdot \Pr_{S \sim P^m} \left[d_P(\Gamma(A(S)), \mathcal{C}^*) \leq \eta \right] + 1 \cdot \Pr_{S \sim P^m} \left[d_P(\Gamma(A(S)), \mathcal{C}^*) > \eta \right] \right) \\
&\leq 2 \left(\eta + \Pr_{S \sim P^m} \left[R(P, A(S)) > \mathrm{opt}_R(P) + \epsilon \right] \right) \\
&\leq 2 \left(\eta + \delta \right)
\end{aligned}$$

Since η, δ are arbitrary, $\lim_{m \to \infty} \mathrm{instab}(A, P, m) = 0$. ■

4.2 Stability of k-means and k-medians

We apply the Stability Theorem to a particular instantiation of the k-means and k-medians optimizations schemes and the ERM optimization algorithms B and B' from Section 3.2. The instantiation we have in mind is the case when the domain X is a compact convex subset of an Euclidean space $\mathbb{R}^N$ with non-empty interior. The domain is endowed with σ-algebra of Borel sets and Euclidean metric ℓ, defined for any $x = (x_1, x_2, \ldots, x_N), x' = (x'_1, x'_2, \ldots, x'_N) \in X$ as

$$\ell(x, x') = \|x - x'\|_2 = \sqrt{\sum_{j=1}^{N} (x_j - x'_j)^2} \ ,$$

and $\mathfrak{P}$ is the set of all absolutely continuous probability distributions over X, densities of which are bounded. Precisely, P belongs to $\mathfrak{P}$ if and only if there exists a measurable function $f : X \to \mathbb{R}^+_0$ such that for any Borel set $M \subset X$, $P(M)$ is defined as the Lebesgue integral

$$P(M) = \int_M f(x) \, dx \ ,$$

and there exists a number G such that for all $x \in X$, $f(x) \leq G$.[1] Our goal is to prove that

[1] Note that G depends on P.

Theorem 4.2 (Stability of k-means and k-medians)**.** *Let X be a convex compact subset of $\mathbb{R}^N$ with non-empty interior and let $k \geq 2$ be an integer. Let $(Y, \mathfrak{P}, R, \Gamma)$ ($(Y, \mathfrak{P}, R', \Gamma)$) be the instantiation of the k-means (k-medians) optimization scheme over the domain X where $\mathfrak{P}$ is the set of absolutely continuous probability distributions over X with bounded densities. Let d be the Hamming clustering distance. For any instance $P \in \mathfrak{P}$, if there exists up to permutation of the centers unique risk minimizer y^* of R (R'), then the clustering algorithm $\Gamma \circ B$ ($\Gamma \circ B'$) is stable on P.*

Let us clarify what is meant by saying that y^* is unique risk minimizer of R up to permutation of the centers. If we write out $y^* = (c_1^*, c_2^*, \ldots, c_k^*)$, then we say that y^* is unique up to permutation of the centers if and only if

$$\{y' \in Y \,:\, R(P, y') = \mathrm{opt}_R(P)\} = \{(c^*_{\pi(1)}, c^*_{\pi(2)}, \ldots, c^*_{\pi(k)}) \,:\, \pi \in S_k\}$$

where S_k denotes the set of all permutations on $\{1, 2, \ldots, k\}$. The definition for R' is the same. In the proof below we use the notation $\pi(y)$ to denote $(c_{\pi}(1), c_{\pi}(2), \ldots, c_{\pi}(k))$ where $y = (c_1, c_2, \ldots, c_k) \in Y$ and $\pi \in S_k$.

Proof. We prove the theorem for k-means. The proof for k-medians is analogous. Consider any probability distribution $P \in \mathfrak{P}$. By Theorem 3.7 the algorithm B is risk converging. Therefore, we need to verify the inverse continuity condition.

For that purpose we define the metric D on $Y = X^k$ as

$$D(y, y') = D((c_1, c_2, \ldots, c_k), (c'_1, c'_2, \ldots, c'_k)) = \sqrt{\sum_{i=1}^{k} \|c_i - c'_i\|_2^2}\,.$$

that is, D is simply the Euclidean metric on a subset of $\mathbb{R}^{Nk}$. Since X is compact, the metric space (Y, D) is also compact. Recall that P has a density function which is bounded. Also recall that X is compact and hence its has finite diameter. These two facts imply that the function $R(P, \cdot) : Y \to \mathbb{R}$ is a continuous function on the metric space (Y, D).

The first claim that we make is that the following "inverse continuity" holds:

$$\forall\, \zeta > 0\ \exists\, \epsilon > 0\ \forall y' \in Y \quad \left[R(P, y') \leq \mathrm{opt}_R(P) + \epsilon \implies (\exists \pi \in S_k\ D(y', \pi(y^*)) \leq \zeta)\right]\,.$$

The claim follows from continuity of $R(P, \cdot) : Y \to \mathbb{R}$, compactness of (Y, D) and uniqueness of y^*. To give a more detailed proof of the claim, consider any $\zeta > 0$. Since Y is compact it can be covered by finitely many closed balls $b_1, b_2, \ldots, b_n$ each of radius $\zeta/2$. Each ball b_j is itself compact and hence $R(P, \cdot)$ attains minimum $m_j = \min_{y \in b_j} R(P, y)$. Let $m^* = \min\{m_1, m_2, \ldots, m_j\} = \mathrm{opt}_R(P)$ be the minimum of the minima and let $m' = \min(\{m_1, m_2, \ldots, m_n\} \setminus \{m^*\})$ be the second smallest minimum. In the case when $m_1 = m_2 = \cdots = m_n$ we define $m' = m^* + 1$. Let $\epsilon = (m' - m^*)/2$. Consider any $y' \in Y$ that satisfies $R(P, y') \leq \mathrm{opt}_R(P) + \epsilon$. If y' lies in a ball b_j for some j, then $m_j = m^*$ since otherwise $R(P, y') \geq m' = \mathrm{opt}_R(P) + \epsilon/2 > \mathrm{opt}_R(P) + \epsilon$ which is a contradiction. Consider an arbitrary ball b_j in which y' lies. (Since the balls cover Y, there exists at least one such ball.) Since $m_j = m^*$ and b_j is compact, there exists a minimizer $y'^* \in b_j$, $R(P, y'^*) = \mathrm{opt}_R(P)$. Since y'^* and y' lie in the same ball of radius $\zeta/2$, $D(y', y'^*) \leq \zeta$. Finally, since y^* is unique

up to permutation of the centers, there exists $\pi \in S_k$ such that $y'^* = \pi(y^*)$ and the claim follows.

The second claim that we make is that

$$\forall \eta > 0 \; \exists \zeta > 0 \; \forall y' \in Y \; ((\exists \pi \in S_k \; D(y', \pi(y^*)) \leq \zeta) \implies d_P(\Gamma(y'), \Gamma(y^*)) \leq \eta) \; .$$

The claim follows from boundedness of density of P, boundedness of X and that the centers $(c_1^*, c_2^*, \ldots, c_k^*)$ of y^* have certain positive separation

$$\Delta = \min_{1 \leq i < j \leq k} \|c_i^* - c_j^*\|_2 \; .$$

Note that if the separation Δ was not positive, then two centers would be collocated. But in such case one of the centers could be moved, provided that $|X| > 1$, without increasing the risk $R(P, \cdot)$, which would contradict uniqueness of y^*. (If $|X| = 1$, the claim trivially follows.)

To prove the second claim, let $\eta > 0$ and consider $\zeta \leq \Delta/2$ which we specify later. Let $y' = (c_1', c_2', \ldots, c_k')$ be any solution such that $D(y', \pi(y^*)) \leq \zeta$ for some $\pi \in S_k$. For notational convenience and without loss of generality assume that π is the identity permutation. The minimum distance $\Delta' = \min_{1 \leq i < j \leq k} \|c_i' - c_j'\|_2$ between centers of y' is at least $\Delta - \eta \geq \Delta/2$. Consider, for any $i \neq j$, $1 \leq i, j \leq k$, the "bisecting" halfspaces

$$\begin{aligned} h_{i,j}^* &= \{x \in \mathbb{R}^N \; : \; (c_i^* - c_j^*)^T x \geq (c_i^* - c_j^*)^T (c_i^* + c_j^*)/2\} \\ h_{i,j}' &= \{x \in \mathbb{R}^N \; : \; (c_i' - c_j')^T x \geq (c_i' - c_j')^T (c_i' + c_j')/2\} \; . \end{aligned}$$

The i-th cells of Voronoi$(c_1^*, c_2^*, \ldots, c_k^*)$ and Voronoi$(c_1', c_2', \ldots, c_k')$ can be written as

$$C_i^* = \bigcap_{\substack{j=1\ldots k \\ j \neq i}} (X \cap h_{i,j}^*) \qquad \text{and} \qquad C_i' = \bigcap_{\substack{j=1\ldots k \\ j \neq i}} (X \cap h_{i,j}')$$

respectively. Let $S_{i,j}$ be the symmetric difference of $(X \cap h_{i,j}^*)$ and $(X \cap h_{i,j}')$. Since $\|c_i' - c_i^*\|_2 \leq \eta$, $\|c_j' - c_j^*\|_2 \leq \eta$, $\|c_i' - c_j'\|_2 \geq \Delta/2$, $\|c_i^* - c_j^*\|_2 \geq \Delta$ and X has finite diameter, the constant ζ can be chosen sufficiently small so that the Euclidean volume of $S_{i,j}$ is arbitrarily small. Since P is absolutely continuous and has bounded density, ζ can also be chosen small enough so that the probability mass $P(S_{i,j})$ is arbitrarily small. More precisely, ζ can be chosen so that for any $P(S_{i,j}) \leq \eta/k^2$ for any $i \neq j$

$$\begin{aligned} d_P(\Gamma(y'), \Gamma(y^*)) &\leq \Pr_{\substack{x \sim P \\ x' \sim P}} \left[x \in \bigcup_{1 \leq i < j \leq k} S_{i,j} \; \vee \; x' \in \bigcup_{1 \leq i < j \leq k} S_{i,j} \right] \\ &\leq 2P \left(\bigcup_{1 \leq i < j \leq k} S_{i,j} \right) \\ &\leq \eta \; . \end{aligned}$$

Since y' was arbitrary, the claim is proved.

Combining the two claims in an obvious way implies the inverse continuity condition in the Stability Theorem. Therefore by the Stability Theorem the clustering algorithm $\Gamma \circ B$ is stable on P. ■

Chapter 5

Instability Because of Symmetry

So far we have looked at conditions which guarantee that a clustering algorithm is asymptotically stable. Now we investigate conditions which imply that a clustering algorithm is unstable. One such condition is symmetry.

Before we dive into technical details we illustrate the idea on a simple example. Consider the output of the k-means ERM algorithm B from Section 3.2 on samples coming from a distribution P over the Euclidean plane $\mathbb{R}^2$ depicted on Figure 5.1. For $k = 2$ there are, up to permutation of the centers, exactly two risk minimizers (c_1, c_2) and (c_1', c_2') of the k-means risk. The risk minimizers are depicted on Figure 5.2. Notice that P as well as the risk minimizers are symmetric with respect the group generated by the rotation of the plane by 90° around the mean of the distribution. It is not hard to see that as $m \to \infty$ with probability approaching $1/2$ the algorithm B outputs a solution approximately equal to (c_1, c_2) and with probability approaching $1/2$ it outputs a solution approximately equal to (c_1', c_2'). Thus because of symmetry B for $k = 2$ is unstable on P.

Symmetries are certain types of transformations. Recall that if $T : X \to X'$ is a measurable transformation between two measurable spaces $(X, \mathfrak{M})$ and $(X', \mathfrak{M}')$, then for any probability distribution P over X, T induces a probability measure over X'. The induced measure is denoted by $P \circ T^{-1}$ and defined for any measurable set $M' \subseteq X'$ as

$$(P \circ T^{-1})(M') = P(T^{-1}(M')) \,.$$

In a similar fashion, for any clustering $\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$ of X, T induces a clustering of X'. We denote the *induced clustering* by $T(\mathcal{C})$ and define it as

$$T(\mathcal{C}) = \{T(C_1), T(C_2), \ldots, T(C_k)\} \,.$$

Likewise, for any sample $S = (x_1, x_2, \ldots, x_m) \in X^m$, T induces a sample $T(S) \in (X')^m$ which we define as $T(S) = (T(x_1), T(x_2), \ldots, T(x_m))$. Finally, for an event $E \subseteq X^m$, T induces an event $T(E)$ which we define as $T(E) = \{T(S) \,:\, S \in E\}$.

Definition 5.1 (Transformation Invariance)**.** *Let $(Y, \mathfrak{P}, R, \Gamma)$ be an optimization scheme over a domain X, B be an optimization algorithm for the scheme, $d : \mathfrak{P} \times \mathfrak{C} \times \mathfrak{C} \to [0, 1]$ be a clustering distance. Let $T : X \to X$ be a measurable bijection and $U : Y \to Y$ be a bijection on the set of solutions.*

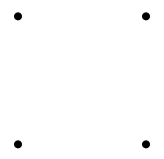

Figure 5.1: The picture shows a discrete probability distribution P over $\mathbb{R}^2$ concentrated on 4 vertices of a square. Each vertex has probability mass $1/4$.

Figure 5.2: The picture shows two risk minimizers (c_1, c_2) and (c'_1, c'_2) of the k-means risk, $k = 2$, for the distribution P shown on Figure 5.1. Up to permutation of the centers these are the only risk minimizers.

- *We say that* d *is* T-invariant *if for any probability distribution* $P \in \mathfrak{P}$ *also* $P \circ T^{-1}$ *lies in* $\mathfrak{P}$*, and furthermore for any clusterings* $\mathcal{C}_1, \mathcal{C}_2 \in \mathfrak{C}$

$$d_{P \circ T^{-1}}(T(\mathcal{C}_1), T(\mathcal{C}_2)) = d_P(\mathcal{C}_1, \mathcal{C}_2) \ .^1$$

- *We say that the risk function* R *is* (T, U)-invariant *if for any instance* $P \in \mathfrak{P}$

$$R(P \circ T^{-1}, U(y)) = R(P, y) \ .$$

- *We say that* B *is* (T, U)-invariant *when* $B(T(S)) = U(B(S))$ *asymptotically almost surely*[2]*. Formally,*

$$\lim_{m \to \infty} \Pr_{S \sim P^m} [B(T(S)) = U(B(S))] = 1 \ .$$

- *We say that* T *and* U commute *with respect to* Γ *if for any* $P \in \mathfrak{P}$ *and any solution* $y \in Y$

$$d_P(T(\Gamma(y)), \Gamma(U(y))) = 0 \ .$$

For example, if X is any domain, $T : X \to X$ is any measurable bijection, $\mathfrak{P}$ is the family of all distributions over X, then the Hamming clustering d is T-invariant. In essence, it means that the Hamming clustering distance does not depend on the identity of the domain points.

[1] Note the non-trivial fact that the clusters of $T(\mathcal{C}_1)$ and $T(\mathcal{C}_2)$ are measurable.

[2] Asymptotically almost sure convergence is used in the theory of random graphs. Usually it is abbreviated as a.a.s.

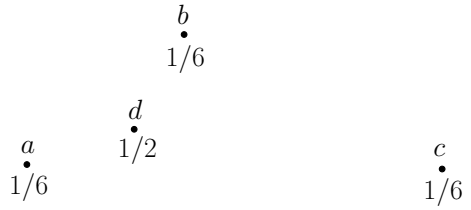

Figure 5.3: The figure shows a discrete probability distribution P over $\mathbb{R}^2$ concentrated on four points $\{a, b, c, d\}$. The probability masses are $P(\{a\}) = P(\{b\}) = P(\{c\}) = 1/6$ and $P(\{d\}) = 1/2$. Any measurable bijection $T : \mathbb{R}^2 \to \mathbb{R}^2$ such that $T(\{a, b, c\}) = \{a, b, c\}$ and $T(d) = d$ is a symmetry of P. Note that the geometric positions of the points a, b, c, d are *not* symmetric in any way whatsoever.

The risks R and R' of the generic k-means and k-medians optimization schemes over a metric space (X, ℓ) are (T, U)-invariant if T is an isometry of (X, ℓ), that is, when

$$\forall x, x' \in X \qquad \ell(T(x), T(x')) = \ell(x, x') \ ,$$

and $U : Y \to Y$ is defined in terms of T for any $(c_1, c_2, \ldots, c_k) \in Y$ as

$$U(c_1, c_2, \ldots, c_k) = (T(c_1), T(c_2), \ldots, T(c_k)) \ .$$

When $X = \mathbb{R}^d$, $\mathfrak{P}$ is the set of absolutely continuous distributions, then T and U commute with respect to Γ. In this case, the ERM algorithms B and B' are (T, U)-invariant. Note that, however, $B(T(S)) = U(B(S))$ might fail for certain samples S. For example, the equality might fail to hold if $R(P_S, \cdot)$ has multiple optimal solutions. For this reason the definition requires only $B(T(S)) = U(B(S))$ to hold only asymptotically almost surely.

Definition 5.2 (Measure-preserving Symmetry)**.** *Let P be a probability distribution over a domain X. A measurable bijection $T : X \to X$ is said to be a* symmetry *of P if $P \circ T^{-1} = P$. That is, T is a* symmetry *of P if for any measurable $M \subset X$, $T^{-1}(M)$ is also measurable and $P(T^{-1}(M)) = P(M)$.*

An example of a measure-preserving symmetry is shown on Figure 5.3.

Now, we describe the conditions under which symmetry leads to instability.

Theorem 5.3 (Instability Because of Symmetry)**.** *Let* $(Y, \mathfrak{P}, R, \Gamma)$ *be an optimization scheme over* X*. Let* B *be a risk converging algorithm for the optimization scheme. Let* $d : \mathfrak{P}\times\mathfrak{C}\times\mathfrak{C} \to [0,1]$ *be clustering distance. Let* $P \in \mathfrak{P}$ *be an instance.*

Let G *be a finite set of measurable transformations* $T : X \to X$ *which forms a group under function composition and has order at least* 2*. Let* ϕ *be a function that maps any element* $T \in G$ *to a bijection* $U : Y \to Y$ *and* ϕ *is a group homomorphism, that is, for any* $T, T' \in G$*,* $\phi(T \circ T') = \phi(T) \circ \phi(T')$*.*

Suppose for any $T \in G$*,*

- d *is* T*-invariant,*
- R *is* $(T, \phi(T))$*-invariant,*
- B *is* $(T, \phi(T))$*-invariant,*
- T *and* $\phi(T)$ *commute with respect to* Γ*,*
- T *is a symmetry of* P*.*

Let y^* *be a risk minimizer of* $R(P, \cdot)$*. If for every non-indentity element* $T \in G$

$$d_P(\Gamma(y^*), T(\Gamma(y^*))) > 0 \ .$$

and the following generalized inverse continuity *condition holds*

$$\forall \eta > 0 \ \exists \epsilon > 0 \ \forall y' \in Y \ \left[R(P, y') \le \mathrm{opt}_R(P) + \epsilon \implies (\exists T \in G \ d_P(\Gamma(y'), T(\Gamma(y^*))) \le \eta)\right]$$

then the induced algorithm $\Gamma \circ B$ *is unstable on* P*.*

Proof. Let $\Delta = \min\{d_P(\Gamma(y^*), T(\Gamma(y^*)) \ : \ T \in G\}$. The generalized inverse continuity condition guarantees that there exists $\epsilon > 0$ such that for all $y' \in Y$

$$R(P, y') \le \mathrm{opt}_R(P) + \epsilon \implies \exists T \in G \ d_P(\Gamma(y'), T(\Gamma(y^*))) \le \Delta/3 \ . \tag{5.1}$$

For any sample size m we define the events[3]

$$\begin{aligned} E_1^T &= \{S \in X^m \ : \ B(T(S)) = \phi(T)(B(S))\} && \text{for any } T \in G \\ E_1 &= \bigcap_{T \in G} E_1^T \\ E_2 &= \left\{S \in X^m \ : \ R(P, B(S)) \le \mathrm{opt}_R(P) + \epsilon\right\} \end{aligned}$$

To keep our notation simple we do not write explicitly the dependence of E_1 and E_2 on m.

Now, since B is $(T, \phi(T))$-invariant for every $T \in G$, there exists m_0 such that for all $m \ge m_0$, $\Pr[E_1] \ge 0.9$. Since B is risk converging, there exists m_0' such that for all $m \ge m_0'$, $\Pr[E_2] \ge 0.9$. Then by the union bound, for $m \ge \max\{m_0, m_0'\}$,

$$\Pr\left[E_1 \cap E_2\right] = 1 - \Pr[E_1^c \cup E_2^c] \ge 0.8 \ .$$

[3]Both E_1 and E_2 need to be measurable for any m. We ignore this issue here.

For every $T \in G$ define the event

$$H_T = \{S \in X^m \ : \ d_P(\Gamma(B(S)), T(\Gamma(y^*))) \leq \Delta/3\} \ .$$

First, because of (5.1), H_T is a superset of E_2. Second, if $T \neq T'$,

$$\begin{aligned}
&d_P(T(\Gamma(y^*)), T'(\Gamma(y^*))) \\
&\quad = d_{P\circ T}(\Gamma(y^*), (T^{-1} \circ T')(\Gamma(y^*))) && \text{(by T-invariance of d)} \\
&\quad = d_P(\Gamma(y^*), (T^{-1} \circ T')(\Gamma(y^*))) && \text{(since } T^{-1} \text{ is a symmetry of P)} \\
&\quad \geq \Delta
\end{aligned}$$

and thus by triangle inequality for any $S \in H_T$ and any $S' \in H_{T'}$

$$\begin{aligned}
&d_P(\Gamma(B(S)), \Gamma(B(S'))) \\
&\geq d_P(T(\Gamma(y^*)), T'(\Gamma(y^*))) - d_P(\Gamma(B(S)), T(\Gamma(y^*))) - d_P(\Gamma(B(S')), T'(\Gamma(y^*))) \\
&\geq \Delta - \Delta/3 - \Delta/3 \\
&= \Delta/3
\end{aligned}$$

and therefore H_T and $H_{T'}$ are disjoint. Therefore, $\{H_T \cap E_1 \cap E_2 \ : \ T \in G\}$ is a partition of $E_1 \cap E_2$ and hence for any $m \geq \max\{m_0, m_0'\}$ there exists $T_0 \in G$ such that

$$\Pr[H_{T_0} \cap E_1 \cap E_2] \geq 0.8/|G| \ .$$

Let T_1 be any non-identity element of G. We claim that

$$T_1(H_{T_0} \cap E_1 \cap E_2) = H_{T_1 \circ T_0} \cap E_1 \cap E_2 \ .$$

To see that, we write $T_1(H_{T_0} \cap E_1 \cap E_2)$ as

$$\begin{aligned}
T_1(H_{T_0} \cap E_1 \cap E_2) = \Big\{ T_1(S) \in X^m \ : \ & d_P(\Gamma(B(S)), T_0(\Gamma(y^*))) \leq \Delta/3, \\
& \forall T \in G \quad B(T(S)) = \phi(T)(B(S)), \\
& R(P, B(S)) \leq \mathrm{opt}_R(P) + \epsilon \Big\}
\end{aligned}$$

Equivalently,

$$\begin{aligned}
T_1(H_{T_0} \cap E_1 \cap E_2) = \Big\{ S \in X^m \ : \ & d_P(\Gamma(B(T_1^{-1}(S))), T_0(\Gamma(y^*))) \leq \Delta/3, \\
& \forall T \in G \quad B(T(T_1^{-1}(S))) = \phi(T)(B(T_1^{-1}(S))), \\
& R(P, B(T_1^{-1}(S))) \leq \mathrm{opt}_R(P) + \epsilon \Big\}
\end{aligned}$$

Consider the three conditions on S defining the set. Since ϕ is a homomorphism and G is a group, the second condition is equivalent to

$$\forall T \in G \quad B(T(S)) = \phi(T)(B(S)) \ . \tag{5.2}$$

Therefore the left side of the third condition can be written as

$$\begin{aligned}
&R(P, B(T_1^{-1}(S))) \\
&\quad = R(P, \phi(T_1^{-1})(B(S))) && \text{(by (5.2))} \\
&\quad = R(P \circ T_1^{-1}, \phi(T_1)(\phi(T_1)^{-1}(B(S)))) && \text{(since } R \text{ is } (T_1, \phi(T_1)) \text{ invariant)} \\
&\quad = R(P \circ T_1^{-1}, B(S)) && \text{(since } \phi \text{ is homomorphism)} \\
&\quad = R(P, B(S)) && \text{(since } T_1 \text{ is a symmetry of } P)
\end{aligned}$$

and the left hand side of the first condition can be written as

$$\begin{aligned}
&d_P(\Gamma(B(T_1^{-1}(S))), T_0(\Gamma(y^*))) \\
&\quad = d_P(\Gamma(\phi(T_1^{-1})(B(S))), T_0(\Gamma(y^*))) && \text{(by (5.2))} \\
&\quad = d_P(T_1^{-1}(\Gamma(B(S))), T_0(\Gamma(y^*))) && \text{(by } (T^{-1}, \phi(T^{-1}))\text{-commutativity of } \Gamma) \\
&\quad = d_{P \circ T_1^{-1}}(\Gamma(B(S))), (T_1 \circ T_0)(\Gamma(y^*))) && \text{(by } T_1\text{-invariance of } d) \\
&\quad = d_P(\Gamma(B(S))), (T_1 \circ T_0)(\Gamma(y^*))) && \text{(since } T_1 \text{ is a symmetry of } P)
\end{aligned}$$

and thus

$$\begin{aligned}
T_1(H_{T_0} \cap E_1 \cap E_2) = \Big\{ S \in X^m \,:\, & d_P(\Gamma(B((S))), (T_1 \circ T_0)(\Gamma(y^*))) \le \Delta/3, \\
& \forall T \in G \quad B(T(S)) = \phi(T)(B(S)), \\
& R(P, B(S)) \le \mathrm{opt}_R(P) + \epsilon \Big\} = H_{T_1 \circ T_0} \cap E_1 \cap E_2
\end{aligned}$$

and the claim is proven. Since T_1^{-1} is a symmetry of P and therefore also of P^m

$$P^m(H_{T_1 \circ T_0} \cap E_1 \cap E_2) = P^m(T_1(H_{T_0} \cap E_1 \cap E_2)) = P^m(H_{T_0} \cap E_1 \cap E_2) \ge 0.8/|G| \,.$$

We lower bound stability for any $m \ge \max\{m_0, m_0'\}$ as

$$\begin{aligned}
\mathrm{instab}(\Gamma \circ B, P, m) &= \mathop{\mathbf{E}}_{\substack{S_1 \sim P^m \\ S_2 \sim P^m}} d_P(\Gamma(B(S_1)), \Gamma(B(S_2))) \\
&\ge \Delta/3 \cdot \Pr[S_1 \in H_{T_0} \cap E_1 \cap E_2] \cdot \Pr[S_2 \in H_{T_1 \circ T_0} \cap E_1 \cap E_2] \\
&\ge \Delta/3 \cdot (0.8/|G|)^2
\end{aligned}$$

Thus $\Gamma \circ B$ is unstable on P. ■

5.1 Instability Because of Symmetry for k-means and k-medians

We apply Theorem 5.3 to k-means and k-medians. For a solution $y = (c_1, c_2, \ldots, c_k) \in X^k$ and $T : X \to X$ we use $T(y)$ to denote $(T(c_1), T(c_2), \ldots, T(c_k))$.

Corollary 5.4. *Let* $(Y, \mathfrak{P}, R, \Gamma)$ *(or* $(Y, \mathfrak{P}, R', \Gamma)$*) be a* k*-means (or* k*-medians) scheme over domain* X *which is a convex compact subset of* $\mathbb{R}^d$ *and* $\mathfrak{P}$ *is the set of absolutely continuous distributions over* X *with bounded densities. Let* $d : \mathfrak{P} \times \mathfrak{C} \times \mathfrak{C} \to [0, 1]$ *be the Hamming clustering distance. Let* $P \in \mathfrak{P}$ *be an instance and* G *be a finite group of order at least 2 of isometries of* X *such that each elemnent of* G *is also a symmetry of* P*. Suppose there exists a risk minimizer* y^* *for* P *such that for any risk minimizer* y'^* *there exists* $T \in G$ *such that* $y'^* = T(y^*)$*. Suppose further for every non-identity element* $T \in G$*,* $d_P(\Gamma(y^*), T(\Gamma(y^*))) > 0$*. Then, the clustering algorithm induced by* B *(or* B'*) is unstable on* P*.*

Proof. We prove the Corollary for k-means only. The proof for k-medians is the same. We define $\phi(T) = T$, that is, $(\phi(T))(c_1, c_2, \ldots, c_k) = (T(c_1), T(c_2), \ldots, T(c_k))$ for any $(c_1, c_2, \ldots, c_k) \in Y$. Obviously ϕ is an homomorphism. Clearly, for any isometry T, d is T-invariant, R and B are $(T, \phi(T))$ invariant, T and $\phi(T)$ commute with respect to Γ. The generalized inverse continuity is verified similarly as in Theorem 4.2; we only need to realize that $\{T(y^*) : T \in G\}$ is the set of all risk minimizers. ■

Chapter 6

Instability of k-means

In this chapter we prove that for a large class of probability distributions over a Euclidean space $\mathbb{R}^N$ the clustering algorithm induced by the k-means ERM algorithm is unstable whenever there exist multiple k-means risk minimizers. The class of distributions for which we prove the result consists of all probability distributions with finite support. To avoid trivial degenerate cases we exclude the distributions with size of the support less than k.

Recall that the support of a probability distribution P over $\mathbb{R}^N$ is the set

$$\operatorname{support}(P) = \left\{ x \in \mathbb{R}^N \ : \ \forall r > 0 \quad P(B(x, r)) > 0 \right\}$$

where $B(x, r)$ denotes the open ball of radius r centered at x. It can be shown that $P(\operatorname{support}(P)) = 1$; see Devroye et al. [13, Lemma A.1]. Thus a probability distribution over $\mathbb{R}^N$ with a finite support is such for which there exists a finite set F such that $P(F) = 1$.

Theorem 6.1 (Instability of k-means)**.** *Let $k \geq 2$ be an integer. Let $(Y, \mathfrak{P}, R, \Gamma)$ be the instantiation of the k-means optimization scheme over the Euclidean space $\mathbb{R}^N$ where $\mathfrak{P}$ is the family of all probability distributions with finite support of size at least k. Let d be the Hamming clustering distance. Let $P \in \mathfrak{P}$ be any instance. If*

$$\left| \operatorname*{argmin}_{y \in Y} R(P, y) \right| > k! \ , \qquad [1]$$

then the clustering algorithm $\Gamma \circ B$ induced by the k-means ERM algorithm B is unstable on P.

An example of a probability distribution satisfying the hypotheses of the theorem is shown in Figure 6.1.

The proof of the theorem is lengthy and technical. For this reason, we provide an explanatory outline in the next section. The outline provides glue for the technical lemmas that are contained in Section 6.2, as well as, it sets up the common notation.

[1] Recall that $\operatorname{argmin}_{y \in Y} = \{(c_1, c_2, \ldots, c_k) \ : \ R(P; c_1, c_2, \ldots, c_k) = \operatorname{opt}_R(P)\}$ and thus the size of this set counts the permutations as well.

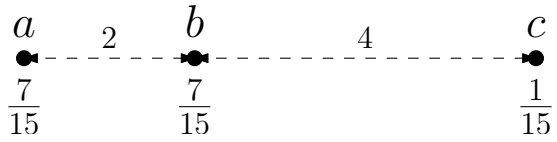

Figure 6.1: A distribution over the real line $\mathbb{R}$ with finite support. The support consists of three points a, b, c with probability masses and relative distances as shown. For $k = 2$ there are two k-means optimal solutions both with k-means cost $14/15$. In the first solution, a, b lie in one cluster and c lies in a separate cluster. In the second solution, b, c lie in one cluster and a lies in separate cluster.

6.1 Outline of the Proof

We view the clustering problem as a mapping from samples to partitions of the support. We denote by

$$\mathcal{C}_1, \mathcal{C}_2, \ldots \mathcal{C}_h$$

the "optimal" partitions of the support i.e. the partitions induced by the risk minimizers $\operatorname{argmin}_{y \in Y} R(P, y)$. Since there are only finitely many partitions of the support and because of the risk convergence, for large enough samples, with high probability, the partition output by the algorithm is among the optimal partitions. We show that the probability that the algorithm outputs any of the optimal partitions is bounded away from one. Or in other words, for sample sizes approaching infinity, at least two of the optimal partitions will be output by the algorithm with probability bounded away from zero. This implies instability, since having two (or more) different optimal partitions of the support, each with non-zero probability, implies a non-zero expectation of the Hamming distance between the outputs of the algorithm.

To analyze which sample leads to which partition of the support, we use elementary finite-dimensional calculus. A probability distribution with finite support can be viewed as a finite-dimensional weight vector μ which has one coordinate for each point of the support and the value of the coordinate is the probability of that point. Likewise, a sample can be viewed as a frequency vector w where value of each coordinate of w is the empirical frequency of the corresponding point of the support. Abusing notation somewhat, we consider the functions

$$R(w, \mathcal{C}_1), R(w, \mathcal{C}_2), \ldots, R(w, \mathcal{C}_h),$$

which assign risks to the optimal partitions and the frequency vector w. We view these functions as functions of w and we use Taylor expansion to analyze their behavior in the neighborhood of μ. We show that for any pair $\mathcal{C}_i, \mathcal{C}_j$, $i \neq j$, in an arbitrarily small neighborhood of μ there are vectors w for which $R(w, \mathcal{C}_i) > R(w, \mathcal{C}_j)$ and, vice versa, there are vectors w in the neighborhood for which the inequality is reversed. Via the multidimensional central limit theorem, this property of the risk functions translates to the property that the probability that the empirical risk of $\mathcal{C}_i$ is smaller than the empirical risk of $\mathcal{C}_j$ is bounded away from zero and, vice-versa, the probability that the empirical

risk of $\mathcal{C}_i$ is greater than the empirical risk of $\mathcal{C}_j$ is bounded away from zero. From that we derive that the probability that the algorithm outputs any particular optimal partition is bounded away from one.

In the above two paragraphs, we have essentially outlined the proof of the theorem. To explain further details, we start with the notation that will be used in the technical lemmas. We then re-iterate and we explain the outline one more time in much more detail, and we point on individual lemmas in which each of the pieces is proved.

First, let $F = \{x_1, x_2, \ldots, x_n\}$ be the support of P. In the rest of the proof we call F simply as *the support*. Note that $P(\{x_i\}) > 0$ for all $1 \leq i \leq n$[2] and let $\mu_i = P(\{x_i\})$ and let

$$\mu = (\mu_1, \mu_2, \ldots, \mu_n).$$

Note that $\mu_1 + \mu_2 + \cdots + \mu_n = 1$.

For a sample $S \in F^m$, we denote the number of occurrences of the point x_i in S by m_i, and use $w_i = m_i/m$ to denote the empirical frequency (*weight*) of the point x_i in the sample. The sample is completely determined by the vector of weights

$$w = (w_1, w_2, \ldots, w_n) \ .$$

Note that $w_1 + w_2 + \ldots w_n = 1$. In the all the proofs, we identify w, the sample S and its empirical distribution P_S. More generally, we associate w with any probability distribution over F. Having this in mind and abusing the notation a little, we define

$$R(w, y) = R(P_S, y) \ .$$

Note that $R(\mu, y) = R(P, y)$.

Consider the set $\operatorname{argmin}_{y \in Y} R(P, y)$ of risk minimizers. This set can be partitioned into $h \geq 2$ equivalence classes, each of size $k!$, of the form[3] $\{\pi(y^*) \ : \ \pi \in S_k\}$. Let $y_1^*, y_2^*, \ldots, y_h^* \in \mathbb{R}^{Nk}$ be representatives of the equivalence classes, one from each class. According to Lemma 6.2, each y_i^*, $1 \leq i \leq h$, has the property that each point of the support has a unique closest center in y_i^*. Let $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_h$ the partitions of the support induced by $y_1^*, y_2^*, \ldots, y_h^*$ respectively. Formally, each $\mathcal{C}_i$, $1 \leq i \leq h$, is the restriction of the clustering $\Gamma(y_i^*)$ to F. We call these partitions *optimal*. The mentioned property of $y_1^*, y_2^*, \ldots, y_h^*$ guarantees that $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_h$ are well defined. It is also easy to see that the optimality of y_i^*, $1 \leq i \leq h$ ensures that $\mathcal{C}_i^*$ consists of k non-empty sets.

We now associate risk with any partition $\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$ of the support into k non-empty sets and any weight vector $w \in \mathbb{R}_+^n$. Abusing a notation a bit further, we define

$$R(w, \mathcal{C}) = \sum_{i=1}^{k} \sum_{x_t \in C_i} w_t \left\| x_t - \frac{\sum\limits_{x_s \in C_i} w_s x_s}{\sum\limits_{x_s \in C_i} w_s} \right\|_2^2 \ . \tag{6.1}$$

[2]Note that N and n are two different parameters. The upper case N is the dimension of the domain. The lower case n is the size of the support.

[3]The symbol S_k and notation $\pi(y)$ were defined in Section 4.2.

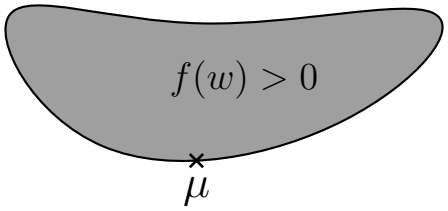

Figure 6.2: The gray colored area represents the set $Q = \{w \in \mathbb{R}^n_+ \mid R(w, \mathcal{C}) < R(w, \mathcal{D})\}$. The boundary of Q, colored in black, are the points w where $f(w) = 0$. The point μ is marked with a cross.

Note that we allow arguments w with the sum of the coordinates $w_1 + w_2 + \cdots + w_n$ not necessarily equal to one. This will greatly simplify our analysis. Also, note that R is homogeneous in w, that is, for any real number $\alpha > 0$, $R(\alpha w, \mathcal{C}) = \alpha R(w, \mathcal{C})$.

Proposition 6.3 explains the connection of (6.1) to the k-means risk. In particular, the proposition gives us an equivalent characterization of the k-means ERM algorithm: Assuming that the vector w of the input sample is an ϵ-neighborhood of μ, the k-means ERM algorithm can be viewed as picking an optimal partition $\mathcal{C}$ which minimizes $R(w, \mathcal{C})$.

In this view, consider a pair of distinct optimal partitions $\mathcal{C}$ and $\mathcal{D}$.[4] The k-means ERM algorithm prefers $\mathcal{C}$ over $\mathcal{D}$ when $R(w, \mathcal{C}) < R(w, \mathcal{D})$. We consider the set of weight vectors

$$Q = \{w \in \mathbb{R}^n_+ \mid R(w, \mathcal{C}) < R(w, \mathcal{D})\} .$$

Step 1: We analyze the set Q in a small neighborhood of μ. In Lemma 6.9, we show that Q contains an *open cone* T *with peak at* μ. The proof of the Lemma consists of several smaller steps.

(a) We first define the function $f : \mathbb{R}^n \to \mathbb{R}$, $f(w) = R(w, \mathcal{D}) - R(w, \mathcal{C})$. In this notation $Q = \{w \mid f(w) > 0\}$. Note the important fact that $f(\mu) = 0$. We analyze the behavior of f near μ. See Figure 6.2.

(b) In Lemma 6.4, we compute for any partition $\mathcal{C}$ the Taylor expansion of $R(w, \mathcal{C})$ at the point μ. This way we put our hands on the Taylor expansion of f.

(c) In Lemma 6.7 we show that the first non-zero term in the Taylor expansion of f attains both positive and negative values, and thus f itself attains both positive and negative values arbitrarily close to μ. (As it turns out, there are only two options. The first non-zero term is either the gradient or the Hessian.)

(d) We show that, since f is rational and hence analytic in the neighborhood of μ, it follows that Q contains a cone T whose peak is at μ. See Figure 6.3.

Step 2: Consider the hyperplane

$$H = \{w \in \mathbb{R}^n \mid w_1 + w_2 + \cdots + w_n = 1\}$$

[4] We use $\mathcal{C}$ and $\mathcal{D}$ instead of $\mathcal{C}_i$ and $\mathcal{C}_j$ so that we can use indices i, j for better purposes.

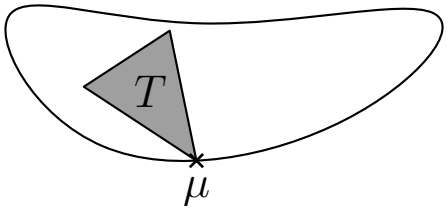

Figure 6.3: The gray colored area represents the cone T contained in the set Q. The peak of the cone is at μ.

in which the weights actually lie. In Lemma 6.10 we show that $Q \cap H$ contains an $(n-1)$-dimensional open cone Y.

Step 3: The distribution of the random vector w describing the sample is a multinomial distribution with m trials. From central limit theorem it follows that as the sample size m approaches infinity the probability distribution of w can be approximated by a multivariate Gaussian distribution lying in H. The Gaussian distribution concentrates near its mean value μ as the sample size increases. The shape of Q near μ determines the probability that the algorithm prefers partition $\mathcal{C}$ over $\mathcal{D}$. Formally, in Lemma 6.11 we show that $\lim_{m\to\infty} \Pr[w \in Y] > 0$; hence $\lim_{m\to\infty} \Pr[w \in Q] > 0$.

Step 4: For sufficiently large sample sizes the partition of F output by the algorithm is, with high probability, one of the optimal partitions. From the previous step it follows that with non-zero probability any optimal partition has lower empirical risk than any other optimal partition. Hence, there exist at least two optimal partitions of F, such that each of them is empirically optimal for a sample with non-zero probability. These two partitions cause instability of the algorithm. A precise argument is presented in Lemma 6.12.

6.2 The Technical Lemmas

Lemma 6.2 (No Ties)**.** *Let $w \in \mathbb{R}^n_+$ be any weight vector with sum of coordinates $w_1 + w_2 + \cdots + w_n = 1$. Suppose $y^* = (c^*_1, c^*_2, \ldots, c^*_k) \in \mathbb{R}^{Nk}$ is a minimizer of $R(w, \cdot)$. Then, for any point $x \in F$, the center of y^* closest to x is unique.*

Proof. Suppose by contradiction there exists a point $x \in F$ to which two or more different centers of y^* are the closest. Let us define two different partitions of the support, $\{C'_1, C'_2, \ldots, C'_k\}$ and $\{C''_1, C''_2, \ldots, C''_k\}$, in which for every i the clusters C'_i, C''_i consists of the points of the support closest to c^*_i, however the ties are broken differently for the point x.

Consider the solutions $y' = (c'_1, c'_2, \ldots, c'_k)$ and $y'' = (c''_1, c''_2, \ldots, c''_k)$ where the centers are defined as the means of the clusters of the partitions $\{C'_1, C'_2, \ldots, C'_k\}$ and $\{C''_1, C''_2, \ldots, C''_k\}$.

Formally, for $1 \leq i \leq k$, we define

$$c_i' = \frac{\sum_{x_t \in C_i'} w_t x_t}{\sum_{x_t \in C_i'} w_t} \qquad \text{and} \qquad c_i'' = \frac{\sum_{x_t \in C_i''} w_t x_t}{\sum_{x_t \in C_i''} w_t} .$$

The risks $R(w, y')$ and $R(w, y'')$ are no more than $R(w, y^*)$, since the centers y^* were replaced by the cluster means. Furthermore, since $y' \neq y''$, either $y^* \neq y'$ or $y^* \neq y''$ and thus either $R(w, y')$ or $R(w, y'')$ is strictly smaller than $R(w, y^*)$. This contradicts the optimality of y^*. ■

Proposition 6.3. *Let $\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$ be an optimal partition. For any $w \in \mathbb{R}_+^n$ define a k-tuple of centers $y(w) = (c_1(w), c_2(w), \ldots, c_k(w))$ as*

$$c_i(w) = \frac{\sum_{x_t \in C_i} w_t x_t}{\sum_{x_t \in C_i} w_t} .$$

There exists $\epsilon > 0$ such that if $\|w - \mu\| < \epsilon$ and $\sum_{i=1}^n w_i = 1$, then

$$R(w, \mathcal{C}) = R(w, y(w)) \quad [5]$$

and the restriction of $\Gamma(y(w))$ to the support equals $\mathcal{C}$.

Proof. From optimality of $\mathcal{C}$ it follows that there exist a risk minimizer $y^* = (c_1^*, c_2^*, \ldots, c_k^*)$ inducing $\mathcal{C}$. Optimality of y^* implies that $y^* = y(\mu)$. From Lemma 6.2 it follows that $y(\mu)$ has the property that each point of F has a unique closest center in $y(\mu)$. Then there exists $\Delta > 0$ such that for any point $x \in F$ the difference of between the distances to the closest and to the second closest center is at least Δ. By continuity of $y(w)$ there exists some $\epsilon > 0$ such that for any w in the ϵ-neighborhood of μ,

$$\|c_i(w) - c_i(\mu)\| < \Delta/2 \qquad \text{for all } i = 1, 2, \ldots, k.$$

Hence, for any w in the ϵ-neighborhood the restriction of $\Gamma(y(w))$ to the support is $\mathcal{C}$. The equality $R(w, \mathcal{C}) = R(w, y(w))$ follows by substituting the $y(w)$ into the definition of the k-means risk $R(P, y)$. ■

Lemma 6.4 (Derivatives of f)**.** *Let $\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$ be a partition of the support. The first two derivatives of the risk function $R(w, \mathcal{C})$ with respect to w at μ are as follows.*

1. *The p-th entry of the gradient is*

$$(\nabla R(\mu, \mathcal{C}))_p = \left.\frac{\partial R(w, \mathcal{C})}{\partial w_p}\right|_{w=\mu} = \|c_\ell - x_p\|_2^2 ,$$

assuming that x_p lies in the cluster C_ℓ.

[5] The right hand side is defined only if $\sum_{i=1}^n w_i = 1$.

2. *The* (p, q)*-th entry of the Hessian matrix*

$$\left(\nabla^2 R(\mu, \mathcal{C})\right)_{p,q} = \left.\frac{\partial^2 R(w, \mathcal{C})}{\partial w_p \partial w_q}\right|_{w=\mu}$$

equals to

$$-2\frac{(c_\ell - x_p)^T (c_\ell - x_q)}{\sum_{x_s \in C_\ell} \mu_s}$$

if x_p, x_q *lie in a common cluster* C_ℓ*, and is zero otherwise.*

Here, $c_1, c_2, \ldots, c_k$ *are the optimal centers*

$$c_i = \frac{\sum_{x_t \in C_i} \mu_t x_t}{\sum_{x_t \in C_i} \mu_t}\,.$$

Proof. For brevity, let us denote $\hat{c}_i := \hat{c}_i(w)$, $\hat{c}_i : \mathbb{R}^n \to \mathbb{R}^N$, the center of mass of the cluster C_i with respect to the empirical weights w. That is,

$$\hat{c}_i(w) = \frac{\sum_{x_s \in C_i} w_s x_s}{\sum_{x_s \in C_i} w_s}\,.$$

Plainly, $\hat{c}_i(\mu) = c_i$.

Suppose that x_p lies in cluster C_ℓ. The p-th component of the gradient $\nabla R(w, \mathcal{C})$ is

$$\begin{aligned}
\frac{\partial R(w, \mathcal{C})}{\partial w_p} &= \frac{\partial}{\partial w_p}\left(\sum_{i=1}^{k} \sum_{x_t \in C_i} w_t \left\|x_t - \hat{c}_i\right\|_2^2\right) \\
&= \left\|x_p - \hat{c}_\ell\right\|_2^2 + \sum_{x_t \in C_\ell} w_t \frac{\partial \left\|x_t - \hat{c}_\ell\right\|_2^2}{\partial w_p} \\
&= \left\|x_p - \hat{c}_\ell\right\|_2^2 - 2 \sum_{x_t \in C_\ell} w_t (x_t - \hat{c}_\ell)^T \frac{\partial \hat{c}_\ell}{\partial w_p} \\
&= \left\|x_p - \hat{c}_\ell\right\|_2^2 - 2 \underbrace{\left(\sum_{x_t \in C_\ell} w_t x_t - w_t \hat{c}_\ell\right)^T}_{=0} \frac{\partial \hat{c}_\ell}{\partial w_p} \\
&= \left\|x_p - \hat{c}_\ell\right\|_2^2
\end{aligned}$$

and at μ it is $(\nabla R(\mu, \mathcal{C}))_p = \|c_\ell - x_p\|_2^2$.

The (p, q)-th entry of the Hessian matrix is

$$\frac{\partial^2 R(w, \mathcal{C})}{\partial w_p \partial w_q} = \frac{\partial}{\partial w_q} \|x_p - \hat{c}_\ell\|_2^2 = 2(\hat{c}_\ell - x_p)^T \frac{\partial \hat{c}_\ell}{\partial w_q} . \tag{6.2}$$

If x_q does not lie in the cluster C_ℓ, then $\partial \hat{c}_\ell / \partial w_q = 0$ and hence also (p, q)-th entry of the Hessian matrix is zero. Otherwise

$$\begin{aligned}
\frac{\partial \hat{c}_\ell}{\partial w_q} &= \frac{\partial}{\partial w_q} \left(\frac{\sum\limits_{x_s \in C_\ell} w_s x_s}{\sum\limits_{x_s \in C_\ell} w_s} \right) \\
&= \frac{x_q \left(\sum\limits_{x_s \in C_\ell} w_s \right) - \left(\sum\limits_{x_s \in C_\ell} w_s x_s \right)}{\left(\sum\limits_{x_s \in C_\ell} w_s \right)^2} \\
&= \frac{x_q - \hat{c}_\ell}{\sum\limits_{x_s \in C_\ell} w_s} .
\end{aligned} \tag{6.3}$$

Substituting (6.3) into (6.2) we finally get (p, q)-th entry of the Hessian matrix

$$\frac{\partial^2 R(w, \mathcal{C})}{\partial w_p \partial w_q} = -2 \frac{(\hat{c}_\ell - x_p)^T (\hat{c}_\ell - x_q)}{\sum\limits_{x_s \in C_\ell} w_s} .$$

■

Lemma 6.5 (Weights of Clusters)**.** *For any subset E of domain, define its* weight *as $\mu(E) = \sum_{x_t \in E} \mu_t$. Let $\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$ and $\mathcal{D} = \{D_1, D_2, \ldots, D_k\}$ be two partitions of the support into k non-empty sets. For every point $x_t \in F$, consider the indices i, j such that $x_t \in C_i$ and $x_t \in D_j$, and define the weights $a_t = \mu(C_i)$ and $b_t = \mu(D_j)$. The following holds:*

- *Either, $a_t = b_t$ for all points $x_t \in F$.*
- *Or, there exist two points x_t, x_s such that $a_t > b_t$ and $a_s < b_s$.*

Proof. Consider the two sums

$$\sum_{t=1}^{n} \frac{\mu_t}{a_t} \qquad \text{and} \qquad \sum_{t=1}^{n} \frac{\mu_t}{b_t} .$$

It easy to see that the sums are equal,

$$\sum_{t=1}^{n} \frac{\mu_t}{a_t} = \sum_{i=1}^{k} \sum_{x_t \in C_i} \frac{\mu_t}{a_t} = k = \sum_{i=1}^{k} \sum_{x_t \in D_i} \frac{\mu_t}{b_t} = \sum_{t=1}^{n} \frac{\mu_t}{b_t} .$$

Either all the corresponding summands μ_t/a_t and μ_t/b_t in the two sums are equal and hence $a_t = b_t$ for all t. Or, there exist points x_t and x_s such that $\mu_t/a_t < \mu_t/b_t$ and $\mu_s/a_s > \mu_s/b_s$, and hence $a_t > b_t$ and $a_s < b_s$. ■

Lemma 6.6 (Hessian Determines Clustering)**.** *Let $\mathcal{C}, \mathcal{D}$ be optimal partitions of the support. If the Hessians of $R(w, \mathcal{C}), R(w, \mathcal{D})$ coincide at μ, then $\mathcal{C} = \mathcal{D}$.*

Proof. For the sake of brevity, let

$$H_{p,q} := \left(\nabla^2 R(\mu, \mathcal{C})\right)_{p,q} .$$

It suffices to show that the centers $c_1, c_2, \ldots, c_k$ of the partition $\mathcal{C}$ are uniquely determined by the matrix H. To this end, we view H as the adjacency matrix of a graph G with vertex set F, where nodes x_p, x_q are connected by an edge if and only if $H_{p,q} \neq 0$. Let $K_1, K_2, \ldots, K_\ell$ be the connected components of G. Note that there is an edge between x_p and x_q only if p and q belong to the same cluster in $\mathcal{C}$. Thus, the connected components of G represent a refinement of the partition $\mathcal{C}$.

For the sake of brevity, for any subset $E \subseteq F$, define the weight $\mu(E) = \sum_{x_t \in E} \mu_t$. Consider a fixed cluster C_j in $\mathcal{C}$ with center c_j. Recall that

$$c_j = \frac{1}{\mu(C_j)} \sum_{x_t \in C_j} \mu_t x_t \tag{6.4}$$

Let $K \subseteq C_j$ be any connected component of G that is contained in C_j and let $K' = C_j \setminus K$. We claim that

$$c_j = \frac{1}{\mu(K)} \sum_{x_t \in K} \mu_t x_t \ , \tag{6.5}$$

that is, c_j is determined by any component $K \subseteq C_j$. Since this is obvious for $K = C_j$, we assume that $K \subsetneq C_j$. We can rewrite (6.4) as

$$0 = \left(\sum_{x_t \in K} \mu_t (x_t - c_j) \right) + \left(\sum_{x_s \in K'} \mu_s (x_s - c_j) \right) \ . \tag{6.6}$$

Pick any pair s, t such that $x_t \in K$ and $x_s \in K'$. Since x_t and x_s are not neighbors in G, $H_{t,s} = 0$, which means that $x_t - c_j$ is orthogonal to $x_s - c_j$. Thus the vector represented by the first sum in (6.6) is orthogonal to the vector represented by the second sum. It follows that both sums yield zero, respectively. Rewriting this for the first sum, we obtain (6.5). ■

Lemma 6.7 (Indefinitness)**.** *Let $\mathcal{C}$ and $\mathcal{D}$ be any two optimal partitions. Let $f(w) = R(w, \mathcal{D}) - R(w, \mathcal{C})$. Consider the Taylor expansion of f around μ. Then, $\nabla f(\mu) \neq 0$ or the Hessian, $\nabla^2 f(\mu)$, is indefinite.*[6]

[6] A matrix is *indefinite* if it is neither positively semi-definite, nor negatively semi-definite.

Proof. We denote by $C_1, C_2, \ldots, C_k \subseteq F$ the clusters of $\mathcal{C}$ and by $D_1, D_2, \ldots, D_k \subseteq F$ the clusters of $\mathcal{D}$. We denote by $c_1, c_2, \ldots, c_k$ the optimal centers for $\mathcal{C}$, and by $d_1, d_2, \ldots, d_k$ the optimal centers for $\mathcal{D}$. That is, the center c_i is the center of mass of C_i, and d_j is the center of mass of D_j.

Consider the Taylor expansion of f at μ. Lemma 6.6 implies that the Hessian, $\nabla^2 f(\mu)$, is not zero. Assuming $\nabla f(\mu) = 0$ i.e. $\nabla R(\mu, \mathcal{C}) = \nabla R(\mu, \mathcal{D})$, we need to show that $\nabla^2 f(\mu)$ is indefinite.

For any point $x_p \in F$ we define three numbers e_p, a_p and b_p as follows. Suppose $x_p \in C_\ell$ and $x_p \in D_{\ell'}$. The first part of the Lemma 6.4 and $\nabla R(\mu, \mathcal{C}) = \nabla R(\mu, \mathcal{D})$ imply that the distance between x_p and c_ℓ equals to the distance between x_p and $d_{\ell'}$; denote this distance by e_p. Denote by a_p the weight of the cluster C_ℓ, that is, $a_p = \sum_{x_t \in C_\ell} \mu_t$. Likewise, let b_p be the weight of the cluster $D_{\ell'}$, that is, $b_p = \sum_{x_t \in D_{\ell'}} \mu_t$.

Consider the diagonal entries of Hessian matrix of f. Using the notation we had just introduced, by the second part of the Lemma 6.4 the (p,p)-th entry is

$$(\nabla^2 f(\mu))_{p,p} = \left(\frac{\partial^2 R(w, \mathcal{D})}{\partial w_p^2} - \frac{\partial^2 R(w, \mathcal{C})}{\partial w_p^2} \right) \Bigg|_{w=\mu} = 2e_p^2 \left(\frac{1}{a_p} - \frac{1}{b_p} \right) .$$

We claim that if $e_p = 0$, *then* $a_p = b_p$. Let $x_p \in C_\ell \cap D_{\ell'}$, and suppose without loss of generality that $a_p > b_p$. Since $e_p = 0$, we have $x_p = c_\ell = d_{\ell'}$. Since $a_p > b_p$ there is another point x_q that causes the decrease of the weight the cluster C_ℓ. Formally, $x_q \in C_\ell$, $x_q \notin D_{\ell'}$, but $x_q \in D_{\ell''}$. This means that in $\mathcal{D}$ the point x_q is closest to both $d_{\ell'}$ and $d_{\ell''}$. By Lemma 6.2, a tie can not happen in an optimal partition, which is a contradiction.

By Lemma 6.5, either (a) for all indices p, $a_p = b_p$, or (b) there are indices i, j such that $a_i > b_i$ and $a_j < b_j$. In the subcase (a), all the diagonal entries of Hessian matrix are zero. Since the Hessian matrix is non-zero, there must exist a non-zero entry off the diagonal making the matrix indefinite. In the subcase (b), the above claim implies that the indices i, j for which $a_i > b_i$ and $a_j < b_j$ are such that $e_i, e_j > 0$. Hence, the (i,i)-th diagonal entry of the Hessian matrix is negative, and the (j,j)-the diagonal entry of the Hessian matrix is positive. Therefore the Hessian matrix is indefinite. ■

Corollary 6.8. *There exists arbitrarily small* $\delta \in \mathbb{R}^n$ *such that* $f(\mu + \delta) > 0$. *(Similarly, there exists arbitrarily small* δ' *such that* $f(\mu + \delta') < 0$.*)*

Proof. Consider the Taylor expansion of f at μ and its lowest order term $P(x - \mu)$ that does not vanish (according to Lemma 6.7, either the gradient or the Hessian). Since P can take values of positive and of negative sign (obvious for the gradient, and obvious from Lemma 6.7 for the Hessian), we can pick a vector $x = \mu + \delta$ such that $P(x - \mu) = P(\delta) > 0$. Since P is homogeneous in δ, $P(\lambda\delta) > 0$ for every $\lambda > 0$. If λ is chosen sufficiently small, then $f(\mu + \lambda\delta)$ has the same sign as $P(\lambda\delta)$. The considerations for negative sign are symmetric. ■

Lemma 6.9 (Existence of a Positive Open Cone)**.** *There exist positive real numbers* ϵ *and* δ, *and a unit vector* $u \in \mathbb{R}^n$ *such that the open cone*

$$T = \left\{ w \in \mathbb{R}^n_+ \;\middle|\; 0 < \|w - \mu\|_2 < \epsilon,\ \frac{u^T(w - \mu)}{\|w - \mu\|_2} > 1 - \delta \right\}$$

is contained in Q*, the set of weights for which* $R(w, \mathcal{C}) < R(w, \mathcal{D})$.

Proof. This lemma is a refinement of Corollary 6.8 above. Let h be the order of the first non-zero term in the Taylor expansion of f around μ. Let this term be $P(u) = P(w - \mu)$, that is, P is a multivariate polynomial in $u = w - \mu$ and all its terms have the same degree h. Intuitively speaking, P determines the behavior of f in a small neighborhood of μ.

To prove existence of the cone, we proceed as follows. For a unit vector $u \in R^n$ and some $\epsilon > 0$ consider the line segment

$$S_{u,\epsilon} = \{\lambda u \,:\, \lambda \in (0, \epsilon]\} .$$

We say that P *super-dominates* on $S_{u,\epsilon}$ whenever for any $v \in S_{u,\epsilon}$ we have $|P(v)| > 0.9|f(\mu + v)|$. Similarly, we say that P *dominates* on $S_{u,\epsilon}$ if for any $v \in S_{u,\epsilon}$ we have $|P(v)| > 0.8|f(\mu + v)|$. Note that, in particular, if P dominates or super-dominates on $S_{u,\lambda}$, then for any $v \in S_{u,\epsilon}$ the signs of $P(v)$ and $f(\mu + v)$ are the same.

Clearly, there exists $\epsilon > 0$ and a unit vector u such that P super-dominates on $S_{u,\epsilon}$. Since P attains both positive and negative values, there exists a point $x \in \mathbb{R}^n$ such that $P(x) > 0$. All terms of P have the same degree h and hence $P(x/\|x\|_2) > 0$. (Note that $x \neq 0$ since $P(0) = 0$.) Thus we can choose $u = x/\|x\|_2$. After u is chosen, we can choose $\epsilon > 0$ small enough so that P super-dominates on $S_{u,\epsilon}$.

Now think of replacing u with another unit vector u'. If u and u' are "close", then P still "at least" dominates on $S_{u',\eta}$. This follows by from the fact that f is rational and hence analytic. We measure closeness of u and u' by their dot product. Thus, there exists some small enough $\delta > 0$ such that for any unit vector $u' \in R^n$, if $u^T u' > 1 - \delta$, then P dominates on $S_{u',\epsilon}$.

The choices of u, ϵ and δ determine the cone T in the statement of the lemma. The fact that $R(w, \mathcal{C}) < R(w, \mathcal{D})$ for any $w \in T$ directly follows from that P dominates on $S_{u',\epsilon}$ for any unit vector u' such that $u^T u' > 1 - \delta$. Simply substitute $u' = (w - \mu)/\|w - \mu\|_2$ and $\lambda = \|w - \mu\|$. ■

Lemma 6.10 (Existence of a Positive Open Cone II)**.** *There exists positive real numbers* ϵ, δ *and a unit vector* $u \in \mathbb{R}^n$ *with sum of coordinates,* $u_1 + u_2 + \cdots + u_n$*, equal to zero, such that the* $(n-1)$*-dimensional open cone*

$$Y = \left\{ w \in H \cap \mathbb{R}^n_+ \;\middle|\; 0 < \|w - \mu\|_2 < \epsilon,\ \frac{u^T(w - \mu)}{\|w - \mu\|_2} > 1 - \delta \right\}$$

is contained in $Q \cap H$.

Proof. We use the projection $\phi : \mathbb{R}^n_+ \to H$, $\phi(w) = w/(w_1 + w_2 + \cdots + w_n)$. Note that for the k-means cost function, for every partition $\mathcal{C}$ and every positive constant λ, $R(\lambda w, \mathcal{C}) = \lambda R(w, \mathcal{C})$. It follows that the projection ϕ does not affect the sign of f. That is, $\text{sign}(f(w)) = \text{sign}(f(\phi(w)))$. Therefore $Q \cap H = \phi(Q) \subset Q$. The projection $\phi(T)$ clearly contains an $(n-1)$-dimensional open cone Y of the form as stated in the Lemma. More precisely, there exists positive numbers ϵ, δ and unit vector u (the direction of the axis of the cone), such that the cone

$$Y := Y_{\epsilon,\delta,u} = \left\{ w \in H \cap \mathbb{R}^n_+ \;\middle|\; 0 < \|w - \mu\|_2 < \epsilon,\ \frac{u^T(w - \mu)}{\|w - \mu\|_2} > 1 - \delta \right\}$$

is contained in $\phi(T)$. Since the cone Y lies in H, the direction of the axis, u, can be picked in such way that the sum of its coordinates $u_1 + u_2 + \cdots + u_n$ is zero. Since $T \subseteq Q$, we get $Y \subset \phi(T) \subset \phi(Q) = Q \cap H$. ■

Lemma 6.11 (Instability)**.** *Let $\mathcal{C}$ and $\mathcal{D}$ be distinct optimal partitions of the support. Let Q be the set of weights where the k-means ERM algorithm prefers $\mathcal{C}$ over $\mathcal{D}$. If w are the weights for an i.i.d. sample of size m, then*

$$\lim_{m\to\infty} \Pr[w \in Q] > 0 \,.$$

Proof. Let $Y \subset (Q \cap H)$ be an $(n-1)$-dimensional open cone (as implied by lemma 6.10) lying in the hyperplane H defined by the equation $w_1 + w_2 + \cdots + w_n = 1$. We show that,

$$\lim_{m\to\infty} \Pr[w \in Y] > 0 \,,$$

which implies the claim.

We have

$$\begin{aligned}\Pr[w \in Y] &= \Pr\left[\frac{u^T(w-\mu)}{\|w-\mu\|_2} > 1-\delta,\ 0 < \|w-\mu\|_2 < \epsilon\right] \\ &= \Pr\left[\frac{u^T(\sqrt{m}(w-\mu))}{\sqrt{m}\|w-\mu\|_2} > 1-\delta,\ 0 < \sqrt{m}\|w-\mu\|_2 < \epsilon\sqrt{m}\right] \,.\end{aligned}$$

By the central limit theorem $\sqrt{m}(w-\mu)$ weakly converges to a normally distributed random variable $Z \sim N(0,\Sigma)$, where Σ is the covariance matrix.[7] In particular this means that there is a sequence $\{\zeta_m\}_{m=1}^{\infty}$, $\zeta_m \to 0$, such that

$$\left|\Pr\left[\frac{u^T(\sqrt{m}(w-\mu))}{\sqrt{m}\|w-\mu\|_2} > 1-\delta,\ 0 < \sqrt{m}\|w-\mu\|_2 < \epsilon\sqrt{m}\right] - \Pr\left[\frac{u^T Z}{\|Z\|_2} > 1-\delta,\ 0 < \|Z\|_2 < \epsilon\sqrt{m}\right]\right| < \zeta_m$$

Consequently, we can bound the probability $\Pr[w \in Y]$ as

$$\begin{aligned}\Pr[w \in Y] &\geq \Pr\left[\frac{u^T Z}{\|Z\|_2} > 1-\delta,\ 0 < \|Z\|_2 < \epsilon\sqrt{m}\right] - \zeta_m \\ &\geq 1 - \Pr\left[\frac{u^T Z}{\|Z\|_2} < 1-\delta\right] - \Pr\left[\|Z\|_2 \geq \epsilon\sqrt{m}\right] - \Pr[\|Z\|_2 = 0] - \zeta_m \,.\end{aligned}$$

Take the limit $m \to \infty$. The last three terms in the last expression vanish. Since u has sum of its coordinates zero and $Z \sim N(0,\Sigma)$ is normally distributed, the term $\lim_{m\to\infty} \Pr\left[\frac{u^T Z}{\|Z\|_2} < 1-\delta\right]$ lies in $(0,1)$. ■

[7] $\Sigma = \operatorname{diag}(\mu_1, \mu_2, \ldots, \mu_n) - \mu\mu^T$, the rank of Σ is $n-1$, and its rows (or columns) span the $(n-1)$-dimensional vector space $\{u \in \mathbb{R}^n \mid u_1 + u_2 + \cdots + u_n = 0\}$.

Lemma 6.12 (Multiple Optimal Partitions)**.** *If there are* $h \geq 2$ *optimal partitions of the support, then the clustering algorithm* $\Gamma \circ B$ *induced by the* k*-means ERM algorithm* B *is unstable on* P*.*

Proof. Let $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_h$, $h \geq 2$, be the optimal partitions. Let

$$\pi_i = \lim_{m \to \infty} \Pr\left[\Gamma(B(S))\big|_F = \mathcal{C}_i\right] \ ,$$

where by $\Gamma(B(S))|_F$ we mean the clustering output by the algorithm restricted to F.

Claim: Each number π_i is strictly less than one.

Proof of the claim:

$$\begin{aligned}
\Pr_{S \sim P^m}\left[\Gamma(B(S))\big|_F = \mathcal{C}_i\right] &\leq \Pr\left[R(w, \mathcal{C}_i) \leq \min_{\substack{\ell=1,2,\ldots,h \\ \ell \neq i}} R(w, \mathcal{C}_\ell)\right] \\
&\leq \Pr\left[R(w, \mathcal{C}_i) \leq R(w, \mathcal{C}_j)\right] \\
&= 1 - \Pr\left[R(w, \mathcal{C}_i) > R(w, \mathcal{C}_j)\right]
\end{aligned}$$

Taking limit $m \to \infty$ on both sides of the inequality and applying Lemma 6.11,

$$\lim_{m \to \infty} \Pr[R(w, \mathcal{C}_i) > R(w, \mathcal{C}_j)] > 0$$

the claim follows.

Since k-means ERM algorithm is risk converging, as the sample size increases, with probability approaching one, $\Gamma(B(S))|_F$ is an optimal partition, and hence

$$\pi_1 + \pi_2 + \cdots + \pi_h = 1 \ .$$

Necessarily at least two numbers, say, π_i, π_j are strictly positive. That is, the algorithm outputs two different partitions $\mathcal{C}_i, \mathcal{C}_j$ with non-zero probability for arbitrarily large sample size. The algorithm will be switching between these two partitions. Formally,

$$\text{instab}(\Gamma \circ B, P) \geq d_P(\mathcal{C}_i, \mathcal{C}_j)\pi_i\pi_j$$

and the right hand side is is strictly positive. ∎

Chapter 7

Conclusion

We have theoretically investigated the asymptotic stability of clustering algorithms. However, many questions are left unanswered. In this chapter we discuss consequences of our theorems, their relationship to practice and discuss some of the main open questions.

The stability method was proposed by Ben-Hur et al. [9] and Lange et al. [20, 21]. Their idea was that a clustering is good and meaningful if and only if it is stable. Based on this postulate they design a model selection method. Roughly speaking, the method tries several parameters of the clustering algorithm and for each of them numerically calculates the stability score. The score is calculated based on several repetitions of randomly drawn sample pairs (S_1, S_2). We emphasize that all the samples have some fixed finite size m.

Ben-Hur et al. and Lange et al. do not conduct any theoretical investigations of the proposed method. Instead, they conduct numerical experiments on both artificial and real life data sets. The conclusions from these experiments seem encouraging and according to them it seems that the stability method is able to select a sensible model for the number of clusters.

In the next section we describe a simplified method from Ben-Hur et al. [9]. We examine the method on three simple one-dimensional probability distributions. We will use the k-means and k-median ERM clustering algorithms. We apply our theoretical insight from this thesis to argue about asymptotic (in)stability of the clustering algorithms on these distributions. We will conclude that at least from the asymptotic perspective the stability method is seriously flawed.

7.1 Examples

A simplified and adapted version of the stability method used in the numerical experiments conducted by Ben-Hur et al. [9] can be summarized as follows. For each $k = 2, 3, \ldots, k_{max}$ run the clustering algorithm with parameter that controls the desired number of clusters set to k. For each k, draw independently 100 pairs of independent samples (S_1, S_2). For each sample pair (S_1, S_2), cluster S_1, cluster S_2 and compute the clustering distance between the two samples. From the 100 pairwise clustering distances construct a histogram. In our formalism, the histogram represents the distribution of the random variable $d_P(A_k(S_1), A_k(S_2))$, where A_k is the with parameter that controls the de-

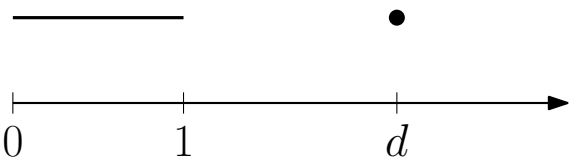

Figure 7.1: The picture shows a mixture of the uniform distribution over $[0, 1]$ and the distribution which concentrates all mass on the point d.

sired number of clusters set to k. The histogram is assessed as "stable" if and only if the probability of $d_P(A_k(S_1), A_k(S_2))$ that is below certain value is large enough.[1] Select, as the "correct" number of clusters, the largest k for which the histogram is assessed as stable.

We demonstrate several natural examples of probability distributions over the real line and we look at the stability of k-means and k-medians ERM algorithms.[2] On these examples, we try to convince the reader that the assumption that stable clustering is not equivalent to a correct one, and that in particular the correct number of clusters can not be detected this way. We emphasize that we do not mathematically formalize what we mean by the correct number of clusters. Instead, for each example, we simply state what the correct number is. We leave the reader to judge whether these number makes sense to him or her. A possible definition, which agrees with the numbers stated by us, is that the correct number of clusters is the number of connected components of the support of the probability distribution.

Our first example is the uniform distribution over the unit interval $[0, 1]$. The correct number of clusters is clearly one. It is not hard to figure out that, for any number of clusters, k, both k-means and k-medians have, up to permutation of the centers, exactly one risk minimizer

$$y^* = (c_1^*, c_2^*, \ldots, c_k^*) = \left(\frac{1}{2k}, \frac{3}{2k}, \frac{5}{2k}, \ldots, \frac{2k-1}{2k}\right)$$

It follows that both the clustering algorithms induced by the k-means and k-medians ERM algorithms are stable on this distribution for any value of k. Therefore, the stability method of Ben-Hur et al. outputs k_{max} for any large enough sample size m and it certainly does not output 1 as the answer.

Generalization of the previous example is the distribution over the union of two intervals $[0, \pi] \cup [10, 11]$. Clearly the correct number of clusters is 2. Yet for any k both the k-means and k-medians algorithms have unique optimum and hence are stable.[3] Therefore, as before, the stability method outputs k_{max} for any large enough sample size m. This example can be further generalized to a distribution over a union of n disjoint intervals in such a way that for any k both k-means and k-medians algorithms are stable.

[1] Ben-Hur et al. choose some ad-hoc unspecified threshold.

[2] We use the standard Euclidean metric on $\mathbb{R}$.

[3] The uniqueness of the optimum follows from the irrationality of the ratio of the lengths of the two intervals.

Figure 7.2: Optimal k-means solutions for $k = 2$ and $k = 3$ are shown. For $k = 2$ there are two optimal solutions. For $k = 3$ the solution is unique.

Our last and perhaps a more interesting is the example of a distribution P over $[0, 1] \cup \{d\}$ where $d \geq 3/2$. The distribution is mixture of two distributions: the uniform distribution over $[0, 1]$ and the distribution which has all mass concentrated on the point d. See Figure 7.1. Clearly, this distribution has two clusters. By appropriately choosing the mixing weights and the location of the point d we can arrange that for $k = 2$ the k-means cost function has two optimal solutions and for $k = 3$ it has unique optimal solution. Hence, for $k = 2$ the k-means ERM algorithm is unstable and for $k = 3$ it is stable. One such choice is $d = 11/6$ and mixing weights $15/16$ and $1/16$. Therefore, on this distribution the stability method never identifies the correct model $k = 2$.

7.2 Mismatch between Theory and Practice

The examples from preceding section are disturbing. They show that for sufficiently large sample size m the stability method fails to recover the correct model.

We can make a more general statement. We can argue that a data set P encountered in practice for all $k \in \{2, 3, \ldots, k_{max}\}$ simultaneously both the k-means and k-medians cost functions have unique minima. Therefore, there exists sample size m such that for all $k = 2, 3, \ldots, k_{max}$ the instability of k-means and k-medians ERM algorithms is below any pre-specified threshold value, say, 0.001. The argument that the optimal solution is unique in practice simply follows from the fact that any small perturbation of a data set P with a unique optimal solution makes it stable.

This general argument is even more disturbing than the examples from the preceding section. Essentially, it says that for any data set the k-means and k-medians clustering algorithms are stable for all k provided that the sample size m is large enough. Thus for sample sizes big enough, the stability method seems to become vacuous. At best, such behavior of clustering stability is unintuitive and undesired, as one would expect that the larger the sample size, the "better" the stability method works and the more "refined picture" of the structure of the data set the stability method reveals. However, as we see, quite the opposite is true.

A natural question to ask is why Ben-Hur et al. [9] and Lange et al. [20, 21] successfully used the method to recover the correct number of clusters. One possible reason is that the algorithm uses that Ben-Hur et al. use hierarchical clustering algorithm and in our thesis we consider cost based algorithms. Another reason seems to be that the sample size they used was "small enough". Currently, however, there is no theoretical answer to what is a "suitably small" sample size.

Shamir and Tishby in their recent work [26] give a partial answer to this problem.

Their answer is, however, an impossibility result. They show that no matter how big sample size we consider, the non-asymptotic instability can be "large", despite the fact that asymptotic instability is 0. More precisely, they show that for any sample size m there exists a probability distribution P such that $\text{instab}(A, P, m) \geq 1/12$ and at the same time $\text{instab}(A, P) = 0$. This might suggest that the asymptotic instability might not be a good measure, and that success of the stability method reported in the numerical experiments crucially depends on the fact that the sample size is finite and small.

It's not clear however which useful property of P is revealed by the numerical value of $\text{instab}(A, P, m)$ for some fixed m. Shamir and Tishby [26] and Ben-David and von Luxburg [7] are trying to relate clustering stability of k-means ERM algorithm and the probability mass that P has near boundaries of clusters of the optimal solution.[4] But both papers miss the point of coming up with some prior assumption on the data set, so that one would have an explicit formula for the sample size m which would guarantee the stability method to work.[5]

7.3 Rates of Convergence and Cluster Boundaries

Shamir and Tishby [26] characterize the rate at which instability of k-means over $\mathbb{R}^d$ approaches zero as a function of the sample size, provided that optimal solution is unique. They assume that the probability distribution has continuous density with respect to the Lebesgue measure. They show that the instability is asymptotically $C/\sqrt{m}$ where C depends only on distribution of the probability density at the cluster boundaries of the optimal solution. The provide an explicit, but very complicated formula for C. Generally speaking, however, the lower density on the boundaries, the smaller the multiplicative constant C.

As a corollary Shamir and Tishby get that for large enough sample size m, for two different choices of the number of clusters, k_1 and k_2, the multiplicative constants C_1 and C_2 can be distinguished based on two pairs of samples (S_1, S_2) and (S_1', S_2') each sample of size m.

This corollary can be viewed as an argument in favor of the stability method as it might explain why the stability method works in practice. Namely, one might postulate that the multiplicative factor C is some inherent quantity that characterizes how well the model fits the data set. And so the stability method might be a way of selecting the model with the smallest multiplicative factor C.

However, we must be very cautious here. The corollary is only an asymptotic statement. In other words, no upper bound on the sample size m is known to guarantee that if, say, $C_1 > C_2$ then $\text{instab}(A_{k_1}, P, m) > \text{instab}(A_{k_2}, P, m)$. It is not clear whether such bound for m exists or a impossibility result, similar to the one in the preceding section, holds.

Ben-David and von Luxburg [7] also study a very similar setup of stability of k-means ERM algorithm on distributions over $\mathbb{R}^d$ under the assumption that the optimal solution is (up to the permutation of the centers) unique. However, they do not assume that

[4] Here, we mean the boundaries of the Voronoi cells.

[5] Ideally, the prior assumption should ensure that the sample size, say, $m = 42$ makes the method work.

the distribution has a density function. In they work they consider the probability mass within a certain distance from the cluster boundaries (boundaries of the Voronoi cells) in the optimal solution. They argue that if $\text{instab}(A, P, m)$ is "large", then the probability mass near cluster boundaries is "large" as well. They provide a counter-example for the converse of the statement, that is, an example where the mass near cluster is large and $\text{instab}(A, P, m)$ quickly converges to zero. Their paper, however, is less quantitative than one would hope for, since the bounds are expressed in terms of quantities for which it is not obvious how to compute them explicitly. For example, the main theorem assumes that for some unknown sample size m with probability at least $1 - \delta$ the boundaries of the Voronoi cells of the empirically optimal solution will not differ from the boundaries of the true optimal solution by more than γ. However, they do not give any explicit nor asymptotic formulas which would relate γ, δ and m. It would be nice if at least the asymptotic rate of $\text{instab}(A, P, m)$, as a function of m, could be explicitly expressed in terms of how the probability distribution looks in the vicinity of the boundaries of the Voronoi cells of the optimal solution, in similar spirit as Shamir and Tishby did for absolutely continuous probability distributions.

7.4 Technical Questions

There remain a couple of open technical questions, which are more directly connected to the work in this thesis. A very concrete problem is to prove that the k-means ERM algorithm is stable on a probability distribution over (a bounded subset of) $\mathbb{R}^N$ if and only if the cost function has one optimal solution. In other words, generalize Theorem 6.1 and Theorem 4.2 to arbitrary distributions. The main technical problem seems to be to prove that if multiple optima exists, the algorithm is unstable. For distributions with infinite support the trick with multinomial distributions no longer works and some more sophisticated technique needs to be applied. Another obstacle might be the issue with infinitely many optimal solutions.

A more general and certainly more challenging question is to find conditions on the loss function $L : X \times Y \to \mathbb{R}$ which would ensure that the ERM algorithm minimizing $R(P_S, y) = \frac{1}{m}\sum_{i=1}^{m} L(x_i, y)$ is stable on P if only if $R(P, Y) = \frac{1}{m}\mathbf{E}_{x\sim P}[L(x, y)]$ has unique optimal solution. In particular, is this true for the k-medians loss function?

Part II

Comparison of Supervised and Semi-Supervised Learning

Chapter 8

Introduction

Classification is one of the most studied machine learning and statistical problems. See the books [13, 29, 1]. In a classification task, we are given a set of unlabeled examples, say, email messages. The examples are manually classified into categories, say, spam or non-spam, and receive corresponding labels. The goal is to automatically construct (learn) from the labeled examples a classifier that can be used to predict labels of future unlabeled examples.

The above approach to classification when the classifier is learned exclusively from labeled examples is called supervised learning, and from a theoretical perspective, it is reasonably well understood. The disadvantage of supervised learning is the need to manually label large quantities of examples, which can be costly and/or laborious. A natural approach to overcome this drawback is to learn from both labeled and unlabeled examples. This approach is called *semi-supervised* learning, and it is theoretically lesser understood than supervised learning; see the recent book edited by Chapelle et al. [12].

A natural question is what is the advantage (or disadvantage) of semi-supervised learning. Sidestepping computational issues, the main hope is the fewer examples need to be labeled manually if one has access to large quantities of unlabeled examples. In this part of the thesis, we propose a simple theoretical framework in which this question can be addressed. The idea of this papt of the thesis comes from a conference paper which I co-authored with Shai Ben-David and Tyler Lu [5]. Tyler Lu's master's thesis [22] is also based on that paper.

Our framework is a utopian extension of the probably approximately correct (PAC) model where the labeled examples are assumed to be generated independently from a fixed probability distribution. In the same manner as in the PAC model, supervised learners are functions that receive a finite sequence of labeled examples, and produce a classifier. Semi-supervised learners, in addition to labeled examples, have access to infinite amount of unlabeled data. And so, we model semi-supervised learners as functions that receive a finite sequence of labeled examples and the distribution of the unlabeled data, and produce a classifier.

Within our framework, we analyze a simple learning problem of learning a threshold (a point on the real line). We place no restrictions or assumptions on the relationship between unlabeled data and the labels. On this problem we demonstrate that access to unlabeled data *at best* roughly halves the number of labeled examples needed. We finish

this part of the thesis by presenting a general definition of a measure which quantifies the relative advantage of having the access to unlabeled data. We call the measure the *semi-supervised learning ratio*.

Chapter 9

Definitions and Notation

In this chapter, we introduce a learning model which will allow us to compare the relative advantage of unlabeled data. Our model is a simple extension of Valiant's probably approximately correct (PAC) model [28]; see also the book by Kearns and Vazirani [19]. Similarly, as in the PAC model the learner receives data an i.i.d. labeled sample $(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)$ generated at random from a probability distribution P over $X \times Y$ where X is some domain and $Y = \{+1, -1\}$ is the set of labels. We consider two types of learners: (a) supervised learners which receive only the sample and (b) semi-supervised learners (SSL) which, in addition to the sample, also receive the marginal distribution P_X i.e. the distribution P "without" the labels. In both cases, the learners output a hypothesis which is a function $f : X \to \{+1, -1\}$. For such a hypothesis f, we are interested in its misclassification probability, that is, the probability that on a random sample (x, y) drawn from P the values $f(x)$ and y disagree. This probability is called error and the learner's task is to output f with as small an error as possible.

The error of the learner (supervised or semi-supervised) is compared with the error of the best function from a class H. The difference between the error of the function output by the learner and the minimum error of a function from the class H is called the excess error. Quite non-standardly, we measure the performance of a learner by the *expected* excess error[1] which we call the learning rate. In the literature, for online learning tasks, the term regret is used; see for example the book [11].

We consider two standard variants of the model. The first variant is the original Valiant's version in which P is such that there exists a function in H with zero error. The second variant is the agnostic version attributed to Haussler [17] where any probability distribution P is allowed.

The fundamental comparison that we want to make is the relative advantage of a semi-supervised learner over a supervised learner. In contrast to the original Valiant's model, we ignore any computational issues such as how the input sample or the output hypotheses are represented and whether the learner runs in polynomial time or not.

We start gently with the basic definitions. However, we postpone our main definition, the definition of semi-supervised learning ratio, to the next chapter.

[1] The common practice is to study the sample size m as function of the probability δ that the excess error is greater than ϵ. We advocate our approach because of its simplicity and more understandable presentation of the results.

Definition 9.1 (Domain)**.** *A* domain *is a measurable space. That is, it is a pair* $(X, \mathfrak{M})$ *where* X *a non-empty set and* $\mathfrak{M}$ *is a* σ*-algebra of subsets of* X*.*

As in the first part of the thesis, we will simply talk about the domain X as long as the σ-algebra is clear from context.

Definition 9.2 (Hypothesis)**.** *Let* $(X, \mathfrak{M})$ *be a domain. A* hypothesis *(over* X*) is a measurable function* $f : X \to \{+1, -1\}$*. We denote by* $\mathfrak{H}$ *the set of all hypotheses i.e. the set of all measurable functions from* X *to* $\{+1, -1\}$

Whenever we write $\mathfrak{H}$, the domain to which $\mathfrak{H}$ refers, will be clear from the context. We use this convention to avoid awkward notations such as $\mathfrak{H}_X$ or $\mathfrak{H}(X)$.

Definition 9.3 (Hypothesis Class)**.** *Let* $(X, \mathfrak{M})$ *be a domain. A* hypothesis class *is a non-empty subset of* $\mathfrak{H}$*.*

Definition 9.4 (Error)**.** *Let* $(X, \mathfrak{M})$ *be a domain and let* P *be a probability distribution over* $X \times \{+1, -1\}$*.*[2] *Let* $h : X \to \{+1, -1\}$ *be a hypothesis. We define the* error *of* h *on* P *as*

$$\mathrm{Err}^P(h) = \Pr_{(x,y)\sim P}[h(x) \neq y] \ .$$

Definition 9.5 (Example and Sample)**.** *Let* X *be a domain. An element of* $X \times \{+1, -1\}$ *is called an* example *or a* labeled example*. The second component of an example is called the* label*. If the label of an example is* $+1$*, we call the example* positive*; otherwise we call the example* negative*. A* sample *or a* labeled sample *is a finite sequence of examples. The* size *of a sample is its length. Formally, a sample is an element of*

$$\bigcup_{m=1}^{\infty} (X \times \{+1, -1\})^m \ .$$

Definition 9.6 (Supervised Learner)**.** *Let* $(X, \mathfrak{M})$ *be a domain. A* supervised learner *is a mapping*

$$A : \bigcup_{m=1}^{\infty} (X \times \{+1, -1\})^m \to \mathfrak{H} \ .$$

Definition 9.7 (Semi-Supervised Learner)**.** *Let* $\mathfrak{D}$ *be a family of probability distributions over a domain* $(X, \mathfrak{M})$*. A* semi-supervised learner *is a mapping*

$$A : \mathfrak{D} \times \left(\bigcup_{m=1}^{\infty} (X \times \{+1, -1\})^m \right) \to \mathfrak{H} \ .$$

An element of $\mathfrak{D}$ *is called an* unlabeled distribution*.*

A learner, without any further qualification, is either a supervised learner or a semi-supervised learner.

[2] P is defined on the σ-algebra generated by the class of sets $\{M \times \{+1\} : M \in \mathfrak{M}\} \cup \{M \times \{-1\} : M \in \mathfrak{M}\}$.

Definition 9.8 (Excess Error)**.** *Let* X *be a domain,* H *be a hypothesis class over* X*, and let* P *be a probability distribution over* $X \times \{+1, -1\}$*. The* excess error *of a hypothesis* $h \in \mathfrak{H}$ *is*

$$\mathrm{Err}^P(h) - \inf_{h \in H} \mathrm{Err}^P(h) \ .$$

Note that, in principle, the excess error of h can be negative.

Definition 9.9 (Unlabeled distribution)**.** *Let* $(X, \mathfrak{M})$ *be a domain and let* P *be a probability distribution over* $X \times \{+1, -1\}$*. The* unlabeled distribution of P *is the marginal probability distribution over* X*. We denote the unlabeled distribution by* P_X*. Formally, for any measurable set* $M \subseteq X$*, it is defined as*

$$P_X(M) = P(M \times \{+1, -1\}) \ .$$

Definition 9.10 (Learning Rate)**.** *Suppose* X *is a domain,* H *is a hypothesis class over* X*,* P *is a probability distribution over* $X \times \{+1, -1\}$*, and* A *is learner and* m *is a positive integer. The* learning rate *is defined as*

$$\mathfrak{L}(A, P, H, m) = \mathop{\mathbf{E}}_{S \sim P^m} \left[\mathrm{Err}^P(A(S)) - \inf_{h \in H} \mathrm{Err}^P(h) \right] \ ,$$

and for semi-supervised learner as

$$\mathfrak{L}(A, P, H, m) = \mathop{\mathbf{E}}_{S \sim P^m} \left[\mathrm{Err}^P(A(P_X, S)) - \inf_{h \in H} \mathrm{Err}^P(h) \right] \ .$$

More intuitively, the learning rate of a learner is the function that maps the sample size m to the expected excess error of a hypothesis produced by the learner.

Let us explain the difference between Valiant's and Haussler's model. The former is sometimes called the realizable case and the second the agnostic or unrealizable case. Technically, a probability distribution P over $X \times \{+1, -1\}$ is *realizable* by H if there exists a hypothesis $h \in H$ such that $\mathrm{Err}^P(h) = 0$. Such hypothesis h is called a *target*. If no target exists in H, we say that P is *unrealizable* by H and we talk about the unrealizable or *agnostic* case.

9.1 A Folklore Example

We present a folklore example of a learner for the class of thresholds on the real line in the realizable case. The example will be useful in the next chapter. Consider the domain $X = \mathbb{R}$. The hypothesis class H of thresholds contains for each real number t a hypothesis h_t which is defined as

$$h_t(x) = \begin{cases} -1 & \text{if } x \leq t, \\ +1 & \text{if } x > t. \end{cases}$$

We define a supervised learner A for probability distributions realizable by this class. The learner A, for a given labeled sample $S = ((x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m))$, computes the position of the rightmost negative example

$$\ell = \max\{x_i \ : \ 1 \leq i \leq m, \ y_i = -1\}$$

and outputs h_ℓ. Here, we use the convention that the maximum of the empty set is $-\infty$ and we define $h_{-\infty}(x) = -1$ for every $x \in \mathbb{R}$.

What is the learning rate of A on distributions realizable by H? Let P be any probability distribution over $\mathbb{R} \times \{+1, -1\}$ realizable by a target $h_{t^*} \in H$. We have

$$\begin{aligned}\mathfrak{L}(A, P, H, m) &= \mathop{\mathbb{E}}_{S \sim P^m}\left[\mathrm{Err}^P(A(S))\right] \\ &= \int_0^1 \Pr_{S \sim P^m}\left[\mathrm{Err}^P(A(S)) \geq x\right] dx\end{aligned}$$

To bound $\Pr_{S \sim P^m}\left[\mathrm{Err}^P(A(S)) \geq x\right]$, let

$$t' = \sup\{t \in \mathbb{R} \,:\, P_X([t, t*]) \geq x\}\,.$$

The event $\mathrm{Err}^P(A(S)) \geq x$ occurs precisely when no (negative) example of S falls in the interval $[t', t^*]$. The probability of that event is

$$(1 - P([t', t^*]))^m \leq (1 - x)^m\,.$$

Therefore,

$$\mathfrak{L}(A, P, H, m) \leq \int_0^1 (1 - x)^m\, dx = \frac{1}{m+1}\,.$$

Note that this result holds for any probability distribution P over $X \times \{+1, -1\}$ realizable by H. Note that when P_X is absolutely continuous, the learning rate is exactly $1/(m+1)$.

Chapter 10

The Hypothesis Class of Thresholds

The goal of this chapter is to analyze the relative advantage of the unlabeled data for the hypothesis class of thresholds on the real line introduced in Section 9.1. Our intention is to keep things as simple as possible so that we can demonstrate our point and at the same time provide enough motivation for the definition of the *semi-supervised ratio*. We propose this ratio as a measure of the relative advantage of the unlabeled data. Roughly speaking, the ratio is defined as the ratio between the learning rates of the best supervised and the best semi-supervised learner.

First, in the next section, we construct a semi-supervised learning algorithm for the hypothesis class of thresholds and compute its learning rate on distributions realizable by the hypothesis class of thresholds and with absolutely continuous unlabeled distribution. Second, we give a lower bound on the learning rate of any semi-supervised learner. As it will turn out, the learning rate will be roughly within factor 2 from the learning rate of the supervised learning algorithm given in Section 9.1. Third, we present the definition of *semi-supervised learning ratio* which quantifies what advantage of having access to unlabeled data i.e. advantage of the "knowledge" of the unlabeled distribution.

10.1 Kääriäinen's Algorithm

Consider the hypothesis class H of thresholds as defined in Section 9.1. We construct a semi-supervised learning algorithm B for probability distributions realizable by this class. The idea behind this algorithm was proposed by Kääriäinen [18]. For simplicity, we define the algorithm on the family of (unlabeled) distributions $\mathfrak{D}$ over $\mathbb{R}$ which are absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}$. Given an unlabeled distribution $D \in \mathfrak{D}$ and a labeled sample $S = ((x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m))$, the algorithm first computes the position ℓ of the rightmost negative example and the position r of the leftmost positive example:

$$\ell = \max\{x_i : 1 \leq i \leq m,\ y_i = -1\},$$
$$r = \min\{x_i : 1 \leq i \leq m,\ y_i = +1\}.$$

It computes, then, a point $t \in (\ell, r)$ such that $D([\ell, t]) = D([t, r]) = \frac{1}{2}D([\ell, r])$. If there are multiple such points t, the algorithm outputs any of them, for definitiveness, say

the leftmost such point. Similarly as before, we use the convention that maximum of an empty set is $-\infty$ and minimum of an empty set is $+\infty$.

Theorem 10.1. *Let* P *be a probability distribution over* $X \times \{+1, -1\}$ *realizable by the class* H *of thresholds and suppose its unlabeled distribution* P_X *is absolutely continuous (with respect to Lebesgue measure on* $\mathbb{R}$*). For any* $m \geq 2$*, the learning rate of Kääriäinen's algorithm* B *is*

$$\mathfrak{L}(B, P, H, m) \leq \frac{1}{2(m+1)} .$$

Proof. Let $F : \mathbb{R} \to [0, 1]$ be the distribution function of the unlabeled distribution P_X, that is, $F(s) = P_X((-\infty, s])$. Let $h_{t^*} \in H$ be a target for P and define $T^* = F(t^*)$. Consider the three real numbers ℓ, r, t computed by Kääriäinen's algorithm as random variables (depending on the random choice of the sample S). Define three related random variables $L = F(\ell)$, $R = F(r)$ and $T = F(t)$. Noting that $T = (L + R)/2$ we have

$$\begin{aligned}
\mathfrak{L}(B, P, H, m) &= \mathop{\mathbf{E}}_{S \sim P^m} \left[\mathrm{Err}^P(B(P_X, S)) \right] \\
&= \mathop{\mathbf{E}}_{S \sim P^m} |T^* - T| \\
&= \mathop{\mathbf{E}}_{S \sim P^m} |T^* - (L + R)/2| .
\end{aligned}$$

To compute the last expectation we consider three cases: (a) both ℓ and r are finite i.e. there exists both a positive and a negative example in the sample, (b) only ℓ is finite i.e. all examples in S are negative, (c) only r is finite i.e. all examples in S are positive. We write each of the cases as a Riemann integral and so we have

$$\begin{aligned}
\mathfrak{L}(B, P, H, m) = m(m+1) \int_0^T \int_T^1 |T^* - (L + R)/2| \cdot (1 - (R - L))^{m-2} \, dR \, dL \\
+ m \int_0^T |T^* - (L + 1)/2| \cdot L^{m-1} \, dL \\
+ m \int_T^1 |T^* - R/2| \cdot (1 - R)^{m-1} \, dR .
\end{aligned}$$

Without loss of generality assume that $T \leq 1/2$. These integrals can be explicitly calcu-

lated

$$\begin{aligned}
\mathfrak{L}(B, P, H, m) &= m(m-1)\int_0^T \int_T^{2T-R} (T-(L+R)/2)\cdot(1+L-R)^{m-2}\,dR\,dL \\
&\quad + m(m-1)\int_0^T \int_{2T-x_1}^{1} ((L+R)/2-T)\cdot(1+L-R)^{m-2}\,dR\,dL \\
&\quad + m\int_T^{2T} (T-R/2)\cdot(1-R)^{m-1}\,dR \\
&\quad + m\int_{2T}^{1} (R/2-T)\cdot(1-R)^{m-1}\,dR \\
&\quad + m\int_0^T ((L+1)/2-T)\cdot L^{m-1}\,dL \\
&= \frac{1}{2(m+1)} + \frac{(1-2T)^{m+1}-(1-T)^{m+1}-T^{m+1}}{2(m+1)} \qquad (10.1)
\end{aligned}$$

It is not hard to see that the second fraction is negative for any $T \in [0, 1/2]$ and hence the theorem follows. ■

10.2 Lower Bound

When we compare the supervised learner from Section 9.1 and Kääriäinen's semi-supervised learner from the previous section, we see that if the unlabeled distribution is absolutely continuous, the upper bound on learning rate of Kääriäinen's algorithm is two times better than the upper bound for the supervised learner. A natural question arises whether there exists a semi-supervised learner with a better learning rate. The following theorem answers this question negatively.

Theorem 10.2 (Lower Bound)**.** *Let* B *be any semi-supervised learner on the domain* $X = \mathbb{R}$*. Let* P_X *be any absolutely continuous (unlabeled) probability distribution over* $\mathbb{R}$*. There exists a probability distribution* P *over* $\mathbb{R} \times \{+1, -1\}$ *realizable by* H*, unlabeled distribution of which is* P_X *and such that for any* $m \geq 1$

$$\mathfrak{L}(B, P, H, m) \geq \frac{1}{2(m+2)}\ .$$

Proof. The proof is a simple application of the averaging argument. We choose $t \in \mathbb{R}$ at random from P_X and we let h_t to be the target. This means that the distribution $P = P_t$ itself is random, determined by the random choice of t. Formally, the distribution P_t is defined for any measurable set $M \subseteq \mathbb{R} \times \{+1, -1\}$ as

$$P_t(M) = P_X(\{x \in \mathbb{R} \,:\, (x, -1) \in M, x \leq t\}) + P_X(\{x \in \mathbb{R} \,:\, (x, +1) \in M, x > t\})\ .$$

We consider the average learning rate $\mathbf{E}_t[\mathfrak{L}(B, P_t, H, m)]$ and we will bound it from below by $\frac{1}{2(m+2)}$. The lower bound will imply existence of at least one t such that $\mathfrak{L}(B, P_t, H, m) \geq \frac{1}{2(m+2)}$ and the theorem will follow.

Let U be the random variable denoting the unlabeled part of the sample S. Formally, if the labeled sample S equals $((x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m))$ the unlabeled sample U is $(x_1, x_2, \ldots, x_m)$. Note that S is determined by U and t, namely, $y_i = -1$ if and only if $x_i \leq t$. With this in mind, we can write the average learning rate as

$$\begin{aligned} \mathop{\mathbb{E}}_{t\sim P_X}[\mathfrak{L}(B, P_t, H, m)] &= \mathop{\mathbb{E}}_{t\sim P_X} \mathop{\mathbb{E}}_{S\sim {P_t}^m} \left[\mathrm{Err}^{P_t}(B(P_X, S))\right] \\ &= \mathop{\mathbb{E}}_{U\sim P_X^m} \mathop{\mathbb{E}}_{t\sim P_X} \left[\mathrm{Err}^{P_t}(B(P_X, S))\right] \end{aligned} \tag{10.2}$$

Fix $U = (x_1, x_2, \ldots, x_m)$ and consider the inner term $\mathbb{E}_{t\sim P_X}\left[\mathrm{Err}^{P_t}(B(P_X, S))\right]$. Without loss of generality, assume that the elements of U are sorted so that $x_1 < x_2 < \cdots < x_m$. Now, since if t lies anywhere in the interval (x_i, x_{i+1}), then S and hence also the hypothesis $h_i := B(P_X, S)$ are fixed. (Here, $i = 0, 1, \ldots, m$ and $x_0 = -\infty$ and $x_{m+1} = +\infty$.) Since

$$\mathop{\mathbb{E}}_{t\sim P_X}\left[\mathrm{Err}^{P_t}(B(P_X, S))\right] = \sum_{i=0}^{m} \int_{x_i}^{x_{i+1}} \mathrm{Err}^{P_t}(h_i)\, dP_X(t)\ . \tag{10.3}$$

It remains to lower bound the Lebesgue integral $\int_{x_i}^{x_{i+1}} \mathrm{Err}^{P_t}(h_i)\, dP_X(t)$ for some fixed i. For that purpose, we define an indicator function

$$I(x, t) = \begin{cases} 1 & \text{if } h_i(x) = +1, x \leq t \\ 1 & \text{if } h_i(x) = -1, x > t \\ 0 & \text{if } h_i(x) = -1, x \leq t \\ 0 & \text{if } h_i(x) = +1, x > t \end{cases}$$

that indicates whether h_i makes error on the domain point x when the target is h_t. Using that indicator we can write

$$\begin{aligned} \int_{x_i}^{x_{i+1}} \mathrm{Err}^{P_t}(h_i)\, dP_X(t) &= \int_{x_i}^{x_{i+1}} \int_{-\infty}^{\infty} I(x, t)\, dP_X(x)\, dP_X(t) \\ &= \int_{-\infty}^{\infty} \int_{x_i}^{x_{i+1}} I(x, t)\, dP_X(t)\, dP_X(x) \end{aligned} \tag{10.4}$$

Consider a fixed $x \in (x_i, x_{i+1})$. Then

$$\begin{aligned} \int_{x_i}^{x_{i+1}} I(x, t)\, dP_X(t) &= \begin{cases} P_X((x, x_{i+1})) & \text{if } h_i(x) = -1 \\ P_X((x_i, x)) & \text{if } h_i(x) = +1 \end{cases} \\ &\geq \min\{P_X((x, x_{i+1})),\ P_X((x_i, x))\}\ . \end{aligned}$$

Substituting back into (10.4) we have

$$\begin{aligned} \int_{x_i}^{x_{i+1}} \mathrm{Err}^{P_t}(h_i)\, dP_X(t) &= \int_{-\infty}^{\infty} \int_{x_i}^{x_{i+1}} I(x, t)\, dP_X(t)\, dP_X(x) \\ &\geq \int_{x_i}^{x_{i+1}} \min\{P_X((x, x_{i+1})),\ P_X((x_i, x))\}\, dP_X(x) \\ &= P_X((x_i, x_{i+1}))^2/4\ . \end{aligned}$$

and substituting that into (10.3) we have

$$\mathop{\mathbb{E}}_{t\sim P_X}\left[\mathrm{Err}^{P_t}(B(P_X,S))\right] \geq \sum_{i=0}^{m} P_X((x_i,x_{i+1}))^2/4$$

It thus remains to compute

$$\mathop{\mathbb{E}}_{U\sim P_X^m}\left[\sum_{i=0}^{m} P_X((x_i,x_{i+1}))^2/4\right] .$$

Taking into account that the ordering $x_1 < x_2 < \cdots < x_m$ is one of $m!$ possible, we can write the expectation as an m-folded Riemann integral over a $1/m!$ fraction of the cube $[0,1]^m$:

$$\mathop{\mathbb{E}}_{U\sim P_X^m}\left[\sum_{i=0}^{m} P_X((x_i,x_{i+1}))^2/4\right] = m!\int_0^{z_{m+1}}\int_0^{z_m}\int_0^{z_{m-1}}\cdots\int_0^{z_2}\sum_{i=0}^{m}\frac{(z_{i+1}-z_i)^2}{4}\,dz_1\cdots dz_{m-1}\,dz_m$$

where $z_0 = 0$ and $z_{m+1} = 1$. We consider the last integral as a function of m and the last point z_{m+1} and we write

$$I_m(z_{m+1}) = m!\int_0^{z_{m+1}}\int_0^{z_m}\int_0^{z_{m-1}}\cdots\int_0^{z_2}\sum_{i=0}^{m}\frac{(z_{i+1}-z_i)^2}{4}\,dz_1\cdots dz_{m-1}\,dz_m .$$

Unfolding the integral we get the recurrence

$$\begin{aligned} I_0(z_1) &= z_1^2/4 , \\ I_m(z_{m+1}) &= m\int_0^{z_{m+1}} I_{m-1}(z_m) + (z_m)^{m-1}\cdot(z_{m+1}-z_m)^2/4\,dz_m \qquad \text{for } m \geq 1. \end{aligned}$$

We prove by induction on m that $I_m(z_{m+1}) = \frac{(z_{m+1})^{m+2}}{2(m+2)}$ for any $z_{m+1} \geq 0$. The base case $m = 0$ holds by definition. In the inductive case, $m \geq 1$, we obtain from the recurrence and inductive hypothesis

$$\begin{aligned} I_m(z_{m+1}) &= m\int_0^{z_{m+1}} I_{m-1}(z_m) + (z_m)^{m-1}\cdot(z_{m+1}-z_m)^2/4\,dz_m \\ &= m\int_0^{z_{m+1}} \frac{(z_m)^{m+1}}{2(m+1)} + (z_m)^{m-1}\cdot(z_{m+1}-z_m)^2/4\,dz_m \\ &= m\left(\frac{(z_{m+1})^{m+2}}{2(m+1)(m+2)} + m\frac{(z_{m+1})^{m+2}}{2m(m+1)(m+2)}\right) \\ &= \frac{(z_{m+1})^{m+2}}{2(m+2)} . \end{aligned}$$

Substituting the value $z_{m+1} = 1$ and looking back at (10.2) we obtain the lower bound of the expected learning rate

$$\mathop{\mathbb{E}}_{t\sim P_X}\left[\mathfrak{L}(B,P_t,H,m)\right] \geq I_m(1) = \frac{1}{2(m+2)} .$$

■

10.3 Semi-supervised Learning Ratio

Learning algorithms supervised or semi-supervised learners are intended to be used for an "unknown" distributions over $X \times \{+1, -1\}$. In other words, the learner will used for any distribution coming from a family $\mathfrak{P}$ of probability distributions over $X \times \{+1, -1\}$. Any assumption about the learning problem can be captured by what distributions do we include in (or exclude from) the family $\mathfrak{P}$. (We will talk about this issue later.)

On the family $\mathfrak{P}$, we can compare learning rates of a supervised learning algorithm A and a semi-supervised learning algorithm B. In order to do so, we partition the family $\mathfrak{P}$ into equivalence classes such that within each equivalence class the unlabeled distributions are the same. Formally, we define $\mathfrak{D} = \{P_X : P \in \mathfrak{P}\}$ the family of all unlabeled distributions. And for any unlabeled distribution $D \in \mathfrak{D}$ we define the equivalence class $\mathfrak{P}[D] = \{P : P \in \mathfrak{P}, P_X = D\}$. Using this notation we define the *semi-supervised learning ratio*.

Definition 10.3. *Let X be a domain, let H be a hypothesis class over X, let $\mathfrak{P}$ be a family of probability distributions over $X \times \{+1, -1\}$. Let A be a supervised learner, and let B be a semi-supervised learner. Fro any sample size m, we define* semi-supervised learning ratio *as*

$$\textbf{ssl-ratio}(H, \mathfrak{P}, m, A, B) = \sup_{D \in \mathfrak{D}} \frac{\sup_{P \in \mathfrak{P}[D]} \mathfrak{L}(A, P, H, m)}{\sup_{P \in \mathfrak{P}[D]} \mathfrak{L}(B, P, H, m)}$$

and inherent semi-supervised learning ratio *as*

$$\textbf{ssl-ratio}(H, \mathfrak{P}, m) = \inf_{A} \sup_{B} \textbf{ssl-ratio}(H, \mathfrak{P}, m, A, B) \ .$$

where the infimimum is taken over all supervised learners A and the supremum is taken over all semi-supervised learners B.

The fraction $\frac{\sup_{P \in \mathfrak{P}[D]} \mathfrak{L}(A,P,H,m)}{\sup_{P \in \mathfrak{P}[D]} \mathfrak{L}(B,P,H,m)}$ is the ratio of the worst-case learning rates of A, B on distributions with the same unlabeled distribution D. Worst-case here refers to the suprema in the numerator and the denominator of the fraction. The semi-supervised learning ratio for learners A, B is the supremum of this fraction over all unlabeled distributions. This supremum corresponds to the best-case over the unlabeled distributions in the favor of SSL.

The inherent semi-supervised learning ratio is the min-max value of semi-supervised learning ratio. Perhaps more intuitively, it can be written as

$$\textbf{ssl-ratio}(H, \mathfrak{P}, m) = \inf_{A} \sup_{D \in \mathfrak{D}} \frac{\sup_{P \in \mathfrak{P}[D]} \mathfrak{L}(A, P, H, m)}{\inf_{B} \sup_{P \in \mathfrak{P}[D]} \mathfrak{L}(B, P, H, m)}$$

where the both infima are taken over all supervised learners A, B. This rearrangement follows from the fact that a semi-supervised learner can be seen as collection of supervised learners; one supervised learner for each unlabeled distribution D.

Our results from about thresholds on the real line can expressed as follows. Let $\mathfrak{P}$ be the family of distributions over $\mathbb{R} \times \{+1, -1\}$ realizable by the hypothesis class of thresholds H and having the absolutely continuous unlabeled distribution. If we denote by by A the supervised learner from Section 9.1, and by B Kääriäinen's semi-supervised learner, then

$$\textbf{ssl-ratio}(H, \mathfrak{P}, A, B, m) = 2 + O(1/m) \ .$$

This follows from the note at the end of Section 9.1 and equation (10.1) at the end of the proof of Theorem 10.1. More importantly, the inherent semi-supervised learning ratio is

$$\textbf{ssl-ratio}(H, \mathfrak{P}, m) \leq 2 + O(1/m) \ .$$

as follows from the lower bound in stated in Theorem 10.2. This says intuitively that on the family $\mathfrak{P}$ and the hypothesis class of thresholds the relative advantage of unlabeled data is an improvement of the learning rate roughly by a factor of 2.

Chapter 11

Conclusion

Semi-supervised learning is an active research topic. Our aim was to draw attention to the basic fact that access to unlabeled data itself does not decrease the worst-case sample complexity of learning, if one does not postulate any relationship between unlabeled data and the labels. We have mathematically proven that this is indeed the case of learning thresholds on the real line provided that the unlabeled distribution is absolutely continuous. We stress that in semi-supervised learning in order to utilize the unlabeled data, it is important to make non-trivial prior assumptions on the relationship between unlabeled data and the labels. Our work shows that in order to gain a non-trivial advantage from unlabeled data, making these assumptions is, in fact, necessary.

In retrospect our claim sounds almost obvious. However, it is less obvious mathematically. And so, as in Ben-David et al. [5], we make the following conjecture.

Conjecture 11.1. *Let* H *be any hypothesis class over some domain* X*. Let* $\mathfrak{P}$ *be a family of all probability distributions over* $X \times \{+1, -1\}$ *realizable by* H*. There exists a universal*[1] *constant* c *such that*

$$\textbf{ssl-ratio}(H, \mathfrak{P}, m) \leq c \quad \textit{for all } m \geq 0.$$

A step towards proving the conjecture would be to consider as H the class of thresholds on the real line and as $\mathfrak{P}$ the class of all realizable distributions over $\mathbb{R} \times \{+1, -1\}$ with *finite support*. It seems that these distributions have faster (i.e. lower) learning rates than distributions with absolutely continuous unlabeled distributions. As an extreme example, any consistent learner has learning rate 0 on a distribution concentrated on one point.

Notice that we do not put any restriction on the hypothesis class H. In particular, we do not insist on the Vapnik-Chervonenkis dimension to be finite. For example, consider an uncountable domain X (e.g. the real line) and the class H of all hypotheses h such that either h labels finitely many points $+1$ and all other points by -1, or vice versa h labels finitely many points -1 and all other points by $+1$. This class has infinite VC-dimension. Despite that one can construct a supervised learner which for any realizable distribution has learning rate approaching zero. We conjecture the learning rate of any semi-supervised learner is within constant factor.

A similar conjecture can be made in the unrealizable case.

[1] Universal means that c does not depend on H and X.

Conjecture 11.2. *Let* H *be any hypothesis class over some domain* X*. Let* $\mathfrak{P}$ *be a family of all probability distributions over* $X \times \{+1, -1\}$*. There exists a universal constant* c' *such that*

$$\textbf{ssl-ratio}(H, \mathfrak{P}, m) \leq c' \quad \textit{for all } m \geq 0.$$

A special case of this conjecture was proved by Ben-David et al. [5] and Lu [22]. Namely, the conjecture was proved for the hypothesis class of unions of d intervals on the real line and $\mathfrak{P}$ restricted to distributions which have absolutely continuous unlabeled distributions. (The constant c' is not explicitly computed in these works.)

Bibliography

[1] Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, Cambridge, UK, 1999.

[2] Peter L. Bartlett, Tamás Linder, and Gábor Lugosi. The minimax distortion redundancy in empirical quantizer design. *IEEE Transactions on Information Theory*, 44(5):1802–1813, 1998.

[3] Shai Ben-David. A framework for statistical clustering with a constant time approximation algorithms for k-median clustering. In John Shawe-Taylor and Yoram Singer, editors, *Proceedings of the 17th Annual Conference on Learning Theory (COLT), Banff Canada*, Lecture Notes in Artificial Intelligence, pages 415–426. Springer, July 2004.

[4] Shai Ben-David. A framework for statistical clustering with constant time approximation algorithms for k-median and k-means clustering. *Machine Learning*, 66(2–3), March 2007. Preliminary version as [3].

[5] Shai Ben-David, Tyler Lu, and Dávid Pál. Does unlabeled data provably help? Worst-case analysis of the sample complexity of semi-supervised learning. In Servedio and Zhang [25], pages 33–44. Available at `http://colt2008.cs.helsinki.fi/papers/COLT2008.pdf`.

[6] Shai Ben-David, Dávid Pál, and Hans Ulrich Simon. Stability of k-means clustering. In Nader H. Bshouty and Claudio Gentile, editors, *Proceedings of the 20th Annual Conference on Learning Theory (COLT), San Diego, CA, USA*, Lecture Notes in Artificial Intelligence, pages 20–34. Springer, July 2007.

[7] Shai Ben-David and Ulrike von Luxburg. Relating clustering stability to properties of cluster boundaries. In Servedio and Zhang [25], pages 397–390. Available at `http://colt2008.cs.helsinki.fi/papers/COLT2008.pdf`.

[8] Shai Ben-David, Ulrike von Luxburg, and Dávid Pál. A sober look at clustering stability. In Gábor Lugosi and Hans Ulrich Simon, editors, *Proceedings of the 19th Annual Conference on Learning Theory (COLT), Pittsburgh, PA, USA*, Lecture Notes in Artificial Intelligence, pages 5–19. Springer, 2006.

[9] Asa Ben-Hur, André Elisseeff, and Isabelle Guyon. A stability based method for discovering structure in clustered data. In *Pacific Symposium on Biocomputing (PSB)*, volume 7, pages 6–17, 2002.

[10] Gérard Biau, Luc Devroye, and Gábor Lugosi Biau. On the performance of clustering in hilbert spaces. *IEEE Transactions on Information Theory*, 54(2):781–790, 2008.

[11] Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[12] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien, editors. *Semi-Supervised Learning*. MIT Press, Cambridge, Massachusetts, USA, September 2006.

[13] Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

[14] Luc Devroye and Gábor Lugosi. *Combinatorial Methods in Density Estimation*. Springer, 2001.

[15] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. Willey Interscience, second edition edition, 2001.

[16] Richard M. Dudley. Balls in $\mathbb{R}^k$ do not cut all subsets of $k + 2$ points. *Advances in Mathematics*, 31(3):306–308, 1979.

[17] David Haussler. Decision theoretic generalizations of the pac model for neural net and other learning applications. *Information and computation*, 100(1):78–150, 1992.

[18] Matti Kääriäinen. Generalization error bounds using unlabeled data. In Peter Auer and Ron Meir, editors, *Proceedings of the 18th Annual Conference on Learning Theory (COLT), Bertinoro, Italy*, Lecture Notes in Artificial Intelligence, pages 127–142. Springer, June 2005.

[19] Michael J. Kearns and Umesh V. Vazirani. *An Introduction Computational Learning Theory*. MIT Press, Cambridge, Massachusetts, USA, 1994.

[20] Tilman Lange, Volker Roth, Mikio L. Braun, and Joachim M. Buhmann. Stability-based model selection. In S. Thrun S. Becker and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15 (NIPS)*, pages 617–624, Cambridge, Massachusetts, USA, 2003. MIT Press.

[21] Tilman Lange, Volker Roth, Mikio L. Braun, and Joachim M. Buhmann. Stability-based validation of clustering solutions. *Neural Computation*, 16(6):1299–1323, 2004. Preliminary version appeared as [20].

[22] Tyler Lu. Worst case analysis of semi-supervised learning. Master's thesis, University of Waterloo, Canada, 2009.

[23] David Pollard. Strong consistency of k-means clustering. *Annals of Statistics*, 9(1):135–140, 1981.

[24] Sidney I. Resnik. *A Probability Path*. Birkhäuser, Boston, 1999.

[25] Rocco A. Servedio and Tong Zhang, editors. *Proceedings of the 21st Annual Conference on Learning Theory (COLT), Helsinki, Finland*. Omnipress, July 2008. Available at `http://colt2008.cs.helsinki.fi/papers/COLT2008.pdf`.

[26] Ohad Shamir and Naftali Tishby. Model selection and stability in k-means clustering. In Servedio and Zhang [25], pages 367–378. Available at `http://colt2008.cs.helsinki.fi/papers/COLT2008.pdf`.

[27] Georgi E. Shilov. *Elementary Real and Complex Analysis*. Dover Publications, 1996.

[28] Leslie G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

[29] Vladimir N. Vapnik. *Statistical Learning Theory*. Jonh Wiley & Sons, 1998.

[30] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, second edition, 1999.

[31] Vladimir N. Vapnik and Alexey Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971.