Algorithms for the Bregman k-Median Problem

Dipl.-Math. Dipl.-Inform.
Marcel R. Ackermann

October 17, 2009
(revised version)

A dissertation submitted to the
Department of Computer Science
University of Paderborn

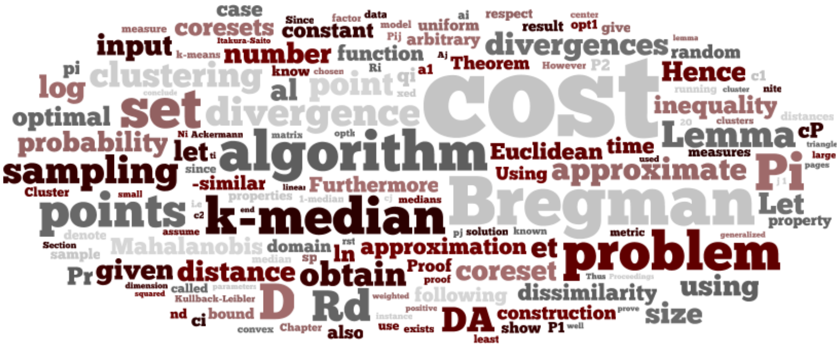
for the degree of
Doktor der Naturwissenschaften
(doctor rerum naturalium)

accepted on the recommendation of
Prof. Dr. Johannes Blömer
University of Paderborn
Prof. Dr. Christian Sohler
Technische Universität Dortmund

defended on
December 4, 2009

"A point is that of which there is no part."

— Euclid's Elements, Book I, Definition 1.



Word cloud of this thesis, created by the `wordle.net` web application, licensed under Creative Commons Attribution 3.0 United States License.

Acknowledgment

First of all, I am forever grateful to my advisor, Prof. Dr. Johannes Blömer, for his guidance and his support through the past years. He gave me this once-in-a-lifetime opportunity to join his research group, and he kept the faith in my research even when the work was progressing much slower than expected. I am also very grateful to Prof. Dr. Christian Sohler for his advice and for introducing me to the field of computational geometry; a field I actually never intended to study before his advice.

The research presented in this thesis was conducted while I was employed at the Department of Computer Science of the University of Paderborn. This research was supported in part by the Deutsche Forschungsgemeinschaft (DFG), grant BL 314/6-1. I am grateful for the funding I received.

I owe many thanks to my colleagues and friends Dr. Valentina Damerow, Sascha Effert, Birgitta Grimm, Dr. Mirko Hessel-von Molo, Dr. Volker Krummel, Daniel Kuntze, Dr. Alexander May, Stefanie Naewe, Dr. Martin Otto, and Jonas Schrieb. They all contributed to this thesis through many small but fruitful discussions and by stirring up new ideas. I am also thankful to Eva Kuntze and Dr. Mirko Hessel-von Molo for proof-reading my thesis and for giving valuable criticisms.

Last but not least, I would have never succeeded in such an ambitious project as this thesis without the enduring support of “my families”: my parents and my sisters who gave me constant encouragement, Josefin, Joey, and Jil, because they truly know how to cheer me up, and all of *The Family* here in Paderborn for the many, many reasons you (hopefully) know, or may not even be aware of. I am also indebted to Gabriele Kappius, because she is the main reason I started all this.

This document has been typeset using $\LaTeX\,2_\varepsilon$ under *openSUSE 11.0*. All images have been created using *MuPAD Pro 4.0.2* and *GIMP 2.4.5*. Thanks for these great products.

Abstract

In this thesis, we study the k-median problem with respect to a dissimilarity measure D_φ from the family of Bregman divergences: Given a finite set P of size n from $\mathbb{R}^d$, our goal is to find a set C of size k such that the sum of error $\mathrm{cost}(P, C) = \sum_{p \in P} \min_{c \in C} \big\{ \mathrm{D}_\varphi(p, c) \big\}$ is minimized. This problem plays an important role in applications from many different areas of computer science, such as information theory, statistics, data mining, and speech processing.

Our main contribution is the development of a general framework of algorithms and techniques that is applicable to (almost) all Bregman divergences. In particular, we give a randomized approximation algorithm for the Bregman k-median problem that computes a $(1 + \varepsilon)$-approximate solution using at most $2^{\tilde{\mathcal{O}}(k/\varepsilon)} n$ arithmetic operations, including evaluations of Bregman divergence D_φ. In doing so, we give the first approximation algorithm known for this problem that provides any provable approximation guarantee. We also give a fast, practical, randomized approximation algorithm that computes an $\mathcal{O}(\log k)$-approximate solution for arbitrary input instances, or even an $\mathcal{O}(1)$-approximate solution for certain, well separated input instances.

In addition to that, we study the use of coresets in the context of Bregman k-median clusterings. In a nutshell, a coreset is a small (weighted) set that features the same clustering behavior as the original input set. We show how classical coreset constructions for the Euclidean k-means problem can be adapted to a special subfamily of the Bregman divergences, namely the class of Mahalanobis distances. We also give a new, randomized coreset construction for the Mahalanobis k-median problem in low dimensional spaces that has several practical advantages. Furthermore, by introducing the notion of weak coresets, we give the first coreset construction applicable to (almost) all Bregman k-median clustering problems. Using these weak coresets, we are able to give the currently asymptotically fastest $(1 + \varepsilon)$-approximation algorithm known for the Bregman k-median problem. This algorithm uses at most $\mathcal{O}(kn) + 2^{\tilde{\mathcal{O}}(k/\varepsilon)} \log^{k+2}(n)$ arithmetic operations, including evaluations of Bregman divergence D_φ.

Zusammenfassung

Thema dieser Dissertationsschrift ist das k-Median Problem unter Verwendung eines Abstandsmaßes D_φ aus der Familie der Bregman-Divergenzen: Zu einer gegebenen Eingabemenge P der Größe n aus dem $\mathbb{R}^d$ ist eine Zentrenmenge C der Größe k gesucht, welche die Zielfunktion $\operatorname{cost}(P, C) = \sum_{p \in P} \min_{c \in C} \big\{ \mathrm{D}_\varphi(p, c) \big\}$ minimiert. Dieses Problem ergibt sich in Anwendungen aus den verschiedensten Teilbereichen der Informatik, etwa in der Informationstheorie, in der Statistik, beim Durchsuchen großer Datenbestände oder bei der Verarbeitung von Sprachsignalen.

Das Hauptresultat dieser Arbeit besteht in der Entwicklung einer Sammlung von Algorithmen und Techniken, die sich auf (nahezu) alle Bregman-Divergenzen anwenden lassen. Insbesondere präsentieren wir einen randomisierten Approximationsalgorithmus der Güte $(1 + \varepsilon)$ für das Bregman-k-Median-Problem. Dieser Algorithmus berechnet seine Lösung unter Verwendung von maximal $2^{\tilde{\mathcal{O}}(k/\varepsilon)} n$ arithmetischen Operationen, darunter auch Auswertungen des Abstandsmaßes D_φ. Dabei handelt es sich um den ersten für das Bregman-k-Median-Problem anwendbaren Algorithmus, der eine beweisbare Approximationsgüte aufweist. Außerdem präsentieren wir einen effizienten, praktisch relevanten, randomisierten Approximationsalgorithmus, der Lösungen der Güte $\mathcal{O}(\log k)$ berechnet; für spezielle, wohlseparierte Eingabeinstanzen berechnet dieser Algorithmus sogar Lösungen konstanter Güte.

Darüber hinaus untersuchen wir die Anwendung von Kernmengen für das Bregman-k-Median-Problem. Kurz zusammengefasst handelt es sich bei einer Kernmenge um eine kleine (gewichtete) Punktemenge, welche die gleichen Clustering-Eigenschaften wie die ursprüngliche Eingabemenge aufweist. Wir demonstrieren, wie sich klassische Kernmengenkonstruktionen des euklidischen k-Mittelwert-Problems auf eine spezielle Teilmenge der Bregman-Divergenzen verallgemeinern lassen, nämlich auf die Klasse der so genannten Mahalanobis-Distanzen. Wir präsentieren ferner eine neue, praktisch vorteilhafte, randomisierte Kernmengenkonstruktion für das Mahalanobis-k-Median-Problem in niedrigdimensionalen Räumen. Zudem greifen wir das Konzept der schwachen Kernmengen auf und präsentieren damit die erste Kernmengenkonstruktion, die sich für (fast) alle Bregman-Divergenzen anwenden läßt. Unter Anwendung dieser schwachen Kernmengen erhalten wir den derzeit asymptotisch effizientesten $(1+\varepsilon)$-Approximationsalgorithmus für das Bregman-k-Median-Problem. Dieser Algorithmus benötigt maximal $\mathcal{O}(kn) + 2^{\tilde{\mathcal{O}}(k/\varepsilon)} \log^{k+2}(n)$ arithmetische Operationen, darunter auch Auswertungen des Abstandsmaßes D_φ.

Contents

Some notes on notation

$\emptyset$	empty set
$\mathbb{N}$	set of the natural numbers $1, 2, 3, \ldots$
$\mathbb{N}_0$	set of the natural numbers, including 0
$\mathbb{R}$	set of the reals x with $-\infty < x < \infty$
$\mathbb{R}_{\geq 0}$	set of the non-negative reals x with $0 \leq x < \infty$
$\mathbb{R}_+$	set of the positive reals x with $0 < x < \infty$
$[a, b]$	closed interval of the reals x with $a \leq x \leq b$

$\|M\|$	cardinality of set M
$M \cup N$	union of sets M and N
$M \cap N$	intersection of sets M and N
$M \setminus N$	difference set M minus N
$M \times N$	Cartesian product of sets M and N
M^d	set of d-dimensional column vectors with entries from set M
$M^{d_1 \times d_2}$	set of $(d_1 \times d_2)$-matrices with entries from set M

I_d	identity matrix from $\mathbb{R}^{d \times d}$
$A^\top$,	transpose of matrix $A \in \mathbb{R}^{d_1 \times d_2}$
$v^\top$	transpose of column vector $v \in \mathbb{R}^{d \times 1}$ or row vector $v \in \mathbb{R}^{1 \times d}$
$v^\top w$	inner product of column vectors $v, w \in \mathbb{R}^d$
$\\|v\\|$	Euclidean norm of column vector $v \in \mathbb{R}^d$

$\lceil x \rceil$	smallest integer n with $x \leq n$
$\lfloor x \rfloor$	largest integer n with $x \geq n$
$n!$	factorial of positive integer n
$\binom{n}{k}$	binomial coefficient n over k

π	ratio of a circle's circumference to its diameter, i.e., $\pi \approx 3.141\ldots$
e	base of the natural logarithm, i.e., $e \approx 2.718\ldots$
$\exp(x)$	exponential function, i.e., $\exp(x) = e^x$
$\ln(x)$	natural logarithm of $x > 0$
$\log(x)$	binary logarithm of $x > 0$

$f = \mathcal{O}(g)$	$f, g : \mathbb{N} \to \mathbb{R}_{\geq 0}$ with $\lim_{n\to\infty} f(n)/g(n) < \infty$
$f = o(g)$	$f, g : \mathbb{N} \to \mathbb{R}_{\geq 0}$ with $\lim_{n\to\infty} f(n)/g(n) = 0$
$f = \Omega(g)$	$f, g : \mathbb{N} \to \mathbb{R}_{\geq 0}$ with $\lim_{n\to\infty} f(n)/g(n) > 0$
$f = \omega(g)$	$f, g : \mathbb{N} \to \mathbb{R}_{\geq 0}$ with $\lim_{n\to\infty} f(n)/g(n) = \infty$
$f = \Theta(g)$	$f, g : \mathbb{N} \to \mathbb{R}_{\geq 0}$ with $f = \mathcal{O}(g)$ and $f = \Omega(g)$
$f = \tilde{\mathcal{O}}(g)$	$f, g : \mathbb{N} \to \mathbb{R}_{\geq 0}$ with $f = \mathcal{O}\big(g \log^k(g)\big)$ for some $k \in \mathbb{N}$
$\inf_{x\in M} f(x)$	largest real number y with $y \leq f(x)$ for all $x \in M$
$\sup_{x\in M} f(x)$	smallest real number y with $y \geq f(x)$ for all $x \in M$
$\min_{x\in M} f(x)$	minimal value of $f(x)$ for all $x \in M$
$\max_{x\in M} f(x)$	maximal value of $f(x)$ for all $x \in M$
$\arg\min_{x\in M} f(x)$	element $x^* \in M$ with $f(x^*) \leq f(x)$ for all $x \in M$
$\arg\max_{x\in M} f(x)$	element $x^* \in M$ with $f(x^*) \geq f(x)$ for all $x \in M$
$\frac{\partial}{\partial x_i} f(x)$	i-th partial derivative of function f at point x
$\nabla f(x)$	gradient vector of function f at point x
$\nabla^2 f(x)$	Hessian matrix of function f at point x
$\mathrm{ri}(M)$	relative interior of set $M \subseteq \mathbb{R}^d$
$\mathrm{vol}_d(M)$	d-dimensional volume of set $M \subseteq \mathbb{R}^d$
$\Pr[A]$	probability of event A
$\mathrm{E}[X]$	expectation of random variable X
$\mathrm{Var}[X]$	variance of random variable X

1 Introduction

One frequent task in computer science is to find a representation of a large number of data objects by a small number of prototypical representatives. The quality of such a representation is usually measured by using the average dissimilarity of an object towards its prototype or, equivalently, by summing up the total dissimilarity of all objects towards their prototype. Hence, to achieve a good representation, the goal is to find a number of prototypes that leads to small, or even minimal, representation error in terms of total dissimilarity. Here, the concrete dissimilarity can be anything, depending on the concrete application, from the perceivable difference in colors in the RGB color space, to the disagreement between the nucleotides of gene sequences, or the semantic (dis-)similarity of webpages in the world wide web.

To obtain a mathematical tractability of this task, data objects are usually assumed to be represented by elements from vector space $\mathbb{R}^d$ with a possibly very large dimension d. Furthermore, it is usually convenient to assume that the dissimilarity between the data objects can be described as a metric, such as the Euclidean metric in $\mathbb{R}^d$. In this case, if the number k of prototypical center points is fixed, the problem of finding these prototypes is usually called the *Euclidean k-median problem.* This is because it is a natural generalization to the problem of finding the median element of a finite number of elements from $\mathbb{R}$. A solution of the k-median problem defines a clustering of the data: Each data object is identified with its closest prototype, and the set of all points that are assigned to a common prototype form a cluster.

The origin of the k-median problem (and of the related facility location problem) is commonly attributed to the German economist Alfred Weber; although the roots of this problem can be traced back as far as the early 17th century. It has already been considered by illustrious individuals such as the French scholar Pierre de Fermat and the Italian physicist Evangelista Torricelli (see [Wesolowsky, 1993] for an in-depth report on the rich history of the k-median problem). In his seminal work [Weber, 1909], Weber studies the minimization of the transport cost from an industrial facility to its clients. Here, the clients and the facility are assumed to be located in the plane, and the transport cost is assumed to be proportional to the distance between the clients and the facility, scaled by an individual weight for each client. In Weber's work (or, more precisely, in its mathematical appendix, written by Austrian mathematician Georg Pick) a method is described for constructing the location of the spatial median of merely three weighted points in the plane geometrically, using compass and straightedge. However, this construction already fails for more than three points. In fact, from a computational perspective, it turns out that finding the exact 1-median of a finite subset of $\mathbb{R}^d$ is indeed an infeasible problem (cf. [Bajaj, 1988]), known as the *Fermat-Weber problem* in literature.

A problem very similar to the Euclidean k-median problem can be stated as follows. Given a finite number of elements from the Euclidean space $\mathbb{R}^d$, find a number of k prototypical center points such that the sum of the squared distances of all input points towards their prototype is minimal. This problem is usually called the *Euclidean k-means problem*, since it is a natural generalization to the problem of finding the arithmetic mean of a finite number of elements from $\mathbb{R}$. The use of squared distances has several practical advantages. Most importantly, the 1-means problem avoids the inherent infeasibility of the 1-median problem: The arithmetic mean of a finite point set from $\mathbb{R}^d$ can always be computed efficiently. Nevertheless, both the Euclidean k-median problem and the Euclidean k-means problem do share a number of combinatorial properties. In fact, the Euclidean k-means problem can also be seen as a k-median problem using the (non-metric) squared Euclidean distance as dissimilarity measure.

With the growing interest of the theoretical computer science community in the field of computational geometry, the Euclidean k-median problem and the Euclidean k-means problem have received a lot of attention in the recent decades. In particular, the computational hardness of solving these problems exactly has been studied, and a large number of efficient approximation algorithms have been developed. Unfortunately, these results rely on the geometrical properties of the Euclidean distance, such as symmetry and the triangle inequality. Hence, prior to our work, almost no clustering algorithms were known for the k-median problem using arbitrary non-metric distances, or even an asymmetric dissimilarity measure. This is contrary to the fact that there are a large number of applications where the k-median problem with respect to a non-metric dissimilarity measures is considered. For instance, in the spectral analysis of speech signals, k-median clustering by Itakura-Saito divergence is used to quantize speech signals, and in information theory and statistics, k-median clustering with respect to the Kullback-Leibler divergence is used to estimate the parameters of a maximum likelihood model from multinomially distributed data. Both of these dissimilarity measures are neither a metric, nor a symmetric distance function.

In this thesis, we take a first step towards closing this unfortunate gap between theory and application. We study the k-median problem with respect to a large class of non-metric dissimilarity measures, that includes prominent, symmetric instances such as the squared Euclidean distance and the Mahalanobis distances, as well as asymmetric instances like the Itakura-Saito divergences and the Kullback-Leibler divergences: the family of *Bregman divergences*. In particular, we seek to obtain our results in a general framework that works with only minimal assumptions regarding the concrete nature of the Bregman divergence used. This is achieved by focussing on the combinatorial and statistical properties of a whole class of dissimilarity functions instead of relying on concrete geometric properties of single, well-natured instances. Hence, the theory provided in this thesis yields algorithms and techniques that are applicable for (almost) all instances of the Bregman k-median problem.

To avoid the inherent hardness of finding optimal solutions of the k-median problem, we will rely on two standard techniques of algorithm design to obtain efficient algorithms: *approximation* and *randomization*. That is, we do not try to solve the Bregman k-median problem exactly. Rather, we concentrate on finding good, approximate solutions with a total dissimilarity that is guaranteed to be within a certain factor from the optimal k-median cost of the input instance. In fact, most of the time, we seek to give approximations that come arbitrarily close to the quality of an optimal solution. Furthermore, in our algorithms, we make use of random sampling, i.e., the selection of a random element from a given finite set. To this end, we always assume that the data is given in an appropriate data structure that supports random sampling of a single element according to a given probability distribution in constant time. We obtain algorithms and techniques for the Bregman k-median problem that are guaranteed to yield the desired result with at least constant probability, or even with a high probability that comes arbitrarily close to 1.

1.1 State of the art

Euclidean k-median problem. In [Megiddo and Supowit, 1984], it has been shown that the Euclidean k-median problem is $\mathcal{NP}$-hard in any dimension $d \geq 2$. Furthermore, in general metric spaces, the k-median problem of n input points can not be approximated in polynomial time within a factor of 1.73 unless $\mathcal{NP} \subseteq \mathcal{DTIME}(n^{\mathcal{O}(\log\log n)})$ (cf. [Jain et al., 2002]).

However, this non-approximability result is no longer valid in the case of the d-dimensional Euclidean space. In [Arora et al., 1998] a first randomized polynomial time approximation scheme for the Euclidean k-median problem in the plane has been given, computing a $(1+\varepsilon)$-approximate solution in time $kn^{\mathcal{O}(1/\varepsilon)}$. This result was later improved to arbitrary dimension d in time $2^{\mathcal{O}(1/\varepsilon^d)} n \log(n) \log(k)$ in [Kolliopoulos and Rao, 1999]. Using so called coresets, [Har-Peled and Mazumdar, 2004] improved the running time of the algorithm from [Kolliopoulos and Rao, 1999] further to $\mathcal{O}(dn) + 2^{\mathcal{O}(1/\varepsilon^d)} k^5 \log^9(n)$, that is, to a running time that is merely linear in the number of points n, but still doubly exponential in d.

In [Bădoiu et al., 2002], a randomized algorithm has been given that computes a $(1+\varepsilon)$-approximate solution in time $d^{\mathcal{O}(1)} 2^{(k/\varepsilon)^{\mathcal{O}(1)}} n \log^{\mathcal{O}(k)}(n)$. Thereby, the authors gave the first approximation scheme for the Euclidean k-median problem with a running time polynomial in n and d. A further improvement has been made by the randomized $(1+\varepsilon)$-approximation algorithm given in [Kumar et al., 2005] that runs in time $d2^{(k/\varepsilon)^{\mathcal{O}(1)}} n$, which is merely linear in n and d. This algorithm was improved in [Chen, 2006] by combining it with a new coreset construction that leads to a $(1+\varepsilon)$-approximation algorithm with running time $\mathcal{O}(dkn) + d^2 2^{(k/\varepsilon)^{\mathcal{O}(1)}} \log^{k+2}(n)$.

Euclidean k-means problem. While an optimal solution of the Euclidean k-means problem of n points from $\mathbb{R}^d$ can always be found in time $n^{\mathcal{O}(dk)}$ due to an observation of [Inaba et al., 1994], this problem is known to be $\mathcal{NP}$-hard if the dimension d or the number of clusters k is unbounded (see [Dasgupta, 2007], [Aloise et al., 2009], and [Mahajan et al., 2009]).

One particular heuristic that has been used intensively since the late 1950s to find solutions for the Euclidean k-means problem is Lloyd's algorithm (cf. [Lloyd, 1982]). In fact, this algorithm has become so popular among practitioners that it has been commonly named *the* k-means algorithm. Unfortunately, the deterministic textbook version of this algorithm can obtain arbitrary bad clusterings. In [Arthur and Vassilvitskii, 2007], Lloyd's algorithm has been combined with a new, randomized seeding technique to compute $\mathcal{O}(\log k)$-approximate solutions. In addition, it has been shown that a variant of this seeding technique computes a constant factor approximate solution, provided that the input set consists of k well separated clusters (cf. [Ostrovsky et al., 2006]). Also, the exact running time of Lloyd's algorithm has been unknown for a long time. This was resolved not before 2009, when it was proven that in any dimension $d \geq 2$ there are worst-case instances of input points that require at least $2^{\Omega(n)}$ operations (cf. [Vattani, 2009]). On the other hand, the smoothed complexity of the running time has been proven to be merely polynomial in n, k and d (see [Arthur et al., 2009]), which confirms the observable speed of Lloyd's method on real world data sets.

Considering $(1+\varepsilon)$-approximations, [Inaba et al., 1994] proposed a randomized $\mathcal{O}(\varepsilon^{-d}n)$-time algorithm for the case of $k = 2$ clusters. This was later improved in [Matoušek, 2000] where a deterministic $(1+\varepsilon)$-approximation algorithm for arbitrary k with a running time of $\varepsilon^{-k^2 d} n \log(n)$ has been given. Using coresets, [Har-Peled and Mazumdar, 2004] improved the running time of this algorithm to $\mathcal{O}(dn) + k^{\mathcal{O}(k)}\varepsilon^{-\mathcal{O}(dk)} \log^{k+1}(n)$, that is, to a running time that is merely linear in n. Furthermore, the algorithm from [Fernandez de la Vega et al., 2003] achieves a running time of $d^{\mathcal{O}(1)} 2^{(k/\varepsilon)^{\mathcal{O}(1)}} n \log^k(n)$, thereby giving the first randomized approximation scheme with a running time polynomial in both n and d.

Today, a number of approximation schemes are known whose running time is linear in n and d but exponential in k. The first of these algorithms was given in [Kumar et al., 2004] and achieves a running time of $d\, 2^{(k/\varepsilon)^{\mathcal{O}(1)}} n$. Using the coresets from [Chen, 2006], this algorithm was later improved to $\mathcal{O}(dkn) + d^2 2^{(k/\varepsilon)^{\mathcal{O}(1)}} \log^{k+2}(n)$ in [Chen, 2009], and to

$\mathcal{O}(dkn)+d\,2^{\tilde{\mathcal{O}}(k/\varepsilon)}\log^{k+2}(n)$ in [Ackermann and Blömer, 2009]. The asymptotically fastest $(1+\varepsilon)$-approximation algorithm currently known is given in [Feldman et al., 2007]. This algorithm combines the algorithm from [Kumar et al., 2004] with so called weak coresets to achieve a running time of $\mathcal{O}(dkn) + d(k/\varepsilon)^{\mathcal{O}(1)} + 2^{\tilde{\mathcal{O}}(k/\varepsilon)}$.

Bregman k-median problem. Prior to the work presented in this thesis, relatively little has been known about the complexity and geometry of the general Bregman k-median problem. Heuristic methods for k-median clustering by Kullback-Leibler divergence were first suggested in [Pereira et al., 1993]. In [Baker and McCallum, 1998], a simple agglomerative greedy strategy has been proposed which turns out to perform surprisingly well in empirical tests. Independently, a similar algorithm was stated in [Slonim and Tishby, 1999]. [Dhillon et al., 2003] proposed a local improvement heuristic for clustering by Kullback-Leibler divergence which is an adaptation of Lloyd's k-means algorithm. Earlier, [Buzo et al., 1980] already proposed the use of Lloyd's algorithm for k-median clustering by Itakura-Saito divergence in the context of vector quantization. In a breakthrough result in [Banerjee et al., 2005b], Lloyd's algorithm has been generalized to the whole class of all Bregman divergences. Hence, the authors gave a unified explanation for the earlier observations regarding the Itakura-Satito divergence and the the Kullback-Leibler divergence. However, all these recent strategies lack any provable approximation ratio and rely solely on empirical evaluation.

A first $(1+\varepsilon)$-approximation algorithm applicable to the Bregman k-median problem has been proposed in [Ackermann et al., 2008]. This result generalizes an earlier algorithm from [Kumar et al., 2004] for the squared Euclidean distances to a large number of Bregman divergences and some other dissimilarity measures under the assumption that the dissimilarity measure used satisfies a certain statistical property. The algorithm from [Ackermann et al., 2008] uses at most $2^{\tilde{\mathcal{O}}(k/\varepsilon)}n$ arithmetic operations, including evaluations of Bregman divergence D_φ. Using the coresets from [Chen, 2006], this running time was later improved to using at most $\mathcal{O}(kn)+2^{\tilde{\mathcal{O}}(k/\varepsilon)}\log^{k+2}(n)$ arithmetic operations in [Ackermann and Blömer, 2009].

Furthermore, the seeding technique from [Arthur and Vassilvitskii, 2007] has been generalized to the class of Bregman divergences independently in at least three different publications (see [Ackermann and Blömer, 2009], [Nock et al., 2008], and [Sra et al., 2008]). In addition, it has been shown in [Ackermann and Blömer, 2010] that this seeding technique computes a constant factor approximate solution, provided that the input set consists of k well separated clusters.

Recently, a polynomial time reduction from the Kullback-Leibler k-median problem to the Euclidean k-means problem has been proposed in [Chaudhuri and McGregor, 2008]. This reduction leads to a polynomial time $\mathcal{O}(\log n)$-approximation algorithm. The interesting aspect of this result is that it does not rely on any of the assumptions made to achieve the results above or the results given in this thesis.

1.2 Outline and main results

This thesis is organized as follows.

Chapter 2: We start by giving a formal definition of the family of Bregman divergences. We study some common basic properties and give several examples of concrete, well-known Bregman divergences. Furthermore, we introduce our notion of μ-similar Bregman divergences, which is fundamental to many of the results given in this thesis. In a nutshell, μ-similar Bregman divergences are a subclass of the family of Bregman divergences that feature some quasi-metric properties. We also argue that, to some extend, all Bregman divergences that are used in practice are μ-similar, provided that their domain avoids certain singularities.

Chapter 3: In this chapter, we introduce our generalized formulation of the k-median problem. We investigate properties of the generalized k-median problem, as well as properties of the k-median problem using a Bregman divergence as dissimilarity measure. The results proven in this chapter are valid regardless of whether or not the dissimilarity measure used employs any (approximate) metric properties. Instead, the proofs in this chapter are based on the combinatorial properties of k-median clusterings. We conclude Chapter 3 by stating a simple algorithm solving the 1-dimensional Bregman k-median clustering problem optimally in time $\mathcal{O}(kn^3)$, where n denotes the number of input points. This algorithm is given to demonstrate the existence of efficient k-median clustering algorithms that do not rely on metric properties such as symmetry or the triangle inequality.

Chapter 4: In Chapter 4, we present the first main result of this thesis. In detail, we give a randomized $(1+\varepsilon)$-approximation algorithm for the generalized k-median problem using at most $n2^{\mathcal{O}(mk\log(mk/\varepsilon))}$ arithmetic operations (including evaluations of dissimilarity measure D), under the assumption that D satisfies a certain statistical sampling property. Here, n denotes the number of input points and m is a constant that depends only on ε and D. The analysis of this algorithm is purely combinatorial, and does not rely on metric properties such as symmetry or triangle inequality. Hence, our algorithm is well-suited for the k-median problem using metric as well as non-metric dissimilarity measures.

In addition to the result above, we show that the necessary sampling property is satisfied by a large number of dissimilarity measures, including the squared Euclidean distance, the Mahalanobis distances, the Kullback-Leibler divergence, and all other μ-similar Bregman divergences. We also show that the sampling property is satisfied for all metrics with bounded doubling dimension as well, provided that finding the 1-median of a given set is a feasible problem. These sampling results are the second main contribution of this chapter. In doing so, we obtain the first $(1+\varepsilon)$-approximation algorithm known for a large number of non-metric dissimilarity measures, such as the Kullback-Leibler divergence, the Itakura-Saito divergence, and all other μ-similar Bregman divergences.

The results presented in this chapter have been published previously in [Ackermann et al., 2008] and [Ackermann et al., 2010a].

Chapter 5: Unfortunately, the algorithm from Chapter 4 turns out to be not very practical due to the huge constants involved in the running time. Hence, in this chapter, we state and analyze a practical, randomized $\mathcal{O}(\log k)$-approximate algorithm applicable to the k-median problem using an arbitrary μ-similar Bregman divergence as dissimilarity measure. This algorithm uses at most $\mathcal{O}(kn)$ arithmetic operations, including evaluations of Bregman divergence D_φ, and can be implemented to run quite fast in practice. This result has already appeared in [Ackermann and Blömer, 2009].

Furthermore, empirical evaluation of the algorithm from this chapter on real world data indicates a better approximation ratio than the theoretically proven bound of $\mathcal{O}(\log k)$. To give a theoretical justification for this observation, we analyze our algorithm in the practically relevant case of input instances that consist of a number of k well separated input instances. In detail, we show that in this case with constant probability the algorithm computes an $\mathcal{O}(1)$-approximate solution of the μ-similar Bregman k-median problem. This result has also been published in [Ackermann and Blömer, 2010].

Chapter 6: In the recent years, the use of so called coresets has become a standard technique in computational geometry. A (k, ε)-coreset for an input set is a small (weighted) set such that for any set of k cluster centers the (weighted) k-median clustering cost of the coreset is an approximation for the clustering cost of the original set with relative error at most ε. The goal of a good coreset construction is to set up coresets that are significantly smaller than the original input set. These coresets can be used to speed up existing approximation algorithms, especially if the running time of these algorithms depends strongly on the number of input points.

In Chapter 6, we show how coresets for the Euclidean k-means problem can be generalized to the case of Mahalanobis k-median clustering. In particular, we give a generalization of the deterministic coreset construction from [Har-Peled and Mazumdar, 2004] to Mahalanobis distances. These (k, ε)-coresets are of size $2^{\mathcal{O}(d \log d)} \varepsilon^{-d} k \log n$, where n denotes the number of points of the original set and d the dimension.

In addition, we also give a new, randomized coreset construction for the Mahalanobis k-median problem. This new construction is rather easy to implement and relies solely on random sampling. We prove that, with high probability, our construction yields (k, ε)-coresets of size $2^{\mathcal{O}(d \log d)} \varepsilon^{-d} k \log(n) \log^{d/2}\big(\varepsilon^{-1} k \log(n)\big)$.

A preliminary version of the results from this chapter can be found in [Ackermann et al., 2010b].

Chapter 7: Up to date, there is no coreset construction known for the Bregman k-median problem using Bregman divergences other than Mahalanobis distances. However, it turns out that the classical definition of (strong) coresets as given in Chapter 6 is unnecessary strict: If a set of cluster centers will never turn up as output of the algorithm we use, we don't care whether or not the clustering cost of the original set and the coreset are approximately the same. Thus, in

Chapter 7, we consider a relaxed notion of coresets where the center points only have to come from a fixed and finite set Γ, which we call a Γ-weak coreset.

In this chapter, we give a randomized coreset construction that computes Γ-weak coresets for the μ-similar Bregman k-median problem. With high probability, we obtain that our construction yields Γ-weak (k, ε)-coresets of size $\mathcal{O}\left(\varepsilon^{-2} k \log(n) \log(k|\Gamma|^k \log n)\right)$. Hence, we are able to give the first construction of weak coresets that is applicable to all μ-similar Bregman divergences.

Moreover, we show how these weak coresets can be used to speed up existing approximation algorithms. We use this approach to improve the running time of the $(1+\varepsilon)$-approximation algorithm from Chapter 4. In doing so, we give the asymptotically fastest algorithm currently known for the k-median problem with respect to an arbitrary μ-similar Bregman divergence D_φ which requires at most $\mathcal{O}(kn) + 2^{\mathcal{O}(k/\varepsilon \log(k/\varepsilon))} \log^{k+2}(n)$ arithmetic operations, including evaluations of D_φ.

The results we give in this chapter have already been published in [Ackermann and Blömer, 2009].

Chapter 8: In the previous chapters, we presented randomized approximation algorithms for the Bregman k-median problem that make use of the sampling of a constant number of elements uniformly at random from a given point set. To obtain our approximation guarantees, we always assumed that the Bregman divergence used is μ-similar. The question arises whether this assumption is actually necessary to make use of the uniform sampling technique. In this chapter, by analyzing the common analytical properties of the family of Bregman divergences, we provide strong evidence that this is indeed the case. Hence, we conjecture that in the case of a Bregman divergence, μ-similarity is indeed necessary to make use of the sampling techniques employed in this thesis.

In addition, we show that this intuition can be made explicit by taking into account the concrete analytical properties of a given Bregman divergence. More precisely, we prove that in the case of the Kullback-Leibler divergence and the Itakura-Saito divergence, the assumption of μ-similarity is indeed necessary to make use of the uniform sampling technique.

Appendix A: In this short appendix, we turn our attention to several practical applications that benefit from efficient Bregman clustering algorithms. We present three practical scenarios from the fields of statistical inference, data compression, and speech processing, respectively. Using a minimum of formalism, we show how these three different applications are related to solving a k-median problem with respect to three different Bregman divergences.

Appendix B: In this mathematical appendix, we give a brief overview of the mathematical fundamentals that are assumed to be common knowledge throughout this thesis. These fundamentals include well-known, basic results from the areas of vector algebra, calculus, and probability theory. The facts in this appendix are stated as concise as possible, and without any formal proof.

1.3 Bibliographic notes

Parts of the work given in this thesis have been published earlier, although some of the results in this thesis are presented in a more general setting. These earlier publications are:

[Ackermann et al., 2008] Ackermann, M. R., Blömer, J., and Sohler, C. (2008). Clustering for metric and non-metric distance measures. In *Proceedings of the 19th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '08)*, pages 799–808. Society for Industrial and Applied Mathematics.

[Ackermann and Blömer, 2009] Ackermann, M. R. and Blömer, J. (2009). Coresets and approximate clustering for Bregman divergences. In *Proceedings of the 20th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '09)*, pages 1088–1097. Society for Industrial and Applied Mathematics.

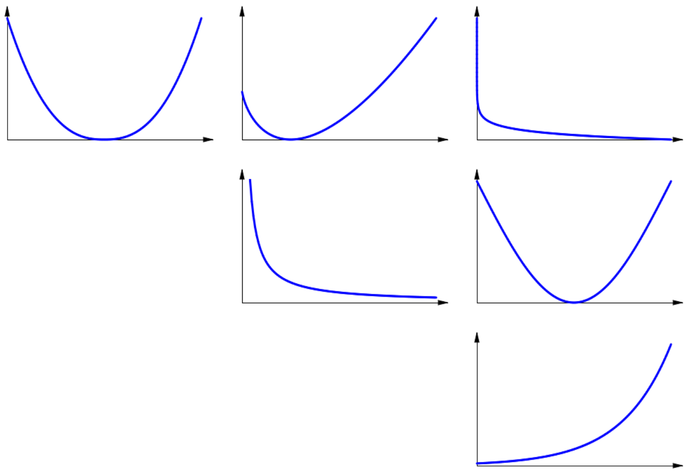
[Ackermann and Blömer, 2010] Ackermann, M. R. and Blömer, J. (2010). Bregman clustering for separable instances. In *Proceedings of the 12th Scandinavian Symposium and Workshop on Algorithm Theory (SWAT '10)*. Springer. To appear.

[Ackermann et al., 2010a] Ackermann, M. R., Blömer, J., and Sohler, C. (2010a). Clustering for metric and non-metric distance measures.

ACM Transactions on Algorithms. Special issue on SODA '08. To appear.

[Ackermann et al., 2010b] Ackermann, M. R., Lammersen, C., Märtens, M., Raupach, C., Sohler, C., and Swierkot, K. (2010b). StreamKM++: A clustering algorithm for data streams. In *Proceedings of the Workshop on Algorithm Engineering and Experiments (ALENEX '10)*, pages 175–187. Society for Industrial and Applied Mathematics.

2 Bregman divergences



The dissimilarity measures known as Bregman divergences were first proposed in 1967 by Lev M. Bregman (cf. [Bregman, 1967]). Bregman divergences were originally introduced as generalized distances in the context of solving convex optimization problems. In Bregman's method, an arbitrary initial solution is iteratively projected[1] on one of the convex sets that correspond to the given convex constraints. A number of sufficient conditions for a dissimilarity function were formulated such that these iterated projections converge to a common point in the intersection of these sets. The term "Bregman distance" was later coined in [Censor and Lent, 1981] to describe the family of dissimilarity measure satisfying these properties. For an in-depth study of the optimization problems involving Bregman divergences, the reader is directed to [Censor and Zenios, 1997].

[1] In this generalized sense, a *projection* of a point x onto a convex set S with respect to a given dissimilarity measure is the (unique) point from S that is closest to x.

Intuitively, a Bregman divergence arises as an error function when approximating a strictly convex function by a tangent hyperplane. In the context of clustering problems, the interesting aspect of these divergences is that they give a generalized description of a wide array of dissimilarity functions that are frequently used in practice. In addition, the family of Bregman divergences shares a number of analytical and combinatorial properties, as is discussed in the following sections of this chapter.

This chapter is organized as follows. First, we give a formal definition of the family of Bregman divergences in Section 2.1. We also state and prove a number of basic properties common for all Bregman divergences. In addition, we present several concrete examples of well-known Bregman divergences, such as the squared Euclidean distance, the Kullback-Leibler divergence, and the Itakura-Saito divergence.

In Section 2.2, we introduce an important subclass of the class of Bregman divergences, namely the class of Mahalanobis distances. We argue that, to some extend, Mahalanobis distances are exactly the instances of Bregman divergences that feature certain well-natured geometric properties. In addition, we also identify a very large subclass of the Bregman divergences that share approximately the same geometric properties as the Mahalanobis distances do. We call these Bregman divergences μ-similar. We also provide evidence that most of the Bregman divergences used in practice are μ-similar. The notion of μ-similarity plays an important role throughout the rest of this thesis.

2.1 Definition

A Bregman divergence D_φ is defined with respect to a strictly convex function $\varphi : \mathrm{ri}(\mathbb{X}) \to \mathbb{R}$ on the relative interior $\mathrm{ri}(\mathbb{X})$ of convex domain $\mathbb{X}$.

Definition 2.1. *Let $\mathbb{X} \subseteq \mathbb{R}^d$ be a convex and non-singleton set. A function $\varphi : \mathrm{ri}(\mathbb{X}) \to \mathbb{R}$ is called a* (Bregman) generating function *if the following conditions are satisfied:*

a) φ is strictly convex on $\mathrm{ri}(\mathbb{X})$.

b) φ has continuous first-order partial derivatives on $\mathrm{ri}(\mathbb{X})$.

Intuitively, the Bregman divergence with respect to φ from point p towards point q can be seen as the error when approximating $\varphi(p)$ by using the tangent hyperplane of φ at point q (see Figure 2.1). We will use the

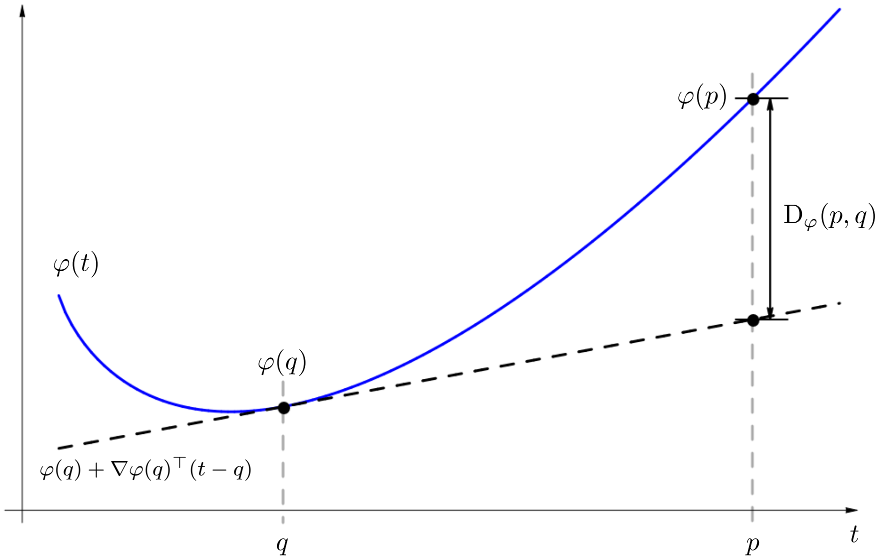

$\varphi(p)$
$\mathrm{D}_\varphi(p,q)$
$\varphi(t)$
$\varphi(q)$
$\varphi(q) + \nabla\varphi(q)^\top(t-q)$
q
p
t

Figure 2.1: A geometric interpretation of the Bregman divergence D_φ with convex generating function φ. A linear approximation of $\varphi(t)$ is given by $\varphi(q) + \nabla\varphi(q)^\top(t-q)$, i.e., by the tangent hyperplane of the graph of φ at point $\big(q, \varphi(q)\big)$. The Bregman divergence $\mathrm{D}_\varphi(p,q)$ is obtained as the error when approximating $\varphi(p)$ by this tangent hyperplane.

following formal definition.

Definition 2.2. *Let $\mathbb{X} \subseteq \mathbb{R}^d$ be a convex and non-singleton set, and let $\varphi : \mathrm{ri}(\mathbb{X}) \to \mathbb{R}$ be a generating function. For $t = (t_1, t_2, \ldots, t_d)^\top \in \mathrm{ri}(\mathbb{X})$ let $\nabla\varphi(t)$ denote the* gradient *of φ at point t, i.e.,*

$$\nabla\varphi(t) = \begin{pmatrix} \frac{\partial}{\partial t_1}\varphi(t) \\ \frac{\partial}{\partial t_2}\varphi(t) \\ \vdots \\ \frac{\partial}{\partial t_d}\varphi(t) \end{pmatrix} . \tag{2.1}$$

The Bregman divergence *with respect to φ is defined as*

$$\mathrm{D}_\varphi(p,q) = \varphi(p) - \varphi(q) - \nabla\varphi(q)^\top(p-q) \tag{2.2}$$

for all $p, q \in \mathrm{ri}(\mathbb{X})$.

The generating function φ and the Bregman divergence D_φ are defined on the relative interior of $\mathbb{X}$. However, it is convenient to extend these definitions to the whole domain $\mathbb{X}$. That is, we assume that the extension $\varphi : \mathbb{X} \to \mathbb{R} \cup \{\infty\}$ is well defined by

$$\varphi(t) = \lim_{\lambda \to 0^+} \varphi\big(\lambda x + (1-\lambda)t\big) \ , \tag{2.3}$$

for all $t \in \mathbb{X} \setminus \mathrm{ri}(\mathbb{X})$ and an arbitrary $x \in \mathrm{ri}(\mathbb{X})$. Furthermore, we assume that the extension $\mathrm{D}_\varphi : \mathbb{X} \times \mathbb{X} \to \mathbb{R} \cup \{\infty\}$ is well defined by

$$\mathrm{D}_\varphi(p,q) = \lim_{\lambda \to 0^+} \mathrm{D}_\varphi\big(\lambda x + (1-\lambda)p, \lambda x + (1-\lambda)q\big) \tag{2.4}$$

for $p, q \in \mathbb{X}$ with $p \notin \mathrm{ri}(\mathbb{X})$ or $q \notin \mathrm{ri}(\mathbb{X})$ and an arbitrary $x \in \mathrm{ri}(\mathbb{X})$. In doing so, we potentially introduce points $p, q \in \mathbb{X}$ such that $\mathrm{D}_\varphi(p,q) = \infty$. That is, D_φ may possess *singularities* if at least one of the arguments lies on the boundary of $\mathbb{X}$. Also note that we have $\mathrm{D}_\varphi(p,q) < \infty$ as long as $p, q \in \mathrm{ri}(\mathbb{X})$.

2.1.1 Basic properties

In this section, we prove a number of basic properties common for all Bregman divergences, such as non-negativity, convexity, or the ambiguity of the generating function. All the following properties are well-known in literature and have been discussed before (for instance, see [Bregman, 1967], [Csiszár, 1991], or [Banerjee et al., 2005b]). Some of these properties such as the non-negativity or the Lagrange form of Bregman divergences are crucial for the techniques employed in this thesis. Other properties are included for the sake of completeness.

It is easy to see that, in general, Bregman divergences are asymmetric and do not satisfy the triangle inequality (see the examples in Section 2.1.2 below). By allowing the partial derivatives $\frac{\partial}{\partial q_i}\varphi(q)$ to approach $\pm\infty$ for q on the boundary of $\mathbb{X}$, we obtain that D_φ may even possess singularities, that is, points $p, q \in \mathbb{X}$ such that $\mathrm{D}_\varphi(p,q) = \infty$.

However, from the strict convexity of the generating function φ follows that D_φ is non-negative and that $\mathrm{D}_\varphi(p,q) = 0$ if and only if $p = q$, as is stated in the following lemma. Intuitively, this lemma is equivalent to the observation that for a strictly convex function φ the first-order Taylor expansion of φ at point q always underestimates $\varphi(p)$.

Lemma 2.3 (non-negativity)**.** *For all Bregman divergences* D_φ *on domain* $\mathbb{X}$ *and for all* $p, q \in \mathrm{ri}(\mathbb{X})$ *we have*

$$\mathrm{D}_\varphi(p,q) \geq 0 \ . \tag{2.5}$$

Furthermore, $\mathrm{D}_\varphi(p, q) = 0$ *if and only if* $p = q$.

Proof. To prove the lemma we show that for all $p, q \in \mathrm{ri}(\mathbb{X})$ we have

$$\varphi(p) \geq \varphi(q) + \nabla\varphi(q)^\top (p - q) \ , \tag{2.6}$$

and that (2.6) holds with equality if and only if $p = q$.

First, we consider the case of dimension $d = 1$. Let $p, q \in \mathrm{ri}(\mathbb{X}) \subseteq \mathbb{R}$ be with $p \neq q$. Since $\mathbb{X}$ is convex we have $\lambda p + (1 - \lambda)q \in \mathrm{ri}(\mathbb{X})$ for any $0 < \lambda < 1$. Hence, using the convexity of φ we obtain

$$\lambda\varphi(p) + (1 - \lambda)\varphi(q) \geq \varphi\big(\lambda p + (1 - \lambda)q\big) = \varphi\big(q + \lambda(p - q)\big) \ . \tag{2.7}$$

Dividing both sides of (2.7) by λ leads to

$$\varphi(p) + \frac{1}{\lambda}\varphi(q) - \varphi(q) \geq \frac{1}{\lambda}\varphi\big(q + \lambda(p - q)\big) \ , \tag{2.8}$$

or, equivalently,

$$\varphi(p) \geq \varphi(q) + \frac{\varphi\big(q + \lambda(p - q)\big) - \varphi(q)}{\lambda(p - q)} (p - q) \ . \tag{2.9}$$

Note that φ is differentiable on $\mathrm{ri}(\mathbb{X})$, and that the quotient in the right hand side of inequality (2.9) gives the difference quotient of φ when $q + \lambda(p - q)$ approaches q. Hence, from elementary calculus it follows that by taking the limit $\lambda \to 0$ we obtain

$$\varphi(p) \geq \varphi(q) + \varphi'(q)(p - q) \ , \tag{2.10}$$

which proves inequality (2.6) in the case of dimension $d = 1$.

Now, we use the result for the one-dimensional case to prove inequality (2.6) in arbitrary dimensions $d \geq 1$. Again, let $p, q \in \mathrm{ri}(\mathbb{X})$ be with $p \neq q$. Let

$$\overline{pq} = \{x \in \mathbb{X} \,|\, \exists 0 \leq \lambda \leq 1 : x = \lambda p + (1 - \lambda)q\} \subseteq \mathrm{ri}(\mathbb{X}) \tag{2.11}$$

denote the line segment passing through p and q. Since $p, q \in \mathrm{ri}(\mathbb{X})$ we know that there exists an $\varepsilon > 0$ such that the extended line segment

$$\overline{pq}^\varepsilon = \{x \in \mathbb{X} \,|\, \exists {-\varepsilon} \leq \lambda \leq 1 + \varepsilon : x = \lambda p + (1 - \lambda)q\} \tag{2.12}$$

is still completely contained in $\mathrm{ri}(\mathbb{X})$. Furthermore, let $\psi : [-\varepsilon, 1+\varepsilon] \to \mathbb{R}$ with

$$\psi(\lambda) = \varphi\big(\lambda p + (1-\lambda)q\big) \tag{2.13}$$

be the restriction of φ on $\overline{pq}^{\varepsilon}$. Since φ is strictly convex on $\mathbb{X}$ it follows that ψ is also strictly convex on $[-\varepsilon, 1+\varepsilon]$. Also, we find that ψ is differentiable on $[0,1] \subseteq \mathrm{ri}([-\varepsilon, 1+\varepsilon])$. Hence, from inequality (2.10) we obtain

$$\psi(1) \geq \psi(0) + \psi'(0)(1-0) = \psi(0) + \psi'(0) \ . \tag{2.14}$$

Obviously, we have $\psi(1) = \varphi(p)$ and $\psi(0) = \varphi(q)$. In addition to that, for all $\lambda \in [0,1]$ we know that

$$\begin{aligned}\psi'(\lambda) &= \frac{\partial}{\partial(p-q)}\varphi\big(\lambda p + (1-\lambda)q\big) &(2.15)\\ &= \nabla\varphi\big(\lambda p + (1-\lambda)q\big)^{\top}(p-q) \ . &(2.16)\end{aligned}$$

Using inequality (2.14), we conclude

$$\varphi(p) \geq \varphi(q) + \nabla\varphi(q)^{\top}(p-q) \ , \tag{2.17}$$

and we obtain $\mathrm{D}_{\varphi}(p,q) \geq 0$ for all $p, q \in \mathrm{ri}(\mathbb{X})$.

Obviously, if $p = q$ we have

$$\mathrm{D}_{\varphi}(p,q) = \varphi(p) - \varphi(q) - \nabla\varphi(q)^{\top}(p-q) = 0 \ . \tag{2.18}$$

Now, assume for the sake of contradiction that there exist distinct points $p, q \in \mathrm{ri}(\mathbb{X})$ with $\mathrm{D}_{\varphi}(p,q) = 0$. Using the same notation as above, this leads to

$$\psi(1) - \psi(0) - \psi'(0) = \mathrm{D}_{\varphi}(p,q) = 0 \ . \tag{2.19}$$

Hence,

$$\psi(1) - \psi(0) = \psi'(0) \ . \tag{2.20}$$

On the other hand, by the mean value theorem we know that there exists a $0 < \xi < 1$ such that

$$\psi'(\xi) = \frac{\psi(1) - \psi(0)}{1-0} = \psi(1) - \psi(0) = \psi'(0) \ . \tag{2.21}$$

But this is a contradiction since function ψ is not strictly convex unless the first order derivate ψ' is strictly increasing. Hence, we find $\mathrm{D}_{\varphi}(p,q) = 0$ if and only if $p = q$. □

Furthermore, a Bregman divergence is always a convex function in its first argument, but not necessarily in its second argument.

Lemma 2.4 (convexity of first argument)**.** *Let* D_φ *be a Bregman divergence on domain* $\mathbb{X}$*. For all* $p, q, r \in \mathbb{X}$ *and* $0 \leq \lambda \leq 1$ *we have*

$$\mathrm{D}_\varphi\big(\lambda p + (1-\lambda)q, r\big) \leq \lambda\, \mathrm{D}_\varphi(p,r) + (1-\lambda)\, \mathrm{D}_\varphi(q,r)\ . \tag{2.22}$$

Proof. Using the convexity of φ and the bi-linearity of the inner product we obtain

$$\begin{aligned}
&\mathrm{D}_\varphi\big(\lambda p + (1-\lambda)q, r\big) \\
&\quad = \varphi\big(\lambda p + (1-\lambda)q\big) - \varphi(r) - \nabla\varphi(r)^\top\big(\lambda p + (1-\lambda)q - r\big) && (2.23)\\
&\quad \leq \lambda\varphi(p) + (1-\lambda)\varphi(q) - \varphi(r) - \nabla\varphi(r)^\top\big(\lambda p + (1-\lambda)q - r\big) && (2.24)\\
&\quad = \lambda\varphi(p) - \lambda\varphi(r) - \lambda\nabla\varphi(r)^\top(p-r) \\
&\qquad + (1-\lambda)\varphi(q) - (1-\lambda)\varphi(r) - (1-\lambda)\nabla\varphi(r)^\top(q-r) && (2.25)\\
&\quad = \lambda\, \mathrm{D}_\varphi(p,r) + (1-\lambda)\, \mathrm{D}_\varphi(q,r)\ . && (2.26)
\end{aligned}$$

□

It is an immediate consequence of Lemma 2.4 that for all points $c \in \mathbb{X}$ and all $r \geq 0$ the level set

$$U_\varphi(c,r) = \{x \in \mathbb{X} \mid \mathrm{D}_\varphi(x,c) \leq r\} \tag{2.27}$$

is a convex set. $U_\varphi(c,r)$ is occasionally called the *Bregman ball* with center c and radius r.

A Bregman divergence is well defined by a generating function φ and its domain $\mathbb{X}$. However, this definition is not unique. More precisely, if two generating functions differ only in affine terms, they define the same Bregman divergence. We obtain the following lemma.

Lemma 2.5 (ambiguity of the generating function)**.** *Let* $\varphi : \mathbb{X} \to \mathbb{R}$ *with* $\mathbb{X} \subseteq \mathbb{R}^d$ *be a strictly convex and differentiable function. Furthermore, let* $a \in \mathbb{R}^d$*,* $b \in \mathbb{R}$*, and let* $\psi : \mathbb{X} \to \mathbb{R}$ *be given by*

$$\psi(t) = \varphi(t) + a^\top t + b\ . \tag{2.28}$$

Then we have $\mathrm{D}_\varphi = \mathrm{D}_\psi$*.*

Proof. Since φ is strictly convex and differentiable on $\mathbb{X}$, we find that ψ is also strictly convex and differentiable on $\mathbb{X}$. In particular, for $t = (t_1, t_2, \dots, t_d) \in \mathbb{X}$ and $a = (a_1, a_2, \dots, a_d) \in \mathbb{R}^d$ we have

$$\frac{\partial}{\partial t_i}\psi(t) = \frac{\partial}{\partial t_i}\big(\varphi(t) + a^\top t + b\big) = \frac{\partial}{\partial t_i}\varphi(t) + a_i \ . \tag{2.29}$$

Hence,

$$\nabla\psi(t) = \nabla\varphi(t) + a \ . \tag{2.30}$$

Therefore, for arbitrary $p, q \in \mathbb{X}$ we obtain

$$\begin{aligned} \mathrm{D}_\psi(p,q) &= \psi(p) - \psi(q) - \nabla\psi(q)^\top(p-q) && (2.31)\\ &= \varphi(p) + a^\top p + b - \varphi(q) - a^\top q - b - \nabla\varphi(t)^\top(p-q) - a^\top(p-q) && (2.32)\\ &= \varphi(p) - \varphi(q) - \nabla\varphi(q)^\top(p-q) && (2.33)\\ &= \mathrm{D}_\varphi(p,q) \ . && (2.34) \end{aligned}$$

□

Furthermore, we find that the operator mapping φ on its corresponding Bregman divergence D_φ is a linear operator.

Lemma 2.6 (linearity of the Bregman operator)**.** *Let $\varphi, \psi : \mathbb{X} \to \mathbb{R}$ be strictly convex and differentiable functions, and let $\alpha, \beta > 0$ be arbitrary. Then we have*

$$\mathrm{D}_{\alpha\varphi+\beta\psi} = \alpha\,\mathrm{D}_\varphi + \beta\,\mathrm{D}_\psi \ . \tag{2.35}$$

Proof. Note that a positive linear combination of strictly convex and differentiable functions is also strictly convex and differentiable. In particular, for $t = (t_1, t_2, \dots, t_d) \in \mathbb{X}$ we have

$$\frac{\partial}{\partial t_i}\big(\alpha\psi + \beta\psi\big)(t) = \alpha\frac{\partial}{\partial t_i}\varphi(t) + \beta\frac{\partial}{\partial t_i}\psi(t) \ . \tag{2.36}$$

Hence,

$$\nabla\big(\alpha\varphi + \beta\psi\big)(t) = \alpha\nabla\varphi(t) + \beta\nabla\psi(t) \ . \tag{2.37}$$

Therefore, for arbitrary $p, q \in \mathbb{X}$ we obtain

$$\begin{aligned} \mathrm{D}_{\alpha\psi+\beta\psi}(p,q) &= \alpha\varphi(p) + \beta\varphi(p) - \alpha\varphi(q) - \beta\psi(q) \\ &\quad - \alpha\nabla\varphi(t)^\top(p-q) + \beta\nabla\psi(t)^\top(p-q) && (2.38) \\ &= \alpha\,\mathrm{D}_\varphi(p,q) + \beta\,\mathrm{D}_\psi(p,q)\ . && (2.39) \end{aligned}$$

□

Finally, in the following lemma we give formal proof to an elementary yet crucial observation: Since $\mathrm{D}_\varphi(p,q)$ equals the remainder term of the first-order Taylor expansion of $\varphi(p)$ at point q, the Bregman divergence D_φ can be expressed in terms of the Hessian matrix of φ. While this fact has been known for some time, it is the keystone to a novel interpretation of Bregman divergences which is fundamental to many of the results obtained in this thesis. This interpretation is discussed in detail in Section 2.2.

Lemma 2.7 (Lagrange form of Bregman divergences). *Let D_φ be a Bregman divergence on domain $\mathbb{X}$ with a twice differentiable generating function $\varphi : \mathbb{X} \to \mathbb{R}$. Furthermore, for $t = (t_1, t_2, \ldots, t_d)^\top \in \mathbb{X}$ let*

$$\nabla^2\varphi(t) = \begin{pmatrix} \frac{\partial^2}{\partial t_1^2}\varphi(t) & \frac{\partial^2}{\partial t_1 t_2}\varphi(t) & \cdots & \frac{\partial^2}{\partial t_1 t_d}\varphi(t) \\ \frac{\partial^2}{\partial t_2 t_1}\varphi(t) & \frac{\partial^2}{\partial t_2^2}\varphi(t) & \cdots & \frac{\partial^2}{\partial t_2 t_d}\varphi(t) \\ \vdots & & & \\ \frac{\partial^2}{\partial t_d t_1}\varphi(t) & \frac{\partial^2}{\partial t_d t_2}\varphi(t) & \cdots & \frac{\partial^2}{\partial t_d^2}\varphi(t) \end{pmatrix} \qquad (2.40)$$

denote the Hessian matrix of φ at point t. Then for all $p, q \in \mathbb{X}$ there exists a point ξ on the line segment through p and q such that

$$\mathrm{D}_\varphi(p,q) = \frac{1}{2}(p-q)^\top\nabla^2\varphi(\xi)\,(p-q)\ . \qquad (2.41)$$

Proof. Consider the first-order Taylor expansion of $\varphi(p)$ at point q, that is,

$$\varphi(p) = \varphi(q) + \nabla\varphi(q)(p-q) + R_1(p)\ , \qquad (2.42)$$

where $R_1(p)$ denotes the remainder term of the first-order Taylor expansion. Using the Lagrange form of the remainder term we obtain that there exists a point ξ on the line segment through p and q such that

$$\mathrm{D}_\varphi(p,q) = R_1(p) = \frac{1}{2}(p-q)^\top\nabla^2\varphi(\xi)\,(p-q)\ . \qquad (2.43)$$

□

2.1.2 Examples of Bregman divergences

The class of all Bregman divergences includes a number of prominent dissimilarity measures. Here, we give three important examples: the squared Euclidean distance, the Kullback-Leibler divergence and the Itakura-Saito divergence. An overview of more Bregman divergences can be found in Figure 2.2.

Example 1: Squared Euclidean distance. Most notably, the *square of the Euclidean distance*

$$\mathrm{D}_{\ell_2^2}(p,q) = \|p-q\|^2 = \sum_{i=1}^{d}(p_i - q_i)^2 \tag{2.44}$$

with $p = (p_1, p_2, \ldots, p_d)^\top \in \mathbb{R}^d$ and $q = (q_1, q_2, \ldots, q_d)^\top \in \mathbb{R}^d$ is a Bregman divergence, as is stated in the following lemma.

Lemma 2.8. *The square of the Euclidean distance* $\mathrm{D}_{\ell_2^2}$ *on domain* $\mathbb{R}^d$ *is a Bregman divergence by means of generating function*

$$\varphi_{\ell_2^2}(t) = \|t\|^2 \ . \tag{2.45}$$

Proof. Let $t = (t_1, t_2, \ldots, t_d)^\top \in \mathbb{R}^d$. The function $\varphi_{\ell_2^2} = \sum_{i=1}^{d} t_i^2$ has continuous and differentiable first order partial derivatives

$$\frac{\partial}{\partial t_i}\varphi_{\ell_2^2}(t) = 2t_i \tag{2.46}$$

for all $i = 1, 2, \ldots, d$. Furthermore, the function $\varphi_{\ell_2^2}(t)$ has constant second order partial derivatives

$$\frac{\partial^2}{\partial t_i \partial t_j}\varphi_{\ell_2^2}(t) = \begin{cases} 2 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \ . \tag{2.47}$$

Hence, for all $t \in \mathbb{R}^d$ we have

$$\nabla^2 \varphi_{\ell_2^2}(t) = \begin{pmatrix} 2 & 0 & \cdots & 0 \\ 0 & 2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 2 \end{pmatrix} = 2I_d \ . \tag{2.48}$$

domain $\mathbb{X}$	$\varphi(t)$	$\mathrm{D}_\varphi(p,q)$
$\mathbb{R}^d$	squared ℓ_2-norm $\lVert t \rVert_2^2$	squared Euclidean distance $\lVert p-q \rVert_2^2$
$\mathbb{R}^d$	generalized norm $t^\top A t$	Mahalanobis distance $(p-q)^\top A(p-q)$
$\mathbb{R}^d_{\geq 0}$	neg. Shannon entropy $\sum t_i \ln(t_i) - t_i$	Kullback-Leibler divergence $\sum p_i \ln(\frac{p_i}{q_i}) - p_i + q_i$
$\mathbb{R}^d_{\geq 0}$	Burg entropy $-\sum \ln(t_i)$	Itakura-Saito divergence $\sum \frac{p_i}{q_i} - \ln(\frac{p_i}{q_i}) - 1$
$[-\frac{\pi}{2}, \frac{\pi}{2}]^d$	negative cosine $-\sum \cos t_i$	trigonometric divergence $\sum \cos(q_i) - \cos(p_i) - (p_i - q_i)\sin(q_i)$
$\mathbb{R}^d$	harmonic ($\alpha > 0$) $\sum \frac{1}{t_i^\alpha}$	harmonic divergence ($\alpha > 0$) $\sum \frac{1}{p_i^\alpha} - \frac{\alpha+1}{q_i^\alpha} + \frac{\alpha p_i}{q_i^{\alpha+1}}$
$\mathbb{R}^d$	norm-like ($\alpha \geq 2$) $\sum t_i^\alpha$	norm-like divergence ($\alpha \geq 2$) $\sum p_i^\alpha + (\alpha-1) q_i^\alpha - \alpha p_i q_i^{\alpha-1}$
$\mathbb{R}^d$	exponential $\sum \exp(t_i)$	exponential divergence $\sum \exp(p_i) - (p_i - q_i + 1)\exp(q_i)$
$\mathbb{R}^d$	reciprocal exponential $\sum \exp(-t_i)$	reciprocal exponential divergence $\sum \exp(-p_i) - (p_i - q_i + 1)\exp(-q_i)$
$[0,1]^d$	bit entropy $\sum t_i \ln t_i + (1-t_i)\ln(1-t_i)$	logistic loss $\sum p_i \ln \frac{p_i}{q_i} + (1-p_i)\ln \frac{1-p_i}{1-q_i}$
$\mathbb{R}^d$	dual bit entropy $\sum \ln\big(1 + \exp(t_i)\big)$	dual logistic loss $\sum \ln \frac{1+\exp(p_i)}{1+\exp(q_i)} - (p_i - q_i)\frac{\exp(q_i)}{1+\exp(q_i)}$
$[-1,1]^d$	Hellinger-like $-\sum \sqrt{1-t_i^2}$	Hellinger-like divergence $\sum \frac{1-p_i q_i}{\sqrt{1-q_i^2}} - \sqrt{1-p_i^2}$

Figure 2.2: An overview of some Bregman divergences.

Therefore, for each $t \in \mathbb{R}^d$ and for each $x = (x_1, x_2, \ldots, x_d)^\top \in \mathbb{R}^d \setminus \{0\}$ we obtain

$$x^\top \nabla^2 \varphi_{\ell_2^2}(t)\, x = 2x^\top x = 2 \sum_{i=1}^{d} x_i^2 > 0 \ . \tag{2.49}$$

Thus, $\nabla^2 \varphi_{\ell_2^2}(t)$ is a symmetric positive definite matrix for each $t \in \mathbb{R}^d$, and we conclude that $\varphi_{\ell_2^2}$ is strictly convex. Hence, $\varphi_{\ell_2^2}$ is a generating function. Furthermore, for each $p, q \in \mathbb{R}^d$ with $p = (p_1, p_2, \ldots, p_d)^\top$ and $q = (q_1, q_2, \ldots, q_d)^\top$ and for the Bregman divergence using $\varphi_{\ell_2^2}$ as generating function we have

$$\begin{aligned}
\mathrm{D}_{\varphi_{\ell_2^2}}(p, q) &= \varphi_{\ell_2^2}(p) - \varphi_{\ell_2^2}(q) - \nabla \varphi_{\ell_2^2}(q)^\top (p - q) && (2.50)\\
&= \sum_{i=1}^{d} p_i^2 - q_i^2 - 2q_i(p_i - q_i) && (2.51)\\
&= \sum_{i=1}^{d} p_i^2 - 2p_i q_i + q_i^2 && (2.52)\\
&= \sum_{i=1}^{d} (p_i - q_i)^2 && (2.53)\\
&= \|p - q\|^2 \ . && (2.54)
\end{aligned}$$

□

The squared Euclidean distance is symmetric (by virtue of the Euclidean distance being a metric) but does not satisfy the triangle inequality. For instance, in dimension $d = 1$ for points $0, 1, 2 \in \mathbb{R}$ we have

$$\mathrm{D}_{\ell_2^2}(0, 2) = 4 > 2 = \mathrm{D}_{\ell_2^2}(0, 1) + \mathrm{D}_{\ell_2^2}(1, 2) \ . \tag{2.55}$$

However, the squared Euclidean distance always satisfies the triangle inequality within a factor of two.

Lemma 2.9. *For all $p, q, r \in \mathbb{R}^d$ we have*

$$\mathrm{D}_{\ell_2^2}(p, q) \le 2\,\mathrm{D}_{\ell_2^2}(p, r) + 2\,\mathrm{D}_{\ell_2^2}(r, q) \ . \tag{2.56}$$

Proof. Using the triangle inequality of the Euclidean distance, we obtain

$$\begin{aligned}
\mathrm{D}_{\ell_2^2}(p,q) &= \|p-q\|^2 && (2.57)\\
&\le \big(\|p-r\|+\|r-q\|\big)^2 && (2.58)\\
&= \|p-r\|^2+\|r-q\|^2+2\|p-r\|\|r-q\| \; . && (2.59)
\end{aligned}$$

Now, note that

$$\begin{aligned}
0 &\le \big(\|p-r\|-\|r-q\|\big)^2 && (2.60)\\
&= \|p-r\|^2+\|r-q\|^2-2\|p-r\|\|r-q\| && (2.61)
\end{aligned}$$

which leads to

$$2\|p-r\|\|r-q\| \le \|p-r\|^2+\|r-q\|^2 \; . \qquad (2.62)$$

Hence, using inequality (2.59) and (2.62) we conclude

$$\begin{aligned}
\mathrm{D}_{\ell_2^2}(p,q) &\le 2\|p-r\|^2+2\|r-q\|^2 && (2.63)\\
&= 2\,\mathrm{D}_{\ell_2^2}(p,r)+2\,\mathrm{D}_{\ell_2^2}(r,q) \; . && (2.64)
\end{aligned}$$

□

Furthermore, the squared Euclidean distance does not possess singularities on $\mathbb{R}^d$, that is, for all $p,q \in \mathbb{R}^d$ we have $\mathrm{D}_{\ell_2^2}(p,q) < \infty$. Hence, the squared Euclidean distance is one of the more geometrically tractable dissimilarity measures among the whole class of Bregman divergences.

Example 2: Kullback-Leibler divergence. An important distance measure that has many applications in information theory and statistics is the *Kullback-Leibler divergence* (see [Kullback and Leibler, 1951]), which is also known as the relative entropy or the I-divergence. The (generalized) Kullback-Leibler divergence for points $p = (p_1, p_2, \ldots, p_d)^\top \in \mathbb{R}^d_{\ge 0}$ and $q = (q_1, q_2, \ldots, q_d)^\top \in \mathbb{R}^d_{\ge 0}$ is defined as

$$\mathrm{D}_{\mathrm{KL}}(p,q) = \sum_{i=1}^{d} \left(p_i \ln \frac{p_i}{q_i} - p_i + q_i \right) \; . \qquad (2.65)$$

Here, we use the common convention that for $q_i = 0$, we have

$$p_i \ln \frac{p_i}{q_i} = \begin{cases} 0 & \text{if } p_i = 0 \\ \infty & \text{if } p_i > 0 \end{cases} \; . \qquad (2.66)$$

It is known that the Kullback-Leibler divergence is also a Bregman divergence.

Lemma 2.10. *The Kullback-Leibler divergence* D_{KL} *on domain* $\mathbb{R}^d_{\geq 0}$ *is a Bregman divergence by means of generating function*

$$\varphi_{KL}(t) = \sum_{i=1}^{d} t_i \ln t_i - t_i \ . \tag{2.67}$$

Proof. Let $t = (t_1, t_2, \dots, t_d)^\top \in \mathbb{R}^d_{\geq 0}$. The function φ_{KL} has continuous and differentiable first order partial derivatives

$$\frac{\partial}{\partial t_i} \varphi_{\mathrm{KL}}(t) = \ln t_i \tag{2.68}$$

for all $i = 1, 2, \dots, d$. Furthermore, the function $\varphi_{\mathrm{KL}}(t)$ has second order partial derivatives

$$\frac{\partial^2}{\partial t_i \partial t_j} \varphi_{\mathrm{KL}}(t) = \begin{cases} \frac{1}{t_i} & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} . \tag{2.69}$$

Hence, for all $t \in \mathbb{R}^d_{\geq 0}$ we have

$$\nabla^2 \varphi_{\mathrm{KL}}(t) = \begin{pmatrix} \frac{1}{t_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{t_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{t_d} \end{pmatrix} . \tag{2.70}$$

Therefore, for each $t \in \mathbb{R}^d_{\geq 0}$ and for each $x = (x_1, x_2, \dots, x_d)^\top \in \mathbb{R}^d \setminus \{0\}$ we obtain

$$x^\top \nabla^2 \varphi_{\mathrm{KL}}(t)\, x = \sum_{i=1}^{d} \frac{x_i^2}{t_i} > 0 \ . \tag{2.71}$$

Thus, $\nabla^2 \varphi_{\mathrm{KL}}(t)$ is a symmetric positive definite matrix for each $t \in \mathbb{R}^d_{\geq 0}$, and we conclude that φ_{KL} is strictly convex. Hence, φ_{KL} is a generating function. Furthermore, for each $p, q \in \mathbb{R}^d_{\geq 0}$ with $p = (p_1, p_2, \dots, p_d)^\top$ and $q = (q_1, q_2, \dots, q_d)^\top$ and for the Bregman divergence using φ_{KL} as generating function we have

$$\begin{aligned}
\mathrm{D}_{\varphi_{\mathrm{KL}}}(p,q) &= \varphi_{\mathrm{KL}}(p) - \varphi_{\mathrm{KL}}(q) - \nabla\varphi_{\mathrm{KL}}(q)^\top(p-q) && (2.72)\\
&= \sum_{i=1}^{d} \left(p_i \ln p_i - p_i - q_i \ln q_i + q_i - \ln(q_i)(p_i - q_i)\right) && (2.73)\\
&= \sum_{i=1}^{d} \left(p_i \ln p_i - p_i \ln q_i - p_i + q_i\right) && (2.74)\\
&= \sum_{i=1}^{d} \left(p_i \ln \frac{p_i}{q_i} - p_i + q_i\right) \ . && (2.75)
\end{aligned}$$

□

It can easily be seen that, in general, the Kullback-Leibler divergence is asymmetric. Even worse, the discrepancy between $\mathrm{D}_{\mathrm{KL}}(p,q)$ and $\mathrm{D}_{\mathrm{KL}}(q,p)$ can be arbitrarily large. To see this, assume that we are given two points $p,q \in \mathbb{R}^2_{\geq 0}$ by $p = (\frac{1}{2}, \frac{1}{2})$ and $q = (\varepsilon, 1-\varepsilon)$ for a small positive $\varepsilon < \frac{1}{2}$. Since $\varepsilon < \frac{1}{2}$ and $1 - \varepsilon < 1$ we obtain

$$\mathrm{D}_{\mathrm{KL}}(q,p) = \varepsilon \ln(2\varepsilon) + (1-\varepsilon)\ln\big(2(1-\varepsilon)\big) \leq \varepsilon \ln 1 + \ln 2 = \ln 2 \ . \quad (2.76)$$

That is, $\mathrm{D}_{\mathrm{KL}}(q,p)$ is bounded from above by a constant, while the lower bound of

$$\mathrm{D}_{\mathrm{KL}}(p,q) = \frac{1}{2}\ln\frac{1}{2\varepsilon} + \frac{1}{2}\ln\frac{1}{2(1-\varepsilon)} = \frac{1}{2}\ln\frac{1}{4\varepsilon(1-\varepsilon)} \geq \frac{1}{2}\ln\frac{1}{4\varepsilon} \quad (2.77)$$

approaches infinity for $\varepsilon \to 0$. A very similar observation can be made with respect to the triangle inequality.

The reason behind this is that the Kullback-Leibler divergence on $\mathbb{R}^d_{\geq 0}$ possesses singularies. While $\varphi_{\mathrm{KL}}(t) < \infty$ for all $t \in \mathbb{R}^d_{\geq 0}$, the partial derivatives $\frac{\partial}{\partial t_i}\varphi_{\mathrm{KL}}(t)$ approach $-\infty$ for $t_i \to 0$. Hence, for points $p,q \in \mathbb{R}^d_{\geq 0}$ with $p = (p_1, p_2, \ldots, p_d)^\top$ and $q = (q_1, q_2, \ldots, q_d)^\top$ we obtain $\mathrm{D}_{\mathrm{KL}}(p,q) = \infty$ if and only if there exists an index i with $p_i > 0$ and $q_i = 0$.

Example 3: Itakura-Saito divergence. The dissimilarity measure known as *Itakura-Saito divergence* (originally defined in [Itakura and Saito, 1968]) has applications in the context of speech analysis and sound processing.

For points $p = (p_1, p_2, \dots, p_d)^\top \in \mathbb{R}^d_{\geq 0}$ and $q = (q_1, q_2, \dots, q_d)^\top \in \mathbb{R}^d_{\geq 0}$ the discrete version of the Itakura-Saito divergence is defined as

$$\mathrm{D}_{\mathrm{IS}}(p, q) = \sum_{i=1}^{d} \left(\frac{p_i}{q_i} - \ln \frac{p_i}{q_i} - 1 \right) \ . \tag{2.78}$$

Here, we use the convention that for $q_i = 0$, we have

$$\frac{p_i}{q_i} - \ln \frac{p_i}{q_i} = \begin{cases} 1 & \text{if } p_i = 0 \\ \infty & \text{if } p_i > 0 \end{cases} \ . \tag{2.79}$$

It is known that the Itakura-Saito divergence is also a Bregman divergence, as is given in the following lemma.

Lemma 2.11. *The Itakura-Saito divergence* D_{IS} *on domain* $\mathbb{R}^d_{\geq 0}$ *is a Bregman divergence by means of generating function*

$$\varphi_{IS}(t) = \sum_{i=1}^{d} \ln \frac{1}{t_i} \ . \tag{2.80}$$

Proof. Let $t = (t_1, t_2, \dots, t_d)^\top \in \mathbb{R}^d_{\geq 0}$. The function $\varphi_{\mathrm{IS}}(t)$ has continuous and differentiable first order partial derivatives

$$\frac{\partial}{\partial t_i} \varphi_{\mathrm{IS}}(t) = -\frac{1}{t_i} \tag{2.81}$$

for all $i = 1, 2, \dots, d$. Furthermore, the function $\varphi_{\mathrm{IS}}(t)$ has second order partial derivatives

$$\frac{\partial^2}{\partial t_i \partial t_j} \varphi_{\mathrm{IS}}(t) = \begin{cases} \frac{1}{t_i^2} & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \ . \tag{2.82}$$

Hence, for all $t \in \mathbb{R}^d_{\geq 0}$ we have

$$\nabla^2 \varphi_{\mathrm{IS}}(t) = \begin{pmatrix} \frac{1}{t_1^2} & 0 & \cdots & 0 \\ 0 & \frac{1}{t_2^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{t_d^2} \end{pmatrix} \ . \tag{2.83}$$

Therefore, for each $t \in \mathbb{R}^d_{\geq 0}$ and for each $x = (x_1, x_2, \ldots, x_d) \in \mathbb{R}^d \setminus \{0\}$ we obtain

$$x^\top \nabla^2 \varphi_{\mathrm{IS}}(t)\, x = \sum_{i=1}^{d} \left(\frac{x_i}{t_i}\right)^2 > 0 \ . \tag{2.84}$$

Thus, $\nabla^2 \varphi_{\mathrm{IS}}(t)$ is a symmetric positive definite matrix for each $t \in \mathbb{R}^d_{\geq 0}$, and we conclude that φ_{IS} is strictly convex. Hence, φ_{IS} is a generating function. Furthermore, for each $p, q \in \mathbb{R}^d_{\geq 0}$ with $p = (p_1, p_2, \ldots, p_d)^\top$ and $q = (q_1, q_2, \ldots, q_d)^\top$ and for the Bregman divergence using φ_{IS} as generating function we have

$$\begin{aligned} \mathrm{D}_{\varphi_{\mathrm{IS}}}(p, q) &= \varphi_{\mathrm{IS}}(p) - \varphi_{\mathrm{IS}}(q) - \nabla\varphi_{\mathrm{IS}}(q)^\top (p - q) && (2.85)\\ &= \sum_{i=1}^{d} \left(\ln \frac{1}{p_i} - \ln \frac{1}{q_i} + \frac{1}{q_i}(p_i - q_i) \right) && (2.86)\\ &= \sum_{i=1}^{d} \left(\frac{p_i}{q_i} - \ln \frac{p_i}{q_i} - 1 \right) \ . && (2.87) \end{aligned}$$

□

It is easy to see that the Itakura-Saito divergence on $\mathbb{R}^d_{\geq 0}$ is asymmetric and does not satisfy the triangle inequality. As in the case of the Kullback-Leibler divergence, the ratio $\mathrm{D}_{\mathrm{IS}}(p, q)/\,\mathrm{D}_{\mathrm{IS}}(q, p)$ is unbounded. For instance, by choosing points $p, q \in \mathbb{R}_{\geq 0}$ with $p = 1$ and $q = 1/\alpha$ for some large $\alpha > 1$ we obtain

$$\mathrm{D}_{\mathrm{IS}}(q, p) = \frac{1}{\alpha} - \ln \frac{1}{\alpha} - 1 \leq \ln \alpha \ . \tag{2.88}$$

On the other hand, there exists a constant $c > 0$ such that for large enough α we have

$$\mathrm{D}_{\mathrm{IS}}(p, q) = \alpha - \ln \alpha - 1 \geq c\alpha \ . \tag{2.89}$$

Hence, the quotient

$$\frac{\mathrm{D}_{\mathrm{IS}}(p, q)}{\mathrm{D}_{\mathrm{IS}}(q, p)} \geq \frac{c\alpha}{\ln(\alpha)} \to \infty \tag{2.90}$$

as α approaches infinity, that is, $q = 1/\alpha \to 0$. A similar observation holds for the ratio $\mathrm{D}_{\mathrm{IS}}(p, q)/\big(\mathrm{D}_{\mathrm{IS}}(p, r) + \mathrm{D}_{\mathrm{IS}}(r, q)\big)$ considering the triangle inequality.

As in the case of the Kullback-Leibler divergence, this behavior is due to the fact that the Itakura-Saito divergence features singularities on $\mathbb{R}^d_{\geq 0}$. More precisely, for points $p, q \in \mathbb{R}^d_{\geq 0}$ with $p = (p_1, p_2, \ldots, p_d)^\top$ and $q = (q_1, q_2, \ldots, q_d)^\top$ we obtain $\mathrm{D}_{\mathrm{IS}}(p, q) = \infty$ if and only if there exists an index i with $p_i > 0$ and $q_i = 0$.

2.2 μ-similarity

In this section we introduce our notion of μ-similar Bregman divergences, originally introduced in [Ackermann et al., 2008]. This notion is fundamental to many of the results obtained in this thesis. To this end, we first introduce the class of Mahalanobis distances. Mahalanobis distances will turn out to be the prototypical subclass of Bregman divergences with tractable geometric properties. Second, we show that any Bregman divergence D_φ can be related to a Mahalanobis distance as long as the domain of D_φ avoids certain singularities. In particular, we show that for any Bregman divergence D_φ approximately the same geometric properties like the properties of a Mahalanobis distance can be derived from bounds to the second order derivative of φ.

2.2.1 Mahalanobis distances

Among the Bregman divergences, one particular class of dissimilarity measures plays an important role to the approach presented in this thesis. For a symmetric positive definite matrix $A \in \mathbb{R}^{d \times d}$ the *Mahalanobis distance* with respect to A is defined as

$$\mathrm{D}_A(p, q) = (p - q)^\top A \, (p - q) \tag{2.91}$$

for $p, q \in \mathbb{R}^d$. The Mahalanobis distance was introduced in 1936 by the Indian statistician P. C. Mahalanobis based on the inverse of the covariance matrix of two random variables (cf. [Mahalanobis, 1936]). All Mahalanobis distances are Bregman divergences, as is stated by the following lemma.

Lemma 2.12. *The Mahalanobis distance* D_A *on domain* $\mathbb{R}^d$ *with respect to symmetric positive definite matrix* $A \in \mathbb{R}^{d \times d}$ *is a Bregman divergence by means of generating function*

$$\varphi_A(t) = t^\top A \, t \ . \tag{2.92}$$

Proof. Let $A = (a_{ij})_{0 \le i,j \le d}$ and $t = (t_1, t_2, \dots, t_d)^\top \in \mathbb{R}^d$. The function

$$\varphi_A(t) = \sum_{i=1}^{d} \sum_{j=1}^{d} a_{ij} t_i t_j \tag{2.93}$$

has continuous and differential first order partial derivatives

$$\frac{\partial}{\partial t_i} \varphi_A(t) = 2 \sum_{j=1}^{d} a_{ij} t_j \tag{2.94}$$

for all $i = 1, 2, \dots, d$. Hence,

$$\nabla \varphi_A(t) = 2A\,t \ . \tag{2.95}$$

Furthermore, the function φ_A has constant second order partial derivatives

$$\frac{\partial^2}{\partial t_i \partial t_j} \varphi_A(t) = 2a_{ij} \ . \tag{2.96}$$

Hence, for all $t \in \mathbb{R}^d$ we have

$$\nabla^2 \varphi_A(t) = 2A \ . \tag{2.97}$$

Thus, since A is symmetric positive definite, we find that $\nabla^2 \varphi_A(t)$ is also a symmetric positive definite matrix for all $t \in \mathbb{R}^d$. We obtain that φ_A is strictly convex. Hence, φ_A is a generating function. Furthermore, for arbitrary $p, q \in \mathbb{R}^d$, using the bi-linearity of the inner product we conclude

$$\begin{aligned}
\mathrm{D}_{\varphi_A}(p, q) &= \varphi_A(p) - \varphi_A(q) - \nabla \varphi_A(q)^\top (p - q) && (2.98)\\
&= p^\top A\,p - q^\top A\,q - 2q^\top A\,(p - q) && (2.99)\\
&= p^\top A\,p + q^\top A\,q - 2q^\top A\,p && (2.100)\\
&= (p - q)^\top A\,p + q^\top A\,(q - p) && (2.101)\\
&= (p - q)^\top A\,p - (p - q)^\top A\,q && (2.102)\\
&= (p - q)^\top A\,(p - q) \ . && (2.103)
\end{aligned}$$

□

In many ways, a Mahalanobis distance D_A can be seen as a generalization of the square of the Euclidean distance. In particular, the squared

Euclidean distance is a Mahalanobis distance with respect to the identity matrix $A = I_d$. Moreover, Mahalanobis distances exhibit many of the geometrical properties of the squared Euclidean distance. In fact, there exists a linear mapping such that the squared Euclidean distance of the images under this mapping equals the Mahalanobis distance of the preimages.

Lemma 2.13. *Let* D_A *be a Mahalanobis distance with respect to symmetric positive definite matrix* $A \in \mathbb{R}^{d\times d}$*. Then there exists a non-singular matrix* $B \in \mathbb{R}^{d\times d}$ *such that for each* $p, q \in \mathbb{R}^d$ *we have*

$$\mathrm{D}_A(p,q) = \|Bp - Bq\|^2 \ . \tag{2.104}$$

Proof. Since A is a symmetric positive definite matrix, it is a well-known fact from linear algebra that there exists a non-singular matrix B with

$$A = B^\top B \ . \tag{2.105}$$

For instance, such a matrix B is given by the Cholesky decomposition of matrix A (cf. [Trefethen and Bau, 1997]). Hence, we obtain

$$\begin{aligned} \mathrm{D}_A(p,q) &= (p-q)^\top B^\top B\,(p-q) & (2.106)\\ &= (Bp - Bq)^\top (Bp - Bq) & (2.107)\\ &= \|Bp - Bq\|^2 \ . & (2.108) \end{aligned}$$

□

In a certain sense, the family of Mahalanobis distances is the subclass of well-natured dissimilarity measures among the Bregman divergences. The reason behind this is that the Mahalanobis distances in general, like the squared Euclidean distance in particular, are the square of a metric.

Lemma 2.14. *Each Mahalanobis distance* D_A *is the square of a metric.*

Proof. Let $p, q, r \in \mathbb{R}^d$. From Lemma 2.13 we know that there is a non-singular matrix B with $\mathrm{D}_A(p,q) = \|Bp - Bq\|^2$. We obtain

$$\sqrt{\mathrm{D}_A(p,q)} = \|Bp - Bq\| = \|Bq - Bp\| = \sqrt{\mathrm{D}_A(q,p)} \tag{2.109}$$

and

$$\begin{aligned} \sqrt{\mathrm{D}_A(p,q)} &= \|Bp - Bq\| & (2.110)\\ &\le \|Bp - Br\| + \|Br - Bq\| & (2.111)\\ &= \sqrt{\mathrm{D}_A(p,r)} + \sqrt{\mathrm{D}_A(r,q)} \ . & (2.112) \end{aligned}$$

Hence, $\sqrt{\mathrm{D}_A(\cdot,\cdot)}$ is symmetric and obeys the triangle inequality. □

As an immediate consequence, we obtain that Mahalanobis distances feature some convenient geometrical properties. First of all, unlike other Bregman divergences, Mahalanobis distances are symmetric. In fact, one can show that Mahalanobis distances are the only symmetric Bregman divergences (cf. [Nielsen et al., 2007], Lemma 2). Furthermore, all Mahalanobis distances satisfy the following *double triangle inequality*.

Lemma 2.15. *For all Mahalanobis distances* D_A *and for all* $p, q, r \in \mathbb{R}^d$ *we have*

$$\mathrm{D}_A(p,q) \leq 2\,\mathrm{D}_A(p,r) + 2\,\mathrm{D}_A(r,q)\ . \tag{2.113}$$

Proof. Using Lemma 2.9 and Lemma 2.13 we obtain that there exists a matrix B such that

$$\begin{aligned} \mathrm{D}_A(p,q) &= \|Bp - Bq\|^2 && (2.114)\\ &\leq 2\|Bp - Br\|^2 + 2\|Br - Bq\|^2 && (2.115)\\ &= 2\,\mathrm{D}_A(p,r) + 2\,\mathrm{D}_A(r,q)\ . && (2.116) \end{aligned}$$

□

2.2.2 μ-similar Bregman divergences

Our interest in Mahalanobis distances is due to the observation that, to some extent, Mahalanobis distances are prototypical for all Bregman divergences. To understand this connection, recall that by the Lagrange form of a Bregman divergence D_φ as given in Lemma 2.7 we have

$$\mathrm{D}_\varphi(p,q) = \frac{1}{2}(p-q)^\top \nabla^2\varphi(\xi)\,(p-q) \tag{2.117}$$

for all points $p, q \in \mathbb{X}$ and for some point ξ on the line segment through p and q. It is easy to see that if the Hessian matrix $\nabla^2\varphi(t)$ is constant for all $t \in \mathbb{X}$, then the Bregman divergence D_φ is a Mahalanobis distance with respect to matrix $A = \frac{1}{2}\nabla^2\varphi(t)$. In this case, D_φ exhibits the same geometrical properties as any Mahalanobis distance, such as symmetry and the triangle inequality within a factor of 2. If, however, $\nabla^2\varphi(t)$ is nonconstant, yet varies only slightly for all $t \in \mathbb{X}$, we would expect that D_φ is still close to having these well-natured metric properties, maybe within a small margin of error. On the other hand, if $\nabla^2\varphi(t)$ varies greatly for $t \in \mathbb{X}$

we may have that the behavior of D_φ is very unlike to the behavior of a Mahalanobis distance. Hence, D_φ may be far off from metric properties.

Thus, we expect that the the closeness to properties like symmetry and the double triangle inequality can be parameterized by the similarity of a Bregman divergence D_φ towards any Mahalanobis distance. This intuition is formalized in the following notion of μ-similarity.

Definition 2.16. *A Bregman divergence D_φ on domain $\mathbb{X} \subseteq \mathbb{R}^d$ is called* μ-similar *for a positive constant $0 < \mu \leq 1$ if there exists a symmetric positive definite matrix $A \in \mathbb{R}^{d\times d}$ such that for the Mahalanobis distance D_A and for all $p, q \in \mathbb{X}$ we have*

$$\mu\,\mathrm{D}_A(p,q) \leq \mathrm{D}_\varphi(p,q) \leq \mathrm{D}_A(p,q)\ . \tag{2.118}$$

The notion of μ-similar Bregman divergences has already been used in [Ackermann et al., 2008] and [Ackermann and Blömer, 2009]. To the best of this author's knowledge, all Bregman divergences D_φ that are used in practice are μ-similar when restricted to a domain $\mathbb{X}$ that avoids the singularities of D_φ. More precisely, consider a Bregman divergence D_φ on domain $\mathbb{X}$ with twice differentiable generating function φ. One can show that D_φ is μ-similar as long as domain $\mathbb{X}$ avoids points where the quadratic form given by $x^\top\nabla^2\varphi(t)\,x$ is either zero or infinity for any $t \in \mathbb{X}$, as is implied by the following lemma. Note that since φ is strictly convex, we know that $\nabla^2\varphi(t)$ is a symmetric positive semi-definite matrix for all $t \in \mathbb{X}$, and that $\nabla^2\varphi(t)$ is symmetric positive definite for almost all $t \in \mathbb{X}$. More precisely, the set of all $t \in \mathbb{X}$ with $x^\top\nabla^2\varphi(t)\,x = 0$ is merely a discrete subset of $\mathbb{X}$, if it exists at all. Also note that that $x^\top\nabla^2\varphi(t)\,x$ may only be infinity if $t \notin \mathrm{ri}(\mathbb{X})$. Otherwise, using Lemma 2.7, we would be able to find $p, q \in \mathrm{ri}(X)$ with $\mathrm{D}_\varphi(p,q) = x^\top\nabla^2\varphi(t)\,x = \infty$ for $x = p - q$, which stands in contradiction to the fact that $\mathrm{D}_\varphi(p,q) < \infty$ for all $p, q \in \mathrm{ri}(X)$.

Lemma 2.17. *Let φ be a twice differentiable generating function on domain $\mathbb{X}$ such that the Hessian $\nabla^2\varphi(t)$ is symmetric positive definite for all $t \in \mathbb{X}$. Furthermore, let*

$$\mu(\varphi, \mathbb{X}) = \inf_{\substack{p,q\in\mathbb{X}\\ \xi,\zeta\in\overline{pq}}} \frac{(p-q)^\top\nabla^2\varphi(\xi)\,(p-q)}{(p-q)^\top\nabla^2\varphi(\zeta)\,(p-q)}\ , \tag{2.119}$$

where $\overline{pq} = \{x \in \mathbb{X} \mid \exists 0 \leq \lambda \leq 1 : x = \lambda p + (1-\lambda)q\}$ denotes the line segment through p and q. If $\mu(\varphi, \mathbb{X}) > 0$ then D_φ is a $\mu(\varphi, \mathbb{X})$-similar Bregman divergence on $\mathbb{X}$.

Proof. Fix any distinct $p, q \in \mathbb{X}$. Note that by definition of $\mu(\varphi, \mathbb{X})$ we have

$$\mu(\varphi, \mathbb{X}) \leq \frac{\min_{\xi \in \overline{pq}} (p-q)^\top \nabla^2 \varphi(\xi)\,(p-q)}{\max_{\zeta \in \overline{pq}} (p-q)^\top \nabla^2 \varphi(\zeta)\,(p-q)} \tag{2.120}$$

for all $p, q \in \mathbb{X}$. Let

$$\zeta^* = \arg\max_{\zeta \in \overline{pq}} (p-q)^\top \nabla^2 \varphi(\zeta)\,(p-q) \tag{2.121}$$

and let

$$A = \frac{1}{2} \nabla^2 \varphi(\zeta^*)\ . \tag{2.122}$$

Since the Hessian matrix $\nabla^2 \varphi(t)$ is a symmetric positive definite matrix for all $t \in \mathbb{X}$, so is matrix A. Hence, Mahalanobis distance D_A is well defined. From the Lagrange form of Lemma 2.7 we know that

$$\mathrm{D}_\varphi(p, q) = \frac{1}{2} (p-q)^\top \nabla^2 \varphi(\xi^*)\,(p-q) \tag{2.123}$$

for some $\xi^* \in \overline{pq}$. We conclude

$$\begin{aligned} \mathrm{D}_\varphi(p, q) &\geq \frac{1}{2} \min_{\xi \in \overline{pq}} (p-q)^\top \nabla^2 \varphi(\xi)\,(p-q) && (2.124)\\ &\geq \frac{1}{2} \mu(\varphi, \mathbb{X}) \max_{\zeta \in \overline{pq}} (p-q)^\top \nabla^2 \varphi(\zeta)\,(p-q) && (2.125)\\ &= \frac{1}{2} \mu(\varphi, \mathbb{X}) (p-q)^\top \nabla^2 \varphi(\zeta^*)\,(p-q) && (2.126)\\ &= \mu(\varphi, \mathbb{X})\, \mathrm{D}_A(p, q) && (2.127) \end{aligned}$$

and

$$\begin{aligned} \mathrm{D}_\varphi(p, q) &\leq \frac{1}{2} \max_{\zeta \in \overline{pq}} (p-q)^\top \nabla^2 \varphi(\zeta)\,(p-q) && (2.128)\\ &= \frac{1}{2} (p-q)^\top \nabla^2 \varphi(\zeta^*)\,(p-q) && (2.129)\\ &= \mathrm{D}_A(p, q)\ . && (2.130) \end{aligned}$$

□

As expected, we obtain that μ-similar Bergman divergences feature some approximate metric properties. In particular, μ-similar Bergman divergences are approximately symmetric within a factor of $\mathcal{O}(1/\mu)$ and satisfy the triangle inequality within a factor of $\mathcal{O}(1/\mu)$, as is stated in the following lemma.

Lemma 2.18. *Let* D_φ *be a* μ*-similar Bregman divergence on domain* $\mathbb{X}$*. For all* $p, q, r \in \mathbb{X}$ *we have*

$$\mathrm{D}_\varphi(p,q) \leq \frac{1}{\mu}\,\mathrm{D}_\varphi(q,p)\ , \tag{2.131}$$

$$\mathrm{D}_\varphi(p,q) \leq \frac{2}{\mu}\,\mathrm{D}_\varphi(p,r) + \frac{2}{\mu} D_\varphi(r,q)\ , \tag{2.132}$$

$$\mathrm{D}_\varphi(p,q) \leq \frac{2}{\mu}\,\mathrm{D}_\varphi(p,r) + \frac{2}{\mu} D_\varphi(q,r)\ , \tag{2.133}$$

$$\mathrm{D}_\varphi(p,q) \leq \frac{2}{\mu}\,\mathrm{D}_\varphi(r,p) + \frac{2}{\mu} D_\varphi(r,q)\ , \tag{2.134}$$

$$\mathrm{D}_\varphi(p,q) \leq \frac{2}{\mu}\,\mathrm{D}_\varphi(r,p) + \frac{2}{\mu} D_\varphi(q,r)\ . \tag{2.135}$$

Proof. Let D_φ be μ-similar with respect to Mahalanobis distance D_A. Using the μ-similarity of D_φ and the symmetry of D_A we get

$$\mathrm{D}_\varphi(p,q) \leq \mathrm{D}_A(p,q) = \mathrm{D}_A(q,p) \leq \frac{1}{\mu}\,\mathrm{D}_\varphi(q,p)\ . \tag{2.136}$$

This proves inequality (2.131). Furthermore, using the μ-similarity of D_φ and the double triangle inequality of D_A from Lemma 2.15 we obtain

$$\begin{aligned} \mathrm{D}_\varphi(p,q) &\leq \mathrm{D}_A(p,q) && (2.137)\\ &\leq 2\,\mathrm{D}_A(p,r) + 2\,\mathrm{D}_A(r,q) && (2.138)\\ &\leq \frac{2}{\mu}\,\mathrm{D}_\varphi(p,r)\frac{2}{\mu}\,\mathrm{D}_\varphi(r,q)\ . && (2.139) \end{aligned}$$

This proves inequality (2.132). Using the symmetry of D_A, inequalities (2.133) to (2.135) follow analogously. □

2.2.3 Examples of μ-similar Bregman divergences

We show that all Bregman divergences considered so far are, indeed, μ-similar Bregman divergences. An overview of more μ-similar Bregman divergences can be found in Figure 2.2.

domain $\mathbb{X}$	$\mathrm{D}_\varphi(p,q)$	μ	A
$\mathbb{R}^d$	squared Euclidean distance $\\|p-q\\|_2^2$	1	I_d
$\mathbb{R}^d$	Mahalanobis distance $(p-q)^\top A(p-q)$	1	A
$[\lambda,\upsilon]^d \subseteq \mathbb{R}_+^d$	Kullback-Leibler divergence $\sum p_i \ln(\frac{p_i}{q_i}) - p_i + q_i$	$\frac{\lambda}{\upsilon}$	$\frac{1}{2\lambda} I_d$
$[\lambda,\upsilon]^d \subseteq \mathbb{R}_+^d$	Itakura-Saito divergence $\sum \frac{p_i}{q_i} - \ln(\frac{p_i}{q_i}) - 1$	$\frac{\lambda^2}{\upsilon^2}$	$\frac{1}{2\lambda^2} I_d$
$[-\upsilon,\upsilon]^d \subseteq (-\frac{\pi}{2},\frac{\pi}{2})^d$	trigonometric divergence $\sum \cos(q_i) - \cos(p_i) - (p_i - q_i)\sin(q_i)$	$\cos(\upsilon)$	$\frac{1}{2} I_d$
$[\lambda,\upsilon]^d \subseteq \mathbb{R}_+^d$	harmonic divergence ($\alpha > 0$) $\sum \frac{1}{p_i^\alpha} - \frac{\alpha+1}{q_i^\alpha} + \frac{\alpha p_i}{q_i^{\alpha+1}}$	$\frac{\lambda^{\alpha+2}}{\upsilon^{\alpha+2}}$	$\frac{\alpha(\alpha-1)}{2\lambda^{\alpha+2}} I_d$
$[\lambda,\upsilon]^d \subseteq \mathbb{R}_+^d$	norm-like divergence ($\alpha \geq 2$) $\sum p_i^\alpha + (\alpha-1)q_i^\alpha - \alpha p_i q_i^{\alpha-1}$	$\frac{\lambda^{\alpha-2}}{\upsilon^{\alpha-2}}$	$\frac{\alpha(\alpha-1)}{2} \upsilon^{\alpha-2} I_d$
$[\lambda,\upsilon]^d \subseteq \mathbb{R}^d$	exponential divergence $\sum \exp(p_i) - (p_i - q_i + 1)\exp(q_i)$	$\exp(\lambda - \upsilon)$	$\frac{\exp(\upsilon)}{2} I_d$
$[\lambda,\upsilon]^d \subseteq \mathbb{R}^d$	reciprocal exponential divergence $\sum \exp(-p_i) - (p_i - q_i + 1)\exp(-q_i)$	$\exp(\lambda - \upsilon)$	$\frac{1}{2\exp(\lambda)} I_d$
$[\lambda,\upsilon]^d \subseteq (0,1)^d$	logistic loss $\sum p_i \ln \frac{p_i}{q_i} + (1-p_i) \ln \frac{1-p_i}{1-q_i}$	$\frac{\lambda(1-\upsilon)(1-\lambda+\upsilon)}{\upsilon(1-\lambda)(1-\upsilon+\lambda)}$	$\frac{1-\upsilon+\lambda}{2\lambda(1-\upsilon)} I_d$
$[-\upsilon,\upsilon]^d \subseteq \mathbb{R}^d$	dual logistic loss $\sum \ln \frac{1+\exp(p_i)}{1+\exp(q_i)} - (p_i - q_i) \frac{\exp(q_i)}{1+\exp(q_i)}$	$\frac{4\exp(\upsilon)}{\big(1+\exp(\upsilon)\big)^2}$	$\frac{1}{8} I_d$
$[-\upsilon,\upsilon]^d \subseteq (-1,1)^d$	Hellinger-like divergence $\sum \frac{1-p_i q_i}{\sqrt{1-q_i^2}} - \sqrt{1-p_i^2}$	$(1-\upsilon^2)^{3/2}$	$\frac{1}{2(1-\upsilon^2)^{3/2}} I_d$

Figure 2.3: An overview of some μ-similar Bregman divergences.

Example 1: Mahalanobis distances. As a trivial observation, note that all Mahalanobis distances D_A on domain $\mathbb{R}^d$, such as the squared Euclidean distance, are 1-similar Bregman divergences.

Example 2: Kullback-Leibler divergence. We show that the generalized Kullback-Leibler divergence

$$\mathrm{D}_{\mathrm{KL}}(p,q) = \sum_{i=1}^{d} \left(p_i \ln \frac{p_i}{q_i} - p_i + q_i \right) \tag{2.140}$$

is μ-similar when restricted to domain $[\lambda, \upsilon]^d \subseteq \mathbb{R}^d_+$ with $0 < \lambda < \upsilon$.

Lemma 2.19. *Let $0 < \lambda < \upsilon$. D_{KL} on domain $[\lambda, \upsilon]^d \subseteq \mathbb{R}^d_+$ is a μ-similar Bregman divergence with $\mu = \frac{\lambda}{\upsilon}$ and $A = \frac{1}{2\lambda} I_d$.*

Proof. Let $p, q \in [\lambda, \upsilon]^d$ with $p = (p_1, p_2, \ldots, p_d)^\top$ and $q = (q_1, q_2, \ldots, q_d)^\top$. Recall that generating function

$$\varphi_{\mathrm{KL}}(t) = \sum_{i=0}^{d} (t_i \ln t_i - t_i) \tag{2.141}$$

has a Hessian matrix given by

$$\nabla^2 \varphi_{\mathrm{KL}}(t) = \begin{pmatrix} \frac{1}{t_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{t_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{t_d} \end{pmatrix} \tag{2.142}$$

for all $t = (t_1, t_2, \ldots, t_d)^\top \in [\lambda, \upsilon]^d$. Furthermore, by the Lagrange form of Lemma 2.7 we know that

$$\mathrm{D}_{\mathrm{KL}}(p,q) = \frac{1}{2}(p-q)^\top \nabla^2 \varphi_{\mathrm{KL}}(\xi)(p-q) \tag{2.143}$$

$$= \frac{1}{2} \sum_{i=1}^{d} \frac{1}{\xi_i}(p_i - q_i)^2 \tag{2.144}$$

for some $\xi = (\xi_1, \xi_2, \ldots, \xi_d)^\top \in [\lambda, \upsilon]^d$. Since we have $\frac{1}{\upsilon} \leq \frac{1}{\xi_i} \leq \frac{1}{\lambda}$ for all

$i = 1, 2, \ldots, d$ we obtain

$$\mathrm{D}_{\mathrm{KL}}(p, q) \leq \frac{1}{2\lambda} \sum_{i=1}^{d} (p_i - q_i)^2 \tag{2.145}$$

$$= (p-q)^\top \left(\frac{1}{2\lambda} I_d \right) (p-q) \tag{2.146}$$

$$= \mathrm{D}_A(p, q) \tag{2.147}$$

and

$$\mathrm{D}_{\mathrm{KL}}(p, q) \geq \frac{1}{2\upsilon} \sum_{i=1}^{d} (p_i - q_i)^2 \tag{2.148}$$

$$= \frac{\lambda}{\upsilon} (p-q)^\top \left(\frac{1}{2\lambda} I_d \right) (p-q) \tag{2.149}$$

$$= \frac{\lambda}{\upsilon} \mathrm{D}_A(p, q) \ . \tag{2.150}$$

□

Example 3: Itakura-Saito divergence. The Itakura-Saito divergence

$$\mathrm{D}_{\mathrm{IS}}(p, q) = \sum_{i=1}^{d} \left(\frac{p_i}{q_i} - \ln \frac{p_i}{q_i} - 1 \right) \ . \tag{2.151}$$

is also a μ-similar Bregman divergence when restricted to domain $[\lambda, \upsilon]^d$.

Lemma 2.20. *Let $0 < \lambda < \upsilon$. D_{IS} on domain $[\lambda, \upsilon]^d \subseteq \mathbb{R}_+^d$ is a μ-similar Bregman divergence with $\mu = \frac{\lambda^2}{\upsilon^2}$ and $A = \frac{1}{2\lambda^2} I_d$.*

Proof. This lemma can be shown in analogy to Lemma 2.19. That is, let $p, q \in [\lambda, \upsilon]^d$ with $p = (p_1, p_2, \ldots, p_d)^\top$ and $q = (q_1, q_2, \ldots, q_d)^\top$. Recall that generating function

$$\varphi_{\mathrm{IS}}(t) = \sum_{i=0}^{d} \ln \frac{1}{t_i} \tag{2.152}$$

has a Hessian matrix given by

$$\nabla^2 \varphi_{\mathrm{IS}}(t) = \begin{pmatrix} \frac{1}{t_1^2} & 0 & \cdots & 0 \\ 0 & \frac{1}{t_2^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{t_d^2} \end{pmatrix} \tag{2.153}$$

for all $t = (t_1, t_2, \dots, t_d)^\top \in [\lambda, \upsilon]^d$. Furthermore, by the Lagrange form of Lemma 2.7 we know that

$$\begin{aligned} \mathrm{D}_{\mathrm{IS}}(p,q) &= \frac{1}{2}(p-q)^\top \nabla^2 \varphi_{\mathrm{IS}}(\xi)(p-q) && (2.154)\\ &= \frac{1}{2}\sum_{i=1}^{d} \frac{1}{\xi_i^2}(p_i - q_i)^2 && (2.155) \end{aligned}$$

for some $\xi = (\xi_1, \xi_2, \dots, \xi_d)^\top \in [\lambda, \upsilon]^d$. Since $\frac{1}{\upsilon^2} \leq \frac{1}{\xi_i^2} \leq \frac{1}{\lambda^2}$ for all $i = 1, 2, \dots, d$ we obtain

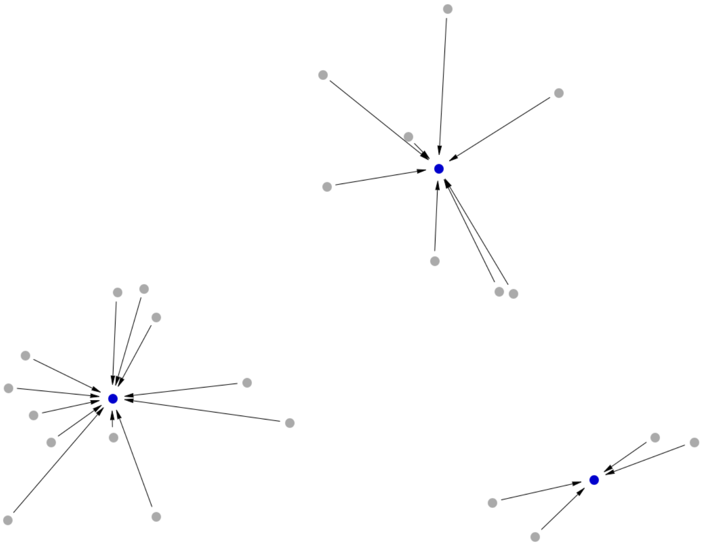
$$\begin{aligned} \mathrm{D}_{\mathrm{IS}}(p,q) &\leq \frac{1}{2\lambda^2}\sum_{i=1}^{d}(p_i - q_i)^2 && (2.156)\\ &= (p-q)^\top \left(\frac{1}{2\lambda^2} I_d\right)(p-q) && (2.157)\\ &= \mathrm{D}_A(p,q) && (2.158) \end{aligned}$$

and

$$\begin{aligned} \mathrm{D}_{\mathrm{IS}}(p,q) &\geq \frac{1}{2\upsilon^2}\sum_{i=1}^{d}(p_i - q_i)^2 && (2.159)\\ &= \frac{\lambda^2}{\upsilon^2}(p-q)^\top \left(\frac{1}{2\lambda^2} I_d\right)(p-q) && (2.160)\\ &= \frac{\lambda^2}{\upsilon^2}\,\mathrm{D}_A(p,q)\ . && (2.161) \end{aligned}$$

□

3 The k-median problem



This thesis is mainly dedicated to the study of the k-median problem using a Bregman divergence D_φ (as introduced in Chapter 2) as dissimilarity measure. In this case, we call the problem the *Bregman k-median problem*. This notion captures a large number of clustering problems that have been intensively studied by the scientific community, such as the *Euclidean k-means problem* (when using the squared Euclidean distance as dissimilarity measure), *information theoretic clustering* (using the Kullback-Leibler divergence or similar entropy based divergences), or the *vector quantization problem* (using either the squared Euclidean distance or the Itakura-Saito divergence). These and other related median-type clustering problems go by a variety of different names in literature, even sometimes when addressing the same problem. In this thesis, we will address them in a unified manner as is provided in this chapter. A small overview of three different practical applications of the Bregman k-median problem is given later in Appendix A.

The rest of this chapter is organized as follows. In Section 3.1 we introduce our terms and our notation of a general formulation of the k-median problem. We also prove some useful properties common to all instances of the generalized k-median problem. After that, in Section 3.2, we study the Bregman k-median problem. In particular, we investigate geometrical and combinatorial aspects of optimal Bregman k-median clusterings that are valid even in the absence of convenient metric properties of the given Bregman divergence. We conclude this chapter by giving a simple polynomial time algorithm for the Bregman k-median problem in dimension $d = 1$. This algorithm serves as a toy example to demonstrate the feasibility of giving Bregman clustering algorithms that do not rely on metric properties such as symmetry or the triangle inequality.

3.1 The generalized k-median problem

In this section we introduce our generalized formulation of the k-median problem. This notion captures a large number of well-known clustering problems and has already been used in [Ackermann et al., 2008]. We also prove some basic properties of the generalized k-median problem.

3.1.1 Definitions and notation

We are given an arbitrary *domain* $\mathbb{X} \subseteq \mathbb{R}^d$. Usually, the elements from $\mathbb{X}$ are called *points*. On domain $\mathbb{X}$ an arbitrary *dissimilarity measure*

$$\mathrm{D} : \mathbb{X} \times \mathbb{X} \to \mathbb{R}_{\geq 0} \cup \{\infty\} \tag{3.1}$$

is defined. $\mathrm{D}(x, y)$ specifies the directed dissimilarity from point $x \in \mathbb{X}$ towards point $y \in \mathbb{X}$. For a point $x \in \mathbb{X}$ and a finite subset $C \subseteq \mathbb{X}$ we sometimes write

$$\mathrm{D}(x, C) = \min_{c \in C} \mathrm{D}(x, c) \tag{3.2}$$

to specify the dissimilarity from point x towards the closest point from set C. We make no assumption on the nature of the dissimilarity measure D other than $\mathrm{D}(x, y) = 0$ if and only if $x = y$. In particular, D may be *asymmetric* (i.e, there may be $x, y \in \mathbb{X}$ with $\mathrm{D}(x, y) \neq \mathrm{D}(y, x)$) and does not necessarily satisfy the *triangle inequality* (i.e, there may exist $x, y, z \in \mathbb{X}$ with $\mathrm{D}(x, z) > \mathrm{D}(x, y) + \mathrm{D}(y, z)$). We also allow D to have *singularities*, that is, there may be points $x, y \in \mathbb{X}$ with $\mathrm{D}(x, y) = \infty$.

For finite sets $P, C \subseteq \mathbb{X}$ with $|C| = k$ we denote the *k-median cost* of point set P towards C with respect to D by

$$\mathrm{cost}^{\mathrm{D}}(P, C) = \sum_{p \in P} \mathrm{D}(p, C) \ , \tag{3.3}$$

which is the total dissimilarity of all points from P towards their closest point from set C. In this context, the points from C are called *centers.* If a single center point $c \in \mathbb{X}$ is used we also write

$$\mathrm{cost}^{\mathrm{D}}(P, c) = \sum_{p \in P} \mathrm{D}(p, c) \ . \tag{3.4}$$

If point set P is associated with a weight function $w : P \to \mathbb{R}_{\geq 0}$, the *weighted k-median cost* is given by the weighted sum of the dissimilarity of all points from P towards their closest point from set C, that is,

$$\mathrm{cost}^{\mathrm{D}}_{w}(P, C) = \sum_{p \in P} w(p) \, \mathrm{D}(p, C) \ . \tag{3.5}$$

A center point $c \in \mathbb{X}$ that minimizes $\mathrm{cost}^{\mathrm{D}}(P, c)$ is called a *median* of P. If well defined and unique, we denote such a median by

$$\mathrm{med}^{\mathrm{D}}(P) = \arg\min_{c \in \mathbb{X}} \mathrm{cost}^{\mathrm{D}}(P, c) \ . \tag{3.6}$$

The points of a set $C \subseteq \mathbb{X}$ of size $|C| = k$ that minimizes $\mathrm{cost}^{\mathrm{D}}(P, C)$ are called *k-medians* of P. The cost of such a set of medians is denoted by

$$opt^{\mathrm{D}}_{k}(P) = \min_{\substack{C \subseteq \mathbb{X}, \\ |C| = k}} \mathrm{cost}^{\mathrm{D}}(P, C) \ . \tag{3.7}$$

The *k-median problem* with respect to dissimilarity measure D is defined as follows.

Problem 3.1 (generalized k-median problem)**.** *Let* D *be a dissimilarity measure on domain* $\mathbb{X} \subseteq \mathbb{R}^d$ *and let* $k \in \mathbb{N}$*. Given a finite set* $P \subseteq \mathbb{X}$*, find a set* $C \subseteq \mathbb{X}$ *of size* $|C| = k$ *such that* $\mathrm{cost}^{\mathrm{D}}(P, C)$ *is minimized.*

For any finite $P \subseteq \mathbb{X}$, a *k-clustering* of P is given by a partition of P into k non-empty sets. Alternatively, given a set $C \subseteq \mathbb{X}$ of $|C| = k$ center points, a k-clustering of P is induced by assigning each point from P to its closest center point in C, breaking ties arbitrarily. Here, for each $c \in C$ the set of points from P that are assigned to a common center point form one set of the partition of P.

A partition $P_1, P_2, \dots, P_k$ of P is called an *optimal k-median clustering* of P if the set $C = \{\mathrm{med}^{\mathrm{D}}(P_i) \mid i = 1, 2, \dots, k\}$ of the medians of all P_i achieves minimal k-median cost, that is,

$$\mathrm{cost}^{\mathrm{D}}(P, C) = opt_k^{\mathrm{D}}(P) \ . \tag{3.8}$$

Furthermore, for an $\alpha > 1$, a partition $P_1, P_2, \dots, P_k$ is called an *α-approximate k-median clustering* of P if $C = \{\mathrm{med}^{\mathrm{D}}(P_i) \mid i = 1, 2, \dots, k\}$ satisfies

$$\mathrm{cost}^{\mathrm{D}}(P, C) \leq \alpha \, opt_k^{\mathrm{D}}(P) \ . \tag{3.9}$$

Throughout this thesis, if the dissimilarity measure used is unambiguous, we omit the superscript D and simply write cost, cost_w, med, and opt_k instead.

3.1.2 Properties

In this section we study some basic properties of the generalized k-median problem. These properties are valid regardless of which concrete dissimilarity measure D is used. We will make intensive use of these properties throughout the rest of this thesis.

For the remainder of this section, let $P \subseteq \mathbb{X}$ and let $C = \{c_1, c_2, \dots, c_k\} \subseteq \mathbb{X}$ be an arbitrary set of k centers. Furthermore, let $P_1, P_2, \dots, P_k$ denote the k-clustering of P induced by C, where P_i denotes the points from P closest to center c_i, i.e., $p \in P_i$ if and only if $c_i = \arg\min_{c \in C} \mathrm{D}(p, c)$.

Our first lemma addresses the simple observation that due to the additive definition of the k-median cost function, the total cost of P towards a set of centers C is given by the sum of the 1-median cost of each individual cluster.

Lemma 3.2. *We have*

$$\mathrm{cost}(P, C) = \sum_{i=1}^{k} \mathrm{cost}(P_i, c_i) \tag{3.10}$$

and

$$opt_k(P) = \sum_{i=1}^{k} opt_1(P_i) \ . \tag{3.11}$$

Proof. Trivially, by definition of k-median cost function we obtain

$$\mathrm{cost}(P, C) = \sum_{p \in P} \min_{c \in C} \mathrm{D}(p, c) = \sum_{i=1}^{k} \sum_{p \in P_i} \mathrm{D}(p, c_i) = \sum_{i=1}^{k} \mathrm{cost}(P_i, c_i) \ . \tag{3.12}$$

This proves equation (3.10). Equation (3.11) follows immediately from equation (3.10) if C denotes the set of optimal k-medians of P. □

We also find that a larger number of centers can not increase the optimal k-median cost of a point set P, as is stated in the following lemma.

Lemma 3.3. *For all* $k \geq 2$ *and* $P \subseteq \mathbb{X}$ *we have*

$$opt_k(P) \leq opt_{k-1}(P) \ . \tag{3.13}$$

Proof. Let $C' = \{c'_1, c'_2, \ldots, c'_{k-1}\}$ denote the optimal $(k-1)$-medians of P, i.e., $\mathrm{cost}(P, C') = opt_{k-1}(P)$. Furthermore, let $P'_1, P'_2, \ldots, P'_{k-1}$ denote the optimal $(k-1)$-clustering of P induced by C'. From Lemma 3.2 we know

$$opt_{k-1}(P) = \sum_{i=1}^{k-1} opt_1(P'_i) = \sum_{i=1}^{k-1} \mathrm{cost}(P'_i, c'_i) \ . \tag{3.14}$$

Now, pick an arbitrary cluster of this partition, for instance, P'_1. Let $P'_1 = P_1^{(1)} \cup P_1^{(2)}$ be an arbitrary partition of P'_1 and let $c_1^{(1)}, c_1^{(2)}$ denote the optimal medians of $P_1^{(1)}, P_1^{(2)}$. Then we have $\mathrm{cost}(P_1^{(1)}, c_1^{(1)}) \leq \mathrm{cost}(P_1^{(1)}, c'_1)$ and $\mathrm{cost}(P_1^{(2)}, c_1^{(2)}) \leq \mathrm{cost}(P_1^{(2)}, c'_1)$, and we obtain

$$opt_{k-1}(P) = \mathrm{cost}(P'_1, c'_1) + \sum_{i=2}^{k-1} \mathrm{cost}(P'_i, c'_i) \tag{3.15}$$

$$\geq \mathrm{cost}(P_1^{(1)}, c_1^{(1)}) + \mathrm{cost}(P_1^{(2)}, c_1^{(2)}) + \sum_{i=2}^{k-1} \mathrm{cost}(P'_i, c'_i) \tag{3.16}$$

$$\geq \mathrm{cost}(P, \tilde{C}) \ , \tag{3.17}$$

where $\tilde{C} = \{c_1^{(1)}, c_1^{(2)}, c'_2, \ldots, c'_{k-1}\}$. Since $|\tilde{C}| = k$ we conclude

$$opt_{k-1}(P) \geq \mathrm{cost}(P, \tilde{C}) \geq opt_k(P) \ . \tag{3.18}$$

□

Finally, we observe that the generalized k-median problem satisfies the following optimal substructure property: Every optimal solution to the k-median problem contains an optimal solution to a $(k-1)$-median problem. More precisely, if $P_1, P_2, \ldots, P_k$ denote the clusters of an optimal k-median clustering of P and we remove any cluster P_i completely from the point set, then the remaining $k-1$ clusters form an optimal $(k-1)$-median clustering of the remaining points $P \setminus P_i$. The optimal substructure property is an important ingredient that allows the application of generic algorithmic strategies, such as the divide-and-conquer method or dynamic programming (cf. [Cormen et al., 2009]). We will make use of this property in the simple optimal algorithm from Section 3.3, as well as in the approximation algorithm from Chapter 4.

Lemma 3.4 (optimal substructure property). *Let C be a set of optimal k-medians of P, that is,* $\mathrm{cost}(P, C) = opt_k(P)$. *Then for all $i = 1, 2, \ldots, k$ we have*

$$\mathrm{cost}\big(P \setminus P_i, C \setminus \{c_i\}\big) = opt_{k-1}(P \setminus P_i) \ . \tag{3.19}$$

Proof. Let $P' = P \setminus P_i$ denote the remaining point set. Furthermore, assume for the sake of contradiction that $opt_{k-1}(P') < \mathrm{cost}\big(P', C \setminus \{c_i\}\big)$. That is, there exists an optimal partition $P'_1, P'_2, \ldots, P'_{k-1}$ and a corresponding set of $(k-1)$-medians $C' = \{c'_1, c'_2, \ldots, c'_{k-1}\}$ such that

$$\mathrm{cost}(P', C') < \mathrm{cost}\big(P', C \setminus \{c_i\}\big) \ . \tag{3.20}$$

But this leads to

$$\begin{aligned} \mathrm{cost}(P, C' \cup \{c_i\}) &\leq \mathrm{cost}(P_i, c_i) + \mathrm{cost}(P', C') && (3.21)\\ &< \mathrm{cost}(P_i, c_i) + \mathrm{cost}\big(P', C \setminus \{c_i\}\big) && (3.22)\\ &= \mathrm{cost}(P, C) \ , && (3.23) \end{aligned}$$

which stands in contradiction to the optimality of C as the k-medians of P. Hence, the lemma follows. □

3.2 Bregman k-median clustering

In this section, we study special instances of the generalized k-median problem using Bregman divergences as dissimilarity measure. In particular, we discuss the geometrical and combinatorial properties of optimal Bregman k-median clusterings. The properties discussed in this section do not rely on any metric or approximate metric properties of the given Bregman divergence, and, in fact, lead to a polynomial time algorithm solving the Bregman k-median in fixed dimension and for a fixed number of clusters.

In the sequel, let D_φ denote a Bregman divergence on domain $\mathbb{X}$. For every finite $P \subseteq \mathbb{X}$ of size $|P| = n$ we denote by

$$c_P = \frac{1}{n} \sum_{p \in P} p \tag{3.24}$$

the *centroid* of P. The centroid plays an important role in the context of Bregman k-median clustering as is observed by the following two lemmas due to [Banerjee et al., 2005b]. First, it is known that for all Bregman divergences the following *central identity* holds. This identity is of crucial importance for the techniques employed throughout this paper.

Lemma 3.5 ([Banerjee et al., 2005b], proof of Proposition 1)**.** *Let $P \subseteq \mathbb{X}$ be of size $|P| = n$. For all $q \in \mathbb{X}$ we have*

$$\mathrm{cost}(P, q) = \mathrm{cost}(P, c_P) + n\,\mathrm{D}_\varphi(c_P, q)\ . \tag{3.25}$$

Proof. We have

$$\begin{aligned}
&\mathrm{cost}(P, q) - \mathrm{cost}(P, c_P) \\
&= \sum_{p \in P} \big(\mathrm{D}_\varphi(p, q) - \mathrm{D}_\varphi(p, c_P)\big) && (3.26)\\
&= \sum_{p \in P} \big(\varphi(c_p) - \varphi(q) - \nabla\varphi(q)^\top (p - q) + \nabla\varphi(c_P)^\top (p - c_P)\big) && (3.27)\\
&= n\varphi(c_P) - n\varphi(q) - \sum_{p \in P} \nabla\varphi(q)^\top (p - q) + \sum_{p \in P} \nabla\varphi(c_P)^\top (p - c_P)\ . && (3.28)
\end{aligned}$$

Using the bi-linearity of the inner product we find

$$\sum_{p \in P} \nabla\varphi(q)^\top (p - q) = n\,\nabla\varphi(q)^\top \left(\frac{1}{n}\sum_{p \in P} p - q\right) = n\,\nabla\varphi(q)^\top\,(c_P - q)\ . \tag{3.29}$$

Analogously, we obtain

$$\sum_{p \in P} \nabla\varphi(c_P)^\top (p - c_P) = n\,\nabla\varphi(c_P)^\top \left(\frac{1}{n}\sum_{p \in P} p - c_P\right) = 0\ . \tag{3.30}$$

We conclude

$$\begin{aligned}\operatorname{cost}(P,q)-\operatorname{cost}(P,c_P) &= n\big(\varphi(c_P)-\varphi(q)-\nabla\varphi(q)^\top(c_P-q)\big) &(3.31)\\ &= n\,\mathrm{D}_\varphi(c_P,q)\ . &(3.32)\end{aligned}$$

□

Second, it is an immediate consequence of the central identity of Lemma 3.5 that for any Bregman divergence the centroid c_P is, indeed, the unique median of a cluster P.

Lemma 3.6 ([Banerjee et al., 2005b], Proposition 1)**.** *Let $P \subseteq \mathbb{X}$ be finite. Then the centroid c_P is the unique optimal 1-median of P, i.e.,*

$$c_P = \operatorname{med}(P)\ . \tag{3.33}$$

Proof. From Lemma 2.3 we know that $\mathrm{D}_\varphi(\cdot,\cdot)$ and $\operatorname{cost}(\cdot,\cdot)$ are non-negative, and that $\mathrm{D}_\varphi(x,y)=0$ if and only if $x=y$. Since $\operatorname{cost}(P,c_P)$ is constant for any fixed $P\subseteq\mathbb{X}$, using Lemma 3.5 we conclude that

$$\operatorname{cost}(P,q)=\operatorname{cost}(P,c_P)+n\,\mathrm{D}_\varphi(c_P,q) \tag{3.34}$$

is minimal if and only if $q=c_P$. □

Furthermore, a k-clustering $P_1,P_2,\dots,P_k$ of P is called *linearly separable* if every two distinct clusters P_i,P_j are separated by a hyperplane H, i.e., for each i,j with $i\neq j$ there exist some $a,b\in\mathbb{R}^d$ such that

$$P_i\subseteq H^+=\{x\in\mathbb{R}^d\,|\,a^\top x\le b\} \tag{3.35}$$

and

$$P_j\subseteq H^-=\{x\in\mathbb{R}^d\,|\,a^\top x> b\}\ . \tag{3.36}$$

Intuitively, linear separability assures that the convex hulls of different clusters do not overlap. It is an important observation that any optimal Bregman k-median clustering is linearly separable, as is given by the following lemma.

Lemma 3.7. *Let $P\subseteq\mathbb{X}$ and let $P_1,P_2,\dots,P_k$ be an optimal Bregman k-median clustering of P. Then $P_1,P_2,\dots,P_k$ is linearly separable.*

Proof. For all $i = 1, 2, \dots, k$ let $c_i = \mathrm{med}(P_i)$ denote the centroid of P_i. Then for each $p \in P_i$ we have $\mathrm{D}_\varphi(p, c_i) \leq \mathrm{D}_\varphi(p, c_j)$ for all $i \neq j$, since otherwise swapping point p from cluster P_i to cluster P_j would decrease the k-median cost, which contradicts the optimality of $P_1, P_2, \dots, P_k$.

Fix any two distinct indices i, j. We prove the lemma by showing that P_i, P_j are separated by the hyperplane

$$H_{ij} = \{x \in \mathbb{R}^d \,|\, a^\top x = b\} \tag{3.37}$$

where

$$a = \nabla\varphi(c_j) - \nabla\varphi(c_i) \ , \tag{3.38}$$

$$b = \nabla\varphi(c_j)^\top c_j - \nabla\varphi(c_i)^\top c_i - \varphi(c_j) + \varphi(c_i) \ . \tag{3.39}$$

To this end, note that if there are any $p \in P_i \cup P_j$ with $\mathrm{D}_\varphi(p, c_i) = \mathrm{D}_\varphi(p, c_j)$ we may assume without loss of generality that all these points are from P_i. This is valid since swapping these points from P_j to P_i can not increase the k-median clustering cost.

Hence, in the case $p \in P_j$ we have $\mathrm{D}_\varphi(p, c_i) > \mathrm{D}_\varphi(p, c_j)$, and we obtain

$$\varphi(x) - \varphi(c_i) - \nabla\varphi(c_i)^\top (x - c_i) > \varphi(x) - \varphi(c_j) - \nabla\varphi(c_j)^\top (x - c_j) \tag{3.40}$$

or, equivalently,

$$\big(\nabla\varphi(c_j) - \nabla\varphi(c_i)\big)^\top x > \nabla\varphi(c_j)^\top c_j - \nabla\varphi(c_i)^\top c_i - \varphi(c_j) + \varphi(c_i) \ . \tag{3.41}$$

This leads to $p \in H_{ij}^- = \{x \in \mathbb{R}^d \,|\, a^\top x > b\}$.

On the other hand, in the case $p \in P_i$ we have $\mathrm{D}_\varphi(p, c_i) \leq \mathrm{D}_\varphi(p, c_j)$. Analogously to the case above we obtain $p \in H_{ij}^+ = \{x \in \mathbb{R}^d \,|\, a^\top x \leq b\}$. We conclude that for any pair of distinct indices i, j there is a hyperplane that separates P_i and P_j. Thus, the k-clustering $P_1, P_2, \dots, P_k$ is linearly separable. □

The linear separability of optimal Bregman k-median clusterings has several important implications. First of all, this property allows the use of Voronoi-type diagrams in the context of Bregman k-median clustering (for an in-depth study of Bregman-Voronoi diagrams see [Nielsen et al., 2007]). Second, it allows to give a non-trivial size bound on the search space for an optimal solution of the Bregman k-median problem. Trivially, the number of distinct clusterings of a k-median problem is bounded by the number of partitions of P into k subsets. That is, there are at most k^n feasible clusterings to a general k-median problem. However, in case of linearly separable optimal clusterings, we learn that the number of potential optimal solutions can be much smaller, as is stated in the following theorem.

Lemma 3.8. *Let $P \subseteq \mathbb{R}^d$ be of size $|P| = n$. The number of linearly separable k-clusterings of P is bounded by $n^{d(k-1)k}$.*

Proof. Let $P_1, P_2, \ldots, P_k$ be an arbitrary linearly separable k-clustering of P. Furthermore, fix any index i. Since the k-clustering $P_1, P_2, \ldots, P_k$ is linearly separable we know that P_i is separated from the other $k-1$ clusters by at most $k-1$ oriented hyperplanes $H_1, H_2, \ldots, H_{k-1}$. Hence there exist at most $k-1$ halfspaces $H_1^+, H_2^+, \ldots, H_{k-1}^+$ such that

$$P_i = P \cap \bigcap_{j=1}^{k} H_j^+ = \bigcap_{j=1}^{k} \left(P \cap H_j^+\right) \ . \tag{3.42}$$

Now, consider a single hyperplane H_j. In d-dimensional space $\mathbb{R}^d$, there exists another hyperplane G_j that contains d points from P such that

$$P \cap H_j^+ = P \cap G_j^+ \ . \tag{3.43}$$

Hence, $P \cap H_j^+$ is properly determined by selecting d points from P. This argument can be repeated for each of the $k-1$ hyperplanes $H_1, H_2, \ldots, H_{k-1}$. Therefore, cluster P_i is properly determined by selecting at most $d(k-1)$ points from P.

It follows that there are at most $n^{d(k-1)}$ different subsets P_i that may occur in a linearly separable k-clustering of P. Thus, there are at most $n^{d(k-1)k}$ collections of k such subsets, and the lemma follows. □

Hence, if dimension d and number of clusters k are constant, there are merely a polynomial number of potential optimal solutions. Furthermore, as described in the proof of Lemma 3.8, these optimal partitions can be enumerated explicitly by constructing all $\mathcal{O}(dk^2)$-tupels of points from P, building the corresponding hyperplanes, and partitioning the point set in time $\mathcal{O}(d^2 n)$ for each partition. Thus, by enumerating all Bregman-Voronoi partitions of P and returning the partition with minimal k-median cost, we obtain from Lemma 3.8 the following straightforward generalization of a classical result from [Boros and Hammer, 1989] and [Hasegawa et al., 1993].

Corollary 3.9. *Let D_φ be an arbitrary Bregman divergence on domain $\mathbb{X} \subseteq \mathbb{R}^d$. Furthermore, let $P \subseteq \mathbb{X}$ be of size $|P| = n$. Then the k-median problem with respect to D_φ and input instance P can be solved optimally using at most $n^{\mathcal{O}(dk^2)}$ arithmetic operations, including evaluation of D_φ.*

It has to be mentioned that in [Inaba et al., 1994] the bound from Lemma 3.8 has been improved to $n^{\mathcal{O}(dk)}$ for the case of the Euclidean k-means clustering. Using the connection between the squared Euclidean distance and any Mahalanobis distance from Lemma 2.13, we find that this strengthened bound also applies to k-median clustering using Mahalanobis distances. Unfortunately, the proof from [Inaba et al., 1994] does not generalize to the case of arbitrary Bregman k-median problems.

In addition to the observations from this section, we make use of the linear separability in the simple optimal algorithm given in the next section.

3.3 A simple optimal algorithm for $d = 1$

In this section we give a simple algorithm for solving the Bregman k-median problem in dimension $d = 1$ optimally in polynomial time. This algorithm demonstrates that it is possible to give efficient k-median clustering algorithms that do not rely on metric properties such as symmetry or the triangle inequality. Rather, the algorithm given in this section relies on the combinatorial properties of optimal Bregman k-median clusterings.

The algorithm given in this section is assumed to be folklore to the scientific community, although its origin seems to be unknown. To the best of this author's knowledge, the earliest reference to this algorithm has been given in [Brucker, 1977] in the context of Euclidean k-means clustering.

The algorithm we give for the Bregman k-median problem relies on the following two properties:

(a) Every two distinct clusters of an optimal k-median clustering of P are separated by a hyperplane. This is guaranteed by the *linear separability property* from Lemma 3.7.

(b) Every selection of $k - 1$ clusters of an optimal k-median clustering of P forms an optimal $(k - 1)$-median clustering of the points from these clusters. This is given by the *optimal substructure property* from Lemma 3.4.

Hence, we can use the following simple, recursive strategy to solve the Bregman k-median problem optimally: First, find any one optimal cluster and remove all of its points from the input point set. Then, recursively, solve the $(k-1)$-median problem on the remaining point set. The optimal substructure property guarantees that the solution found this way is optimal.

Of course, we do not know the clusters of an optimal k-median clustering in advance. Here, the linear separability property comes into play. In the following, let D_φ be a Bregman divergence on domain $\mathbb{X} \subseteq \mathbb{R}$, and let $P \subseteq \mathbb{X}$ be finite, i.e., P is the input set of a Bregman k-median problem in dimension $d = 1$. Without loss of generality, we may assume that the input points are given in non-decreasing order, that is, $P = \{p_1, p_2, \ldots, p_n\}$ with $p_1 \leq p_2 \leq \ldots \leq p_n$. Now, assume we want to find the optimal cluster $P' \subseteq P$ that contains the largest input point p_n. From the linear separability property we know that $P' = \{p_t, p_{t+1}, \ldots, p_n\}$ for some index $1 \leq t \leq n$ (or, more precisely, $k \leq t \leq n$ since we assume the remaining $k-1$ clusters to be non-empty). Note that the number of all possible linearly separable subsets that contain p_n is merely linear in n, and independent of k. Thus, to find an optimal k-median clustering, all we have to do is to try all values of t, solve the $(k-1)$-median problem on $P \setminus P' = \{p_1, p_2, \ldots, p_{t-1}\}$ recursively, and return the best clustering obtained this way.

This approach is summarized in the following recursive algorithm on input $P = \{p_1, p_2, \ldots, p_n\} \subseteq \mathbb{X}$ and $k \in \mathbb{N}$:

1. If $k = 1$ return $\mathrm{med}(P)$ as the optimal 1-median clustering of P and terminate. Otherwise continue with step 2.
2. For each t with $k \leq t \leq n$, repeat:
    a) Let $P' = \{p_t, p_{t+1}, \ldots, p_n\}$.
    b) Recursively, solve the $(k-1)$-clustering problem optimally for input set $\{p_1, p_2, \ldots, p_{t-1}\}$. Let $P_1, P_2, \ldots, P_{k-1}$ denote the optimal $(k-1)$-median clustering found this way.
    c) Compute the k-median cost of the k-clustering $P_1, \ldots, P_{k-1}, P'$ to store the k-median clustering with minimal cost seen so far.
3. Finally, return the k-clustering with minimal cost.

However, a straight-forward recursive implementation of this strategy fails to achieve a running time that is polynomial in n and k. This problem is avoided using a dynamic programming implementation of the recursive

```
SimpleCluster1D$(P, k)$:
 $P$   ordered set of input points $\{p_i\}_i$ with $p_1 \leq p_2 \leq \ldots \leq p_n$
 $k$   number of medians to be found with $k \leq n$

for $i = 1, 2, \ldots, k$ do
    $c \leftarrow \mathrm{med}\big(\{p_1. \ldots, p_i\}\big)$
    $B[i, 1] \leftarrow \mathrm{cost}\big(\{p_1. \ldots, p_i\}, c\big)$
    $C[i, 1] \leftarrow \{c\}$
end for
for $j = 2, 3, \ldots, k$ do
    for $i = j, j + 1, \ldots, n$ do
        $B[i, j] \leftarrow \infty$
        $C[i, j] \leftarrow \emptyset$
        for $t = j, j + 1, \ldots, i$ do
            $c \leftarrow \mathrm{med}\big(\{p_t. \ldots, p_i\}\big)$
            $b \leftarrow B[t - 1, j - 1] + \mathrm{cost}\big(\{p_t. \ldots, p_i\}, c\big)$
            if $b < B[i, j]$ then
                $B[i, j] \leftarrow b$
                $C[i, j] \leftarrow C[t - 1, j - 1] \cup \{c\}$
            end if
        end for
    end for
end for
return $C[n, k]$
```

Figure 3.1: A simple, optimal k-median clustering algorithm for $d = 1$.

approach, as is given in detail in Figure 3.1. We obtain the following result for the 1-dimensional Bregman k-median problem.

Theorem 3.10. *Let* D_φ *be a Bregman divergence on domain* $\mathbb{X} \subseteq \mathbb{R}$. *Algorithm* SimpleCluster1D *computes an optimal solution to the* k*-median problem with respect to* D_φ *for input instance* P *of size* n *using at most* $\mathcal{O}(kn^3)$ *arithmetic operations, including evaluations of* D_φ.

Unfortunately, the strategy of this simple algorithm fails to achieve a running time polynomial in n and k for any fixed dimension $d \geq 2$. To see this, recall that in the case $d = 1$ there exists a distinguished point $p \in P$ such that the number of linearly separable subsets of P that contain p is merely linear in n and independent of k. In general, this is already no longer true in the case $d = 2$, as we easily see using the following counterexample. Let $\mathbb{X} \subseteq \mathbb{R}^2$ and assume that $\mathbb{X}$ contains the 2-dimensional unit circle

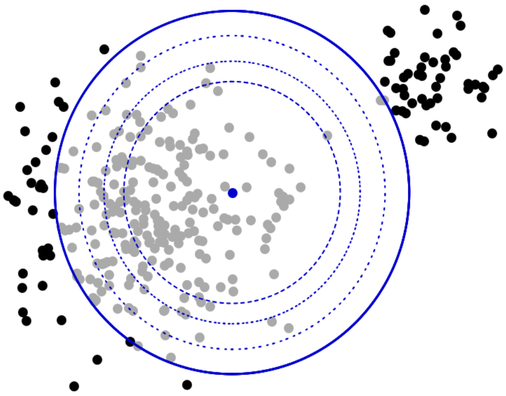
$S = \{(x_1, x_2)^\top \in \mathbb{R}^2 \,|\, x_1^2 + x_2^2 = 1\}$. We consider the input set P consisting of n points placed evenly on the unit circle S. Fix an arbitrary $p \in P$ and let us derive a lower bound on the number of subsets of P that contain p and that are separated from the remaining points of P by up to $k-1$ hyperplanes. To this end, since P lies on the unit circle S, we observe that for each $q \in P \setminus \{p\}$ there exists an oriented hyperplane H_q and a halfspace H_q^+ defined by H_q such that $P \cap H_q^+ = P \setminus \{q\}$. Therefore, we obtain a different, linearly separable subset of P containing p for each selection of $k-1$ points $q_1, q_2, \ldots, q_{k-1}$ from $P \setminus \{p\}$. Thus, the number of such subsets is bounded from below by

$$\binom{n-1}{k-1} = \frac{k}{n}\binom{n}{k} \geq \frac{n^{k-1}}{k^{k-1}} \,. \tag{3.44}$$

Hence, for a reasonably small number of clusters of, say, $k \leq n^{1-\varepsilon}$ with an arbitrarily small constant $\varepsilon > 0$, we obtain that the number of linearly separable subsets of P containing any fixed point p is at least $n^{\Omega(k)}$.

Thus, the strategy of this simple algorithm provides little help if we seek to solve the Bregman k-median problem optimally in arbitrary dimension d. However, as we learn in Chapter 4, the general idea of this strategy still leads to an efficient, randomized approximation scheme in arbitrary dimension d for the generalized k-median problem with respect to a large number of metric and non-metric dissimilarity measures, including all μ-similar Bregman divergences.

4 $(1+\varepsilon)$-approximate clustering by uniform sampling



In this chapter we study the generalized k-median problem with respect to an arbitrary dissimilarity measure D. We give an asymptotically fast linear time algorithm that relies on the usefulness of uniform sampling.

Our main result can be roughly stated as follows. For every dissimilarity measure exists a linear time $(1+\varepsilon)$-approximation algorithm for the k-median problem, provided that the 1-median problem can be approximated within a factor of $(1+\varepsilon)$ by taking a random sample of constant size and solving the 1-median problem on the sample exactly. In this way we show an interesting connection between sampleability and clusterability: For the existence of good approximation algorithms it is sufficient to guarantee that a constant sized sample set contains enough information to approximate the median of the original set. The interesting aspect of this characterization is that it needs no further assumption whatsoever on the properties of the dissimilarity measure. Hence, it is well-suited for arbitrary, non-metric k-median problems (e.g., the Bregman k-median problem).

Stated in detail, we formulate the following sampling property.

Property 4.1. *Let $\gamma > 0$ and $0 < \delta < 1$ be arbitrary constants. We say a dissimilarity measure* D *satisfies the* (strong) $[\gamma, \delta]$-sampling property *if the following conditions are satisfied.*

(a) There exists an algorithm that for every finite subset $S \subseteq \mathbb{X}$ computes an optimal 1-median $\mathrm{med}(S)$ *of S in time depending only on $|S|$.*

(b) There exists a constant $m_{\gamma,\delta} \in \mathbb{N}$ such that for every subset $P \subseteq \mathbb{X}$ of size n and for every uniform sample multiset $S \subseteq P$ of size $m_{\gamma,\delta}$ an optimal 1-median $\mathrm{med}(S) \in \mathbb{X}$ *satisfies*

$$\Pr\left[\mathrm{cost}\big(P, \mathrm{med}(S)\big) \leq (1+\gamma)\, opt_1(P)\right] \geq 1-\delta \ . \tag{4.1}$$

Here, condition (a) captures mainly the fact that the problem is well-posed, i.e. an optimal 1-median $\mathrm{med}(P)$ is computable. Condition (b) requires that, with high probability, the median a constant size uniform sample is a good approximate solution for the 1-median problem of the original set. Using this property, in Section 4.1, we show the following result.

Theorem 4.2. *Let $k \in \mathbb{N}$ and let $0 < \varepsilon, \delta < 1$ be arbitrary constants. Let* D *be a dissimilarity measure on domain $\mathbb{X}$ satisfying the $[\varepsilon/3, \delta]$-sampling property. Then there exists an algorithm that, with constant probability, computes a $(1+\varepsilon)$-approximate solution of the k-median problem with respect to* D *for any input instance P of size n. Furthermore, this solution can be found using at most $2^{\mathcal{O}(mk\log(mk/\varepsilon))}n$ operations, including evaluations of* D*, where m is a constant that depends only on ε, δ, and* D.

Using this characterization, we obtain linear time $(1+\varepsilon)$-approximation algorithms for the k-median problem with respect to a number of metric and non-metric distance measures, such as:

- k-median clustering in $\mathbb{R}^d$ with respect to an arbitrary Mahalanobis distance.

- k-median clustering in $[\lambda, \upsilon]^d \subseteq \mathbb{R}_+^d$ with respect to the Kullback-Leibler divergence, where λ, υ with $\lambda < \upsilon$ are arbitrary positive constants. This is the first approximation algorithm for k-median clustering using the Kullback-Leibler divergence that provides *any* non-trivial approximation ratio.

- k-median clustering in $[\lambda, \upsilon]^d \subseteq \mathbb{R}_+^d$ with respect to the Itakura-Saito divergence, where λ, υ with $\lambda < \upsilon$ are arbitrary positive constants. This is the first approximation algorithm for k-median clustering using the Itakura-Saito divergence that provides *any* non-trivial approximation ratio.

- k-median clustering with respect to any μ-similar Bregman divergence defined on domain $\mathbb{X} \subseteq \mathbb{R}^d$. This is the first approximation algorithm that provides *any* non-trivial approximation ratio for a large number of Bregman k-median problems.

- k-median clustering in $\mathbb{R}_{\geq 0}^d$ with respect to the Hellinger distance.

- k-median clustering in an arbitrary metric space $(\mathbb{X}, \mathrm{D})$ with bounded doubling dimension, provided that D satisfies condition (a).

In addition to that, a previously known result from [Kumar et al., 2004] states that there exists a linear time $(1 + \varepsilon)$-approximation algorithm for the Euclidean k-means problem. Using our characterization, we obtain the same result (as an instance of the k-median problem with Mahalanobis distances) in a simplified manner. We also confirm an observation from [Ailon et al., 2006] by using our approach to give a $(1 + \varepsilon)$-approximation algorithm for the k-median problem on the Hamming cube $\{0, 1\}^d$.

To obtain our results for specific dissimilarity measures like the Mahalanobis distances, Kullback-Leibler divergence, Itakura-Saito divergence, etc. we show in Section 4.2 that the optimal median of a constant sized uniform sample set $S \subseteq P$ is an approximate median of P. We also prove such results for arbitrary doubling metrics, the Hamming distance, and non-metric, non-Bregman distance measures such as the Hellinger distance. These sampling results are the second main contribution of this chapter.

However, for some dissimilarity measures like the Euclidean distance on $\mathbb{R}^d$, condition (a) is not satisfied because one can not compute an exact solution to the Euclidean 1-median problem. In literature, this is known as the Fermat-Weber problem (cf. [Weber, 1909] and [Bajaj, 1988]). To deal with these problems, we relax our sampling property in Section 4.3 to what we call the *weak* $[\gamma,\delta]$*-sampling property.* Even under this weaker assumption we are able to obtain linear time $(1+\varepsilon)$-approximation algorithms. Thereby, we show that the Euclidean k-median problem also fits into our framework, and we obtain a previously known result from [Kumar et al., 2005] in a simplified manner.

The results from this chapter (except for the results addressing the Hellinger distance and the Hamming metric) have already been published in [Ackermann et al., 2008] and [Ackermann et al., 2010a].

4.1 Algorithm for $[\gamma,\delta]$-sampleable dissimilarity measures

In this section we describe and analyze our main algorithm. To obtain our main result we give a generalized and improved analysis of algorithm IRRED-k-MEANS from [Kumar et al., 2004]. This algorithm has been generalized to other clustering problems before. In [Kumar et al., 2005], sufficient conditions for dissimilarity measures have been given that allow for the application of the algorithm from [Kumar et al., 2004]. However, symmetry and the triangle inequality are always assumed. Our generalization does not require these assumptions. Instead, we give a purely combinatorial analysis. Therefore, we are able to obtain results for non-metric dissimilarity measures like the Kullback-Leibler divergence, which seems to be impossible using previous results.

Our new approach does not only generalize to non-metric dissimilarity measures, it can also be used to obtain the results for the Euclidean k-median and the Euclidean k-means problem from [Kumar et al., 2004, Kumar et al., 2005]. Moreover, these results are obtained by a significantly simplified analysis.

In the following, let dissimilarity measure D satisfy the $[\gamma,\delta]$-sampling property. Furthermore, let constant $m_{\gamma,\delta}$ and the algorithm computing mapping $\mathrm{med}(\cdot)$ be as required in Theorem 4.2.

4.1.1 Superset sampling

Our algorithm makes use of the *superset sampling technique* introduced in [Kumar et al., 2004]. This technique is used in the following way. For input instance P of size n let $P' \subseteq P$ be a subset of at least a constant fraction of the elements of P, say $|P'| \geq \alpha n$ with constant $\alpha > 0$. We want to draw a uniform sample multiset of size m from P' without knowing P' explicitly. The main observation of the superset sampling technique states that if we take a slightly larger uniform sample from P and inspect all its subsets of size m, then with constant probability we will find a uniform sample set from P' among these subsets.

This technique is an immediate consequence of probabilistic concentration bounds. Since each sampled point comes from P' with probability α, the expected number of sampled points from P' is αm. Hence, oversampling by a factor of $\Theta(1/\alpha)$ will, with constant probability, provide us with at least m points from P'.

Using the superset sampling technique for $[\gamma, \delta]$-sampleable D gives us the following lemma.

Lemma 4.3 (superset sampling lemma)**.** *Let* D *satisfy the* $[\gamma, \delta]$*-sampling property. Let* $P \subseteq \mathbb{X}$ *be of size* n *and let* $P' \subseteq P$ *be with* $|P'| \geq \alpha n$ *for some constant* $\alpha > 0$*. Let* $S \subseteq P$ *be a uniform sample multiset of size at least* $2m_{\gamma,\delta}/\alpha$*. Then there exists with probability at least* $(1-\delta)/5$ *a subset* $S' \subseteq S$ *with* $|S'| = m_{\gamma,\delta}$ *and optimal 1-median* $\mathrm{med}(S')$ *satisfying*

$$\mathrm{cost}\big(P', \mathrm{med}(S')\big) \leq (1+\gamma)\, opt_1(P') \ . \tag{4.2}$$

Proof. Let random variable X denote the number of points from P' contained in sample set S. Obviously, a point $p \in S$ is from P' with probability at least α. Hence,

$$\mathrm{E}[X] = \frac{2}{\alpha} m_{\gamma,\delta} \Pr[p \in S] \geq 2m_{\gamma,\delta} \ . \tag{4.3}$$

Using a Chernoff bound and $m_{\gamma,\delta} \geq 1$ we obtain

$$\Pr[X < m_{\gamma,\delta}] \leq \Pr\left[X < \frac{1}{2}\,\mathrm{E}[X]\right] \leq \exp\left(-\frac{1}{4} m_{\gamma,\delta}\right) \leq 0.78 < \frac{4}{5} \ . \tag{4.4}$$

Hence, with probability at least $\frac{1}{5}$, sample set S contains at least $m_{\gamma,\delta}$ points from P'. Let $S' \subseteq S \cap P'$ be such a subset of size $m_{\gamma,\delta}$, chosen uniformly at random. Then S' is chosen uniformly at random among all $m_{\gamma,\delta}$-sized subsets of P'. By the $[\gamma, \delta]$-sampling property we know that with probability at least $1-\delta$ the median of S' is a $(1+\gamma)$-approximate 1-median of P', and the lemma follows. □

For a sample set $S \subseteq P$ of size $2m_{\gamma,\delta}/\alpha$ let

$$T = \left\{ \mathrm{med}(S') \,\middle|\, S' \subset S,\ |S'| = m_{\gamma,\delta} \right\} \tag{4.5}$$

be the medians of all $m_{\gamma,\delta}$-sized subsets of S. As an immediate consequence of Lemma 4.3, for any fixed subset $P' \subseteq P$ of size $|P'| \geq \alpha|P|$ we find that with constant probability set T contains at least one $(1+\gamma)$-approximate 1-median of P'. We call the elements of T candidates for approximate medians of P'. Note that for constants α and $m_{\gamma,\delta}$ the candidate set T is also of constant size

$$|T| \leq \binom{\frac{2}{\alpha} m_{\gamma,\delta}}{m_{\gamma,\delta}} \leq 2^{\mathcal{O}(m_{\gamma,\delta} k \log(m_{\gamma,\delta} k/\varepsilon))} \ . \tag{4.6}$$

4.1.2 The algorithm

The algorithm we propose in this section bears some resemblance to the initial idea of our simple, optimal algorithm for the case of dimension $d = 1$, given in Section 3.3. For the case of $d = 1$, we were able to give a very simple, recursive strategy for the k-median problem: First, find any one optimal cluster and remove all of its points from the input point set. Then, recursively solve the $(k-1)$-median problem on the remaining point set. We can use this strategy in the one-dimensional case since for $d = 1$ we are able to enumerate all potential optimal clusters efficiently. As we have seen, this property does not hold for the case of $d \geq 2$.

However, algorithm Cluster below captures the spirit of this idea. We adapt our strategy to the arbitrary dimensional case through the use of randomization. Instead of an intractable enumeration of all potential optimal clusters we use uniform sampling to efficiently approximate a median of an optimal cluster. Then, this cluster is removed as accurately as possible from the input set and the $(k-1)$-median problem is solved recursively.

To understand the pits and snares of this adaptation, let us first consider an idealized version of our algorithm Cluster for the case of $k = 2$. Let P_1 and P_2 denote the clusters of an optimal 2-median clustering of input set P, and assume $|P_1| \geq \alpha|P|$. Here $0 < \alpha < 1$ is a constant parameter to be specified later. Our idealized strategy can be stated as follows.

1. Use the superset sampling technique to obtain an approximate median for optimal cluster P_1, that is, a $\tilde{c}_1$ from P with

$$\mathrm{cost}(P_1, \tilde{c}_1) \leq (1+\gamma)\, opt_1(P_1) \ . \tag{4.7}$$

2. Let $N \subseteq P$ be the smallest subset such that
   (i) $\mathrm{D}(p, \tilde{c}_1) \leq \mathrm{D}(q, \tilde{c}_1)$ for all $p \in N$ and $q \in P \setminus N$ and
   (ii) for the remaining points $R = P \setminus N$ we have $|P_2 \cap R| \geq \alpha |R|$.

   Assign N to $\tilde{c}_1$.

3. Use the superset sampling technique again to obtain an approximate median for the points from P_2 within the remaining point set R, that is, a $\tilde{c}_2$ from R with
$$\mathrm{cost}(P_2 \cap R, \tilde{c}_2) \leq (1+\gamma)\, opt_1(P_2 \cap R) \ . \tag{4.8}$$

4. Assign all remaining points to their closest approximate median and return $\{\tilde{c}_1, \tilde{c}_2\}$ as $(1+\gamma)$-approximate solution.

This idealized strategy faces two problems. First, using the superset sampling technique we do not get a single approximate median $\tilde{c}_1$. Instead, we get a set T_1 of candidates for approximate medians. To solve this problem we simply try all possible candidates as approximate median $\tilde{c}_1$ and choose the candidate which leads to minimal cost. Recall that for constants α and $m_{\gamma,\delta}$ the candidate set T_1 is also of constant size $2^{\mathcal{O}(m_{\gamma,\delta} k \log(m_{\gamma,\delta} k/\varepsilon))}$. The same procedure is used for obtaining $\tilde{c}_2$ in step 3.

Second, it is obvious that we do not know the optimal clusters P_1 and P_2. Thus, we do not know how to choose N from step 2 explicitly. To cope with this problem we approximate N by partitioning P into subsets $N^{(1)}, N^{(2)}, \ldots, N^{(\lceil \log n \rceil)}$. Here, $N^{(1)}$ denotes the $\frac{n}{2}$ closest points towards $\tilde{c}_1$, $N^{(2)}$ the next $\frac{n}{4}$ closest points, $N^{(3)}$ the next $\frac{n}{8}$ closest points, and so on. Let
$$R^{(j)} = P \setminus \bigcup_{i=1}^{j} N^{(i)} \tag{4.9}$$
and let ν be the minimal index such that
$$|P_2 \cap R^{(\nu)}| \geq \alpha |R^{(\nu)}| \ . \tag{4.10}$$
Instead of N we will assign the points from $N^{(1)} \cup N^{(2)} \cup \ldots \cup N^{(\nu)}$ to $\tilde{c}_1$.

Of course, we still do not know the index ν. However, we can guess ν by trying all $\Theta(\log n)$ possible values and choosing the value that leads to minimal cost.

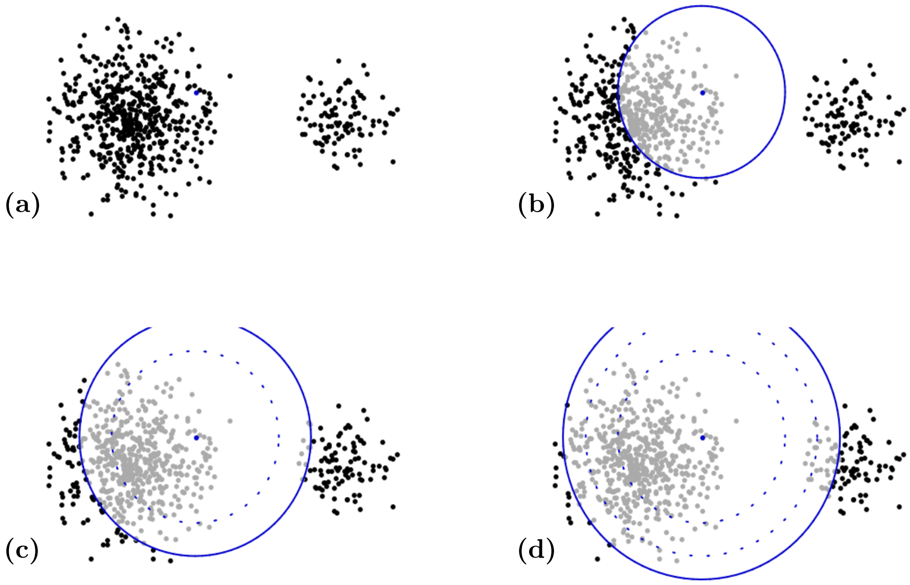


Figure 4.1: An illustration of the strategy of algorithm Cluster. In this example, a point set P has to be partitioned into $k=2$ clusters. **(a)** First, an approximate median $\tilde{c}_1$ for the larger cluster to the left is obtained by the superset sampling approach from Lemma 4.3. **(b)–(d)** Then, step by step, the input points closest to $\tilde{c}_1$ are removed. Eventually, the fraction of the remaining points from the smaller cluster to the right becomes large enough such that the superset sampling technique can be applied to obtain an approximate median of the smaller cluster.

The strategy of our algorithm is depicted in Figure 4.1. Figure 4.2 gives a precise recursive definition of Algorithm Cluster that uses the aforementioned strategy for arbitrary k. The algorithm alternates between two different types of phases: A *sampling phase* and a *pruning phase*. In the sampling phase, new candidates for approximate medians are computed (according to Lemma 4.3) and tried in a recursive manner. In the pruning phase, new values for ν are tried: The next $\frac{1}{2}|R|$ closest points to already computed approximate medians are assigned and discarded from future consideration. Hence, the algorithm computes a set of possible solutions to the k-median problem for all candidates of $\tilde{c}_i$ and all values of ν. In the final step the algorithm chooses the best solution found this way.

```
CLUSTER($R, l, \tilde{C}$):
  $R$   set of remaining input points
  $l$   number of medians yet to be found
  $\tilde{C}$   set of medians already found

if $l = 0$ then return $\tilde{C}$
else
    if $l \geq |R|$ then return $\tilde{C} \cup R$
    else
        /* sampling phase */
        sample a multiset $S$ of size $2m_{\gamma,\delta}/\alpha$ uniformly at random from $R$
        $T \leftarrow \{\mathrm{med}(S') \mid S' \subseteq S,\ |S'| = m_{\gamma,\delta}\}$
        for all $\tilde{c} \in T$ do
            $C^{(\tilde{c})} \leftarrow$ CLUSTER($R, l-1, \tilde{C} \cup \{\tilde{c}\}$)
        end for
        /* pruning phase */
        let $N$ be the set of the $\frac{1}{2}|R|$ minimal points $p \in R$ w.r.t. $\mathrm{D}(p, \tilde{C})$
        $C^* \leftarrow$ CLUSTER($R \setminus N, l, \tilde{C}$)
        return $C^{(\tilde{c})}$ or $C^*$ with minimal cost
    end if
end if
```

Figure 4.2: Algorithm CLUSTER for arbitrary k and fixed positive real constants α, γ, δ.

4.1.3 Analysis for $k = 2$

To simplify notation, we first analyze algorithm CLUSTER for the case of $k = 2$. In the following, let D satisfy the $[\gamma, \delta]$-sampling property.

Theorem 4.4. *Let $\alpha < \frac{1}{4}$ be an arbitrary positive constant. Then algorithm* CLUSTER *started with parameters $(P, 2, \emptyset)$ computes a solution $\tilde{C}$ of the 2-median problem for input instance P of size n satisfying*

$$\Pr\left[\mathrm{cost}(P, \tilde{C}) \leq (1 + 8\alpha)(1 + \gamma)\, opt_2(P)\right] \geq \left(\frac{1-\delta}{5}\right)^2 . \tag{4.11}$$

Proof. Assume for simplicity of notation that n is a power of 2. Furthermore, let P_1 and P_2 denote the clusters of the optimal 2-clustering of P with the optimal set of medians $C = \{c_1, c_2\}$, i.e. $\mathrm{cost}(P, C) = opt_2(P)$ and $\mathrm{cost}(P_i, c_i) = opt_1(P_i)$ for $i = 1, 2$. Assume

$$|P_1| \geq \frac{1}{2}|P| > \alpha|P| . \tag{4.12}$$

Denote by T_1 the candidate set from step 7 during the initial call of the algorithm. By Lemma 4.3 with probability at least $(1-\delta)/5$ we have that set T_1 contains a $\tilde{c}_1$ with

$$\mathrm{cost}(P_1, \tilde{c}_1) \leq (1+\gamma)\,\mathrm{cost}(P_1, c_1)\ . \tag{4.13}$$

We consider two cases. First, we assume that during the execution of algorithm CLUSTER there exists a recursive call with parameters $\big(R, 1, \{\tilde{c}_1\}\big)$ such that $|P_2 \cap R| \geq \alpha|R|$. Later we consider the case when there is no such recursive call.

So let us assume there exists a recursive call with

$$|P_2 \cap R| \geq \alpha|R|\ . \tag{4.14}$$

Let R be the largest input set with that property. Let T_2 be the candidate set from step 7 of this call. Again by Lemma 4.3 with probability $(1-\delta)/5$ set T_2 contains a $\tilde{c}_2$ satisfying

$$\mathrm{cost}(P_2 \cap R, \tilde{c}_2) \leq (1+\gamma)\,\mathrm{cost}(P_2 \cap R, c_2')\ . \tag{4.15}$$

Here c_2' denotes the optimal 1-median of $P_2 \cap R$, i.e. $\mathrm{cost}(P_2 \cap R, c_2') = opt_1(P_2 \cap R)$. Hence, with probability $\big((1-\delta)/5\big)^2$ a set $\tilde{C} = \{\tilde{c}_1, \tilde{c}_2\}$ satisfying inequalities (4.13) and (4.15) is found by algorithm CLUSTER. Thus, $\mathrm{cost}(P, \tilde{C})$ yields an upper bound on the cost of the solution returned by our algorithm.

Let $N = P \setminus R$ denote the neighboring points removed between the sampling of $\tilde{c}_1$ and $\tilde{c}_2$ by step 12 of the algorithm. Note that the sets P_1, $P_2 \cap N$, and $P_2 \cap R$ form a disjoint partition of P. Here the sets P_1 and $P_2 \cap R$ contain the points that the approximate medians $\tilde{c}_1$ and $\tilde{c}_2$ have been sampled from by the superset sampling technique, while the set $P_2 \cap N$ contains the points incorrectly assigned to $\tilde{c}_1$ during the pruning phases of the algorithm. Using

$$\mathrm{cost}(P, \tilde{C}) \leq \mathrm{cost}(P_1, \tilde{c}_1) + \mathrm{cost}(P_2 \cap N, \tilde{c}_1) + \mathrm{cost}(P_2 \cap R, \tilde{c}_2) \tag{4.16}$$

we bound each term of the sum individually. Using Claim 4.5 and 4.6 stated below we conclude

$$\begin{aligned} \mathrm{cost}(P, \tilde{C}) &\leq (1+8\alpha)\,\mathrm{cost}(P_1, \tilde{c}_1) + \mathrm{cost}(P_2 \cap R, \tilde{c}_2) & (4.17)\\ &\leq (1+8\alpha)(1+\gamma)\,\mathrm{cost}(P_1, c_1) + (1+\gamma)\,\mathrm{cost}(P_2, c_2) & (4.18)\\ &\leq (1+8\alpha)(1+\gamma)\, opt_2(P)\ . & (4.19) \end{aligned}$$

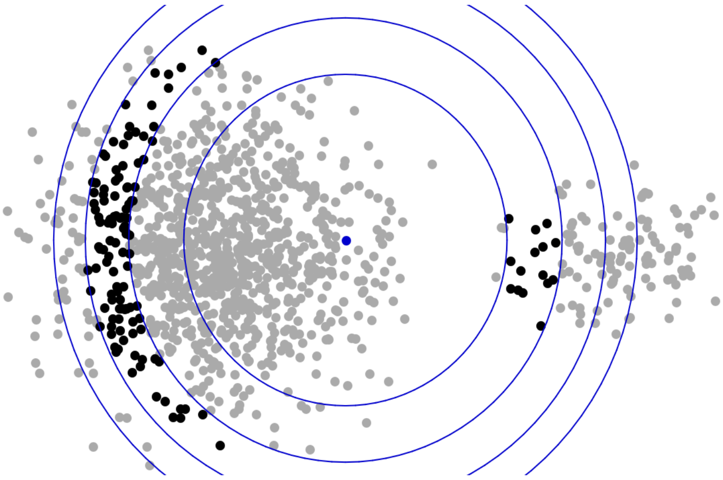


Figure 4.3: An illustration of the charging of the cost of points in the proof of Claim 4.5. The total cost of the few incorrectly assigned points from $N^{(j)}$ (the dark points from the smaller cluster to the right) is negligible when compared to the total cost of the many correctly assigned points from $N^{(j+1)}$ (the dark points from the larger cluster to the left).

Claim 4.5. $\text{cost}(P_2 \cap N, \tilde{c}_1) \leq 8\alpha \, \text{cost}(P_1, \tilde{c}_1)$.

Proof. Assume $N \neq \emptyset$, otherwise the claim is trivially true. Hence, N is the disjoint union of ν different subsets $N^{(j)}$ of size $\frac{n}{2^j}$ which correspond to the neighborhoods removed in step 12 of the algorithm, i.e.,

$$N = N^{(1)} \cup N^{(2)} \cup \ldots \cup N^{(\nu)} \ . \tag{4.20}$$

We prove the claim using the following strategy. First, we show that for each j the set $N^{(j)}$ contains a large number of points from P_1 and only a few points from P_2. Then, for each j we charge the cost of the few and inexpensive points from $N^{(j)} \cap P_2$ against the many and costly points from $N^{(j+1)} \cap P_1$. This strategy is illustrated in Figure 4.3.

To this end, define $R^{(0)} = P$ and $R^{(j)} = R^{(j-1)} \setminus N^{(j)}$. By definition we have $|R^{(j)}| = |N^{(j)}| = \frac{n}{2^j}$. Note that the $R^{(j)}$ have been input sets of recursive calls prior to the call on $R = R^{(\nu)}$ and, hence,

$$\forall j < \nu : \ |P_2 \cap R^{(j)}| < \alpha |R^{(j)}| \ . \tag{4.21}$$

We obtain

$$\forall j \le \nu : \ |P_2 \cap N^{(j)}| \le |P_2 \cap R^{(j-1)}| < \alpha |R^{(j-1)}| = 2\alpha \frac{n}{2^j} \tag{4.22}$$

where the first inequality holds since $N^{(j)} \subseteq R^{(j-1)}$. Using (4.22), we also get

$$\forall j \le \nu : \ |P_1 \cap N^{(j)}| = |N^{(j)}| - |P_2 \cap N^{(j)}| \ge (1-2\alpha)\frac{n}{2^j}. \tag{4.23}$$

Now we show that the cost of assigning $P_2 \cap N$ to $\tilde{c}_1$ is small. By definition of the $N^{(j)}$ we know that for all $j < \nu$ and for $p \in N^{(j)}$ and $p' \in N^{(j+1)}$ we have $\mathrm{D}(p, \tilde{c}_1) \le \mathrm{D}(p', \tilde{c}_1)$. Thus,

$$\forall j < \nu : \ \frac{\mathrm{cost}(P_2 \cap N^{(j)}, \tilde{c}_1)}{|P_2 \cap N^{(j)}|} \le \frac{\mathrm{cost}(P_1 \cap N^{(j+1)}, \tilde{c}_1)}{|P_1 \cap N^{(j+1)}|}. \tag{4.24}$$

Using (4.22) and (4.23) we get

$$\forall j < \nu : \ \frac{2^j}{2\alpha n} \mathrm{cost}(P_2 \cap N^{(j)}, \tilde{c}_1) \le \frac{2^{j+1}}{(1-2\alpha)n} \mathrm{cost}(P_1 \cap N^{(j+1)}, \tilde{c}_1) \tag{4.25}$$

or, equivalently,

$$\forall j < \nu : \ \mathrm{cost}(P_2 \cap N^{(j)}, \tilde{c}_1) \le \frac{4\alpha}{1-2\alpha} \mathrm{cost}(P_1 \cap N^{(j+1)}, \tilde{c}_1) \ . \tag{4.26}$$

We still need an upper bound on $\mathrm{cost}(P_2 \cap N^{(\nu)}, \tilde{c}_1)$. As a lower bound on the size of set $P_1 \cap R^{(\nu)}$ we obtain

$$\begin{aligned} |P_1 \cap R^{(\nu)}| &= |R^{(\nu)}| - |P_2 \cap R^{(\nu)}| && (4.27) \\ &\ge |R^{(\nu)}| - |P_2 \cap R^{(\nu-1)}| && (4.28) \\ &> (1-2\alpha)\frac{n}{2^\nu} \ . && (4.29) \end{aligned}$$

By definition of $N^{(\nu)}$ and $R^{(\nu)}$ we also know that for all $p \in N^{(\nu)}$ and $p' \in R^{(\nu)}$ we have $\mathrm{D}(p, \tilde{c}_1) \le \mathrm{D}(p', \tilde{c}_1)$. Analogously to above, combining (4.22) and (4.29) we conclude

$$\mathrm{cost}(P_2 \cap N^{(\nu)}, \tilde{c}_1) \le \frac{2\alpha}{1-2\alpha} \mathrm{cost}(P_1 \cap R^{(\nu)}, \tilde{c}_1). \tag{4.30}$$

Using (4.26) and (4.30) we obtain

$$\begin{aligned}
&\operatorname{cost}(P_2 \cap N, \tilde{c}_1) \\
&\quad = \sum_{j=1}^{\nu} \operatorname{cost}(P_2 \cap N^{(j)}, \tilde{c}_1) && (4.31)\\
&\quad \le \frac{4\alpha}{1-2\alpha} \sum_{j=1}^{\nu-1} \operatorname{cost}(P_1 \cap N^{(j+1)}, \tilde{c}_1) + \frac{2\alpha}{1-2\alpha} \operatorname{cost}(P_1 \cap R^{(\nu)}, \tilde{c}_1) && (4.32)\\
&\quad \le 8\alpha \sum_{j=1}^{\nu-1} \operatorname{cost}(P_1 \cap N^{(j+1)}, \tilde{c}_1) + 8\alpha \operatorname{cost}(P_1 \cap R^{(\nu)}, \tilde{c}_1) && (4.33)\\
&\quad \le 8\alpha \operatorname{cost}(P_1, \tilde{c}_1) && (4.34)
\end{aligned}$$

since $\frac{2\alpha}{1-2\alpha} \le \frac{4\alpha}{1-2\alpha} \le 8\alpha$ for $\alpha \le \frac{1}{4}$. □

Claim 4.6. $\operatorname{cost}(P_2 \cap R, \tilde{c}_2) \le (1+\gamma) \operatorname{cost}(P_2, c_2)$.

Proof. Let c_2' be the optimal 1-median of $P_2 \cap R$. By choice of $\tilde{c}_2$ we get

$$\begin{aligned}
\operatorname{cost}(P_2 \cap R, \tilde{c}_2) &\le (1+\gamma) \operatorname{cost}(P_2 \cap R, c_2') && (4.35)\\
&\le (1+\gamma) \operatorname{cost}(P_2 \cap R, c_2) && (4.36)\\
&\le (1+\gamma) \operatorname{cost}(P_2, c_2)\ . && (4.37)
\end{aligned}$$

□

Proof of Theorem 4.4 (continued). Finally, we consider the case when there has not been a recursive call on an input set R with $|P_2 \cap R| \ge \alpha|R|$. In this case there is a sequence of recursive calls consecutively using step 13 for $\nu = \lceil \log n \rceil$ times. We end up with a single point $q \in R$. This q can be assigned to its own cluster with median $\tilde{c}_2 = q$. This cluster does not contribute any to $\operatorname{cost}(P, \{\tilde{c}_1, \tilde{c}_2\})$. On the other hand, $\operatorname{cost}(P_2 \cap N, \tilde{c}_1)$ with

$$N = \bigcup_{j=1}^{\log n} N^{(j)} = P \setminus \{q\} \qquad (4.38)$$

is still bounded as given above, thus concluding the proof. □

4.1.4 Analysis for $k \geq 2$

We generalize the analysis of algorithm Cluster to the case $k \geq 2$, leading to the following theorem for D satisfying the $[\gamma, \delta]$-sampling property.

Theorem 4.7. *Let $\alpha < \frac{1}{4k}$ be an arbitrary positive constant. Then algorithm* Cluster *started with parameters $(P, k, \emptyset)$ computes a solution $\tilde{C}$ of the k-median problem for input instance P of size n such that*

$$\Pr\left[\operatorname{cost}(P, \tilde{C}) \leq (1 + 8\alpha k^2)(1+\gamma)\, opt_k(P)\right] \geq \left(\frac{1-\delta}{5}\right)^k . \tag{4.39}$$

Proof. This theorem can be proven analogously to the proof of Theorem 4.4. Generally speaking, during the execution of the algorithm we consider the two superclusters P_1' and P_2', with P_1' consisting of the clusters whose medians have already been approximated by $\tilde{C}_i = \{\tilde{c}_1, \tilde{c}_2, \ldots, \tilde{c}_i\}$ and P_2' consisting of the clusters whose medians have yet to be found. In this case, the goal of algorithm Cluster is to keep pruning points until P_2' becomes large enough to be considered in the sampling phase. Again, it can be shown that by removing neighborhoods in step 12 of the pruning phase only a small fraction of points from P_2' are removed. Eventually, the fraction of points from P_2' in the remaining point set will become large enough, and the superset sampling technique will find an approximate median for one of the at most k clusters in P_2' that have not yet been considered. As in case $k = 2$, the total cost of points incorrectly assigned to $\tilde{C}_i$ in the pruning phases will turn out to be bounded by $8\alpha k \operatorname{cost}(P_1', \tilde{C}_i)$. Summation over all k sampling phases leads to the given bound.

The analysis in the case when despite pruning the point set no new cluster becomes large enough to apply the superset sampling technique follows analogously to the analysis in the case $k = 2$. Therefore, we will from now on concentrate on the case when the k approximate medians $\tilde{c}_1, \tilde{c}_2, \ldots, \tilde{c}_k$ have been successfully found.

Let us assume for simplicity of notation that n is a power of 2. We use the notation as given in the proof of Theorem 4.4. I.e., let $P_1, P_2, \ldots, P_k$ denote the partition of P of the optimal k-clustering with the optimal set of medians $C = \{c_1, c_2, \ldots, c_k\}$. Hence, $\operatorname{cost}(P, C) = opt_k(P)$ and $\operatorname{cost}(P_i, c_i) = opt_1(P_i)$ for $i = 1, \ldots, k$. For the sake of brevity, we write $P_{[i,j]}$ for the disjoint union $\bigcup_{t=i}^{j} P_t$.

We assume that the P_i are numbered in the order their approximate medians $\tilde{c}_i$ are found by the superset sampling technique. That is, let $R_0 = P$ and let $R_1, R_2, \ldots, R_{k-1}$ with $R_i \subseteq R_{i-1}$ be the input sets from a sequence of recursive calls with $|P_i \cap R_{i-1}| \geq \alpha |R_{i-1}|$. Without loss of generality, let the R_i be the largest input sets with this property. By Lemma 4.3, with probability at least $((1-\delta)/5)^k$ we have

$$\mathrm{cost}(P_i \cap R_{i-1}, \tilde{c}_i) \leq (1+\gamma)\, opt_1(P_i \cap R_{i-1}) \tag{4.40}$$

for all $i = 1, 2, \ldots, k$.

We denote the neighboring points removed between two sampling phases by step 12 of the algorithm by $N_i = R_{i-1} \setminus R_i$. It is easy to check that the sets

$$\begin{gathered} P_1 \cap R_0,\ P_2 \cap R_1,\ \ldots,\ P_k \cap R_{k-1}, \\ P_{[2,k]} \cap N_1,\ P_{[3,k]} \cap N_2,\ \ldots, P_{[k,k]} \cap N_{k-1} \end{gathered} \tag{4.41}$$

form a disjoint partition of P. Here set $P_i \cap R_{i-1}$ contains the points that the approximate median $\tilde{c}_i$ has been sampled from by the superset sampling technique, while set $P_{[i+1,k]} \cap N_i$ contains the points incorrectly assigned to $\tilde{C}_i = \{\tilde{c}_1, \tilde{c}_2, \ldots, \tilde{c}_i\}$ during the pruning phases between the sampling of $\tilde{c}_i$ and $\tilde{c}_{i+1}$. Using

$$\mathrm{cost}(P, \tilde{C}) \leq \sum_{i=1}^{k} \mathrm{cost}(P_i \cap R_{i-1}, \tilde{c}_i) + \sum_{i=1}^{k-1} \mathrm{cost}(P_{[i+1,k]} \cap N_i, \tilde{C}_i) \tag{4.42}$$

we bound each term of the sums individually. Using Claim 4.8 stated below we obtain

$$\mathrm{cost}(P, \tilde{C}) \leq \sum_{i=1}^{k} \mathrm{cost}(P_i \cap R_{i-1}, \tilde{c}_i) + 8\alpha k \sum_{i=1}^{k-1} \mathrm{cost}(P_{[1,i]} \cap R_{i-1}, \{\tilde{c}_1, \ldots, \tilde{c}_i\}) \tag{4.43}$$

$$\leq \sum_{i=1}^{k} \mathrm{cost}(P_i \cap R_{i-1}, \tilde{c}_i) + 8\alpha k \sum_{i=1}^{k-1} \sum_{t=1}^{i} \mathrm{cost}(P_t \cap R_{i-1}, \tilde{c}_t)\ . \tag{4.44}$$

Note that since $R_i \subseteq R_{i-1}$ for all i we have $P_t \cap R_{i-1} \subseteq P_t \cap R_{t-1}$ for all $t \leq i$. Hence,

$$\sum_{i=1}^{k-1} \sum_{t=1}^{i} \mathrm{cost}(P_t \cap R_{i-1}, \tilde{c}_t) \leq \sum_{i=1}^{k-1} \sum_{t=1}^{i} \mathrm{cost}(P_t \cap R_{t-1}, \tilde{c}_t) \tag{4.45}$$

$$\leq k \sum_{i=1}^{k-1} \mathrm{cost}(P_i \cap R_{i-1}, \tilde{c}_i)\ , \tag{4.46}$$

and we obtain

$$\operatorname{cost}(P,\tilde{C}) \leq \sum_{i=1}^{k} \operatorname{cost}(P_i \cap R_{i-1}, \tilde{c}_i) + 8\alpha k^2 \sum_{i=1}^{k-1} \operatorname{cost}(P_i \cap R_{i-1}, \tilde{c}_i) \tag{4.47}$$

$$\leq (1+8\alpha k^2) \sum_{i=1}^{k} \mathrm{D}(P_i \cap R_{i-1}, \tilde{c}_i) \ . \tag{4.48}$$

Using Claim 4.9 stated below we conclude

$$\mathrm{D}(P,\tilde{C}) \leq (1+8\alpha k^2)(1+\gamma) \sum_{i=1}^{k} \mathrm{D}(P_i, c_i) \tag{4.49}$$

$$= (1+8\alpha k^2)(1+\gamma)\, opt_k(P) \ . \tag{4.50}$$

□

Claim 4.8. *For every $i = 1, 2, \ldots, k-1$ we have*

$$\operatorname{cost}\big(P_{[i+1,k]} \cap N_i, \tilde{C}_i\big) \leq 8\alpha k \operatorname{cost}\big(P_{[1,i]} \cap R_{i-1}, \tilde{C}_i\big) \ . \tag{4.51}$$

Proof. Let $n_i = |R_{i-1}|$. Assume $N_i \neq \emptyset$, otherwise the claim is trivially true. Hence, N_i is the disjoint union of ν_i different subsets $N_i^{(j)}$ of size $\frac{n_i}{2^j}$ which correspond to the neighborhoods removed in step 12 of the algorithm, i.e.

$$N_i = N_i^{(1)} \cup N_i^{(2)} \cup \ldots \cup N_i^{(\nu_i)} \ . \tag{4.52}$$

We prove the claim using the same strategy as in Claim 4.5. That is, first, we show that for each j the set $N_i^{(j)}$ contains a large number of points from $P_{[1,i]}$ and only a few points from $P_{[i+1,k]}$. Then, for each j we charge the cost of the few and inexpensive points from $N_i^{(j)} \cap P_{[i+1,k]}$ against the many and costly points from $N_i^{(j+1)} \cap P_{[1,i]}$.

To this end, define $R_i^{(0)} = R_{i-1}$ and $R_i^{(j)} = R_i^{(j-1)} \setminus N_i^{(j)}$. By definition $|R_i^{(j)}| = |N_i^{(j)}| = \frac{n_i}{2^j}$. Note that the $R_i^{(j)}$ have been input sets of recursive calls prior to the call on $R_i = R_i^{(\nu_i)}$ and, hence, for any P_t with $t > i$ we have

$$\forall j < \nu_i : \ |P_t \cap R_i^{(j)}| < \alpha |R_i^{(j)}| \ . \tag{4.53}$$

Using (4.53) we get

$$\forall j < \nu_i : \ |P_{[i+1,k]} \cap R_i^{(j)}| = \sum_{t=i+1}^{k} |P_t \cap R_i^{(j)}| < \alpha k |R_i^{(j)}| \ . \tag{4.54}$$

Thus,

$$\forall j \leq \nu_i : \ |P_{[i+1,k]} \cap N_i^{(j)}| \leq |P_{[i+1,k]} \cap R_i^{(j-1)}| < \alpha k |R_i^{(j-1)}| = 2\alpha k \frac{n_i}{2^j} \tag{4.55}$$

where the first inequality holds since $N_i^{(j)} \subseteq R_i^{(j-1)}$. Using (4.55) we also get

$$\forall j \leq \nu_i : \ |P_{[1,i]} \cap N_i^{(j)}| = |N_i^{(j)}| - |P_{[i+1,k]} \cap N_i^{(j)}| \geq (1 - 2\alpha k)\frac{n_i}{2^j} \ . \tag{4.56}$$

Now we show that the cost of assigning $P_{[i+1,k]} \cap N_i$ to $\tilde{C}_i$ is small. By definition of $N_i^{(j)}$ we know that for all $j < \nu_i$ and for $p \in N_i^{(j)}$ and $p' \in N_i^{(j+1)}$ we have $\mathrm{D}(p, \tilde{C}_i) \leq \mathrm{D}(p', \tilde{C}_i)$. Thus,

$$\forall j < \nu_i : \ \frac{\mathrm{cost}(P_{[i+1,k]} \cap N_i^{(j)}, \tilde{C}_i)}{|P_{[i+1,k]} \cap N_i^{(j)}|} \leq \frac{\mathrm{cost}(P_{[1,i]} \cap N_i^{(j+1)}, \tilde{C}_i)}{|P_{[1,i]} \cap N_i^{(j+1)}|} \ . \tag{4.57}$$

Using (4.55) and (4.56) we get

$$\forall j < \nu_i : \\ \frac{2^j}{2\alpha k n_i} \mathrm{cost}(P_{[i+1,k]} \cap N_i^{(j)}, \tilde{C}_i) \leq \frac{2^{j+1}}{(1-2\alpha k) n_i} \mathrm{cost}(P_{[1,i]} \cap N_i^{(j+1)}, \tilde{C}_i) \tag{4.58}$$

or, equivalently,

$$\forall j < \nu_i : \ \mathrm{cost}(P_{[i+1,k]} \cap N_i^{(j)}, \tilde{C}_i) \leq \frac{4\alpha k}{1 - 2\alpha k} \mathrm{cost}(P_{[1,i]} \cap N_i^{(j+1)}, \tilde{C}_i) \ . \tag{4.59}$$

We still need an upper bound on $\mathrm{cost}(P_{[i+1,k]} \cap N_i^{(\nu_i)}, \tilde{C}_i)$. As a lower bound on the size of set $P_{[1,i]} \cap R_i^{(\nu_i)}$ we obtain

$$\begin{align} |P_{[1,i]} \cap R_i^{(\nu_i)}| &= |R_i^{(\nu_i)}| - |P_{[i+1,k]} \cap R_i^{(\nu_i)}| \tag{4.60} \\ &\geq |R_i^{(\nu_i)}| - |P_{[i+1,k]} \cap R_i^{(\nu_i - 1)}| \tag{4.61} \\ &> (1 - 2\alpha k)\frac{n_i}{2^{\nu_i}} \ . \tag{4.62} \end{align}$$

By definition of $N_i^{(\nu_i)}$ and $R_i^{(\nu_i)}$ we also know that for all $p \in N_i^{(\nu_i)}$ and $p' \in R_i^{(\nu_i)}$ we have $\mathrm{D}\big(p, \tilde{C}_i\big) \leq \mathrm{D}\big(p', \tilde{C}_i\big)$. In analogy to above, combining (4.55) and (4.62) we conclude

$$\operatorname{cost}(P_{[i+1,k]} \cap N_i^{(\nu_i)}, \tilde{C}_i) \leq \frac{2\alpha k}{1-2\alpha k} \operatorname{cost}(P_{[1,i]} \cap R_i^{(\nu_i)}, \tilde{C}_i) \ . \tag{4.63}$$

Using (4.59), (4.63), and

$$\operatorname{cost}(P_{[i+1,k]} \cap N_i, \tilde{C}_i) = \sum_{j=1}^{\nu_i} \operatorname{cost}(P_{[i+1,k]} \cap N_i^{(j)}, \tilde{C}_i) \tag{4.64}$$

we obtain

$$\begin{aligned}
&\operatorname{cost}(P_{[i+1,k]} \cap N_i, \tilde{C}_i) \\
&\quad \leq \frac{4\alpha k}{1-2\alpha k} \sum_{j=1}^{\nu_i - 1} \operatorname{cost}(P_{[1,i]} \cap N_i^{(j+1)}, \tilde{C}_i) + \frac{2\alpha k}{1-2\alpha k} \operatorname{cost}(P_{[1,i]} \cap R_i^{(\nu_i)}, \tilde{C}_i) && (4.65) \\
&\quad \leq 8\alpha k \sum_{j=1}^{\nu_i - 1} \operatorname{cost}(P_{[1,i]} \cap N_i^{(j+1)}, \tilde{C}_i) + 8\alpha k \operatorname{cost}(P_{[1,i]} \cap R_i^{(\nu_i)}, \tilde{C}_i) && (4.66)
\end{aligned}$$

since $\frac{2\alpha k}{1-2\alpha k} \leq \frac{4\alpha k}{1-2\alpha k} \leq 8\alpha k$ for $\alpha \leq \frac{1}{4k}$. Thus, using the fact that $N_i^{(2)}, N_i^{(3)}, \ldots, N_i^{(\nu_i)}, R_i^{(\nu_i)}$ are disjoint subsets of R_{i-1} we obtain

$$\operatorname{cost}(P_{[i+1,k]} \cap N_i, \tilde{C}_i) \leq 8\alpha k \operatorname{cost}(P_{[1,i]} \cap R_{i-1}, \tilde{C}_i) \tag{4.67}$$

□

Claim 4.9. *For every $i = 1, 2, \ldots, k$ we have*

$$\operatorname{cost}(P_i \cap R_{i-1}, \tilde{c}_i) \leq (1+\gamma) \operatorname{cost}(P_i, c_i) \ . \tag{4.68}$$

Proof. Let c_i' be the optimal 1-median of $P_i \cap R_{i-1}$. By choice of $\tilde{c}_i$ we get

$$\begin{aligned}
\operatorname{cost}(P_i \cap R_{i-1}, \tilde{c}_i) &\leq (1+\gamma) \operatorname{cost}(P_i \cap R_{i-1}, c_i') && (4.69) \\
&\leq (1+\gamma) \operatorname{cost}(P_i \cap R_{i-1}, c_i) && (4.70) \\
&\leq (1+\gamma) \operatorname{cost}(P_i, c_i) \ . && (4.71)
\end{aligned}$$

□

By choosing parameters $\alpha = \frac{\varepsilon}{16k^2}$ and $\gamma = \frac{\varepsilon}{3}$ we obtain that algorithm Cluster computes a $(1+\varepsilon)$-approximation to the k-median problem with respect to D.

Corollary 4.10. *Let $P \subseteq \mathbb{X}$, $k \in \mathbb{N}$, and $0 < \varepsilon, \delta < 1$. If* D *satisfies the $[\varepsilon/3, \delta]$-sampling property then algorithm* Cluster *started with parameters $(P, k, \emptyset)$ and using fixed positive parameters $\alpha \leq \frac{\varepsilon}{16k^2}$ and $\gamma \leq \frac{\varepsilon}{3}$ computes a set of approximate centers $\tilde{C} \subseteq \mathbb{X}$ satisfying*

$$\Pr\left[\operatorname{cost}(P, \tilde{C}) \leq (1+\varepsilon)\, opt_k(P)\right] \geq \left(\frac{1-\delta}{5}\right)^k . \tag{4.72}$$

Proof. The corollary follows from Theorem 4.7 since

$$\begin{aligned}(1+8\alpha k^2)(1+\gamma) &= 1 + 8\alpha k^2\gamma + 8\alpha k^2 + \gamma & (4.73)\\ &\leq 1 + \frac{\varepsilon^2}{6} + \frac{\varepsilon}{2} + \frac{\varepsilon}{3} & (4.74)\\ &\leq 1 + \varepsilon & (4.75)\end{aligned}$$

for $\alpha \leq \frac{\varepsilon}{16k^2}$, $\gamma \leq \frac{\varepsilon}{3}$ and $\varepsilon < 1$. □

We now give a running time analysis of algorithm Cluster.

Theorem 4.11. *Let $P \subseteq \mathbb{X}$ be of size $|P| = n$, $k \in \mathbb{N}$ and $0 < \varepsilon, \delta < 1$. Furthermore, let* D *satisfy the $[\varepsilon/3, \delta]$-sampling property with $m = m_{\varepsilon/3,\delta}$. Then algorithm* Cluster *started with parameters $(P, k, \emptyset)$ and using fixed parameters $\alpha = \Theta(\varepsilon/k^2)$ and $\gamma = \varepsilon/3$ requires at most $2^{\mathcal{O}(mk \log(mk/\varepsilon))}n$ arithmetic operations, including evaluations of* D.

Proof. The running time analysis of [Kumar et al., 2004] can easily be adapted to algorithm Cluster. To this end, let $T(n, k)$ denote the running time of algorithm Cluster started with n input points and k approximate medians to be found. For $k = 0$ we have already found all cluster centers and we clearly have $T(n, 0) = \mathcal{O}(1)$. On the other hand, if $k > 0$ and we have $n \leq k$ then we just have to assign the n remaining points as cluster centers. In this case, obviously, $T(n, k) = \mathcal{O}(n)$.

Now, assume $n > k \geq 1$. In the sampling phase, the sampling of $\mathcal{O}(mk^2/\varepsilon)$ points and the construction of the $2^{\mathcal{O}(m \log(mk/\varepsilon))}$ candidate centers requires at most $2^{\mathcal{O}(m \log(mk/\varepsilon))}$ arithmetic operations. After that, each of the candidates is tried recursively, taking $2^{\mathcal{O}(m \log(mk/\varepsilon))}T(n, k-1)$ steps. In the pruning phase, set N is obtained by finding the median element of set

$\{\mathrm{D}(p,\tilde{C}) \,|\, p \in P\}$ and by partitioning the points according to this median element. This takes $\mathcal{O}(n)$ operations, including evaluations of dissimilarity function D. Finally, the algorithm is called recursively once for the remaining point set, requiring time $T(n/2,k)$. We obtain that algorithm Cluster has a running time $T(n,k)$ given by the recurrence

$$\begin{aligned} &T(n,k) \\ &\quad = 2^{\mathcal{O}(m\log(mk/\varepsilon))}T(n,k-1) + T\left(\frac{n}{2},k\right) + \mathcal{O}(n) + 2^{\mathcal{O}(m\log(mk/\varepsilon))} && (4.76) \\ &\quad \le 2^{\mathcal{O}(m\log(mk/\varepsilon))}T(n,k-1) + T\left(\frac{n}{2},k\right) + \mathcal{O}(n)\cdot 2^{\mathcal{O}(m\log(mk/\varepsilon))}\ . && (4.77)\end{aligned}$$

Hence, let $c = 2^{\mathcal{O}(m\log(mk/\varepsilon))}$ be a constant large enough such that

$$T(n,k) \le \begin{cases} c & \text{if } k=0 \\ cn & \text{if } k\ge 1 \text{ and } n \le k \\ c\,T(n,k-1) + T\left(\frac{n}{2},k\right) + cn & \text{if } k \ge 1 \text{ and } n > k \end{cases}\ . \qquad (4.78)$$

We can bound recurrences of this type by Claim 4.12 stated below. Using Claim 4.12 with $i=n$ and $j=k$ we obtain

$$T(n,k) \le n\,4^k c^{k+1} = 2^{\mathcal{O}(mk\log(mk/\varepsilon))}n\ . \qquad (4.79)$$

This shows that the running time of the algorithm is linear in n. □

Claim 4.12. *$T(i,j) \le i\,4^j c^{j+1}$ for all $i \ge 1$ and $j \ge 0$.*

Proof. We prove the claim by induction. For $j=0$ we have

$$T(i,0) \le c \le i\,4^0 c^1\ . \qquad (4.80)$$

On the other hand, for $j \ge 0$ and $i \le j$ we obtain

$$T(i,j) \le ci \le i\,4^j c^{j+1}\ . \qquad (4.81)$$

This concludes the inductive base cases. Hence, let $i > j \ge 1$ and assume that the claim holds for all i', j' with $i' < i$ or $j' < j$. Using recurrence (4.78), by induction hypothesis we obtain

$$\begin{aligned} T(i,j) &\le c\cdot i\,4^{j-1}c^j + \frac{1}{2}i\,4^j c^{j+1} + ci && (4.82) \\ &= \left(\frac{1}{4} + \frac{1}{2} + \frac{1}{4^j c^j}\right) i\,4^j c^{j+1} && (4.83) \\ &\le i\,4^j c^{j+1} && (4.84)\end{aligned}$$

since $\frac{1}{4^j c^j} \le \frac{1}{4}$ for all $c, j \ge 1$. □

Our main result, Theorem 4.2, is an immediate consequence of Corollary 4.10 and Theorem 4.11. By running the algorithm multiple times (say, at least $2^{\Theta(k)}$ times) and choosing the best result obtained this way the error probability can be reduced to an arbitrarily small constant without changing the asymptotic running time.

4.1.5 Adaptation for weighted input sets

We have shown that with constant probability, algorithm Cluster computes a $(1+\varepsilon)$-approximation for the k-median problem with respect to a dissimilarity measure satisfying the $[\gamma, \delta]$-sampling property. During our proof, we always assumed an unweighted input set P. However, our algorithm easily generalizes to the case of weighted input set P with integral weight function $w : P \to \mathbb{N}$. In this case, each input point $p \in P$ is associated with weight $w(p)$. We also write $w(P') = \sum_{p \in P'} w(p)$ for the total weight of any subset $P' \subseteq P$.

For weighted input sets, only two slight modifications to algorithm Cluster are necessary: We have to adapt the way points are sampled during the sampling phase, and we have to adapt the way points are discarded during the pruning phase.

First, in the sampling phase, points are no longer chosen uniformly at random. Instead, a point $p \in R$ is sampled from $R \subseteq P$ with a probability proportional to the weight of point p, that is, with probability $\frac{w(p)}{w(R)}$. We say a point p is chosen *at random according to w*. Using sampling according to w, we obtain a variant of the superset sampling lemma for weighted point sets.

Lemma 4.13 (weighted superset sampling lemma). *Let D satisfy the $[\gamma, \delta]$-sampling property. Let $P \subseteq \mathbb{X}$ be finite and let $P' \subseteq P$ be with $w(P') \geq \alpha w(P)$ for some constant $\alpha > 0$. Let $S \subseteq P$ of size $|S| \geq 2m_{\gamma,\delta}/\alpha$ be an (unweighted) multiset obtained by sampling points from P at random according to w. Then there exists with probability at least $(1-\delta)/5$ a subset $S' \subseteq S$ with $|S'| = m_{\gamma,\delta}$ and optimal 1-median $\mathrm{med}(S')$ satisfying*

$$\mathrm{cost}_w\big(P', \mathrm{med}(S')\big) \leq (1+\gamma)\, opt_1(P') \ . \tag{4.85}$$

Proof. Let Q denote the (unweighted) multiset consisting of $w(p)$ copies of each point $p \in P$, and let Q' denote the multiset consisting of $w(p)$ copies of each point $p \in P'$. Obviousely, we have

$$|Q'| = w(P') \geq \alpha w(P) = \alpha |Q| \ . \tag{4.86}$$

```
$\textsc{WeightedCluster}(R, w, l, \tilde{C})$:
  $R$   set of remaining input points of total weight $w(R)$
  $w$   weight function on $R$
  $l$   number of medians yet to be found
  $C$   set of medians already found

if $l = 0$ then return $\tilde{C}$
else
    if $l \geq |R|$ then return $\tilde{C} \cup R$
    else
        /* sampling phase */
        sample a multiset $S$ of size $2m_{\gamma,\delta}/\alpha$ at random according to $w$ from $R$
        $T \leftarrow \{\mathrm{med}(S') \mid S' \subseteq S,\ |S'| = m_{\gamma,\delta}\}$
        for all $\tilde{c} \in T$ do
            $C^{(\tilde{c})} \leftarrow \textsc{WeightedCluster}(R, w, l-1, \tilde{C} \cup \{\tilde{c}\})$
        end for
        /* pruning phase */
        partition $R$ into set $N$ and $R \setminus N$ such that:
            $\circ\ \forall p \in N, q \in R \setminus N : \mathrm{D}(p, \tilde{C}) \leq \mathrm{D}(q, \tilde{C})$ and
            $\circ\ w(N) = w(R \setminus N) = \frac{1}{2}w(R)$ (if necessary, split a point)
        let $\tilde{w}$ be the new weight function on $R \setminus N$
        $C^* \leftarrow \textsc{WeightedCluster}(R \setminus N, \tilde{w}, l, \tilde{C})$
        return $C^{(\tilde{c})}$ or $C^*$ with minimal cost
    end if
end if
```

Figure 4.4: Adaptation of algorithm Cluster for weighted input sets and fixed positive real constants α, γ, δ.

Note that sampling according to w from P corresponds to sampling uniformly at random from Q. Hence, Lemma 4.13 is an immediate consequence of Lemma 4.3 applied to unweighted set Q. □

Second, in the pruning phase, we also change the way the neighborhood N is approximated. That is, we no longer remove the $\frac{n}{2}, \frac{n}{4}, \frac{n}{8}, \ldots$ closest points from the point set R. Instead, the closest points with a total weight of $\frac{1}{2}w(R), \frac{1}{4}w(R), \frac{1}{8}w(R), \ldots$ are pruned. However, from time to time the weight of the closest points will not add up to exactly $\frac{1}{2^i}w(R)$. In this case a single point p has to be replaced by two copies p_1, p_2 with $w(p) = w(p_1) + w(p_2)$ such that we can find a partition with total weight $\frac{1}{2^i}w(R)$.

The pseudocode of the adaptation of algorithm Cluster for weighted input sets is given in Figure 4.4. We call this algorithm WeightedCluster. Note that due to our slight changes to the algorithm, a run of algorithm WeightedCluster with weighted input set P corresponds to a run of algorithm Cluster started with an unweighted multiset Q consisting of $w(p)$ copies of each point $p \in P$. Thus, analogously to the proof of Theorem 4.7, and by choosing fixed parameters $\alpha = \frac{\varepsilon}{16k^2}$ and $\gamma = \frac{\varepsilon}{3}$, we obtain the following result.

Theorem 4.14. *Let dissimilarity measure* D *on* $\mathbb{X}$ *satisfy the* $[\varepsilon/3, \delta]$*-sampling property. Then with probability at least* $((1-\delta)/5)^k$ *algorithm* WeightedCluster *computes a* $(1+\varepsilon)$*-approximate solution of the* k*-median problem for weighted input instance* $P \subseteq \mathbb{X}$*.*

We also analyze the running time of algorithm WeightedCluster. It turns out that our bound on the running time is increased by an additional factor that is merely polylogarithmic in the total weight of P. Here, we have also chosen the fixed parameters to be $\alpha = \frac{\varepsilon}{16k^2}$ and $\gamma = \frac{\varepsilon}{3}$.

Theorem 4.15. *Let* D *satisfy the* $[\varepsilon/3, \delta]$*-sampling property with* $m = m_{\frac{\varepsilon}{3},\delta}$*. Algorithm* WeightedCluster *started with parameters* $(P, w, k, \emptyset)$*, where* P *is of size* n *with total weight* $w(P) = W$*, and fixed parameters* $\alpha = \mathcal{O}(\varepsilon/k^2)$ *and* $\gamma = \varepsilon/3$*, requires at most* $2^{\mathcal{O}(mk\log(mk/\varepsilon))} n \log^k W$ *arithmetic operations, including evaluations of* D*.*

Proof. Let $T(n, k, W)$ denote the running time of algorithm WeightedCluster started with n input points of total weight W and k approximate medians to be found. For $k = 0$ we have already found all cluster centers and we clearly have $T(n, 0, W) = \mathcal{O}(1)$. On the other hand, if $k > 0$ and we have $n \leq k$ then we just have to assign the n remaining points as cluster centers. In this case, obviously, $T(n,k,W) = \mathcal{O}(n)$.

Now, assume $n > k \geq 1$. In the sampling phase, the sampling of $\mathcal{O}(mk^2/\varepsilon)$ points and the construction of the $2^{\mathcal{O}(m\log(mk/\varepsilon))}$ candidate centers requires at most $2^{\mathcal{O}(m\log(mk/\varepsilon))}$ arithmetic operations. After that, each of the candidates is tried recursively, taking $2^{\mathcal{O}(m\log(mk/\varepsilon))}T(n,k-1,W)$ steps. In the pruning phase, set N is obtained by finding the median element of set $\{\mathrm{D}(p,\tilde{C})\,|\,p \in P\}$ and by partitioning the points according to this median element. This takes $\mathcal{O}(n)$ operations, including evaluations of dissimilarity function D. Finally, the algorithm is called recursively once for the remaining point set, requiring time $T(t,k,W/2)$ for some unknown number of remaining points $t \leq n$ with total weight $W/2$. We obtain that algorithm WeightedCluster has a running time $T(n,k,W)$ given by the recurrence

$$\begin{aligned} &T(n,k,W)\\ &= 2^{\mathcal{O}(m\log(mk/\varepsilon))}T(n,k-1,W) + T\left(t,k,\frac{W}{2}\right) + \mathcal{O}(n) + 2^{\mathcal{O}(m\log(mk/\varepsilon))} \qquad (4.87)\\ &\leq 2^{\mathcal{O}(m\log(mk/\varepsilon))}T(n,k-1,W) + T\left(n,k,\frac{W}{2}\right) + \mathcal{O}(n)\cdot 2^{\mathcal{O}(m\log(mk/\varepsilon))}. \qquad (4.88)\end{aligned}$$

Hence, let $c = 2^{\mathcal{O}(m\log(mk/\varepsilon))}$ be a constant large enough such that

$$T(n,k,W) \leq \begin{cases} c & \text{if } k=0\\ cn & \text{if } k\geq 1 \text{ and } n\leq k\\ c\,T(i,j-1,W) + T\left(i,j,\frac{W}{2}\right) + ci & \text{if } k\geq 1 \text{ and } n>k \end{cases}. \qquad (4.89)$$

This recurrence is of the type

$$\tilde{T}(n,k,l) \leq \begin{cases} c & \text{if } k=0\\ ci & \text{if } k\geq 1 \text{ and } n\leq k\\ c\tilde{T}(i,j-1,l) + \tilde{T}(i,j,l-1) + ci & \text{if } k\geq 1 \text{ and } n>k \end{cases} \qquad (4.90)$$

where the third parameter of T is replaced by its binary logarithm. Note that we always have $W \geq n$ and, hence, $l \geq \log(n)$. We can bound recurrences of this type by Claim 4.16 stated below. Using Claim 4.16 with $i = n$, $j = k$, and $l = \log W$ we obtain

$$T(n, k, W) \leq n\, 2^k c^{k+1} (1 + \log W)^k = 2^{\mathcal{O}(mk \log(mk/\varepsilon))} n \log^k W \ . \tag{4.91}$$

□

Claim 4.16. *$\tilde{T}(i, j, l) \leq i\, 2^j c^{j+1} (l+1)^j$ for all $i \geq 1$, $j \geq 0$, and $l \geq \log(i)$.*

Proof. We prove the claim by induction. For $j = 0$ we have

$$T(i, 0, l) \leq c \leq i\, 2^0 c^1 (l + 1)^0 \ . \tag{4.92}$$

On the other hand, for $j \geq 1$ and $i \leq j$ we obtain

$$T(i, j, l) \leq ci \leq i\, 2^j c^{j+1} (l + 1)^j \ . \tag{4.93}$$

Hence, let $j \geq 1$ and $i > j$. In this case we have $i \geq 2$ and $l \geq \log(i) \geq 1$. Assume that the claim holds for all i', j', l' with $i' < i$, $j' < j$, or $l' < l$. By induction hypothesis we have

$$\begin{aligned} \tilde{T}(i, j, l) &\leq c\big(i\, 2^{j-1} c^j (l + 1)^{j-1}\big) + i\, 2^j c^{j+1} l^j + c\, i && (4.94)\\ &\leq i\, 2^j c^{j+1} \left(\frac{1}{2}(l + 1)^{j-1} + l^j + \frac{1}{2} \right) \ . && (4.95) \end{aligned}$$

From Claim 4.17 stated below we know that $\frac{1}{2}(l+1)^{j-1} + l^j + \frac{1}{2} \leq (l+1)^j$. Thus, we have

$$\tilde{T}(l, j) \leq i\, 2^j c^{j+1} (l + 1)^j \ . \tag{4.96}$$

□

Claim 4.17. *$l^j \leq (l + 1)^j - \frac{1}{2}(l + 1)^{j-1} - \frac{1}{2}$ for all $j \geq 1$ and $l \geq 0$.*

Proof. We prove the claim by induction over j for a fixed $l \geq 0$. Obviously, for $j = 1$ we have

$$l^1 = (l + 1)^1 - \frac{1}{2}(l + 1)^0 - \frac{1}{2} \ . \tag{4.97}$$

Hence, let $j \geq 2$ and assume that the claim holds for any $j' < j$. From induction hypothesis we obtain

$$\begin{aligned} l^j &\leq (l + 1) \cdot l^{j-1} && (4.98)\\ &\leq (l + 1) \cdot \left((l + 1)^{j-1} - \frac{1}{2}(l + 1)^{j-2} - \frac{1}{2} \right) && (4.99)\\ &= (l + 1)^j - \frac{1}{2}(l + 1)^{j-1} - \frac{1}{2}(l + 1) \ . && (4.100) \end{aligned}$$

Using $l \geq 0$ we conclude

$$l^j \leq (l+1)^j - \frac{1}{2}(l+1)^{j-1} - \frac{1}{2}\,. \tag{4.101}$$

□

By running algorithm WeightedCluster at least $2^{\Theta(k)}$ times and choosing the best result returned, the error probability can be reduced to an arbitrarily small constant without changing the asymptotic running time.

4.2 Sampling for large classes of dissimilarity measures

We now show how to apply Theorem 4.2 to various dissimilarity measures. We achieve this by showing that, with high probability, the 1-median of a constant sized uniform sample set from P is an approximate 1-median of P. This result does not only hold for Bregman divergences like the Mahalanobis distance and all μ-similar Bregman divergences such as Kullback-Leibler divergence and Itakura-Saito divergence. It also holds in arbitrary metric spaces with bounded doubling dimension, for the Hamming distance, and even for more exotic distance measures, such as the Hellinger distance, which is neither a metric nor a Bregman divergence. Hence, our framework from Theorem 4.2 is indeed well suited to provide a polynomial time approximation scheme to solve the k-median problem for a large variety of metric and non-metric distance measures.

4.2.1 Sampling for Mahalanobis distances

As given in Section 2.2.1, the Mahalanobis distance on $\mathbb{R}^d$ is defined as

$$\mathrm{D}_A(x,y) = (x-y)^\top A\,(x-y) \tag{4.102}$$

with respect to a symmetric, positive definite matrix $A \in \mathbb{R}^{d\times d}$. Note that with $A = I_d$ we obtain that the squared Euclidean distance is a Mahalanobis distance. Furthermore, recall that since the Mahalanobis distance is a Bregman divergence, for all P the centroid

$$c_P = \frac{1}{n}\sum_{x\in P} x \tag{4.103}$$

is the unique optimal 1-median of P, i.e. $c_P = \mathrm{med}(P)$ (cf. Lemma 3.6). We show the following lemma, which is a generalization of an earlier result by [Inaba et al., 1994] with respect to the squared Euclidean distance.

Lemma 4.18. *Let* D_A *be a Mahalanobis distance and let* $P \subseteq \mathbb{R}^d$ *be finite of size* n. *Then a uniform sample multiset* $S \subseteq P$ *of size* $m \geq \frac{1}{\gamma\delta}$ *satisfies*

$$\Pr\big[\mathrm{cost}(P, c_S) \leq (1+\gamma)\, opt_1(P)\big] \geq 1 - \delta \ . \tag{4.104}$$

Proof. For any fixed multiset $S \subseteq P$ we have

$$\begin{aligned} \mathrm{D}_A(c_P, c_S) &= \Big(c_P - \frac{1}{m}\sum_{x \in S} x\Big)^\top A \,\Big(c_P - \frac{1}{m}\sum_{y \in S} y\Big) && (4.105)\\ &= \frac{1}{m^2}\sum_{x \in S}\sum_{y \in S}(c_P - x)^\top A (c_P - y) \ . && (4.106) \end{aligned}$$

Since $\mathrm{E}\left[(c_P - x)^\top A\,(c_P - y)\right] = 0$ for mutually independent $x, y \in S$ it follows

$$\begin{aligned} \mathrm{E}\left[\mathrm{D}_A(c_P, c_S)\right] &= \mathrm{E}\left[\frac{1}{m^2}\sum_{x \in S}\sum_{y \in S}(c_P - x)^\top A\,(c_P - y)\right] && (4.107)\\ &= \frac{1}{m^2}\,\mathrm{E}\left[\sum_{x \in S}(c_P - x)^\top A\,(c_P - x)\right] && (4.108)\\ &= \frac{1}{mn}\sum_{x \in P}\mathrm{D}_A(x, c_P) && (4.109)\\ &= \frac{1}{mn}\, opt_1(P) \ . && (4.110) \end{aligned}$$

By Lemma 3.5, we know that for all $q \in \mathbb{R}^d$

$$\mathrm{cost}(P, q) = opt_1(P) + n\,\mathrm{D}_A(c_P, q) \ . \tag{4.111}$$

Hence, using (4.111) we get

$$\begin{aligned} &\Pr\big[\mathrm{cost}(P, c_S) \leq (1+\gamma)\, opt_1(P)\big]\\ &\qquad = \Pr\big[opt_1(P) + n\,\mathrm{D}_A(c_P, c_S) \leq (1+\gamma)\, opt_1(P)\big] && (4.112)\\ &\qquad = \Pr\left[\mathrm{D}_A(c_P, c_S) \leq \frac{\gamma}{n}\, opt_1(P)\right] \ . && (4.113) \end{aligned}$$

Using (4.110), $m \geq \frac{1}{\gamma\delta}$, and Markov's inequality, we obtain

$$\begin{aligned}
&\Pr\big[\mathrm{cost}(P, c_S) \leq (1+\gamma)\, opt_1(P)\big] \\
&\qquad = \Pr\big[\mathrm{D}_A(c_P, c_S) \leq \gamma m \, \mathrm{E}\left[\mathrm{D}_A(c_P, c_S)\right]\big] && (4.114)\\
&\qquad \geq \Pr\left[\mathrm{D}_A(c_P, c_S) \leq \frac{1}{\delta}\, \mathrm{E}\left[\mathrm{D}_A(c_P, c_S)\right]\right] && (4.115)\\
&\qquad \geq 1 - \delta \ . && (4.116)
\end{aligned}$$

□

Corollary 4.19. *A Mahalanobis distance* D_A *on domain* $\mathbb{R}^d$ *satisfies the* $[\gamma, \delta]$*-sampling property with* $m_{\gamma,\delta} = \frac{1}{\gamma\delta}$ *and* $\mathrm{med}(S) = c_S$.

4.2.2 Sampling for μ-similar Bregman divergences

Let D_φ on domain $\mathbb{X}$ be a μ-similar Bregman divergence, that is, there exists a positive definite matrix A such that for all $x, y, \in \mathbb{X}$

$$\mu\, \mathrm{D}_A(x,y) \leq \mathrm{D}_\varphi(x,y) \leq \mathrm{D}_A(x,y) \ . \tag{4.117}$$

Again, we know by Lemma 3.6 that the centroid c_P is the optimal median of any given set P. Hence, we have

$$\mu\, opt_1^A(P) = \sum_{x\in P} \mu\, \mathrm{D}_A(x, c_P) \leq \sum_{x\in P} \mathrm{D}_\varphi(x, c_P) = opt_1^\varphi(P) \ . \tag{4.118}$$

Since D_φ is μ-similar we can use our sampling result for Mahalanobis distances to show the $[\gamma, \delta]$-sampling property of D_φ.

Lemma 4.20. *Let* D_φ *be* μ*-similar and let* $P \subseteq \mathbb{X}$ *be finite of size* n. *Then a uniform sample multiset* $S \subseteq P$ *of size* $m \geq \frac{1}{\gamma\delta\mu}$ *satisfies*

$$\Pr\big[\mathrm{cost}^\varphi(P, c_S) \leq (1+\gamma)\, opt_1^\varphi(P)\big] \geq 1 - \delta \ . \tag{4.119}$$

Proof. From proof of Lemma 4.18 we know that

$$\mathrm{E}[\mathrm{D}_A(c_P, c_S)] = \frac{1}{mn}\, opt_1^A(P) \ . \tag{4.120}$$

By Lemma 3.5, we know that for all Bregman divergences D_φ and for all $q \in \mathbb{R}^d$ we have

$$\mathrm{cost}^\varphi(P, q) = opt_1^\varphi(P) + n\, \mathrm{D}_\varphi(c_P, q) \ . \tag{4.121}$$

Hence, using (4.121) we get

$$\begin{aligned}
&\Pr\bigl[\mathrm{cost}^{\varphi}(P, c_S) \le (1+\gamma)\, opt_1^{\varphi}(P)\bigr] \\
&\qquad = \Pr\bigl[\mathrm{cost}^{\varphi}(P) + n\, \mathrm{D}_{\varphi}(c_P, c_S) \le (1+\gamma)\, opt_1^{\varphi}(P)\bigr] && (4.122)\\
&\qquad = \Pr\left[\mathrm{D}_{\varphi}(c_P, c_S) \le \frac{\gamma}{n}\, opt_1^{\varphi}(P)\right] . && (4.123)
\end{aligned}$$

Due to the μ-similarity of D_{φ}, from (4.117) and (4.118) we obtain.

$$\Pr\bigl[\mathrm{cost}^{\varphi}(P,\, c_S) \le (1+\gamma)\, opt_1^{\varphi}(P)\bigr] \ge \Pr\left[\mathrm{D}_A(c_P, c_S) \le \frac{\gamma\mu}{n}\, opt_1^{A}(P)\right] . \qquad (4.124)$$

Using (4.120), $m \ge \frac{1}{\gamma\delta\mu}$, and Markov's inequality we get

$$\begin{aligned}
&\Pr\bigl[\mathrm{cost}^{\varphi}(P, c_S) \le (1+\gamma)\, opt_1^{\varphi}(P)\bigr] \\
&\qquad \ge \Pr\bigl[\mathrm{D}_A(c_P, c_S) \le \gamma\mu m\, \mathrm{E}\left[\mathrm{D}_A(c_P, c_S)\right]\bigr] && (4.125)\\
&\qquad \ge \Pr\left[\mathrm{D}_A(c_P, c_S) \le \frac{1}{\delta}\, \mathrm{E}\left[\mathrm{D}_A(c_P, c_S)\right]\right] && (4.126)\\
&\qquad \ge 1 - \delta . && (4.127)
\end{aligned}$$

□

Corollary 4.21. *A μ-similar Bregman divergence D_{φ} on domain $\mathbb{X}$ satisfies the $[\gamma, \delta]$-sampling property with $m_{\gamma,\delta} = \frac{1}{\gamma\delta\mu}$ and* $\mathrm{med}(S) = c_S$.

Lemma 4.20 shows an interesting tradeoff between μ-similarity and sampleability. Assume D_{φ} is a Bregman divergence that is very similar to a Mahalanobis distance (i.e., μ is close to 1). Then, as expected, we obtain essentially the same sampling result as in Lemma 4.18. On the other hand, assume D_{φ} is very unlike a Mahalanobis distance, but still μ-similar with respect to a small constant μ. Then, we still obtain that the median of a constant sized uniform sample set is a good approximation to the median of the superset. The only difference is that the constant number of sampled points necessary scales linearly in $\frac{1}{\mu}$.

This dependency of sample size m on μ might seem awkward at first. However, as we learn later in Chapter 8, this tradeoff arises quite naturally for Bregman divergences and, in fact, seems to be inevitable. More precisely, from Lemma 4.20 we know that any μ-similar Bregman divergence is also sampleable for some constant sample size m. In Section 8.1, we provide strong evidence that if a Bregman divergence satisfies the sampling property for any constant sample size m, then it also has to be μ-similar for some small constant μ. Hence, there seems to exist a one-to-one correspondence between sampleability and μ-similarity. The reader is directed to Chapter 8 for a discussion on the limits of the use of uniform sampling for the Bregman k-median problem.

As an immediate consequence of Corollary 4.21 we know that algorithm Cluster is applicable for the k-median problem with respect to measures such as the Kullback-Leibler divergence $\mathrm{D_{KL}}$ or the Itakura-Saito divergence $\mathrm{D_{IS}}$, as long as the input set P comes from a domain that avoids the singularities of $\mathrm{D_{KL}}$ or $\mathrm{D_{IS}}$ on $\mathbb{R}^d$.

Example 1: Kullback-Leibler divergence. We can apply this result to the generalized Kullback-Leibler divergence

$$\mathrm{D_{KL}}(p,q)=\sum_{i=1}^{d}\left(p_i\ln\frac{p_i}{q_i}-p_i+q_i\right) \tag{4.128}$$

for $p,q\in\mathbb{R}^d_{\geq 0}$ with $p=(p_1,\dots,p_d)^\top$ and $q=(q_1,\dots,q_d)^\top$. By Lemma 2.19 we know that $\mathrm{D_{KL}}$ on domain $[\lambda,\upsilon]^d$ is a $\frac{\lambda}{\upsilon}$-similar Bregman divergence. Hence, from Corollary 4.21 we obtain the following corollary.

Corollary 4.22. *D_{KL} on domain $[\lambda,\upsilon]^d\subseteq\mathbb{R}^d_+$ satisfies the $[\gamma,\delta]$-sampling property with $m_{\gamma,\delta}=\frac{\upsilon}{\gamma\delta\lambda}$ and* $\mathrm{med}(S)=c_S$.

Example 2: Itakura-Saito divergence. Similar to the case of the Kullback-Leibler divergence, we can apply Lemma 4.21 to the discrete Itakura-Saito divergence, that is

$$\mathrm{D_{IS}}(p,q)=\sum_{i=1}^{d}\left(\frac{p_i}{q_i}-\ln\frac{p_i}{q_i}-1\right) \tag{4.129}$$

for all $p,q\in\mathbb{R}^d_{\geq 0}$ with $p=(p_1,\dots,p_d)^\top$ and $q=(q_1,\dots,q_d)^\top$. From Lemma 2.20 we already know that $\mathrm{D_{IS}}$ on domain $[\lambda,\upsilon]^d$ is a $\left(\frac{\lambda}{\upsilon}\right)^2$-similar Bregman divergence. Using Corollary 4.21 again we find the following corollary.

Corollary 4.23. *D_{IS} on domain $[\lambda,\upsilon]^d\subseteq\mathbb{R}^d_+$ satisfies the $[\gamma,\delta]$-sampling property with $m_{\gamma,\delta}=\frac{\upsilon^2}{\gamma\delta\lambda^2}$ and* $\mathrm{med}(S)=c_S$.

4.2.3 Sampling for the Hellinger distance

The Hellinger distance is another statistical distance measure on $\mathbb{R}^d_{\geq 0}$. It is named after the German mathematician Ernst Hellinger since its continuous form is expressed in terms of a Hellinger integral. The discrete Hellinger distance on $\mathbb{R}^d_{\geq 0}$ is defined as

$$\mathrm{D}_{\mathrm{He}}(p,q) = \frac{1}{2}\sum_{i=1}^{d}\left(\sqrt{p_i} - \sqrt{q_i}\right)^2 \tag{4.130}$$

for $p, q \in \mathbb{R}^d_{\geq 0}$ with $p = (p_1, \ldots, p_d)^\top$ and $q = (q_1, \ldots, q_d)^\top$. Although D_{He} is symmetric, the Hellinger distance is neither a metric nor a Bregman divergence, but belongs to the class of so-called *Csiszár divergences* (cf. [Csiszár, 1991]). The Hellinger distance and other Csiszár divergences are frequently used as a dissimilarity measure on probability distributions in statistical inference and machine learning.

We show that the Hellinger distance fits into our framework, that is, algorithm Cluster is applicable to the Hellinger k-median problem, since D_{He} satisfies the $[\gamma, \delta]$-sampling property. We prove this result by applying a non-linear transformation to the input points and reuse the sampling results for the squared Euclidean distance from Section 4.2.1. This approach borrows from ideas already used in [Chaudhuri and McGregor, 2008].

In a slight abuse of notation, let $\sqrt{\cdot} : \mathbb{R}^d_{\geq 0} \to \mathbb{R}^d_{\geq 0}$ denote the non-linear mapping $(p_1, \ldots, p_d) \mapsto (\sqrt{p_1}, \ldots, \sqrt{p_d})$. Obviousely, $\sqrt{\cdot}$ is a bijection. Thus, let $(\cdot)^2 : \mathbb{R}^d_{\geq 0} \to \mathbb{R}^d_{\geq 0}$ denote the inverse mapping of $\sqrt{\cdot}$, that is, we have $p = q^2$ if and only if $\sqrt{p} = q$. Furthermore, for $P, Q \subseteq \mathbb{R}^d_{\geq 0}$ we define $\sqrt{P} = \{\sqrt{p} \,|\, p \in P\}$ and $Q^2 = \{q^2 \,|\, q \in Q\}$.

We make use of the following connection between the Hellinger distance of points from P and the squared Euclidean distance of points from $\sqrt{P}$.

Lemma 4.24. *For all $p, q \in \mathbb{R}^d_{\geq 0}$ and any finite $P \subseteq \mathbb{R}^d_{\geq 0}$ we have*

$$\mathrm{D}_{He}(p,q) = \frac{1}{2}\|\sqrt{p} - \sqrt{q}\|^2 \tag{4.131}$$

and

$$\mathrm{cost}^{He}(P,q) = \frac{1}{2}\,\mathrm{cost}^{\ell_2^2}(\sqrt{P}, \sqrt{q}\,)\ . \tag{4.132}$$

Proof. Obviously, by definition we have

$$\mathrm{D}_{\mathrm{He}}(p,q) = \frac{1}{2}\sum_{i=1}^{d}(\sqrt{p_i}-\sqrt{q_i})^2 = \frac{1}{2}\|\sqrt{p}-\sqrt{q}\|^2 \tag{4.133}$$

and, using (4.133),

$$\begin{aligned}\mathrm{cost}^{\mathrm{He}}(P,q) &= \sum_{p\in P}\mathrm{D}_{\mathrm{He}}(p,q) &(4.134)\\ &= \frac{1}{2}\sum_{p\in P}\|\sqrt{p}-\sqrt{q}\|^2 &(4.135)\\ &= \frac{1}{2}\,\mathrm{cost}^{\ell_2^2}(\sqrt{P},\sqrt{q}\,)\ . &(4.136)\end{aligned}$$

□

Furthermore, for all point sets $P \subseteq \mathbb{R}^d_{\geq 0}$ of size $|P| = n$ let

$$c_{\sqrt{P}} = \frac{1}{n}\sum_{p\in P}\sqrt{p} \tag{4.137}$$

denote the centroid of the transformed point set $\sqrt{P}$. We find that for the Hellinger distance a variant of the central identity holds (compare to Lemma 3.5).

Lemma 4.25. *For all $q \in \mathbb{R}^d_{\geq 0}$ and any finite $P \subseteq \mathbb{R}^d_{\geq 0}$ of size n we have*

$$\mathrm{cost}^{He}(P,q) = \mathrm{cost}^{He}\big(P, c^2_{\sqrt{P}}\big) + n\,\mathrm{D}_{He}\big(c^2_{\sqrt{P}}, q\big) \tag{4.138}$$

where $c^2_{\sqrt{P}} = \left(\frac{1}{n}\sum_{p\in P}\sqrt{p}\right)^2$.

Proof. Using the central identity for the squared Euclidean distance and set $\sqrt{P}$ with optimal Euclidean mean $c_{\sqrt{P}}$ we obtain

$$\begin{aligned}\mathrm{cost}^{\mathrm{He}}(P,q) &= \frac{1}{2}\,\mathrm{cost}^{\ell_2^2}(\sqrt{P},\sqrt{q}\,) &(4.139)\\ &= \frac{1}{2}\,\mathrm{cost}^{\ell_2^2}(\sqrt{P},c_{\sqrt{P}}) + \frac{n}{2}\|c_{\sqrt{P}}-\sqrt{q}\|^2 &(4.140)\\ &= \mathrm{cost}^{\mathrm{H}}(P,c^2_{\sqrt{P}}) + n\,\mathrm{D}_{\mathrm{He}}\big(c^2_{\sqrt{P}},q\big)\ . &(4.141)\end{aligned}$$

□

Note that $\mathrm{cost}^{\mathrm{He}}\big(P, c^2_{\sqrt{P}}\big)$ is independent of q, and that by definition the Hellinger distance is non-negative with $\mathrm{D}_{\mathrm{He}}(x, y) = 0$ if, and only if, $x = y$. Hence, as an immediate consequence of Lemma 4.25 we conclude that for all $P \subseteq \mathbb{R}^d_{\geq 0}$

$$\mathrm{med}(P) = c^2_{\sqrt{P}} \tag{4.142}$$

is the unique optimal Hellinger 1-median of P.

Lemma 4.26. *Let $P \subseteq \mathbb{R}^d_{\geq 0}$ be finite of size n. Then a uniform sample multiset $S \subseteq P$ of size $m \geq \frac{1}{\gamma\delta}$ satisfies*

$$\Pr\Big[\mathrm{cost}_{He}\big(P, c^2_{\sqrt{S}}\big) \leq (1+\gamma)\, opt_1^{He}(P)\Big] \geq 1 - \delta\ . \tag{4.143}$$

Proof. Using Lemma 4.24, for a uniform sample multiset $S \subseteq P$ we find

$$\begin{aligned} &\Pr\big[\mathrm{cost}^{\mathrm{He}}\big(P, c^2_{\sqrt{S}}\big) \leq (1+\gamma)\,\mathrm{cost}^{\mathrm{He}}\big(P, c^2_{\sqrt{P}}\big)\big] && (4.144)\\ &\quad= \Pr\Big[\mathrm{cost}^{\ell_2^2}\big(\sqrt{P}, c_{\sqrt{S}}\big) \leq (1+\gamma)\,\mathrm{cost}^{\ell_2^2}\big(\sqrt{P}, c_{\sqrt{P}}\big)\Big]\ . && (4.145)\end{aligned}$$

Since S is a uniform sample multiset taken from P, we have that $\sqrt{S}$ is a uniform sample multiset of size $m \geq \frac{1}{\gamma\delta}$, taken from set $\sqrt{P}$ with Euclidean mean $c_{\sqrt{P}}$. Hence, by using Lemma 4.18 for the squared Euclidean distance, we obtain

$$\begin{aligned} &\Pr\big[\mathrm{cost}^{\mathrm{He}}\big(P,c^2_{\sqrt{S}}\big) \leq (1+\gamma)\, opt_1^{\mathrm{He}}\big(P\big)\big] && (4.146)\\ &\quad= \Pr\Big[\mathrm{cost}^{\ell_2^2}\big(\sqrt{P}, c_{\sqrt{S}}\big) \leq (1+\gamma)\, opt_1^{\ell_2^2}\big(\sqrt{P}\big)\Big] && (4.147)\\ &\quad\geq 1 - \delta\ . && (4.148)\end{aligned}$$

□

Corollary 4.27. *The Hellinger distance D_{He} on domain $\mathbb{R}^d_{\geq 0}$ satisfies the $[\gamma, \delta]$-sampling property with $m_{\gamma,\delta} = \frac{1}{\gamma\delta}$ and $\mathrm{med}(S) = \big(\frac{1}{|S|}\sum_{q\in S}\sqrt{q}\big)^2$.*

4.2.4 Sampling for arbitrary metrics with bounded doubling dimension

Let $(\mathbb{X}, \mathrm{D})$ be a metric space and let

$$\mathrm{diam}(Y) = \sup_{x,y\in Y} \mathrm{D}(x, y) \tag{4.149}$$

denote the *diameter* of $Y \subseteq \mathbb{X}$. A collection $\{Y_1, Y_2, \ldots, Y_\nu\}$ of subsets of Y is called a *β-covering* if $Y = \bigcup_{i=1}^{\nu} Y_i$ and $\operatorname{diam}(Y_i) \leq \beta$. The *covering number* $\mathrm{C}(Y, \beta)$ is the smallest cardinality of a β-covering of Y, i.e.

$$\mathrm{C}(Y,\beta) = \min \left\{ \nu \;\middle|\; \exists Y_1, \ldots, Y_\nu \subseteq Y : \; Y = \bigcup_{i=1}^{\nu} Y_i \;\wedge\; \operatorname{diam}(Y_i) \leq \beta \right\} . \tag{4.150}$$

The following definition is taken from [Gupta et al., 2003].

Definition 4.28 (doubling dimension). *For $Y \subseteq \mathbb{X}$ let be*

$$l(Y) = \mathrm{C}\Big(Y, \frac{1}{2}\operatorname{diam}(Y)\Big) . \tag{4.151}$$

Then the doubling dimension *of* $(\mathbb{X}, \mathrm{D})$ *is defined as*

$$\operatorname{ddim}(\mathbb{X}) = \sup_{Y \subseteq \mathbb{X}} \log\big(l(Y)\big) . \tag{4.152}$$

That is, the doubling dimension gives the binary logarithm of the smallest number l, such that every set in $\mathbb{X}$ can be covered by at most l sets of half the diameter. It can be shown that this definition features several natural properties of a dimension (c.f. [Assouad, 1983, Heinonen, 2001]). For instance, considering an ℓ_p-norm on $\mathbb{R}^d$, we obtain $\operatorname{ddim}(\mathbb{R}^d) = \mathcal{O}(d)$.

In the following, let $c \in \mathbb{X}$ be an optimal 1-median of input instance $P \subseteq \mathbb{X}$ of size n, i.e. $c = \operatorname{med}(P)$. We need the following lemma stating that with high probability all points from sample set S are close to median c. This is an immediate consequence of Markov's inequality and the union bound.

Lemma 4.29. *Let $\delta > 0$. A uniform sample multiset $S \subseteq P$ of size m satisfies*

$$\Pr\left[\exists q \in S : \mathrm{D}(q,c) \geq \frac{1}{\delta n}\operatorname{cost}(P,c)\right] \leq \delta m . \tag{4.153}$$

Proof. For a uniform random point $q \in P$ we have

$$\mathrm{E}[\mathrm{D}(q,c)] = \frac{1}{n}\sum_{p \in P} \mathrm{D}(p,c) = \frac{1}{n}\operatorname{cost}(P,c) . \tag{4.154}$$

Hence, using Markov's inequality we find

$$\Pr\left[\mathrm{D}(q,c) \geq \frac{1}{\delta n}\operatorname{cost}(P,c)\right] = \Pr\left[\mathrm{D}(q,c) \geq \frac{1}{\delta}\operatorname{E}\big[\mathrm{D}(q,c)\big]\right] \leq \delta\ . \tag{4.155}$$

By the union bound and (4.155) we obtain

$$\begin{aligned}\Pr\left[\exists q \in S : \mathrm{D}(q,c) \geq \frac{1}{\delta n}\operatorname{cost}(P,c)\right]&\\ \leq \sum_{q\in S}\Pr\left[\mathrm{D}(q,c) \geq \frac{1}{\delta n}\operatorname{cost}(P,c)\right]& \qquad (4.156)\\ \leq \delta m& \qquad (4.157)\end{aligned}$$

for $|S| = m$. □

We also need the following result which is a small modification of a result from [Indyk and Thorup, 2000] (see also [Thorup, 2005]). This lemma guarantees that points which are no good approximate median of P are very unlikely to be of any use as approximate median of S.

Lemma 4.30. *Let $\gamma \leq 1$ and let $b \in \mathbb{X}$ be an arbitrary point with*

$$\operatorname{cost}(P,b) > \left(1 + \frac{4\gamma}{5}\right)\operatorname{cost}(P,c)\ . \tag{4.158}$$

A uniform sample multiset $S \subseteq P$ of size m satisfies

$$\Pr\left[\operatorname{cost}(S,b) \leq \operatorname{cost}(S,c) + \frac{\gamma m}{5n}\operatorname{cost}(P,c)\right] < \exp\left(-\frac{\gamma^2 m}{144}\right)\ . \tag{4.159}$$

Proof. This proof is loosely based on the proof of Theorem 34 presented in [Thorup, 2005]. For a uniform sample multiset $S \subseteq P$ of size m we consider the random variable

$$\hat{X} = \frac{\operatorname{cost}(S,b) - \operatorname{cost}(S,c) + m\,\mathrm{D}(b,c)}{2\big(\mathrm{D}(b,c) + \frac{\gamma}{5n}\operatorname{cost}(P,c)\big)}\ . \tag{4.160}$$

By the triangle inequality we have

$$\operatorname{cost}(S,c) \leq \sum_{q\in S}\big(\mathrm{D}(q,b) + \mathrm{D}(b,c)\big) = \operatorname{cost}(S,b) + m\,\mathrm{D}(b,c)\ . \tag{4.161}$$

Hence, $\hat{X} \geq 0$. Using the triangle inequality, we also find

$$\operatorname{cost}(S,b) \leq \sum_{q\in S}\bigl(\mathrm{D}(q,c)+\mathrm{D}(b,c)\bigr) = \operatorname{cost}(S,c) + m\,\mathrm{D}(b,c) \tag{4.162}$$

which leads to

$$\operatorname{cost}(S,b) - \operatorname{cost}(S,c) \leq m\,\mathrm{D}(b,c)\ . \tag{4.163}$$

We obtain

$$\hat{X} \leq \frac{\operatorname{cost}(S,b)-\operatorname{cost}(S,c)+m\,\mathrm{D}(b,c)}{2\,\mathrm{D}(b,c)} \leq m\ . \tag{4.164}$$

Thus, $0 \leq \hat{X} \leq m$.

We are interested in the probability of the event $\hat{X} \leq \frac{1}{2}m$ since

$$\Pr\left[\hat{X} \leq \frac{1}{2}m\right] = \Pr\left[\frac{\mathrm{D}(S,b)-\mathrm{D}(S,c)+m\,\mathrm{D}(b,c)}{\mathrm{D}(b,c)+\frac{\gamma}{5n}\,\mathrm{D}(P,c)} \leq m\right] \tag{4.165}$$

$$= \Pr\left[\mathrm{D}(S,b)-\mathrm{D}(S,c) \leq \frac{\gamma m}{5n}\,\mathrm{D}(P,c)\right]\ . \tag{4.166}$$

Hence, event $\hat{X} \leq \frac{1}{2}m$ is equivalent to event $\mathrm{D}(S,b) \leq \mathrm{D}(S,c)+\frac{\gamma m}{5n}\,\mathrm{D}(P,c)$. We use a Chernoff bound to show that this event happens only with small probability.

First, let us estimate the expectation $\mathrm{E}[\hat{X}]$. To this end, note that there are n^m multisets $S \subseteq P$ of size m, so for any fixed $x \in \mathbb{X}$ we have

$$\mathrm{E}\bigl[\operatorname{cost}(S,x)\bigr] = \frac{1}{n^m}\sum_{\substack{S\subseteq P,\\ |S|=m}}\sum_{q\in S}\mathrm{D}(q,x)\ . \tag{4.167}$$

For $q \in P$ let ν_q denote the number of individual occurrences of the term $\mathrm{D}(q,x)$ in the double sum of equation (4.167). Since each $q \in P$ is equally likely to occur, there exists a ν with $\nu = \nu_q$ for all $q \in P$. By a double counting argument we find

$$n\nu = \sum_{q\in P}\nu_q = \sum_{\substack{S\subseteq P,\\ |S|=m}} m = n^m m\ . \tag{4.168}$$

This leads to $\nu = n^{m-1}m$ and

$$\mathrm{E}\bigl[\operatorname{cost}(S,x)\bigr] = \frac{1}{n^m}\sum_{q\in p}\nu\,\mathrm{D}(q,x) = \frac{m}{n}\operatorname{cost}(P,x)\ . \tag{4.169}$$

Using the linearity of expectation and equation (4.169) we obtain

$$\mathrm{E}[\hat{X}] = \frac{\mathrm{E}\big[\mathrm{cost}(S,b)\big] - \mathrm{E}\big[\mathrm{cost}(S,c)\big] + m\,\mathrm{D}(b,c)}{2\big(\mathrm{D}(b,c) + \frac{\gamma}{5n}\,\mathrm{cost}(P,c)\big)} \tag{4.170}$$

$$= \frac{m}{n} \cdot \frac{\mathrm{cost}(P,b) - \mathrm{cost}(P,c) + n\,\mathrm{D}(b,c)}{2\big(\mathrm{D}(b,c) + \frac{\gamma}{5n}\,\mathrm{cost}(P,c)\big)} \ . \tag{4.171}$$

Now, we show $\frac{1}{2}m \leq (1-\rho)\,\mathrm{E}[\hat{X}]$ for some positive $\rho < 1$. To this end, we want to give a lower bound on the right hand side of equation (4.171). By definition of b, we have

$$\mathrm{cost}(P,b) - \mathrm{cost}(P,c) > \frac{4\gamma}{5}\,\mathrm{cost}(P,c) \ . \tag{4.172}$$

Furthermore, since (4.158) is equivalent to $\mathrm{cost}(P,c) < \big(1+\frac{4\gamma}{5}\big)^{-1}\,\mathrm{cost}(P,b)$, we also have

$$\mathrm{cost}(P,b) - \mathrm{cost}(P,c) > \left(1 - \left(1 + \frac{4\gamma}{5}\right)^{-1}\right)\mathrm{cost}(P,b) \tag{4.173}$$

$$= \frac{4\gamma}{5}\left(1 + \frac{4\gamma}{5}\right)^{-1}\mathrm{cost}(P,b) \ . \tag{4.174}$$

Using (4.172) and (4.174), we get

$$\begin{aligned}
&\mathrm{cost}(P,b) - \mathrm{cost}(P,c) \\
&\quad = \frac{3-\frac{4\gamma}{5}}{4}(\mathrm{cost}(P,b) - \mathrm{cost}(P,c)) + \frac{1+\frac{4\gamma}{5}}{4}(\mathrm{cost}(P,b) - \mathrm{cost}(P,c)) && (4.175) \\
&\quad > \frac{3-\frac{4\gamma}{5}}{4} \cdot \frac{4\gamma}{5}\,\mathrm{cost}(P,c) + \frac{1+\frac{4\gamma}{5}}{4} \cdot \frac{\frac{4\gamma}{5}}{1+\frac{4\gamma}{5}}\,\mathrm{cost}(P,b) && (4.176) \\
&\quad = \left(3 - \frac{4\gamma}{5}\right)\frac{\gamma}{5}\,\mathrm{cost}(P,c) + \frac{\gamma}{5}\,\mathrm{cost}(P,b) && (4.177) \\
&\quad = \left(2 - \frac{4\gamma}{5}\right)\frac{\gamma}{5}\,\mathrm{cost}(P,c) + \frac{\gamma}{5}\big(\mathrm{cost}(P,c) + \mathrm{cost}(P,b)\big) \ . && (4.178)
\end{aligned}$$

Using (4.178) and the triangle inequality, we obtain

$$\begin{aligned}
&\operatorname{cost}(P,b)-\operatorname{cost}(P,c)\\
&\qquad > \left(2-\frac{4\gamma}{5}\right)\frac{\gamma}{5}\operatorname{cost}(P,c)+\frac{\gamma}{5}n\,\mathrm{D}(b,c) && (4.179)\\
&\qquad = (2-\gamma)\frac{\gamma}{5}\operatorname{cost}(P,c)+\frac{\gamma}{5}\big(n\,\mathrm{D}(b,c)+\frac{\gamma}{5}\operatorname{cost}(P,c)\big) && (4.180)\\
&\qquad \geq \frac{\gamma}{5}\operatorname{cost}(P,c)+\frac{\gamma}{5}\big(n\,\mathrm{D}(b,c)+\frac{\gamma}{5}\operatorname{cost}(P,c)\big)\ . && (4.181)
\end{aligned}$$

Here, inequality (4.181) is due to $2-\gamma \geq 1$ for $\gamma \leq 1$. Hence, using (4.171) with inequality (4.181) we find

$$\begin{aligned}
\frac{2}{m}\,\mathrm{E}[\hat{X}] &= \frac{\operatorname{cost}(P,b)-\operatorname{cost}(P,c)+n\,\mathrm{D}(b,c)}{n\,\mathrm{D}(b,c)+\frac{\gamma}{5}\operatorname{cost}(P,c)} && (4.182)\\
&\geq \frac{n\,\mathrm{D}(b,c)+\frac{\gamma}{5}\operatorname{cost}(P,c)+\frac{\gamma}{5}\big(n\,\mathrm{D}(b,c)+\frac{\gamma}{5}\operatorname{cost}(P,c)\big)}{n\,\mathrm{D}(b,c)+\frac{\gamma}{5}\operatorname{cost}(P,c)} && (4.183)\\
&= \frac{(1+\frac{\gamma}{5})\big(n\,\mathrm{D}(b,c)+\frac{\gamma}{5}\operatorname{cost}(P,c)\big)}{n\,\mathrm{D}(b,c)+\frac{\gamma}{5n}\operatorname{cost}(P,c)} && (4.184)\\
&= 1+\frac{\gamma}{5}\ . && (4.185)
\end{aligned}$$

So we have

$$\frac{1}{2}m \leq \left(1+\frac{\gamma}{5}\right)^{-1}\mathrm{E}[\hat{X}] \leq \left(1-\frac{\gamma}{6}\right)\mathrm{E}[\hat{X}] \tag{4.186}$$

since $(1+\frac{\gamma}{5})^{-1} \leq 1-\frac{\gamma}{6}$ for $\gamma \leq 1$. Using a Chernoff bound we get

$$\begin{aligned}
\Pr\left[\hat{X} \leq \frac{1}{2}m\right] &\leq \Pr\left[\hat{X} \leq \left(1-\frac{\gamma}{6}\right)\mathrm{E}[\hat{X}]\right] && (4.187)\\
&< \exp\left(-\frac{\gamma^2}{72}\,\mathrm{E}[\hat{X}]\right) && (4.188)\\
&< \exp\left(-\frac{\gamma^2 m}{144}\right)\ . && (4.189)
\end{aligned}$$

□

In the following, for a random sample multiset S let $\tilde{c}$ denote an optimal 1-median of S, i.e., $\tilde{c} = \arg\min_{x\in\mathbb{X}}\operatorname{cost}(S,x)$. We obtain the following sampling result for metrics with bounded doubling dimension.

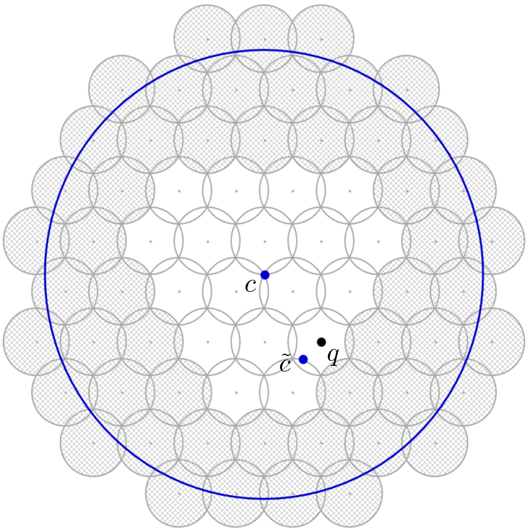


Figure 4.5: A sketch of the proof of Lemma 4.31. The ball U centered at c is covered by a number of smaller balls. Any point q from a small ball that covers $\tilde{c} = \text{med}(S)$ has to be a good approximate median for S. Also, with high probability, such a point hast to be a good approximate median for P (Lemma 4.30 implies that $q \notin N_{\text{bad}}$, i.e., that q is not contained in one of the gray-shaded smaller balls). Using the triangle inequality, we obtain that $\tilde{c}$ is also a good approximate median for P.

Lemma 4.31. *Let* $(\mathbb{X}, \mathrm{D})$ *be a metric space with* $\mathrm{ddim}(\mathbb{X}) \leq B$*. For* $\delta > 0$ *there exists a constant* λ_δ *such that every uniform sample multiset* $S \subseteq P$ *of size* $m \geq \lambda_\delta B \frac{1}{\gamma^2} \log \frac{1}{\gamma}$ *satisfies*

$$\Pr\left[\text{cost}(P, \tilde{c}) \leq (1+\gamma)\, opt_1(P)\right] \geq 1 - \delta \ . \tag{4.190}$$

Proof. Our strategy to prove the lemma is as follows. First, we observe that with high probability median $\tilde{c}$ of S has to lie within a small ball $U \subseteq \mathbb{X}$ which is centered around median c of P. In this case, we define a covering of U by subsets of $\mathbb{X}$ of small diameter. Each covering set is represented by a single point from that set. Since $\tilde{c} \in U$, there is a set covering $\tilde{c}$, and the representative q of that set has to be a good approximate median for S. From Lemma 4.30 we know that bad approximate medians for P can not be good approximate medians for S. Hence, q has to be a good approximate median for P, and since the diameter of the covering sets is small, so has to be $\tilde{c}$. This strategy is illustrated in Figure 4.5.

In the following, let $U \subseteq \mathbb{X}$ be the ball with radius $r = \frac{6m}{\delta n} \operatorname{cost}(P, c)$ and center c, i.e.,

$$U = \{x \in \mathbb{X} \,|\, \mathrm{D}(x, c) \leq r\} \ , \tag{4.191}$$

and let $U' \subseteq \mathbb{X}$ be a smaller ball with radius $r' = \frac{1}{3} r$ and center c, i.e.,

$$U' = \left\{ x \in \mathbb{X} \,\middle|\, \mathrm{D}(x, c) \leq \frac{1}{3} r \right\} \ . \tag{4.192}$$

By Lemma 4.29 we have

$$\Pr\left[\exists q \in S : \mathrm{D}(q, c) \geq \frac{2m}{\delta n} \operatorname{cost}(P, c) \right] \leq \frac{\delta}{2} \ . \tag{4.193}$$

Hence, with probability $1 - \frac{\delta}{2}$ all m sample points lie within U', i.e., $S \subseteq U'$. Now consider an arbitrary $q \in \mathbb{X} \backslash U$. If $S \subseteq U'$, using the triangle inequality for all $x \in S$ we find

$$3r' \leq \mathrm{D}(q, c) \leq \mathrm{D}(x, q) + \mathrm{D}(x, c) \leq \mathrm{D}(x, q) + r' \ . \tag{4.194}$$

Hence, we have $\mathrm{D}(x, q) \geq 2r'$ and

$$\operatorname{cost}(S, q) = \sum_{x \in S} \mathrm{D}(x, q) \geq 2r'm \ . \tag{4.195}$$

However, since $S \subseteq U'$, for center point c we have

$$\operatorname{cost}(S, c) = \sum_{x \in S} \mathrm{D}(x, c) \leq r'm < \operatorname{cost}(S, q) \ . \tag{4.196}$$

Thus, with probability $1 - \frac{\delta}{2}$ we conclude that a point $q \in \mathbb{X} \setminus U$ can not be an optimal 1-median of S. Hence, with probability $1 - \frac{\delta}{2}$, we have $\tilde{c} \in U$. Therefore, from now on we will only consider sample multisets S with 1-medians $\tilde{c}$ contained in U.

We now define a covering of U by small subsets of $\mathbb{X}$. Since $\operatorname{ddim}(\mathbb{X}) \leq B$ we know that every $Y \subseteq \mathbb{X}$ has a $\frac{1}{2} \operatorname{diam}(Y)$-cover of cardinality at most 2^B. Applying this recursively, we obtain that U has an $\frac{r}{2^j}$-covering of cardinality at most 2^{jB} for any $j \in \mathbb{N}$. Thus, for $j = \lceil \log \frac{30m}{\delta\gamma} \rceil$ there exists a $\frac{\gamma}{5n} \operatorname{cost}(P, c)$-cover of cardinality l with

$$l \leq \left(\frac{60m}{\delta\gamma} \right)^B \ . \tag{4.197}$$

Let $\{U_1, U_2, \dots, U_l\}$ be such a cover and let $N = \{x_1, x_2, \dots, x_l\} \subseteq \mathbb{X}$ be an arbitrary set of points with $x_i \in U_i$ for all $i = 1, 2, \dots, l$. Define

$$N_{\text{bad}} = \left\{ b \in N \;\middle|\; \operatorname{cost}(P, b) > \left(1 + \frac{4\gamma}{5}\right) \operatorname{cost}(P, c) \right\} . \tag{4.198}$$

We apply Lemma 4.30 to N_{bad}, using the union bound. That is, for each δ there exists a constant λ_δ such that for $m \geq \lambda_\delta B \frac{1}{\gamma^2} \log \frac{1}{\gamma}$ we have

$$\begin{aligned}
&\Pr\left[\exists\, b \in N_{\text{bad}} : \operatorname{cost}(S, b) \leq \operatorname{cost}(S, c) + \frac{\gamma m}{5n} \operatorname{cost}(P, c)\right] \\
&\quad \leq \sum_{b \in N_{\text{bad}}} \Pr\left[\operatorname{cost}(S, b) \leq \operatorname{cost}(S, c) + \frac{\gamma m}{5n} \operatorname{cost}(P, c)\right] && (4.199) \\
&\quad < \left(\frac{60m}{\delta\gamma}\right)^B \cdot \exp\left(-\frac{\gamma^2 m}{144}\right) && (4.200) \\
&\quad < \frac{\delta}{2} \,. && (4.201)
\end{aligned}$$

So, again with probability $1 - \frac{\delta}{2}$, for all $b \in N_{\text{bad}}$ we have

$$\operatorname{cost}(S, b) > \operatorname{cost}(S, c) + \frac{\gamma m}{5n} \operatorname{cost}(P, c) \,. \tag{4.202}$$

Now consider an optimal 1-median $\tilde{c}$ of S. Since $\tilde{c} \in U$, we know that $\tilde{c}$ is covered by at least one set U_i of the $\frac{\gamma}{5n} \operatorname{cost}(P, c)$-cover of U. Let q be any point from $U_i \cap N$. Since $\tilde{c}$ and q are from the same set U_i, we have

$$\mathrm{D}(q, \tilde{c}) \leq \frac{\gamma}{5n} \operatorname{cost}(P, c) \,. \tag{4.203}$$

Furthermore, using the triangle inequality and inequality (4.203), we know that for all $b \in N_{\text{bad}}$

$$\begin{aligned}
\operatorname{cost}(S, q) &\leq \operatorname{cost}(S, \tilde{c}) + m\, \mathrm{D}(q, \tilde{c}) && (4.204) \\
&\leq \operatorname{cost}(S, c) + \frac{\gamma m}{5n} \operatorname{cost}(P, c) && (4.205) \\
&< \operatorname{cost}(S, b) \,, && (4.206)
\end{aligned}$$

where the last inequality is due to inequality (4.202). Hence, we know that q is not from N_{bad} and we have

$$\operatorname{cost}(P, q) \leq \left(1 + \frac{4\gamma}{5}\right) \operatorname{cost}(P, c) \,. \tag{4.207}$$

Using the triangle inequality, inequality (4.203), and (4.207) we conclude

$$\begin{aligned}\operatorname{cost}(P,\tilde{c}) &\leq \operatorname{cost}(P,q) + n\,\mathrm{D}(q,\tilde{c}) &(4.208)\\ &\leq \left(1+\frac{4\gamma}{5}\right)\operatorname{cost}(P,c) + \frac{\gamma}{5}\operatorname{cost}(P,c) &(4.209)\\ &= (1+\gamma)\operatorname{cost}(P,c)\ . &(4.210)\end{aligned}$$

This event happens with probability at least $(1-\frac{\delta}{2})^2 > 1-\delta$. □

Corollary 4.32. *An arbitrary metric space* $(\mathbb{X},\mathrm{D})$ *with* $\operatorname{ddim}(\mathbb{X}) \leq B$ *satisfies the* $[\gamma,\delta]$*-sampling property with* $m_{\gamma,\delta} = \lambda_\delta B \frac{1}{\gamma^2}\log\frac{1}{\gamma}$*, provided that we have access to an algorithm that computes* $\operatorname{med}(S) = \arg\min_{x\in\mathbb{X}}\operatorname{cost}(S,x)$ *in time depending only on* $|S|$*.*

4.2.5 Sampling for the Hamming distance

A distance measure that has many applications in coding theory and text mining is the *Hamming distance.* The Hamming distance measures the dissimilarity between bit strings of a fixed length d. These bit strings are given as elements from the discrete set $\{0,1\}^d \subseteq \mathbb{R}^d$, which is also called the *Hamming cube* of dimension d. For bit strings $x = (x_1, x_2, \ldots, x_d)^\top \in \{0,1\}^d$ and $y = (y_1, y_2, \ldots, y_d)^\top \in \{0,1\}^d$, the Hamming distance on $\{0,1\}^d$ is defined as the number of coordinates on that x and y disagree, that is,

$$\mathrm{D}_{\mathrm{Ha}}(x,y) = \left|\{1 \leq j \leq d \,|\, x_j \neq y_j\}\right| \ . \tag{4.211}$$

In the Hamming k-median problem, given a finite set $P \subseteq \{0,1\}^d$, we are interested in finding k center points $C = \{c_1, c_2, \ldots, c_k\}$ on the Hamming cube $\{0,1\}^d$ that minimize $\operatorname{cost}^{\mathrm{Ha}}(P,C)$. It is known that the Hamming distance on $\{0,1\}^d$ is a metric with a doubling dimension bounded by $\mathcal{O}(d)$. Hence, in the light of the results from Section 4.2.4, the Hamming k-median problem can be seen as an instance of the metric k-median problem with bounded doubling dimension. However, in this section, we present a more concrete result that leads to better bounds.

In order to study the sampleability of the Hamming k-median problem we make use of the following connection between the Hamming distance and the squared Euclidean distance on the Hamming cube.

Lemma 4.33. *For all* $x, y \in \{0,1\}^d$ *we have* $\mathrm{D}_{Ha}(x,y) = \|x-y\|^2$*.*

Proof. Obviously, for $x = (x_1, x_2, \dots, x_d)^\top$ and $y = (y_1, y_2, \dots, y_d)^\top$ with $x_j, y_j \in \{0,1\}$ we have $x_j \neq y_j$ if and only if $(x_j - y_j)^2 = 1$. In addition, we have $x_j = y_j$ if and only if $(x_j - y_j)^2 = 0$. Hence, we obtain

$$\mathrm{D}_{\mathrm{Ha}}(x,y) = \sum_{j=1}^{d} (p_j - q_j)^2 = \|p - q\|^2 \ . \tag{4.212}$$

□

For any finite number of bits $b_1, b_2, \dots, b_n \in \{0,1\}$ let $\mathrm{maj}\{b_1, b_2, \dots, b_n\}$ denote the majority function of the bits $b_1, b_2, \dots, b_n$, that is, the element from $\{0,1\}$ that occurs more frequently than the other. In case that the majority is ambiguous, i.e., $b_1, b_2, \dots, b_n$ consists of the same number of zeros and ones, we assume $\mathrm{maj}(b_1, b_2, \dots, b_n) = 0$. Furthermore, for a point set $P \subseteq \{0,1\}^d$ and an index $1 \leq j \leq d$, where p_j denotes the j-th coordinate of point $p \in P$, we define

$$\mathrm{maj}_j(P) = \mathrm{maj}\{p_j \,|\, p \in P\} \ . \tag{4.213}$$

We also write

$$\mathrm{maj}(P) = \big(\mathrm{maj}_1(P), \mathrm{maj}_2(P), \dots, \mathrm{maj}_d(P)\big)^\top \tag{4.214}$$

as the vector consisting of the coordinate-wise majority of the points from P. Note that for each finite $P \subseteq \{0,1\}^d$ we also have $\mathrm{maj}(P) \in \{0,1\}^d$. Moreover, it turns out that $\mathrm{maj}(P)$ is indeed an optimal 1-median of point set P, as is stated in the following lemma.

Lemma 4.34. *Let $P \subseteq \{0,1\}^d$ be of size n. Then* $\mathrm{maj}(P)$ *is an optimal solution to the Hamming 1-median problem of input set P, i.e.,*

$$\mathrm{med}(P) = \mathrm{maj}(P) \ . \tag{4.215}$$

In particular, the 1-median of P is unique if and only if the majority of the j-th coordinate of the points from P is unambiguous for each $1 \leq j \leq d$.

Proof. Let $c = (c_1, c_2, \dots, c_d)^\top \in \{0,1\}^d$ be an optimal 1-median of P. Furthermore, for all $1 \leq j \leq d$ let

$$\nu_j = \sum_{p \in P} p_j \tag{4.216}$$

denote the number of elements $p \in P$ with a 1 as its j-th coordinate. If coordinate $c_j = 0$, due to the optimality of c, for the contribution of the j-th coordinate to $\mathrm{cost}^{\mathrm{Ha}}(P, c)$ we have

$$\nu_j = \sum_{p \in P} (p_j - 0)^2 \leq \sum_{p \in P} (p_j - 1)^2 = d - \nu_j, \tag{4.217}$$

or, equivalently, $\nu_j \leq d/2$. Hence, $c_j = 0$ is a majority element of the j-th coordinate of the points from P. On the other hand, if $c_j = 1$ then we have

$$d - \nu_j = \sum_{p \in P} (p_j - 1)^2 \leq \sum_{p \in P} (p_j - 0)^2 = \nu_j, \tag{4.218}$$

or, equivalently, $\nu_j \geq d/2$. Hence, we find that $c_j = 1$ is a majority element of the j-th coordinate of the points from P.

We conclude that $c \in \{0,1\}^d$ is an optimal 1-median of P if and only if the coordinates of c consist of majority elements of the coordinates of the points from P. Hence, $\mathrm{maj}(P)$ is an optimal 1-median of P, and if the majority of the j-th coordinate is unambiguous for all $1 \leq j \leq d$, then this 1-median of P is unique. □

Next, we show that in the case of the Hamming distance, constant sized sampling from point set P leads to a good approximation of the 1-median of P. This observation has already been mentioned in [Ailon et al., 2006]. However, no formal proof has been given.

Lemma 4.35. *Let $P \subseteq \{0,1\}^d$ be of size n. Then a uniform sample multiset $S \subseteq P$ of size $m \geq \frac{4(2+\gamma)}{\gamma^2 \delta}$ satisfies*

$$\Pr\big[\mathrm{cost}(P, \mathrm{maj}(S)) \leq (1+\gamma)\, opt_1(P)\big] \geq 1 - \delta\ . \tag{4.219}$$

Proof. In the following, let $c = \mathrm{maj}(P)$ and let $\tilde{c} = \mathrm{maj}(S)$. While median c is constant for any fixed set P, the approximate median $\tilde{c}$ is a random element from Hamming cube $\{0,1\}^d$, depending on the random choice of uniform sample multiset S. To prove the lemma, we use probabilistic concentration bounds to show that, with high probability, for any fixed coordinate $0 \leq j \leq d$ the (expected) contribution of the j-th coordinate to $\mathrm{cost}^{\mathrm{Ha}}(P, \tilde{c})$ is very close to the contribution of the j-th coordinate to $\mathrm{cost}^{\mathrm{Ha}}(P, c)$.

To this end, let $c = (c_1, c_2, \dots, c_d)^\top$ and $\tilde{c} = (\tilde{c}_1, \tilde{c}_2, \dots, \tilde{c}_d)^\top$. Furthermore, for all $1 \le j \le d$, the fraction of elements $p \in P$ with a 1 as its j-th coordinate is given by

$$\zeta_j = \frac{1}{n} \sum_{p \in P} p_j \ , \tag{4.220}$$

where p_j denotes the j-th coordinate of a point $p \in P$. Without loss of generality, throughout this proof we assume $c_j = \mathrm{maj}_j(P) = 0$ and, hence, $0 \le \zeta_j \le \frac{1}{2}$. Otherwise, the same argumentation as given below is valid by replacing 0 with 1, ζ_j with $1 - \zeta_j$, and so on.

According to Lemma 4.33, for each index j, the contribution of the j-th coordinate to the difference $\mathrm{cost}^{\mathrm{Ha}}(P, \tilde{c}) - \mathrm{cost}^{\mathrm{Ha}}(P, c)$ is given by

$$D_j = \sum_{p \in P} (p_j - \tilde{c}_j)^2 - \sum_{p \in P} (p_j - c_j)^2 \ . \tag{4.221}$$

Obviously, if $\tilde{c}_j = c_j$ then we have $D_j = 0$. Hence, from now on we may assume $\tilde{c}_j = 1 - c_j = 1$. In this case, we obtain

$$D_j = \sum_{p \in P} (1 - p_j) - \sum_{p \in P} p_j = (1 - 2\zeta_j) n = \frac{1 - 2\zeta_j}{\zeta_j} \sum_{p \in P} (p_j - c_j)^2 \ , \tag{4.222}$$

since for $c_j = 0$ we have

$$\sum_{p \in P} (p_j - c_j)^2 = \sum_{p \in P} p_j = \zeta_j n \ . \tag{4.223}$$

In the following, we consider the partition

$$J_1 = \left\{ 1 \le j \le d \,\middle|\, \frac{1}{2 + \gamma} < \zeta_j \le \frac{1}{2} \right\} \tag{4.224}$$

$$J_2 = \left\{ 1 \le j \le d \,\middle|\, 0 \le \zeta_j \le \frac{1}{2 + \gamma} \right\} \tag{4.225}$$

of the set of indices $\{1, 2, \dots, d\}$. We give a bound on $\sum_{j=1}^{d} D_j$ by considering the cases $j \in J_1$ and $j \in J_2$ individually.

First, let us assume $j \in J_1$, that is, ζ_j is close to $\frac{1}{2}$. In this case, obviously, the contributions of the j-th coordinate do not differ greatly when using either 0 or 1 as the j-th coordinate of a center point. More precisely, from equation (4.222) for $j \in J_1$ we obtain

$$D_j \leq \gamma \sum_{p \in P} (p_j - c_j)^2 \tag{4.226}$$

since $\frac{1-2\zeta_j}{\zeta_j} < \gamma$ for $\zeta_j > \frac{1}{2+\gamma}$. Note that the bound given by inequality (4.226) is valid regardless of the random choices made by sampling S. Hence, by summing up over all indices $j \in J_1$, we find

$$\sum_{j \in J_1} D_j \leq \gamma \sum_{p \in P} \sum_{j \in J_1} (p_j - c_j)^2 \; . \tag{4.227}$$

Thus, in the following, let us assume $j \in J_2$ and that ζ_j is significantly smaller than $\frac{1}{2}$, i.e., $0 \leq \zeta_j \leq \frac{1}{2+\gamma}$. In this case, our goal is to give a bound on the expected difference

$$\mathrm{E}[D_j] = \Pr[\tilde{c}_j = 1] \cdot \frac{1-2\zeta_j}{\zeta_j} \sum_{p \in P} (p_j - c_j)^2 \; . \tag{4.228}$$

To this end, let

$$A_j = |\{q \in S \,|\, q_j = 0\}| \tag{4.229}$$

denote the random variable counting the number of sampled points whose j-th coordinate equals the majority element $c_j = 0$. By definition of ζ_j, we know that for each individual $q \in S$ the event $q_j = 0$ occurs exactly with probability ζ_j. Since the elements of S are chosen independently and identically distributed at random, we obtain that A_j is distributed according to a binomial distribution with

$$\mathrm{E}[A_j] = \zeta_j m \tag{4.230}$$

and

$$\mathrm{Var}[A_j] = \zeta_j (1 - \zeta_j) m \; . \tag{4.231}$$

Furthermore, the event $\tilde{c}_j = 1$ is equivalent to the event $A_j \geq \frac{1}{2} m$. Hence,

$$\begin{align}
\Pr\left[\tilde{c}_j = 1\right] &= \Pr\left[A_j \geq \frac{1}{2} m\right] \tag{4.232}\\
&= \Pr\left[A_j - \zeta_j m \geq \left(\frac{1}{2} - \zeta_j\right) m\right] \tag{4.233}\\
&\leq \Pr\left[\left|A_j - E[A_j]\right| \geq \frac{1}{2}(1 - 2\zeta_j) m\right] \; . \tag{4.234}
\end{align}$$

Using Chebyshev's inequality, we find

$$\Pr\left[\tilde{c}_j = 1\right] \leq \frac{4\ \mathrm{Var}[A_j]}{(1-2\zeta_j)^2 m^2} = \frac{4\,\zeta_j(1-\zeta_j)}{(1-2\zeta_j)^2 m} \ . \tag{4.235}$$

Thus, using equation (4.228) we obtain

$$\mathrm{E}[D_j] \leq \frac{4(1-\zeta_j)}{(1-2\zeta_j)m} \sum_{p\in P} (p_j - c_j)^2 \ . \tag{4.236}$$

Since

$$\frac{1-\zeta_j}{1-2\zeta_j} \leq \frac{1}{1-\frac{2}{2+\gamma}} = \frac{2+\gamma}{\gamma} \tag{4.237}$$

for $0 \leq \zeta_j \leq \frac{1}{2+\gamma}$, and since $m \geq \frac{4(2+\gamma)}{\gamma^2\delta}$, for all $j \in J_2$ we obtain

$$\mathrm{E}[D_j] \leq \frac{4(2+\gamma)}{\gamma m} \sum_{p\in P} (p_j - c_j)^2 \leq \gamma\delta \sum_{p\in P} (p_j - c_j)^2 \ . \tag{4.238}$$

By summing up over all indices $j \in J_2$, we find

$$\mathrm{E}\left[\sum_{j\in J_2} D_j\right] = \sum_{j\in J_2} \mathrm{E}[D_j] \leq \gamma\delta \sum_{p\in P}\sum_{j\in J_2} (p_j - c_j)^2 \ . \tag{4.239}$$

Hence, using Markov's inequality, with probability at least $1-\delta$ we obtain

$$\sum_{j\in J_2} D_j \leq \gamma \sum_{p\in P}\sum_{j\in J_2} (p_j - c_j)^2 \ . \tag{4.240}$$

Therefore, using inequalities (4.227) and (4.240) we conclude

$$\mathrm{cost}^{\mathrm{Ha}}(P,\tilde{c}) - \mathrm{cost}^{\mathrm{Ha}}(P,c) = \sum_{j\in J_1} D_j + \sum_{j\in J_2} D_j \tag{4.241}$$

$$\leq \gamma \sum_{p\in P}\sum_{j=1}^{d} (p_j - c_j)^2 \tag{4.242}$$

$$= \gamma\, \mathrm{cost}^{\mathrm{Ha}}(P,c) \ . \tag{4.243}$$

□

Corollary 4.36. *The Hamming distance* D_{Ha} *on* $\{0,1\}^d$ *satisfies the* $[\gamma,\delta]$*-sampling property with* $m_{\gamma,\delta} = \frac{4(2+\gamma)}{\gamma^2\delta}$ *and* $\mathrm{med}(S) = \mathrm{maj}(S)$*.*

4.3 Generalization of the sampling property

In this section we generalize our result to an even larger family of dissimilarity measures. For the $[\gamma, \delta]$-sampling property as stated in Property 4.1, we require that the optimal 1-median $\mathrm{med}(S)$ of S can be computed in finite time. However, for some dissimilarity measures such as the Euclidean distance, no such algorithm is known. Moreover, it has been shown that in the Euclidean case finding $\mathrm{med}(S)$ requires finding roots of high-order polynomials, which can not be achieved using only radicals [Bajaj, 1988]. So, we can not hope to use the characterization of Theorem 4.2 to show that algorithm Cluster is also applicable to the Euclidean k-median problem.

It turns out that our definition of the $[\gamma, \delta]$-sampling property is far more restrictive than is necessary for our algorithm. In fact, all we have to ensure is that given a constant sized sample set $S \subseteq P$, we can compute a constant sized set of candidates for the approximate 1-median of cluster P. That is, we only need to guarantee that at least one of the elements of the candidate set is indeed a good approximation for $\mathrm{med}(P)$. Finding an optimal 1-median of S is sufficient for the dissimilarities studied in Section 4.2, but in general it is not necessary.

Therefore, we formulate the following relaxation of the $[\gamma, \delta]$-sampling property.

Property 4.37. *We say a dissimilarity measure* D *satisfies the* weak $[\gamma, \delta]$-sampling property *if there exist integer constants $m_{\gamma,\delta}$ and $t_{\gamma,\delta}$ such that for each $P \subseteq \mathbb{X}$ of size n and for each uniform sample multiset $S \subseteq P$ of size $m_{\gamma,\delta}$ a set $T(S) \subseteq \mathbb{X}$ of size at most $t_{\gamma,\delta}$ can be computed satisfying*

$$\Pr\big[\exists \tilde{c} \in T(S):\ \mathrm{cost}(P, \tilde{c}) \leq (1+\gamma)\, opt_1(P)\big] \geq 1 - \delta. \tag{4.244}$$

Furthermore, set $T(S)$ can be computed from S in time depending only on γ, δ, and $|S|$.

Obviously, if D satisfies the (strong) $[\gamma, \delta]$-sampling property it also satisfies the weak property by virtue of $T(S) = \big\{\mathrm{med}(S)\big\}$ and $t_{\gamma,\delta} = 1$.

Algorithm Cluster can be easily adapted to this new property. More precisely, for each subset $S' \subseteq S$ of size $m_{\gamma,\delta}$ instead of a single point a constant sized set $T(S')$ is added to the candidate set. Hence, the number of candidates that have to be tried at each level of the recursion scales merely by a factor of $\mathcal{O}(t_{\gamma,\delta})$. Therefore, the asymptotic running time of algorithm Cluster remains the same, up to a factor of $\mathcal{O}(t^k_{\gamma,\delta})$. We obtain the following result.

Theorem 4.38. *Let* D *be an arbitrary dissimilarity measure on domain* $\mathbb{X}$. *Let* $k \in \mathbb{N}$ *and* $0 < \varepsilon, \delta < 1$, *and let* D *satisfy the weak* $[\varepsilon/3, \delta]$*-sampling property. Then there exists an algorithm that, with constant probability, returns a* $(1+\varepsilon)$*-approximate solution of the* k*-median problem with respect to* D *for input instance* P *of size* n. *Furthermore, this solution can be found using at most* $2^{\mathcal{O}(mk\log(mk/\varepsilon))}t^k n$ *operations, including evaluations of* D, *where* $m = m_{\varepsilon/3,\delta}$ *and* $t = t_{\varepsilon/3,\delta}$ *denote constants that depend only on* ε, δ, *and* D.

Using this new characterization, we can show that algorithm CLUSTER is well suited for the Euclidean k-median problem, that is, the k-median problem using the Euclidean distance

$$\mathrm{D}_{\ell_2}(x, y) = \|x - y\| \tag{4.245}$$

on $\mathbb{R}^d$ as dissimilarity measure. To this end, we make use of the following result from [Kumar et al., 2005].

Theorem 4.39 ([Kumar et al., 2005], Theorem 1)**.** *Let* P *be a set of* n *points in* $\mathbb{R}^d$, *and let* γ *be a constant,* $0 < \gamma < \frac{1}{12}$. *There exists an algorithm which randomly samples a set* S *of* $(1/\gamma)^{\mathcal{O}(1)}$ *points from* P. *Using this sample only, it constructs a set of points* $T(S)$ *of size* $2^{(1/\gamma)^{\mathcal{O}(1)}}$ *such that with constant probability there is a point* $x \in T(S)$ *satisfying*

$$\mathrm{cost}(P, x) \leq \big(1 + \mathcal{O}(\gamma)\big)\, opt_1(P) \ . \tag{4.246}$$

Further, the time taken to construct $T(S)$ *from* S *is* $d2^{(1/\gamma)^{\mathcal{O}(1)}}$.

Corollary 4.40. *The Euclidean distance* D_{ℓ_2} *on domain* $\mathbb{R}^d$ *satisfies the weak* $[\gamma, \delta]$*-sampling property with* $m_{\gamma,\delta} = (1/\gamma)^{\mathcal{O}(1)}$ *and* $t_{\gamma,\delta} = 2^{(1/\gamma)^{\mathcal{O}(1)}}$, *where* $T(S)$ *is computed by the algorithm from Theorem 4.39.*

In [Kumar et al., 2005], it is shown that a linear time $(1 + \varepsilon)$-approximation algorithm exists if dissimilarity measure D satisfies two properties: a so-called "random sampling procedure property" and a "tightness property". Property 4.37 can be seen as a generalization of this "random sampling procedure property". However, since we do not restrict ourselves to symmetric distance measures, our property is valid for a larger family of dissimilarity measures. Moreover, Theorem 4.38 shows that the second requirement from [Kumar et al., 2005] ("tightness property") is not necessary at all to achieve a $(1 + \varepsilon)$-approximation for dissimilarity measures satisfying the $[\gamma, \delta]$-sampling property.

4.4 Discussion

In this chapter, we have shown how to obtain a linear time $(1+\varepsilon)$-approximation algorithm for the k-median problem with respect to an arbitrary dissimilarity measure D, provided that the 1-median problem can be approximated within a factor of $(1+\varepsilon)$ by taking a random sample of constant size. In doing so, we have given a sufficient condition for the existence of $(1+\varepsilon)$-approximation algorithms for the k-median problem which we call the $[\gamma, \delta]$-sampling property. This property makes only minimal assumptions on the dissimilarity measure D. Therefore, it is well-suited for application to arbitrary, non-metric dissimilarity measures. In particular, we have shown that the $[\gamma, \delta]$-sampling property is satisfied for Mahalanobis distances, μ-similar Bregman divergences, the Hellinger distance, arbitrary metrics with bounded doubling dimension, and the Hamming distance. An interesting direction for future research is to find additional dissimilarity measures that satisfy the $[\gamma, \delta]$-sampling property. In addition, it still remains an open problem to give sufficient and necessary conditions for dissimilarity measures such that a $(1+\varepsilon)$-approximate solution of the corresponding k-median problem can be found.

Our flexible tool for finding the k-medians of a given clustering problem is algorithm Cluster, given in Figure 4.2. Algorithm Cluster is the generalization of an earlier algorithm, namely algorithm Irred-k-means from [Kumar et al., 2004]. However, our interpretation and analysis of the algorithm differ significantly from [Kumar et al., 2004]. More precisely, consider the case $k = 2$ and let P_1 and P_2 denote the optimal clusters of point set P. Given an approximate median $\tilde{c}_1$ of cluster P_1, the goal of the analysis in [Kumar et al., 2004] is to show that in the pruning phase only points from P_1 will be assigned to $\tilde{c}_1$. To achieve this goal, the analysis relies heavily on the symmetry and the triangle inequality of the Euclidean distance. Furthermore, the notion of irreducibility is of fundamental importance to the analysis from [Kumar et al., 2004].

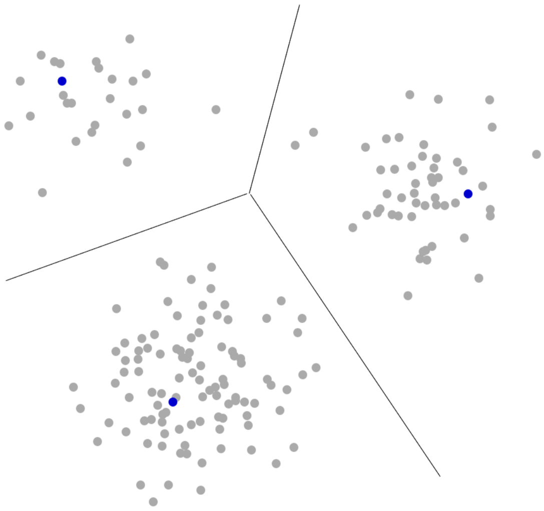
In our analysis, we do not rely on metric properties of the dissimilarity measure, or the irreducibility of the input instance. However, without these assumptions it seems that we are not able to show the same pruning result as Kumar et al. did. Hence, in our analysis, we explicitly allow that points from P_2 are assigned to $\tilde{c}_1$. Our goal is merely to show that most of the points assigned to $\tilde{c}_1$ are from P_1, and that the total cost of the points from P_2 that are incorrectly assigned to $\tilde{c}_1$ is negligible. We use the constant parameter α in algorithm Cluster to control the number of points that are incorrectly assigned. In detail, we show that in each pruning phase the total cost of incorrectly assigned points is always bounded by an $\mathcal{O}(\alpha k)$ fraction of the total cost of the correctly assigned points. By choosing $\alpha = \Theta(\varepsilon/k^2)$ small enough we are able to bound the approximation factor by $1 + \varepsilon$ while the sample size is still constant and only depends on parameters k and ε. This approach is purely combinatorial and requires only minimal assumptions on the dissimilarity measure. In particular, symmetry, triangle inequality, and the notion of irreducible clusterings are no longer needed.

Note that in [Kumar et al., 2005] the methods from [Kumar et al., 2004] have been generalized to a class of metric and Euclidean distance measures. However, symmetry and triangle inequality are always assumed. Our generalized result as stated in Theorem 4.38 is more general than the results in [Kumar et al., 2005] and applies to a broader class of metric and non-metric distance measures.

Unfortunately, algorithm Cluster is not very practical. While the asymptotic running time is linear in n, the constants involved are quite huge even for a relatively harmless choice of parameters. For instance, consider a sampling step for the case of $k = 10$ clusters when we want to obtain a uniform sample set of, say, 10 points from a large cluster P_i with $|P_i| \geq \frac{1}{10}|P|$. Using the superset sampling technique from Section 4.1.1, we have to sample a set of size at least 200 and recursively evaluate all its subsets of size 10. This leads to at least $\binom{200}{10} \approx 2.2 \cdot 10^{16}$ recursive calls — at each level of the recursion! While asymptotically this is a constant number, from a practitioner's viewpoint it is a quite demanding computational task. Hence, in the following chapter we will introduce and analyze a fast approximation algorithm for the Bregman k-median problem that performs quite well in practice.

In addition, in Chapter 7 we will show how the running time of algorithm Cluster can be improved significantly through the use of so called weak coresets. Unfortunately, the constants involved in the running time of this improved algorithm are still huge.

5 A practical $\mathcal{O}(\log k)$-approximate algorithm



In the previous chapter, we have given a linear time algorithm that computes a $(1+\varepsilon)$-approximate solution for the generalized k-median problem with respect to a large number of dissimilarity measures, including all μ-similar Bregman divergences. Unfortunately, this algorithm is not very practical due to the huge constants involved in the running time. One particular algorithm for the Euclidean k-means problem that has appealed to practitioners during the past decades is *Lloyd's k-means algorithm.* This algorithm has been developed by Stuart P. Lloyd as early as the late 1950s, and it has been published later in [Lloyd, 1982]. Starting with an arbitrary set of k center points $c_1, c_2, \ldots, c_k$, Lloyd's local improvement strategy iterates between two steps:

1. Assign each input point to its closest center point to build a partition $P_1, P_2, \ldots, P_k$ of the input points.

2. For each set P_i recompute center point c_i as the centroid of P_i.

These steps are repeated until the partition and the center points become stable. It can easily be seen that the reassignment of partitions and center points can only decrease the Euclidean k-means cost induced by the current set of centers. Since there are only a finite number of partitions it follows that after a finite number of steps Lloyd's algorithm converges to a stable clustering. In addition, a single Lloyd iteration can be implemented to run really fast (see [Kanungo et al., 2002] for the suggestion of an efficient nearest neighbor data structure).

However, it is known that the speed of convergence as well as the quality of the local optimum computed depends strongly on the choice of initial center points. While it has been shown recently that the expected number of iterations is polynomial in n, k, and d in the smoothed complexity model [Arthur et al., 2009], in the worst case, an exponential number of $2^{\Omega(n)}$ iterations are necessary, even in the plane [Vattani, 2009]. Furthermore, there are simple examples of input sets such that a poor choice of initial centers leads to arbitrarily bad clusterings (i.e., the approximation ratio is unbounded, see Figure 5.1). To deal with these problems in practice, the initial center points are usually chosen uniformly at random among the input points. However, no non-trivial approximation guarantees are known for uniform seeding. Recently, in [Arthur and Vassilvitskii, 2007] a new non-uniform initial seeding procedure for the Euclidean k-means problem, named K-MEANS++, has been given. In a breakthrough result it has been shown that this seeding step alone computes an $\mathcal{O}(\log k)$-approximate set of centers. Any following Lloyd iterations only improve this solution. Independently, in [Ostrovsky et al., 2006] the analysis of essentially the same seeding procedure has been given. In this analysis it has been shown, that for certain well-separated input instances the non-uniform seeding step gives an $\mathcal{O}(1)$-approximate set of centers. Hence, this new seeding approach is indeed of high practical relevance.

It has been observed that Lloyd's algorithm is also applicable to other dissimilarity measures. In [Linde et al., 1980], Lloyd's algorithm has been adapted to the Itakura-Saito divergence. In fact, in the context of vector

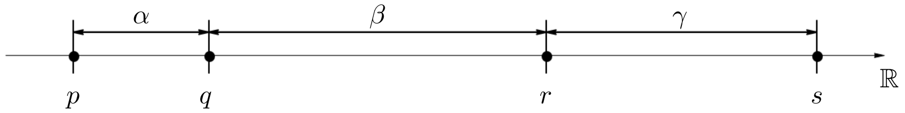


Figure 5.1: An example of Lloyd's algorithm where a poor choice of initial centers leads to an arbitrarily bad clustering for the Euclidean 3-means problem. Let $P = \{p, q, r, s\} \subseteq \mathbb{R}$ be as illustrated above, with $\alpha < \gamma$ and $\beta > \gamma/2$. For the initial set of centers $\{p, q, r\}$, after a single Lloyd iteration we obtain $\tilde{P}_1 = \{p\}$, $\tilde{P}_2 = \{q\}$, and $\tilde{P}_3 = \{r, s\}$ as a stable partition. This clustering has a 3-means cost of $\gamma^2/2$. On the other hand, an optimal 3-means clustering is given by $P_1 = \{p, q\}$, $P_2 = \{r\}$, and $P_3 = \{s\}$ with a cost of $\alpha^2/2$. Hence, by letting γ/α tend to infinity, we obtain arbitrarily bad clusterings.

quantization Lloyd's approach is known as the *Linde-Buzo-Gray algorithm*. In the context of text classification, Lloyd's algorithm has been adapted to k-median clustering by Kullback-Leibler divergence [Dhillon et al., 2003]. These observations have been extended to the whole class of Bregman divergences by the work of Banerjee and others (cf. [Banerjee et al., 2005a] and [Banerjee et al., 2005b]). Moreover, it has been shown that under some mild continuity assumption the class of Bregman divergences is exactly the the class of dissimilarity measures for which Lloyd's algorithm is applicable (see [Banerjee et al., 2005a]). In addition to that, results on the worst-case and smoothed number of iterations have also been generalized to the Bregman k-median problem in [Manthey and Röglin, 2009]. However, approximation guarantees were not known.

In this chapter, we give a generalization of the K-MEANS++ seeding approach from [Arthur and Vassilvitskii, 2007] to the class of Bregman divergences. In particular, in Section 5.1 we show how to construct a factor $\mathcal{O}(\log k)$-approximation for the k-median problem with respect to a μ-similar Bregman divergence D_φ, using a K-MEANS++-like non-uniform sampling approach. Our main result can be summarized as follows.

Theorem 5.1. *Let D_φ be a μ-similar Bregman divergence on domain $\mathbb{X} \subseteq \mathbb{R}^d$. Let $P \subseteq \mathbb{X}$ be of size n. There exists a randomized algorithm using non-uniform sampling that with high probability computes an $\mathcal{O}(\log k)$-approximate solution of the Bregman k-median problem with input instance P. Furthermore, this solution is obtained using at most $\mathcal{O}(kn)$ arithmetic operations, including evaluations of D_φ.*

This result has been published earlier in [Ackermann and Blömer, 2009]. Independently, essentially the same generalization of the sampling result from [Arthur and Vassilvitskii, 2007] has also been given in two other papers from the machine learning community. In [Sra et al., 2008], the authors derive a bound on the curvature of the Hessian $\nabla^2\varphi$ for Bregman divergence D_φ by using spectral information of $\nabla^2\varphi$. They obtain the same result as we do in Theorem 5.3 depending on spectral parameters σ_1, σ_2 of $\nabla^2\varphi$, where $\frac{\sigma_1}{\sigma_2} = \mu$ in our context of μ-similar Bregman divergences. Furthermore, they extend their result to the context of Bregman co-clustering and Tensor clustering. In [Nock et al., 2008], an interesting generalization of Lloyd's k-means method to so-called mixed Bregman clusterings is given. Here a mixed Bregman divergence is a symmetrized version of a Bregman divergences. The optimal 1-median of such a mixed Bregman cluster can be computed by using the Legendre duality of Bregman divergences (cf. [Nielsen et al., 2007]). While Nock et al. tackle the problem from a quite different angle, they obtain the same result as we do in Theorem 5.3 depending on a parameter ρ, where $\rho^2 = \frac{1}{\mu}$ in our context of μ-similar Bregman divergences.

In Section 5.2, we prove that with constant probability the non-uniform seeding from Theorem 5.1 gives a constant factor approximation when restricted to so-called separable input instances. Here, a k-median input instance is called separable if the cost of an optimal $(k-1)$-clustering is by a constant factor larger than the cost of an optimal k-clustering. This notion captures the idea that, in practice, unless the input consists of k well-separated clusters the k-medians will not be a meaningful representation of the data anyway. The main result from Section 5.2 is as follows.

Theorem 5.2. *Let D_φ be a μ-similar Bregman divergence on $\mathbb{X}$ and let $P \subseteq \mathbb{X}$ be a separable input set. Then with probability at least $2^{-\Theta(k)}$ the algorithm from Theorem 5.1 computes a constant factor approximate solution of the Bregman k-median problem with input instance P.*

The notion of separable input instances has been used before to analyze clustering algorithms (cf. [Kanungo et al., 2002], [Kumar et al., 2004], [Ostrovsky et al., 2006]). In [Ostrovsky et al., 2006], it is shown that a non-uniform seeding approach very similar to the seeding approach from [Arthur and Vassilvitskii, 2007] provides a constant factor approximate solution for separable instances in the context of the Euclidean k-means problem. Our result can be seen as a generalization of the result from [Ostrovsky et al., 2006] to the class of Bregman divergences. However, we obtain our result by a significantly simplified proof that focuses on the combinatorial properties of the Bregman k-median problem. Theorem 5.2 has also been published in [Ackermann and Blömer, 2010].

The main algorithm from this chapter will be of use for us in three ways. First of all, the algorithm is highly practical. From a practitioner's viewpoint this might well be the algorithm of choice for the μ-similar Bregman k-median problem. Second, the non-uniform sampling approach will be our main technique for a new randomized coreset construction to obtain strong coresets in the context of the Mahalanobis k-median problem. We will give this coreset construction in Chapter 6. And third, we will also make use of the algorithm given in this chapter to obtain a fast, initial set of centers in the construction of weak coresets for the μ-similar Bregman k-median problem in Chapter 7.

5.1 Algorithm BregMeans++

5.1.1 Non-uniform sampling scheme for μ-similar Bregman divergences

Our approximation algorithm is a generalization of the non-uniform random sampling approach from [Arthur and Vassilvitskii, 2007] as follows. The first of k approximate medians is chosen uniformly at random from input instance P. After that, assume that we have already chosen approximate medians $A_i = \{a_1, a_2, \ldots, a_i\}$. The next approximate median a_{i+1} is chosen from P with probability proportional to $\mathrm{D}_\varphi(a_{i+1}, A_i)$, that is, for all $p \in P$ we have

$$\Pr\Big[p = a_{i+1} \;\Big|\; a_1, a_2, \ldots, a_i \text{ already chosen}\Big] = \frac{\mathrm{D}_\varphi(p, A_i)}{\mathrm{cost}(P, A_i)} \; . \tag{5.1}$$

Note that as long as we have at least $i+1$ different points in P, we have $\mathrm{cost}(P, A_i) > 0$ and this probability distribution is well defined. The sampling scheme is repeated until we have chosen k points. We say set $A = \{a_1, a_2, \ldots, a_k\}$ is chosen *at random according to* D_φ. Algorithm BregMeans++(P, k) realizing this non-uniform sampling strategy is summarized in Figure 5.2.

The k-means++ sampling approach has been originally proposed in the context of Euclidean k-means clustering, as well as for the k-median problem using a t-th power of the ℓ_2-norm as distance measure. In Section 5.1.2

```
BregMeans++($P, k$):
  $P$   set of input points
  $k$   number of medians to be found

1: Choose an initial point $a_1$ uniformly at random from $P$.
2: Let $A$ be the set of points already chosen from $P$. Then element
   $p \in P$ is chosen with probability $\frac{\mathrm{D}_\varphi(p,A)}{\mathrm{cost}(P,A)}$ as next element of $A$.
3: Repeat step 2 until $A$ contains $k$ points.
4: Output set $A = \{a_1, a_2, \ldots, a_k\}$.
```

Figure 5.2: Algorithm BregMeans++ for a Bregman divergence D_φ.

below we prove that the approach is also applicable to μ-similar Bregman k-median clusterings. The following theorem is a generalization of Theorem 3.1 from [Arthur and Vassilvitskii, 2007].

Theorem 5.3. *If D_φ is a μ-similar Bregman divergence and $A \subseteq \mathbb{X}$ with $|A| = k$ is chosen at random according to D_φ, then we have*

$$\mathrm{E}\left[\mathrm{cost}(P,A)\right] \leq \frac{8}{\mu^2}(2 + \ln k)\, opt_k(P) \ . \tag{5.2}$$

From Markov's inequality it follows that with probability at least $1 - \delta$, algorithm BregMeans++ yields a factor $\frac{8}{\delta\mu^2}(2 + \ln k)$ approximation of $opt_k(P)$. For the iterative sampling of A we just have to store the distances from each $p \in P$ to their closest center in A. Since this information can be updated after each new element of A is chosen using at most $\mathcal{O}(n)$ arithmetic operations (including evaluations of D_φ), set A can be obtained using at most $\mathcal{O}(kn)$ arithmetic operations. Hence, Theorem 5.1 follows.

Note that application of a constant number of Lloyd iterations can improve the quality of the solution computed by algorithm BregMeans++ significantly. In fact, a small number of, say, 10 iterations do quite well in practice (cf. [Arthur and Vassilvitskii, 2007] and the empirical results therein). However, no approximation guarantee for this improvement is known, and the theoretically provable approximation factor of $\mathcal{O}(\log k)$ already applies to the solution computed by the seeding step.

5.1.2 Proof of Theorem 5.3

This proof is a straightforward generalization of the proof of Theorem 3.1 from [Arthur and Vassilvitskii, 2007]. That is, first we show in Lemma 5.4

that the optimal median of the cluster from which the first point a_1 is taken from is well approximated by this first uniformly chosen center point a_1. Then, we analyze the iterative choice of points at random according to D_φ. Let us call the optimal clusters from which a center point a_i has been chosen *considered* and the clusters from which no point has been chosen *unconsidered*. We prove that both the optimal clusters that have been considered (Lemma 5.5), as well as the clusters that have not been considered (Lemma 5.6), are well approximated by the sampled points.

In the sequel, let $P_1, P_2, \ldots, P_k$ denote the clusters of an optimal k-median clustering of P and let $C = \{c_1, c_2, \ldots, c_k\}$ be the corresponding optimal k-medians, i.e. $\mathrm{cost}(P, C) = opt_k(P)$ and $\mathrm{cost}(P_i, c_i) = opt_1(P_i)$ for all $i = 1, 2, \ldots, k$.

First, we show that, in expectation, the first, uniformly chosen point a_1 is a good approximate median for its optimal cluster.

Lemma 5.4. *Let $a \in P$ be chosen uniformly at random. For $i = 1, 2, \ldots, k$ we have*

$$\mathrm{E}\left[\mathrm{cost}(P_i, a) \,|\, a \in P_i\right] \leq \left(1 + \frac{1}{\mu}\right) opt_1(P_i) \ . \tag{5.3}$$

Proof. Note that for any index i we have $\Pr[a = q \wedge a \in P_i] = 1/|P|$ for all $q \in P_i$, and $\Pr[a = q \wedge a \in P_i] = 0$ for all $q \notin P_i$. Furthermore, we have $\Pr[a \in P_i] = |P_i|/|P|$. Therefore, for all $i = 1, 2, \ldots, k$ and all $q \in P$ we have

$$\Pr\left[a = q \,|\, a \in P_i\right] = \frac{\Pr\left[a = q \wedge a \in P_i\right]}{\Pr\left[a \in P_i\right]} = \begin{cases} \frac{1}{|P_i|} & \text{if } q \in P_i \\ 0 & \text{if } q \notin P_i \end{cases} \tag{5.4}$$

Hence, using the central identity of Lemma 3.5 we have

$$\begin{aligned} \mathrm{E}\left[\mathrm{cost}(P_i, a) \,|\, a \in P_i\right] &= \sum_{q \in P} \Pr\left[a = q \,|\, a \in P_i\right] \mathrm{cost}(P_i, q) && (5.5) \\ &= \frac{1}{|P_i|} \sum_{q \in P_i} \mathrm{cost}(P_i, q) && (5.6) \\ &= \frac{1}{|P_i|} \sum_{q \in P_i} \big(\mathrm{cost}(P_i, c_i) + |P_i| \, \mathrm{D}_\varphi(c_i, q)\big) && (5.7) \\ &= \mathrm{cost}(P_i, c_i) + \sum_{q \in P_i} \mathrm{D}_\varphi(c_i, q) \ . && (5.8) \end{aligned}$$

Using the approximate symmetry of Lemma 2.18 we conclude

$$\mathrm{E}\left[\mathrm{cost}(P_i, a) \,|\, a \in P_i\right] \leq \mathrm{cost}(P_i, c_i) + \frac{1}{\mu}\,\mathrm{cost}(P_i, c_i) = \left(1 + \frac{1}{\mu}\right) opt_1(P_i) \ . \tag{5.9}$$

□

Next, we show that, in expectation, the set of sampled points yields a constant factor approximation to the cost of the optimal 1-median of any considered cluster P_i.

Lemma 5.5. *Let $B \subseteq P$ be an arbitrary nonempty set. Let $a \in P$ be a point added to B at random according to D_φ. Then, for $i = 1, 2, \ldots, k$ we have*

$$\mathrm{E}\left[\mathrm{cost}(P_i, B \cup \{a\}) \,|\, a \in P_i\right] \leq \frac{8}{\mu^2}\, opt_1(P_i) \ . \tag{5.10}$$

Proof. First, note that for any index i we have $\Pr\left[a = q \wedge a \in P_i\right] = \frac{\mathrm{D}_\varphi(q,B)}{\mathrm{cost}(P,B)}$ for all $q \in P_i$, and $\Pr[a = q \wedge a \in P_i] = 0$ for all $q \notin P_i$. Furthermore, we have $\Pr\left[a \in P_i\right] = \frac{\mathrm{cost}(P_i,B)}{\mathrm{cost}(P,B)}$. Therefore, for all $i = 1, 2, \ldots, k$ and all $q \in P$ we have

$$\Pr\left[a = q \,|\, a \in P_i\right] = \frac{\Pr\left[a = q \wedge a \in P_i\right]}{\Pr\left[a \in P_i\right]} = \begin{cases} \frac{\mathrm{D}_\varphi(q,B)}{\mathrm{cost}(P_i,B)} & \text{if } q \in P_i \\ 0 & \text{if } q \notin P_i \end{cases} \tag{5.11}$$

Hence, we have

$$\mathrm{E}\left[\mathrm{cost}(P_i, B \cup \{a\}) \,|\, a \in P_i\right] = \sum_{q \in P} \Pr\left[a = q \,|\, a \in P_i\right] \mathrm{cost}(P_i, B \cup \{q\}) \tag{5.12}$$

$$= \sum_{q \in P_i} \frac{\mathrm{D}_\varphi(q, B)}{\mathrm{cost}(P_i, B)}\, \mathrm{cost}(P_i, B \cup \{q\}) \ . \tag{5.13}$$

Fix any $q \in P_i$. For $p \in P_i$ let $b_p \in B$ denote the closest point to p within B, that is, $\mathrm{D}_\varphi(p, b_p) = \mathrm{D}_\varphi(p, B)$. By the approximate triangle inequality of Lemma 2.18 we know that for all $p \in P_i$ we have

$$\mathrm{D}_\varphi(q, B) \leq \mathrm{D}_\varphi(q, b_p) \tag{5.14}$$

$$\leq \frac{2}{\mu}\,\mathrm{D}_\varphi(p, q) + \frac{2}{\mu}\,\mathrm{D}_\varphi(p, b_p) \tag{5.15}$$

$$= \frac{2}{\mu}\,\mathrm{D}_\varphi(p, q) + \frac{2}{\mu}\,\mathrm{D}_\varphi(p, B) \ . \tag{5.16}$$

Thus, summing up over all $p \in P_i$ leads to

$$|P_i|\,\mathrm{D}_\varphi(q,B) \leq \sum_{p\in P_i}\left(\frac{2}{\mu}\,\mathrm{D}_\varphi(p,q) + \frac{2}{\mu}\,\mathrm{D}_\varphi(p,B)\right) \tag{5.17}$$

$$= \frac{2}{\mu}\,\mathrm{cost}(P_i,q) + \frac{2}{\mu}\,\mathrm{cost}(P_i,B)\ . \tag{5.18}$$

Using (5.13) and (5.18), we obtain

$$\mathrm{E}\big[\mathrm{cost}\big(P_i, B\cup\{a\}\big)\,\big|\,a\in P_i\big]$$

$$= \frac{1}{|P_i|}\sum_{q\in P_i}\frac{|P_i|\,\mathrm{D}_\varphi(q,B)}{\mathrm{cost}(P_i,B)}\,\mathrm{cost}(P_i,B\cup\{q\}) \tag{5.19}$$

$$\leq \frac{2}{\mu|P_i|}\sum_{q\in P_i}\frac{\mathrm{cost}(P_i,q)}{\mathrm{cost}(P_i,B)}\,\mathrm{cost}(P_i,B\cup\{q\})$$

$$+ \frac{2}{\mu|P_i|}\sum_{q\in P_i}\frac{\mathrm{cost}(P_i,B)}{\mathrm{cost}(P_i,B)}\,\mathrm{cost}(P_i,B\cup\{q\}) \tag{5.20}$$

Now, observe that

$$\mathrm{cost}(P_i,B\cup\{q\}) \leq \mathrm{cost}(P_i,B)\ , \tag{5.21}$$

as well as

$$\mathrm{cost}(P_i,B\cup\{q\}) \leq \mathrm{cost}(P_i,q)\ . \tag{5.22}$$

Using bound (5.21) on the left hand side of the sum in inequality (5.20) and bound (5.22) on the right hand side of the sum in (5.20) we get

$$\mathrm{E}\big[\mathrm{cost}\big(P_i, B\cup\{a\}\big)\,\big|\,a\in P_i\big] \leq \frac{2}{\mu|P_i|}\sum_{q\in P_i}\frac{\mathrm{cost}(P_i,q)}{\mathrm{cost}(P_i,B)}\,\mathrm{cost}(P_i,B)$$

$$+ \frac{2}{\mu|P_i|}\sum_{q\in P_i}\frac{\mathrm{cost}(P_i,B)}{\mathrm{cost}(P_i,B)}\,\mathrm{cost}(P_i,q) \tag{5.23}$$

$$= \frac{4}{\mu}\left(\frac{1}{|P_i|}\sum_{q\in P_i}\mathrm{cost}(P_i,q)\right) \tag{5.24}$$

$$\leq \frac{4}{\mu}\left(1+\frac{1}{\mu}\right)\mathit{opt}_1(P_i)\ . \tag{5.25}$$

Here inequality (5.25) is due to equation (5.6) and Lemma 5.4. Using $\frac{4}{\mu}\big(1+\frac{1}{\mu}\big) \leq \frac{8}{\mu^2}$ for $\mu \leq 1$ concludes the proof. □

We still have to give a bound on the cost in the case that after k points have been sampled according to D_φ some optimal clusters remain unconsidered. This bound is given in Lemma 5.6 below. We use an instance of Lemma 5.6 to proof Theorem 5.3. To this end, consider set $B = \{a_1\}$ of the first point chosen uniformly at random from P. Assume $a_1 \in P_i$. Using Lemma 5.4, in expectation we have

$$\operatorname{cost}(P_i, a_1) \leq \left(1 + \frac{1}{\mu}\right) opt_1(P_i) \leq \frac{8}{\mu^2}\, opt_1(P_i)\ . \tag{5.26}$$

Let $A = \{a_1, a_2, \ldots, a_k\}$ where A has been constructed by adding points $a_2, \ldots, a_k$ to B iteratively at random according to D_φ. For $u = t = k - 1$ we define $P^c = P_i$ as the only cluster considered by B, and $P^u = P \setminus P_i$ as the union of all unconsidered clusters. Then, Lemma 5.6 yields

$$\mathrm{E}\left[\operatorname{cost}(P, A) \,|\, a_1 \in P_i\right] \leq (1 + H_{k-1})\Bigg(\operatorname{cost}(P_i, a_1) + \frac{8}{\mu^2}\, opt_{k-1}(P \setminus P_i)\Bigg) \tag{5.27}$$

$$\leq \frac{8}{\mu^2}(1 + H_{k-1})\Bigg(opt_1(P_i) + opt_{k-1}(P \setminus P_i)\Bigg) \tag{5.28}$$

$$= \frac{8}{\mu^2}(1 + H_{k-1})\, opt_k(P)\ , \tag{5.29}$$

where H_{k-1} denotes the $(k-1)$-th harmonic number. Using

$$\mathrm{E}\left[\operatorname{cost}(P, A)\right] = \sum_{i=1}^{k} \frac{|P_i|}{|P|}\, \mathrm{E}\left[\operatorname{cost}(P, A) \,|\, a_1 \in P_i\right] \leq \frac{8}{\mu^2}(1 + H_{k-1})\, opt_k(P) \tag{5.30}$$

and the well-known fact that $H_{k-1} \leq 1 + \ln k$, Theorem 5.3 follows.

Lemma 5.6. *Let $u \in \mathbb{N}$ with $0 < u < k$ and let $t \in \mathbb{N}_0$ with $t \leq u$. Let P^u be the union of any u different clusters of the optimal k-clustering of P, and let $P^c = P \setminus P^u$. Let $B \subseteq P^c$ be an arbitrary non-empty set of points, and let $A = B \cup \{a_1, a_2 \ldots, a_t\}$ where A is constructed by adding points $a_1, a_2 \ldots, a_t \in P$ to B iteratively at random according to D_φ. Then*

$$\mathrm{E}\Big[\operatorname{cost}(P, A)\Big] \leq (1 + H_t)\Bigg(\operatorname{cost}(P^c, B) + \frac{8}{\mu^2}\, opt_u(P^u)\Bigg) + \frac{u - t}{u} \operatorname{cost}(P^u, B) \tag{5.31}$$

where $H_t = \sum_{i=1}^{t} \frac{1}{i}$ denotes the t-th harmonic number.

Proof. This proof is technically analog to the proof of Lemma 3.3 from [Arthur and Vassilvitskii, 2007], using Lemma 5.4 and 5.5 respectively.

We prove the lemma by induction. For $t = 0$ and any $u > 0$ we have $1 + \mathrm{H}_0 = 1$ and $\frac{u-0}{u} = 1$. Hence, we obtain

$$\begin{aligned}
\mathrm{E}\left[\mathrm{cost}(P, A)\right] &= \mathrm{cost}(P, B) && (5.32)\\
&= \mathrm{cost}(P^c, B) + \mathrm{cost}(P^u, B) && (5.33)\\
&\leq \left(\mathrm{cost}(P^c, B) + \frac{8}{\mu^2}\, opt_u(P^u)\right) + \mathrm{cost}(P^u, B) && (5.34)
\end{aligned}$$

where the right hand side of inequality (5.34) is equal to the right-hand side of the inequality (5.31) in the case of $t = 0$.

For $t = 1$ and $u = 1$ a single point a is sampled according to D_φ. With probability $\frac{\mathrm{cost}(P^u,B)}{\mathrm{cost}(P,B)}$ we choose an $a \in P^u$. According to Lemma 5.5, in this case we have

$$\begin{aligned}
\mathrm{E}\left[\mathrm{cost}(P, A) \,|\, a \in P^u\right] &= \mathrm{E}\left[\mathrm{cost}(P^c, A)\right] + \mathrm{E}\left[\mathrm{cost}(P^u, A)\right] && (5.35)\\
&\leq \mathrm{E}\left[\mathrm{cost}(P^c, B)\right] + \mathrm{E}\left[\mathrm{cost}(P^u, a)\right] && (5.36)\\
&\leq \mathrm{cost}(P^c, B) + \frac{8}{\mu^2}\, opt_1(P^u) \ . && (5.37)
\end{aligned}$$

On the other hand, with probability $\frac{\mathrm{cost}(P^c,B)}{\mathrm{cost}(P,B)}$ we have $a \in P^c$. In this case we obtain

$$\mathrm{E}\left[\mathrm{cost}(P, A) \,|\, a \in P^c\right] \leq \mathrm{E}\left[\mathrm{cost}(P, B)\right] = \mathrm{cost}(P, B) \ . \qquad (5.38)$$

Hence, by the law of total expectation we have

$$\begin{aligned}
\mathrm{E}\left[\mathrm{cost}(P, A)\right] &\leq \frac{\mathrm{cost}(P^u, B)}{\mathrm{cost}(P, B)}\left(\mathrm{cost}(P^c, B) + \frac{8}{\mu^2}\, opt_1(P^u)\right)\\
&\qquad + \frac{\mathrm{cost}(P^c, B)}{\mathrm{cost}(P, B)}\,\mathrm{cost}(P, B) && (5.39)\\
&\leq \mathrm{cost}(P^c, B) + \frac{8}{\mu^2}\, opt_1(P^u) + \mathrm{cost}(P^c, B) && (5.40)\\
&= 2\,\mathrm{cost}(P^c, B) + \frac{8}{\mu^2}\, opt_u(P^u) \ . && (5.41)
\end{aligned}$$

Since $1 + \mathrm{H}_1 = 2$ and $\frac{u-t}{u} = 0$ for $u = t = 1$ we find

$$\mathrm{E}\Big[\mathrm{cost}(P, A)\Big] \leq (1 + H_1)\left(\mathrm{cost}(P^c, B) + \frac{8}{\mu^2}\, opt_u(P^u)\right) \ . \qquad (5.42)$$

This concludes the inductive base cases.

For the inductive step, assume that Lemma 5.6 is true for all parameters (u', t') with $u' < u$ or $t' < t$. We show that (5.31) also holds for parameters (u, t). To this end, consider the first point a_1 of the new points added to B by random sampling. We have one of two cases: Either $a_1 \in P^c$ or $a_1 \in P^u$. Again, we have

$$\Pr[a_1 \in P^c] = \frac{\mathrm{cost}(P^c, B)}{\mathrm{cost}(P, B)} \ , \tag{5.43}$$

and

$$\Pr[a_1 \in P^u] = \frac{\mathrm{cost}(P^u, B)}{\mathrm{cost}(P, B)} \ . \tag{5.44}$$

In the sequel, let

$$E^c = \mathrm{E}\left[\mathrm{cost}(P, A) \,|\, a_1 \in P^c\right] \tag{5.45}$$

denote the expectation of $\mathrm{cost}(P, A)$ in case $a_1 \in P^c$, and let

$$E^u = \mathrm{E}\left[\mathrm{cost}(P, A) \,|\, a_1 \in P^u\right] \tag{5.46}$$

denote the same expectation in case $a_1 \in P^u$. Hence, by the law of total expectation we have

$$\mathrm{E}\left[\mathrm{cost}(P, A)\right] = \frac{\mathrm{cost}(P^c, B)}{\mathrm{cost}(P, B)} \cdot E^c + \frac{\mathrm{cost}(P^u, B)}{\mathrm{cost}(P, B)} \cdot E^u \ . \tag{5.47}$$

We give bounds for E^c and E^u separately in Claim 5.7 and 5.8 below.

Claim 5.7. $E^c \leq (1+H_{t-1})\Big(\mathrm{cost}(P^c, B)+\frac{8}{\mu^2}\, opt_u(P^u)\Big)+\frac{u-t+1}{u}\,\mathrm{cost}(P^u, B)$.

Proof. Define $B' = B \cup \{a_1\}$ and $A' = B' \cup \{a_2, \ldots, a_t\} = A$. Since $a_1 \in P^c$ we have $B' \subseteq P^c$, and we find that B', A' give an instance of Lemma 5.6 with parameters $(u, t-1)$. Thus, by induction hypothesis we have

$$E^c \leq (1 + H_{t-1})\left(\mathrm{cost}(P^c, B') + \frac{8}{\mu^2}\, opt_u(P^u)\right) + \frac{u-t+1}{u}\,\mathrm{cost}(P^u, B') \tag{5.48}$$

$$\leq (1 + H_{t-1})\left(\mathrm{cost}(P^c, B) + \frac{8}{\mu^2}\, opt_u(P^u)\right) + \frac{u-t+1}{u}\,\mathrm{cost}(P^u, B) \ . \tag{5.49}$$

□

Claim 5.8. $E^u \leq (1+H_{t-1})\left(\text{cost}(P^c,B)+\frac{8}{\mu^2}\, opt_u(P^u)\right)+\frac{u-t}{u}\,\text{cost}(P^u,B)$.

Proof. Without loss of generality, assume $P^u = \bigcup_{i=1}^{u} P_i$. Since $a_1 \in P^u$ we know that for each $i = 1, 2, \ldots, u$ we have

$$\Pr[a_1 \in P_i] = \frac{\text{cost}(P_i, B)}{\text{cost}(P^u, B)}\ . \tag{5.50}$$

For each $i = 1, \ldots, u$ let

$$E_i^u = \mathrm{E}\left[\text{cost}(p, A) \,|\, a_1 \in P_i\right] \tag{5.51}$$

denote the expectation of $\text{cost}(P, A)$ in case $a_1 \in P_i$. Therefore, using the law of total expectation we have

$$E^u = \sum_{i=1}^{u} \frac{\text{cost}(P_i, B)}{\text{cost}(P^u, B)} \cdot E_i^u\ . \tag{5.52}$$

Assume that index i with $a_1 \in P_i$ is fixed. We will derive a bound on each E_i^u individually. Define $B' = B \cup \{a_1\}$ and $A' = B' \cup \{a_2, \ldots, a_t\}$. Furthermore, let $P^{u'} = P^u \setminus P_i$ and $P^{c'} = P^c \cup P_i$. Note that $B' \subseteq P^{c'}$. Then $B', A', P^{u'}, P^{c'}$ give an instance of Lemma 5.6 with parameters $(u-1, t-1)$. Hence, in the case of $a_1 \in P_i$, by using the induction hypothesis for $(u-1, t-1)$ we obtain

$$E_i^u \leq (1 + H_{t-1})\left(\text{cost}(P^{c'}, B') + \frac{8}{\mu^2}\, opt_{u-1}(P^{u'})\right) + \frac{u-t}{u-1}\,\text{cost}(P^{u'}, B')\ . \tag{5.53}$$

Furthermore, we know that

$$\begin{aligned} \text{cost}(P^{c'}, B') &= \text{cost}(P^c, B') + \text{cost}(P_i, B') && (5.54)\\ &\leq \text{cost}(P^c, B) + \text{cost}(P_i, B')\ , && (5.55) \end{aligned}$$

$$\begin{aligned} \text{cost}(P^{u'}, B') &\leq \text{cost}(P^{u'}, B) && (5.56)\\ &= \text{cost}(P^u, B) - \text{cost}(P_i, B)\ , && (5.57) \end{aligned}$$

and

$$\begin{aligned} opt_{u-1}(P^{u'}) &= \sum_{\substack{j\le u,\\ j\ne i}} \operatorname{cost}(P_j, c_j) && (5.58)\\ &= \sum_{j\le u} \operatorname{cost}(P_j, c_j) - \operatorname{cost}(P_i, c_i) && (5.59)\\ &= opt_u(P^u) - opt_1(P_i)\ . && (5.60)\end{aligned}$$

Hence, for any fixed index i we obtain

$$\begin{aligned} E_i^u &\le (1+H_{t-1})\Big(\operatorname{cost}(P^c, B) + \operatorname{cost}(P_i, B') + \frac{8}{\mu^2}\, opt_u(P^u) - \frac{8}{\mu^2}\, opt_1(P_i)\Big)\\ &\quad + \frac{u-t}{u-1}\big(\operatorname{cost}(P^u, B) - \operatorname{cost}(P_i, B)\big) && (5.61)\\ &\le (1+H_{t-1})\Big(\operatorname{cost}(P^c, B) + \frac{8}{\mu^2}\, opt_u(P^u)\Big)\\ &\quad + \frac{u-t}{u-1}\big(\operatorname{cost}(P^u, B) - \operatorname{cost}(P_i, B)\big) && (5.62)\end{aligned}$$

since by Lemma 5.5 we have $\operatorname{cost}(P_i, B \cup \{a_1\}) \le \frac{8}{\mu^2}\, opt_1(P_i)$. Using inequality (5.52) and inequality (5.62) we obtain

$$\begin{aligned} E^u &\le (1+H_{t-1}) \sum_{i=1}^{u} \frac{\operatorname{cost}(P_i, B)}{\operatorname{cost}(P^u, B)} \Big(\operatorname{cost}(P^c, B) + \frac{8}{\mu^2}\, opt_u(P^u)\Big)\\ &\quad + \frac{u-t}{u-1} \sum_{i=1}^{u} \frac{\operatorname{cost}(P_i, B)}{\operatorname{cost}(P^u, B)} \big(\operatorname{cost}(P^u, B) - \operatorname{cost}(P_i, B)\big) && (5.63)\\ &= (1+H_{t-1})\Big(\operatorname{cost}(P^c, B) + \frac{8}{\mu^2}\, opt_u(P^u)\Big)\\ &\quad + \frac{u-t}{(u-1)\operatorname{cost}(P^u, B)}\Big(\operatorname{cost}(P^u, B)^2 - \sum_{i=1}^{u} \operatorname{cost}(P_i, B)^2\Big) && (5.64)\end{aligned}$$

Let us concentrate on the right hand side of the sum in (5.64). By Chebyshev's sum inequality we have

$$\sum_{i=1}^{u} \operatorname{cost}(P_i, B)^2 \ge \frac{1}{u}\left(\sum_{i=1}^{u} \operatorname{cost}(P_i, B)\right)^2 = \frac{1}{u} \operatorname{cost}(P^u, B)^2 \qquad (5.65)$$

and

$$\mathrm{cost}(P^u, B)^2 - \sum_{i=1}^{u} \mathrm{cost}(P_i, B)^2 \leq \frac{u-1}{u} \mathrm{cost}(P^u, B)^2 \ . \tag{5.66}$$

Therefore, using (5.64) and (5.66) we obtain

$$E^u \leq (1 + H_{t-1}) \left(\mathrm{cost}(P^c, B) + \frac{8}{\mu^2}\, opt_u(P^u) \right) + \frac{u-t}{u} \mathrm{cost}(P^u, B) \tag{5.67}$$

□

Hence, using inequality (5.47) together with Claim 5.7 and 5.8 we get

$$\begin{aligned} \mathrm{E}\left[\mathrm{cost}(P, A)\right] \leq & \frac{\mathrm{cost}(P^c, B)}{\mathrm{cost}(P, B)}(1 + H_{t-1}) \left(\mathrm{cost}(P^c, B) + \frac{8}{\mu^2}\, opt_u(P^u) \right) \\ & + \frac{\mathrm{cost}(P^c, B)}{\mathrm{cost}(P, B)} \cdot \frac{u-t+1}{u} \mathrm{cost}(P^u, B) \\ & + \frac{\mathrm{cost}(P^u, B)}{\mathrm{cost}(P, B)}(1 + H_{t-1}) \left(\mathrm{cost}(P^c, B) + \frac{8}{\mu^2}\, opt_u(P^u) \right) \\ & + \frac{\mathrm{cost}(P^u, B)}{\mathrm{cost}(P, B)} \cdot \frac{u-t}{u} \mathrm{cost}(P^u, B) \ . \end{aligned} \tag{5.68}$$

Note that $\frac{\mathrm{cost}(P^c,B)}{\mathrm{cost}(P,B)} + \frac{\mathrm{cost}(P^u,B)}{\mathrm{cost}(P,B)} = 1$. We obtain

$$\begin{aligned} \mathrm{E}\left[\mathrm{cost}(P, A)\right] \leq & (1 + H_{t-1}) \left(\mathrm{cost}(P^c, B) + \frac{8}{\mu^2}\, opt_u(P^u) \right) \\ & + \frac{1}{u} \mathrm{cost}(P^c, B) + \frac{u-t}{u} \mathrm{cost}(P^u, B) \qquad (5.69) \\ \leq & \left(1 + H_{t-1} + \frac{1}{u} \right) \left(\mathrm{cost}(P^c, B) + \frac{8}{\mu^2}\, opt_u(P^u) \right) \\ & + \frac{u-t}{u} \mathrm{cost}(P^u, B) \qquad (5.70) \\ \leq & (1 + H_t) \left(\mathrm{cost}(P^c, B) + \frac{8}{\mu^2}\, opt_u(P^u) \right) + \frac{u-t}{u} cost(P^u, B) \qquad (5.71) \end{aligned}$$

since $H_{t-1} + \frac{1}{u} \leq H_t$ for $t \leq u$. This concludes the proof. □

5.2 Non-uniform sampling on separable instances

5.2.1 Separable input sets

In Section 5.1, we have shown that algorithm BregMeans++ computes an $\mathcal{O}(\log k)$-approximate solution to the μ-similar Bregman k-median problem. However, it turns out that our algorithm yields even better results for a certain type of input instances that are most commonly used in practice.

When using solutions of the Bregman k-median problem in real-world applications, we implicitly assume that these k-medians provide a meaningful representation of the input data. That is, we expect the input set to consist of k well-separated clusters, and that each of the k-medians distinctively characterizes one of these clusters. If this is not the case then, obviously, a different number of medians should be considered.

Therefore, in this section, we concentrate on the practical relevant case of input instances for which the optimal k-medians indeed give a meaningful representation of the input set. This motivates the notion of *separable input sets*: A k-median input instance is called separable if no clustering of a cost within a constant factor of the optimal k-median cost can be achieved by using only $k-1$ or fewer medians. This represents the case where we have agreed on a smallest number of k such that the k-medians are still a meaningful representation of the input points. Throughout this section, we use the following formal definition.

Definition 5.9. *Let $0 < \alpha < 1$. An input instance $P \subseteq \mathbb{X}$ is called* (k, α)-separable *if and only if*

$$opt_k(P) \leq \alpha \, opt_{k-1}(P) \ . \tag{5.72}$$

Another notion frequently used to describe meaningful k-clustering is the notion of *stable clusterings.* Here a clustering is assumed to be stable if a small perturbation of the input points leads to essentially the same optimal partition of the input set into clusters (i.e., the symmetric difference of perturbed and unperturbed optimal clusters is small). However, in [Ostrovsky et al., 2006] it is shown that the notions of separable inputs and stable clusterings are equivalent.

When restricted to separable instances, we find that with constant probability algorithm BregMeans++ computes a factor $\mathcal{O}(1/\mu)$-approximate solution to the μ-similar Bregman k-median problem. This implies that our algorithm has in fact a better approximation guarantee for the practical relevant case as is suggested by the result from Theorem 5.3. In Section 5.2.2 below we will prove the following theorem.

Theorem 5.10. *Let* D_φ *be a* μ*-similar Bregman divergence and let* $P \subseteq \mathbb{X}$ *be* (k,α)*-separable with* $\alpha \leq \mu/8$*. Furthermore, let* A *with* $|A| = k$ *be chosen at random according to* D_φ*. Then with probability at least* $2^{-\Theta(k)}\mu^k$ *we have*

$$\mathrm{cost}(P,A) \leq \left(1 + \frac{2}{\mu}\right) opt_k(P) \ . \tag{5.73}$$

Theorem 5.2 is an immediate consequence of Theorem 5.10. Moreover, we can compute an $\mathcal{O}(1/\mu)$-approximate solution for a separable input instance P of size $|P| = n$ with arbitrary high probability by running algorithm BregMeans++ $2^{\mathcal{O}(k)}\mu^{-k}$ times independently and choosing the best set of centers obtained this way. This leads to a constant factor approximation algorithm for the Bregman k-median problem using at most $2^{\mathcal{O}(k\log(1/\mu))}n$ arithmetic operations, including evaluations of D_φ.

5.2.2 Proof of Theorem 5.10

Let $P_1, P_2, \ldots, P_k$ denote the clusters of an optimal k-median clustering of P and let $C = \{c_1, c_2, \ldots, c_k\}$ be the corresponding optimal k-medians, i.e. $\mathrm{cost}(P,C) = opt_k(P)$ and $\mathrm{cost}(P_i, c_i) = opt_1(P_i)$ for all $1 \leq i \leq k$. Furthermore, for all $1 \leq i \leq k$ we define

$$X_i = \left\{ x \in P_i \,\middle|\, \mathrm{D}_\varphi(c_i, x) \leq \frac{2}{\mu|P_i|}\, opt_1(P_i) \right\} \ , \tag{5.74}$$

and

$$Y_i = P_i \setminus X_i \ . \tag{5.75}$$

Note that in the definition of X_i the optimal median c_i is used as the first argument of Bregman divergence D_φ. From the central identity of Lemma 3.5 we know that the elements $x \in X_i$ are exactly the points from P_i that are $(1+2/\mu)$-approximate medians of P_i since

$$\mathrm{cost}(P_i, x) = opt_1(P_i) + |P_i|\,\mathrm{D}_\varphi(c_i, x) \leq \left(1 + \frac{2}{\mu}\right) opt_1(P_i) \ . \tag{5.76}$$

Analogously, we know that the elements $y \in Y_i$ are exactly the points from P_i that fail to be $(1+2/\mu)$-approximate medians of P_i since

$$\mathrm{cost}(P_i, y) = opt_1(P_i) + |P_i|\, \mathrm{D}_\varphi(c_i, y) > \left(1 + \frac{2}{\mu}\right) opt_1(P_i)\ . \tag{5.77}$$

Let $A = \{a_1, a_2, \ldots, a_k\}$ be the set of points chosen iteratively at random according to D_φ by algorithm BregMeans++. Our strategy to prove Theorem 5.10 is to show that for separable input instance P with probability at least $2^{-\Theta(k)}\mu^k$, set A consists of one point from each set $X_1, X_2, \ldots, X_k$ and no point from any set Y_i. In that case, assuming $a_i \in X_i$ for all $1 \le i \le k$ we conclude

$$\mathrm{cost}(P, A) \le \sum_{i=1}^{k} \mathrm{cost}(P_i, a_i) \le \left(1 + \frac{2}{\mu}\right) \sum_{i=1}^{k} opt_1(P_i) = \left(1 + \frac{2}{\mu}\right) opt_k(P)\ . \tag{5.78}$$

We start by proving that each set X_i is indeed a large subset of P_i. This observation is an immediate consequence of the fact that μ-similar Bregman divergences satisfy the $[\gamma, \delta]$-sampling property from Chapter 4.

Lemma 5.11. *For all $i = 1, 2, \ldots .k$ we have*

$$|X_i| \ge \frac{1}{2}|P_i| \ge |Y_i|\ . \tag{5.79}$$

Proof. Using Lemma 4.20 with $m = 1$, $\delta = 1/2$, and $\gamma = 2/\mu$ we find that a single sample point $s \in P_i$ chosen uniformly at random from P_i satisfies

$$\Pr\left[\mathrm{cost}(P_i, s) \le \left(1 + \frac{2}{\mu}\right) opt_1(P_i)\right] \ge \frac{1}{2}\ . \tag{5.80}$$

Hence, with probability at least $\frac{1}{2}$ point $s \in P_i$ is a $(1+2/\mu)$-approximate median of P_i, that is, $s \in X_i$. Since s is chosen uniformly at random from P_i we find $|X_i| \ge \frac{1}{2}|P_i|$. We also find $|Y_i| = |P_i| - |X_i| \le \frac{1}{2}|P_i|$. □

Let us consider the first, uniformly chosen point a_1. In the sequel, let $P_{[i,j]}$ denote the disjoint union $\bigcup_{t=i}^{j} P_t$, and let $X_{[i,j]}$ denote the disjoint union $\bigcup_{t=i}^{j} X_t$. Using Lemma 5.11 we immediately obtain the following lemma.

Lemma 5.12.

$$\Pr\left[a_1 \in X_{[1,k]}\right] \geq \frac{1}{2} \tag{5.81}$$

Proof. Since a_1 is chosen uniformly at random from P we have

$$\Pr\left[a_1 \in X_i \,|\, a_1 \in P_i\right] = \frac{|X_i \cap P_i|}{|P_i|} = \frac{|X_i|}{|P_i|} \ . \tag{5.82}$$

Using Lemma 5.11 we obtain

$$\begin{aligned} \Pr\left[a_1 \in X_{[1,k]}\right] &= \sum_{i=1}^{k} \Pr\left[a_1 \in X_i \,|\, a_1 \in P_i\right] \cdot \Pr\left[a_1 \in P_i\right] && (5.83) \\ &\geq \frac{1}{2} \sum_{i=1}^{k} \Pr\left[a_1 \in P_i\right] && (5.84) \\ &= \frac{1}{2} \ . && (5.85) \end{aligned}$$

□

Now, let us assume that we have already sampled set $A_j = \{a_1, a_2, \ldots, a_j\}$ with $1 \leq j \leq k-1$ and $a_i \in X_i$ for all $1 \leq i \leq j$. Our goal is to show that with significant probability the next sampled point a_{j+1} is chosen from $X_{[j+1,k]}$. In a first step towards this goal, the next lemma states that with high probability point a_{j+1} is chosen from $P_{[j+1,k]}$. Intuitively, this result relies on the fact that for separable instances, on average, points from a certain cluster have to be far away from the optimal medians of all other clusters. Hence, because of the a_i being approximate medians for clusters $P_1, P_2, \ldots, P_j$, sampling at random according to D_φ prefers the points from clusters $P_{j+1}, P_{j+2}, \ldots, P_k$.

Lemma 5.13.

$$\Pr\left[a_{j+1} \in P_{[j+1,k]} \,\middle|\, a_1 \in X_1, \ldots, a_j \in X_j\right] \geq 1 - \frac{3\alpha}{\mu} \tag{5.86}$$

Proof. For (k, α)-separable P we have

$$\mathrm{cost}(P, A_j) \geq opt_j(P) \geq \frac{1}{\alpha}\, opt_k(P) = \frac{1}{\alpha} \sum_{i=1}^{k} opt_1(P_i) \ . \tag{5.87}$$

From $a_1 \in X_1, \dots, a_j \in X_j$ we know that for all $1 \le i \le j$ we find

$$\mathrm{cost}(P_i, a_i) \le \left(1 + \frac{2}{\mu}\right) opt_1(P_i) \le \frac{3}{\mu}\, opt_1(P_i) \tag{5.88}$$

since $1 + \frac{2}{\mu} \le \frac{3}{\mu}$ for $\mu \le 1$. Using (5.87) and (5.88) we obtain

$$\begin{aligned} \mathrm{cost}(P, A_j) &\ge \frac{1}{\alpha} \sum_{i=1}^{j} opt_1(P_i) && (5.89)\\ &\ge \frac{\mu}{3\alpha} \sum_{i=1}^{j} \mathrm{cost}(P_i, a_i) && (5.90)\\ &\ge \frac{\mu}{3\alpha}\, \mathrm{cost}(P_{[1,j]}, A_j)\ . && (5.91) \end{aligned}$$

Hence, using inequality (5.91) we conclude

$$\Pr\left[a_{j+1} \notin P_{[j+1,k]} \,\middle|\, a_1 \in X_1, \dots, a_j \in X_j\right] = \frac{\mathrm{cost}(P_{[1,j]}, A_j)}{\mathrm{cost}(P, A_j)} \le \frac{3\alpha}{\mu}\ . \tag{5.92}$$

□

Next, we show that if $a_i \in X_i$ for all $1 \le i \le j$ and we have that $a_{j+1} \in P_{[j+1,k]}$, it follows that with significant probability point a_{j+1} is chosen from $X_{[j+1,k]}$.

Lemma 5.14.

$$\Pr\left[a_{i+1} \in X_{[i+1,k]} \,\middle|\, a_1 \in X_1, \dots, a_i \in X_i, a_{i+1} \in P_{[i+1,k]}\right] \ge \frac{\mu}{5}\left(1 - \frac{4\alpha}{\mu}\right) \tag{5.93}$$

Proof. We start with the observation that for a separable instance P and points $A_j \subseteq X_{[1,j]}$, set A_j is indeed a poor choice as approximate medians for $P_{j+1}, P_{j+2}, \dots, P_k$. More precisely, for (k, α)-separable P from inequalities (5.87) and (5.88) above we know

$$\begin{aligned} \mathrm{cost}(P, A_j) &\ge \frac{1}{\alpha} \sum_{i=1}^{j} \mathrm{cost}(P_i, c_i) + \frac{1}{\alpha} \sum_{i=j+1}^{k} \mathrm{cost}(P_i, c_i) && (5.94)\\ &\ge \frac{\mu}{3\alpha} \sum_{i=1}^{j} \mathrm{cost}(P_i, a_i) + \frac{1}{\alpha} \sum_{i=j+1}^{k} \mathrm{cost}(P_i, c_i)\ . && (5.95) \end{aligned}$$

Using $\alpha \leq \frac{\mu}{8}$ we have $\frac{\mu}{3\alpha} > 1$, and we obtain

$$\begin{aligned} \mathrm{cost}(P, A_j) &\geq \sum_{i=1}^{j} \mathrm{cost}(P_i, a_i) + \frac{1}{\alpha} \sum_{i=j+1}^{k} \mathrm{cost}(P_i, c_i) && (5.96) \\ &\geq \mathrm{cost}(P_{[1,j]}, A_j) + \frac{1}{\alpha}\, opt_{k-j}(P_{[j+1,k]}) \ . && (5.97) \end{aligned}$$

Hence,

$$\mathrm{cost}(P_{[j+1,k]}, A_j) \geq \frac{1}{\alpha}\, opt_{k-j}(P_{[j+1,k]}) \ . \tag{5.98}$$

Now, we make use of bound (5.98) to show that the cost of $X_{[j+1,k]}$ towards A_j is at least a significant fraction of the cost of $P_{[j+1,k]}$ towards A_j. To this end, fix an index $i > j$. Let D_U be a Mahalanobis distance such that

$$\mu\, \mathrm{D}_U(p,q) \leq \mathrm{D}_\varphi(p,q) \leq \mathrm{D}_U(p,q) \tag{5.99}$$

for all $p, q \in \mathbb{X}$. Using the double triangle inequality of D_U (Lemma 2.15) we deduce that for all $x \in X_i$ and all $y \in Y_i$ we have

$$\begin{aligned} \mathrm{D}_\varphi(y, A_j) &\leq \mathrm{D}_\varphi(y, a^*) && (5.100) \\ &\leq \mathrm{D}_U(y, a^*) && (5.101) \\ &\leq 4\big(\mathrm{D}_U(y, c_i) + \mathrm{D}_U(x, c_i) + \mathrm{D}_U(x, a^*)\big) && (5.102) \\ &\leq \frac{4}{\mu}\big(\mathrm{D}_\varphi(y, c_i) + \mathrm{D}_\varphi(x, c_i) + \mathrm{D}_\varphi(x, A_j)\big) \ , && (5.103) \end{aligned}$$

where $a^* = \arg\min_{a \in A_j} \mathrm{D}_\varphi(x, a)$. Furthermore, due to Lemma 5.11 we know that $|X_i| \geq |Y_i|$. Hence, there exists an injective mapping $\sigma : Y_i \to X_i$ such that inequality (5.103) can be applied to each $y \in Y_i$ using a different intermediate point $\sigma(y) \in X_i$. Therefore, by summing up over all $y \in Y_i$ we obtain

$$\begin{aligned} \mathrm{cost}(Y_i, A_j) &\leq \frac{4}{\mu}\Big(\mathrm{cost}(Y_i, c_i) + \mathrm{cost}\big(\sigma(Y_i), c_i\big) + \mathrm{cost}\big(\sigma(Y_i), A_j\big)\Big) && (5.104) \\ &\leq \frac{4}{\mu}\big(\mathrm{cost}(Y_i, c_i) + \mathrm{cost}(X_i, c_i) + \mathrm{cost}(X_i, A_j)\big) && (5.105) \\ &= \frac{4}{\mu}\, opt_1(P_i) + \frac{4}{\mu}\, \mathrm{cost}(X_i, A_j) \ . && (5.106) \end{aligned}$$

Hence,

$$\operatorname{cost}(P_i, A_j) \le \frac{4}{\mu}\, opt_1(P_i) + \left(\frac{4}{\mu} + 1\right) \operatorname{cost}(X_i, A_j) \tag{5.107}$$

$$\le \frac{4}{\mu}\, opt_1(P_i) + \frac{5}{\mu} \operatorname{cost}(X_i, A_j) \tag{5.108}$$

since $\frac{4}{\mu} + 1 \le \frac{5}{\mu}$ for $\mu < 1$. Summing up over all indices $i > j$ and using (5.98) leads to

$$\operatorname{cost}(P_{[j+1,k]}, A_j) \le \frac{4}{\mu}\, opt_{k-j}(P_{[j+1,k]}) + \frac{5}{\mu} \operatorname{cost}(X_{[j+1,k]}, A_j) \tag{5.109}$$

$$\le \frac{4\alpha}{\mu} \operatorname{cost}(P_{[j+1,k]}, A_j) + \frac{5}{\mu} \operatorname{cost}(X_{[j+1,k]}, A_j)\ . \tag{5.110}$$

Thus,

$$\left(1 - \frac{4\alpha}{\mu}\right) \operatorname{cost}(P_{[j+1,k]}, A_j) \le \frac{5}{\mu} \operatorname{cost}(X_{[j+1,k]}, A_j)\ . \tag{5.111}$$

Using inequality (5.111) we conclude

$$\Pr\left[a_{j+1} \in X_{[j+1,k]} \,\middle|\, a_1 \in X_1, \ldots, a_j \in X_j, a_{j+1} \in P_{[j+1,k]}\right]$$
$$= \frac{\operatorname{cost}(X_{[j+1,k]}, A_j)}{\operatorname{cost}(P_{[j+1,k]}, A_j)} \ge \frac{\mu}{5}\left(1 - \frac{4\alpha}{\mu}\right)\ . \tag{5.112}$$

□

Finally, we use Lemmas 5.12 to 5.14, as well as the law of conditional probability, to prove that with probability at least $2^{-\Theta(k)}\mu^k$, set A obtained by sampling according to D_φ consists exactly of one point from each set $X_1, X_2, \ldots, X_k$. Lemma 5.15 together with inequality (5.78) concludes the proof of Theorem 5.10.

Lemma 5.15.

$$\Pr\left[\forall\, 1 \le i \le k :\ A \cap X_i \ne \emptyset\right] \ge \frac{1}{2}\left(\frac{\mu}{20}\right)^{k-1}\ . \tag{5.113}$$

Proof. In the following, let ν_j denote the number of sets X_i that have been considered by the first j sampled points $A_j = \{a_1, a_2, \ldots, a_j\}$, that is,

$$\nu_j = \left|\{i \,|\, A_j \cap X_i \ne \emptyset\}\right|\ . \tag{5.114}$$

We prove this lemma inductively by showing that for all $1 \leq j \leq k$ we have

$$\Pr\big[\nu_j = j\big] \geq \frac{1}{2}\left(\frac{\mu}{20}\right)^{j-1} . \tag{5.115}$$

From Lemma 5.12 we know that with probability at least $\frac{1}{2}$ we have $a_1 \in X_{[1,k]}$. Since the X_i form a disjoint partition of $X_{[1,k]}$, in this case we have $\nu_1 = 1$. This proves the inductive base case of $j = 1$.

Now, assume that inequality (5.115) holds for j with $1 \leq j < k$. That is, with probability at least $\frac{1}{2}\left(\frac{\mu}{20}\right)^{j-1}$ and without loss of generality we may assume $a_i \in X_i$ for all $1 \leq i \leq j$. In this case, by using Lemma 5.13 and Lemma 5.14 we deduce

$$\begin{aligned}
&\Pr\big[\nu_{j+1} = j+1 \,\big|\, \nu_j = j\big] \\
&\quad\geq \Pr\big[a_{j+1} \in X_{[j+1,k]} \,\big|\, a_1 \in X_1, \ldots, a_j \in X_j\big] && (5.116)\\
&\quad\geq \Pr\left[a_{j+1} \in X_{[j+1,k]} \,\big|\, a_1 \in X_1, \ldots, a_j \in X_j, a_{j+1} \in P_{[j+1,k]}\right] && (5.117)\\
&\qquad \cdot \Pr\left[a_{j+1} \in P_{[j+1,k]} \,\big|\, a_1 \in X_1, \ldots, a_j \in X_j\right] && (5.118)\\
&\quad\geq \frac{\mu}{5}\left(1 - \frac{3\alpha}{\mu}\right)\left(1 - \frac{4\alpha}{\mu}\right) && (5.119)\\
&\quad\geq \frac{\mu}{20} && (5.120)
\end{aligned}$$

since $\frac{3\alpha}{\mu} \leq \frac{4\alpha}{\mu} \leq \frac{1}{2}$ for $\alpha \leq \frac{\mu}{8}$. Hence, by using induction hypothesis (5.115) and inequality (5.120) we conclude

$$\begin{aligned}
\Pr\big[\nu_{j+1} = j+1\big] &\geq \Pr\left[\nu_{j+1} = j+1 \,\big|\, \nu_j = j\right] \cdot \Pr\big[\nu_j = j\big] && (5.121)\\
&\geq \frac{1}{2}\left(\frac{\mu}{20}\right)^{j} . && (5.122)
\end{aligned}$$

□

5.3 Discussion

In this chapter we have introduced and analyzed a practical approximation algorithm applicable to the μ-similar Bregman k-median problem. In particular, we have shown that with high probability a generalization of the K-MEANS++ approach from [Arthur and Vassilvitskii, 2007] computes an $\mathcal{O}(\log k)$-approximate solution using at most $\mathcal{O}(kn)$ arithmetic operations. We call this generalization algorithm BREGMEANS++. Moreover, we have shown that with probability at least $2^{-\Theta(k)}$ algorithm BregMeans++ yields a constant factor approximation in the practically relevant case of separable input instances.

The sampling technique presented in this chapter is easy to implement and runs quite fast in practice. In [Arthur and Vassilvitskii, 2007] several experiments on real-world input instances for the Euclidean k-means problem have been conducted. It turns out that seeding according to D_φ outperforms the standard implementation of Lloyd's algorithm using uniform seeding both in terms of speed of convergence and cost of the clustering. In fact, if the data set consists of k well separated clusters, the cost of the clustering is improved by orders of magnitude.

It should be noted that the analysis of the approximation guarantee of algorithm BregMeans++ depends considerably on the μ-similarity of D_φ with constant $\mu > 0$. In particular, in the proof of Theorem 5.3 the approximate metric properties of D_φ are necessary, while the proof of Theorem 5.10 relies on the approximate metric properties as well as the $[\gamma, \delta]$-sampleability of a μ-similar Bregman divergence.

As for Theorem 5.3, the dependency on μ seems to be unavoidable. In [Arthur and Vassilvitskii, 2007] it is shown that in the worst-case of the k-means++ algorithm the expected clustering cost is at least a factor $\Omega(\log k)$ larger than the optimal solution of the Euclidean k-means problem. This stands in contrast to the case of the μ-similar Bregman k-median problem where only a relatively weak lower bound on the expected clustering cost is known. In particular, in [Nock et al., 2008] a lower bound on the expected cost of the first, uniformly sampled point is derived. Conversely to Lemma 5.4 the following result is obtained.

Lemma 5.16 ([Nock et al., 2008], Lemma 6)**.** *Let $a \in P$ be chosen uniformly at random. For $i = 1, 2, \ldots, k$ we have*

$$\mathrm{E}\left[\mathrm{cost}(P_i, a) \,|\, a \in P_i\right] \geq \frac{2}{2-\mu}\, opt_1(P_i) \ . \tag{5.123}$$

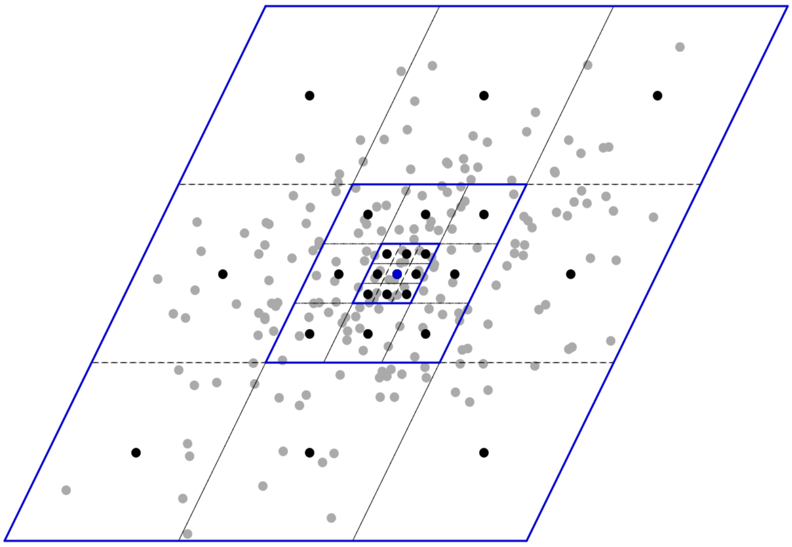
Furthermore, it was shown in [Nock et al., 2008] that the bound from Lemma 5.16 is tight, i.e., there exists a point set such that for a point $a \in P_i$ chosen uniformly at random with high probability $\mathrm{cost}(P_i, a)$ comes arbitrarily close to $\frac{2}{2-\mu}\, opt_1(P_i)$. This matches an earlier observation from [Arthur and Vassilvitskii, 2007] where it has been shown that

$$\mathrm{E}\left[\mathrm{cost}(P_i, a) \,|\, a \in P_i\right] = 2\, opt_1(P_i) \tag{5.124}$$

for the Euclidean k-means problem with $\mu = 1$. However, since the factor of the lower bound approaches 1 for $\mu \to 0$, while the factor of the upper bound from Lemma 5.4 approaches infinity for $\mu \to 0$, this bound is not very meaningful for small μ. Hence, it remains an open problem to prove or disprove the tightness of Theorem 5.3.

As for Theorem 5.10, in contrast to the analysis of [Ostrovsky et al., 2006] our analysis emphasizes the combinatorial structure of the Bregman k-median problem. However, the approximate triangle inequality of μ-similar Bregman divergence D_φ is needed in a single argument in the proof of Lemma 5.14. It remains an open problem to find a proof of the approximation guarantee that relies purely on the combinatorial properties of the Bregman k-median problem.

6 Strong coresets for Mahalanobis distances



Recently, the construction of so called coresets has emerged as a standard technique in computational geometry. Generally speaking, a coreset of a set P is a small (weighted) set S that features the same clustering behavior as the usually much larger original set P. That is, when using the same set of cluster centers, the clustering cost of P and the (weighted) clustering cost of S are approximately the same. Hence, any k center points are as good or as bad as approximate k-medians for coreset S as they are as approximate k-medians for the original set P.

The goal of a coreset construction is to give coresets that are significantly smaller than the original input set. Usually, these coresets are used in two ways. First, a coreset can be used as a smaller input set for any approximation algorithm. Such a preprocessing of the input data can lead to much faster approximation algorithms, especially if the running time of the algorithm of choice depends strongly on the number of input points. Second, coresets play an important role for approximation algorithms in the data streaming model. In the data streaming model, we model the situation when random access to points from P is intractable. For example, this is the case when an input set P is too large to fit into the main memory of a computer. Instead of random access to the points from P, we assume that the input set P is given one point at a time in a single pass. In this context, coresets are used as a dynamic, space-efficient representation of the points seen so far. An example of such a use of coresets is the merge-and-reduce technique from [Har-Peled and Mazumdar, 2004].

Coresets in the context of the Euclidean k-median and k-means problem as well as in the context of metric k-median and k-means problems have been known for some time. Many coreset constructions have been given[1], most notably by [Har-Peled and Kushal, 2005] where the size of the coreset is independent of the size of the input set (but still exponential in dimension d), and by [Chen, 2006] and [Chen, 2009] where the size of the coresets is polylogarithmic in size of the input set and only linear in d. However, until recently, coreset constructions in non-Euclidean and non-metric settings had not been studied. For a survey on coreset methods in computational geometry see [Agarwal et al., 2005].

In this chapter we present two coreset constructions for the Mahalanobis k-median problem. Formally, we will use the following definition of (k, ε)-coresets for the generalized k-median problem.

Definition 6.1. *Let* D *be a dissimilarity measure on domain* $\mathbb{X} \subseteq \mathbb{R}^d$ *and let* $P \subseteq \mathbb{X}$ *be finite. A weighted multiset* $S \subseteq \mathbb{X}$ *with weight function* $w : S \to \mathbb{R}_{\geq 0}$ *such that* $\sum_{p \in S} w(p) = |P|$ *is called a* (k, ε)-coreset *of* P *for the* k*-median problem with respect to* D *if for each* $C \subseteq \mathbb{X}$ *of size* $|C| = k$ *we have*

$$\left|\mathrm{cost}^{\mathrm{D}}(P, C) - \mathrm{cost}_w^{\mathrm{D}}(S, C)\right| \leq \varepsilon \, \mathrm{cost}^{\mathrm{D}}(P, C) \ . \tag{6.1}$$

We also call these coresets *strong coresets* to distinguish them from the relaxed notion of weak coresets which we introduce in Chapter 7.

[1] E.g., [Mishra et al., 2001], [Bădoiu et al., 2002], [Har-Peled and Mazumdar, 2004], [Czumaj and Sohler, 2004], [Frahling and Sohler, 2005].

Our first coreset construction in Section 6.1 is deterministic and gives a generalization of the coresets from [Har-Peled and Mazumdar, 2004] to Mahalanobis distances. The main result of Section 6.1 can be stated as follows.

Theorem 6.2. *Let be $P \subseteq \mathbb{R}^d$ of size n. There exists a (k, ε)-coreset of P of size $2^{\mathcal{O}(d \log d)}\varepsilon^{-d}k \log n$ for the Mahalanobis k-median problem. Furthermore, given a set of medians of a constant factor approximate Mahalanobis k-median clustering of P, such a coreset can be constructed deterministically in time $\mathcal{O}\left(dn \log(k) + d^2 n\right) + 2^{\mathcal{O}(d \log d)}\varepsilon^{-d}k \log(n)$.*

The second construction we give in Section 6.2 is a new randomized construction based on non-uniform sampling. This construction has first been proposed in [Ackermann et al., 2010b] for the Euclidean k-means problem. Here, we give a generalization of this result to the Mahalanobis k-median problem. The main result of Section 6.2 can be summarized as follows.

Theorem 6.3. *Let be $P \subseteq \mathbb{R}^d$ of size n. With high probability, a coreset of size $2^{\mathcal{O}(d \log d)}\varepsilon^{-d}k \log(n) \log^{d/2}\big(\varepsilon^{-1}k \log(n)\big)$ for the Mahalanobis k-median problem can be obtained in time $2^{\mathcal{O}(d \log d)}\varepsilon^{-d}k\, n \log(n) \log^{d/2}\big(\varepsilon^{-1}k \log(n)\big)$ using non-uniform sampling.*

Observe that the deterministic coreset construction from Theorem 6.2 obtains better asymptotic results in terms of coreset size and running time than the randomized construction from Theorem 6.3. However, there are several practical advantages with this randomized construction which will be discussed in Section 6.2. A practical implementation of coresets based on Theorem 6.3 can be given by using a so-called *coreset tree* data structure that enables efficient non-uniform sampling. For a description and an empirical study of this data structure see [Ackermann et al., 2010b].

6.1 Har-Peled-Mazumdar coresets

In this section we present the deterministic coreset construction given by [Har-Peled and Mazumdar, 2004], as originally proposed for the Euclidean k-means problem. We also show how this construction can be generalized to the Mahalanobis k-median problem.

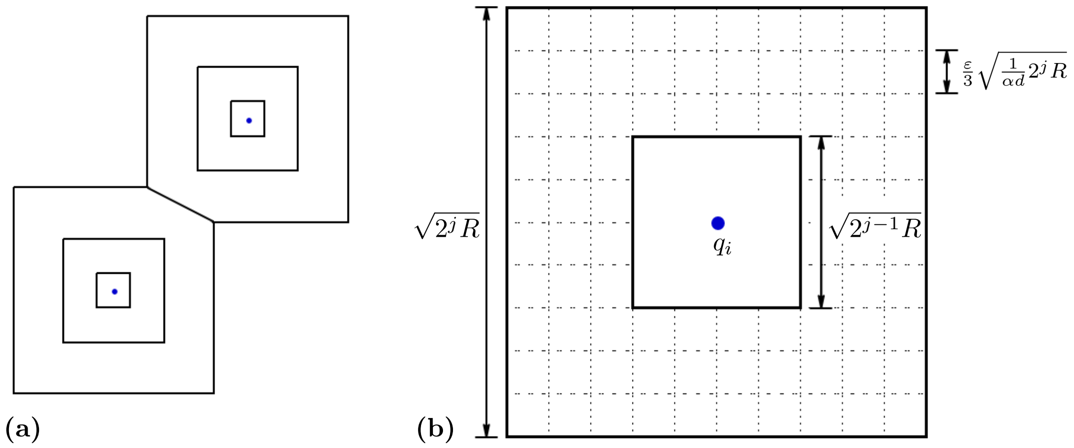


Figure 6.1: An illustration of the deterministic construction of Har-Peled-Mazumdar coresets for the Euclidean k-means problem. **(a)** For each approximate center q_i, the input set P is partitioned by axis-aligned hypercubes of side length $\sqrt{2^j R}$ centered at q_i. **(b)** Then, the sets of this partition are divided into a number of small grid cells of side length $\frac{\varepsilon}{3}\sqrt{\frac{1}{\alpha d}2^j R}$. For each grid cell, one representative point from this cell is added to the coreset with a weight equal to the number of input points from this cell.

6.1.1 Euclidean k-means clustering

In a nutshell, the construction from [Har-Peled and Mazumdar, 2004] for the Euclidean k-means problem is as follows. Using the center points of an arbitrary constant factor approximation, input set P is partitioned by assigning each point to its closest center point. Each subset of the partition is further subdivided using an axis aligned grid of exponentially growing side length. A coreset is obtained by replacing all points from a common grid cell by a single representative from that cell, weighted according to the number of input points within this grid cell. An illustration of this construction is given in Figure 6.1.

We now give the construction in-detail. In the sequel, let input set $P \subseteq \mathbb{R}^d$ be of size $|P| = n$. Let $Q = \{q_1, q_2, \ldots, q_\kappa\}$ be the medians of an $[\alpha, \beta]$-bicriteria approximation for the optimal Euclidean k-means clustering of P, i.e.,

$$\mathrm{cost}^{\ell_2^2}(P, Q) \leq \alpha\, opt_k^{\ell_2^2}(P) \tag{6.2}$$

and

$$\kappa \leq \beta k \tag{6.3}$$

for arbitrary $\alpha, \beta \geq 1$. Note that any fast bicriteria approximation algorithm can be used to obtain the initial constant factor approximate solution Q. For instance, one can use the $[\mathcal{O}(1), 1]$-approximate algorithm from [Har-Peled and Mazumdar, 2004] in time $\mathcal{O}(dn + dk^5 \log^9 n)$, or even our $[\mathcal{O}(\log k), 1]$-approximate algorithm from Chapter 5 in time $\mathcal{O}(dkn)$. Given Q, we show how to construct a coreset S with weight function w.

Let $Q_1, Q_2, \ldots, Q_\kappa$ be the partion of P induced by assigning each $p \in P$ to their closest $q_i \in Q$. That is, $p \in Q_i$ if and only if $q_i = \arg\min_{q \in Q} \|p - q\|$, breaking ties arbitrarily. Furthermore, let

$$R = \frac{1}{\alpha n} \operatorname{cost}^{\ell_2^2}(P, Q)\ . \tag{6.4}$$

Note that $R \leq \frac{1}{n}\, opt_k^{\ell_2^2}(P)$.

For all $i = 1, 2, \ldots, \kappa$ and $j = 1, 2, \ldots, \nu$ where $\nu = \lceil \log(\alpha n) + 2 \rceil$ let $U_{ij} \subseteq \mathbb{R}^d$ denote the axis-parallel cube with side length $\sqrt{2^j R}$ centered at point q_i. Note that each $p \in P$ is contained in at least one cube U_{ij}, since the existence of a point p with $p \notin \bigcup_{i=1}^{\kappa} U_{i\nu}$ leads to

$$\|p - q_i\| > \frac{1}{2}\sqrt{2^\nu R} \geq \frac{1}{2}\sqrt{2^{\log(\alpha n)+2} R} \geq \sqrt{\alpha n R} \tag{6.5}$$

for all $i = 1, 2, \ldots, \kappa$ and, hence,

$$\operatorname{cost}^{\ell_2^2}(P, Q) \geq \min_{i=1,\ldots,\kappa} \|p - q_i\|^2 > \alpha n R = \operatorname{cost}^{\ell_2^2}(P, Q) \tag{6.6}$$

which is a contradiction. Furthermore, let

$$V_{i0} = U_{i0} \tag{6.7}$$

for $i = 1, 2, \ldots, \kappa$ and

$$V_{ij} = U_{ij} \setminus U_{i,j-1} \tag{6.8}$$

for $i = 1, 2, \ldots, \kappa$ and $j = 1, 2, \ldots, \nu$. For each i, j individually, we cover V_{ij} by a grid of side length

$$r_j = \frac{\varepsilon}{3}\sqrt{\frac{1}{\alpha d} 2^j R}\ . \tag{6.9}$$

Note that the number of grid cells necessary to cover V_{ij} is bounded by

$$m \leq \frac{\mathrm{vol}_d(U_{ij})}{r_j^d} = \frac{(2^j R)^{d/2}}{\left(\frac{\varepsilon^2}{9\alpha d} 2^j R\right)^{d/2}} = (9\alpha d)^{d/2} \varepsilon^{-d} \ . \tag{6.10}$$

for all i, j, where $\mathrm{vol}_d(U_{ij})$ denotes the volume of U_{ij} in $\mathbb{R}^d$. For every grid cell that contains points from Q_i we pick an arbitrary point inside the cell as its representative coreset point. Let S_{ij} denote the union of all these representatives. For each point $s \in S_{ij}$ we assign a weight $w(s)$ equal to the number of points from Q_i within its grid cell. Furthermore, let $S = \bigcup_{i,j} S_{ij}$. Since $|S_{ij}| \leq m$ for all i, j we have

$$|S| \leq \kappa \nu m = \mathcal{O}\left(3^d \alpha^{d/2} \beta d^{d/2} \varepsilon^{-d} k \log(\alpha n)\right) \ . \tag{6.11}$$

It has been shown that weighted set S is a (k, ε)-coreset of P, as is summarized in the following theorem. The reader is directed to the original article of Har-Peled and Mazumdar for a formal proof of this theorem.

Theorem 6.4 ([Har-Peled and Mazumdar, 2004], Theorem 3.4)**.** *Given a point set P with $|P| = n$ points, and a point set Q with $|Q| \leq \beta k$ points, such that $\mathrm{cost}^{\ell_2^2}(P, Q) \leq \alpha\, opt_k^{\ell_2^2}(P)$, where $\alpha, \beta \geq 1$. Then, one can compute a (k, ε)-coreset S of size $|S| = \mathcal{O}\left(3^d \alpha^{d/2} \beta d^{d/2} \varepsilon^{-d} k \log(\alpha n)\right)$ for the Euclidean k-means problem. Furthermore, S can be obtained in time $O\big(dn \log(\beta k) + |S|\big)$.*

6.1.2 Mahalanobis k-median clustering

We now show how the construction of [Har-Peled and Mazumdar, 2004] can be used to obtain coresets for the Mahalanobis k-median problem. Recall from Section 2.2.1 that for Mahalanobis distance

$$\mathrm{D}_A(p, q) = (p - q)^\top A\, (p - q) \tag{6.12}$$

with $p, q \in \mathbb{R}^d$ and symmetric positive definite matrix $A \in \mathbb{R}^{d \times d}$ there exists a non-singular matrix $B \in \mathbb{R}^{d \times d}$ with

$$A = B^\top B \tag{6.13}$$

and

$$\mathrm{D}_A(p, q) = \|Bp - Bq\|^2 \ . \tag{6.14}$$

Also note that B can be obtained from A in time $\mathcal{O}(d^3)$ by computing the Cholesky decomposition of A (cf. [Trefethen and Bau, 1997]).

For the remainder of this section, let $x' = Bx$ denote the image of any $x \in \mathbb{R}^d$ under the linear transformation given by B. Furthermore, for any set $P \subseteq \mathbb{R}^d$ we define $P' = \{p' \mid p \in P\}$. Since B is non-singular, the mapping $(\cdot)' : \mathbb{R}^d \to \mathbb{R}^d$ is a bijection and, hence, $|P| = |P'|$. From equation (6.14) we obtain the following lemma.

Lemma 6.5. *Let $P \subset \mathbb{R}^d$ be finite. Then for all $C \subseteq \mathbb{R}^d$ we have*

$$\mathrm{cost}^{D_A}(P, C) = \mathrm{cost}^{\ell_2^2}(P', C') \ . \tag{6.15}$$

In particular,

$$opt_k^{\mathrm{D}_A}(P) = opt_k^{\ell_2^2}(P') \ . \tag{6.16}$$

Proof. We have

$$\begin{aligned}\mathrm{cost}^{D_A}(P, C) &= \sum_{p \in P} \min_{c \in C} \mathrm{D}_A(p, c) &(6.17)\\ &= \sum_{p \in P} \min_{c \in C} \|Bp - Bc\|^2 &(6.18)\\ &= \mathrm{cost}^{\ell_2^2}(P', C') &(6.19)\end{aligned}$$

Furthermore, assume

$$\mathrm{cost}^{D_A}(P, C) = opt_k^{\mathrm{D}_A}(P) < opt_k^{\ell_2^2}(P') \ . \tag{6.20}$$

Then we would have $\mathrm{cost}^{\ell_2^2}(P', C') < opt_k^{\ell_2^2}(P')$, which is a contradiction. Thus, $opt_k^{\mathrm{D}_A}(P) \geq opt_k^{\ell_2^2}(P')$. Analogousely, we find $opt_k^{\mathrm{D}_A}(P) \leq opt_k^{\ell_2^2}(P')$. □

Hence, C is a set of optimal k-medians with respect to D_A for P if and only if C' is a set of optimal k-medians with respect to the squared Euclidean distance for P'. In addition, C is a set of α-approximate k-medians with respect to D_A for P if and only if C' is a set of α-approximate k-medians with respect to the squared Euclidean distance for P'. Also note that the observation from Lemma 6.5 also holds for weighted point sets. In fact, we show the following connection between coresets for the Mahalanobis k-median and the Euclidean k-means problem.

Lemma 6.6. *Weighted multiset S is a (k, ε)-coreset of P with respect to the k-median problem using Mahalanobis distance D_A if and only if weighted multiset S' is a (k, ε)-coreset of P' with respect to the Euclidean k-means problem.*

Proof. Let $C \subseteq \mathbb{R}^d$ be an arbitrary set of size $|C| = k$. Since

$$\mathrm{cost}^{D_A}(P, C) = \mathrm{cost}^{\ell_2^2}(P', C') \tag{6.21}$$

and

$$\mathrm{cost}_w^{D_A}(S, C) = \mathrm{cost}_w^{\ell_2^2}(S', C') \tag{6.22}$$

we have

$$\left|\mathrm{cost}^{D_A}(P, C) - \mathrm{cost}_w^{D_A}(S, C)\right| = \left|\mathrm{cost}^{\ell_2^2}(P', C') - \mathrm{cost}_w^{\ell_2^2}(S', C')\right| . \tag{6.23}$$

Therefore, we have

$$\left|\mathrm{cost}^{D_A}(P, C) - \mathrm{cost}_w^{D_A}(S, C)\right| \leq \varepsilon \, \mathrm{cost}^{D_A}(P, C) \tag{6.24}$$

if and only if we have

$$\left|\mathrm{cost}^{\ell_2^2}(P', C') - \mathrm{cost}_w^{\ell_2^2}(S', C')\right| \leq \varepsilon \, \mathrm{cost}^{\ell_2^2}(P', C') . \tag{6.25}$$

□

Given an input set $P \subseteq \mathbb{R}^d$ of size $|P| = n$ and an $[\alpha, \beta]$-bicriteria approximation $Q \subseteq \mathbb{R}^d$ of the optimal Mahalanobis k-median clustering of P. Then Lemma 6.6 implies that we can obtain a (k, ε)-coreset for the Mahalanobis k-median problem in the following way. First, input set P and initial approximation Q are transformed into sets P' and Q' using at most $\mathcal{O}(d^2 n)$ operations. Then, coreset S' of P' with respect to the squared Euclidean distance is obtained by the construction of Har-Peled and Mazumdar in time $O\big(dn \log(\beta k) + |S'|\big)$ where $|S'| = \mathcal{O}\left(3^d \alpha^{d/2} \beta d^{d/2} \varepsilon^{-d} k \log(\alpha n)\right)$. Finally, S' is tranformed into set S using the inverse mapping B^{-1}. This requires $\mathcal{O}(d^2|S|)$ operations. We obtain the following generalization of Theorem 6.4. Theorem 6.2 is an immediate consequence of Corollary 6.7.

Corollary 6.7. *Let $P \subseteq \mathbb{R}^d$ be of size $|P| = n$. Then S is a (k, ε)-coreset for the Mahalanobis k-median problem of size $\mathcal{O}\left(3^d \alpha^{d/2} \beta d^{d/2} \varepsilon^{-d} k \log(\alpha n)\right)$. Furthermore, S can be obtained in time $O\big(dn \log(\beta k) + d^2 n + d^2 |S|\big)$.*

6.1.3 Properties of Har-Peled-Mazumdar coresets

It is an important observation about Har-Peled-Mazumdar (k, ε)-coresets that the k-median cost of P using S as center points is arbitrarily small when compared to the optimal k-median cost of P. That is,

$$\mathrm{cost}^{\mathrm{D}_A}(P, S) \leq \varepsilon^2 \, opt_k^{\mathrm{D}_A}(P) \ , \tag{6.26}$$

as is proven in detail in Lemma 6.11 below. In fact, inequality (6.26) is sufficient for any set S to be a $\bigl(k, \mathcal{O}(\varepsilon)\bigr)$-coreset for the k-median problem, provided that the dissimilarity measure used is the square of a metric, as is the case for all Mahalanobis distances distance. In the sequel, we show how to use inequality (6.26) to prove that the Har-Peled-Mazumdar construction yields $(k, 6\varepsilon)$-coresets with respect to Mahalanobis distance D_A. We will make use of the observations from this section later in the proof of the new, randomized coreset construction from Section 6.2.

To this end, for all $p' \in P'$ let s'_p denote the representative of p' in S' as given during the coreset construction. Furthermore, let $p \in P$ and $s_p \in S$ denote the corresponding preimages under linear transformation B, and let $C \subseteq \mathbb{R}^d$ be a set of $|C| = k$ arbitrary centers. By the triangle inequality of the reals we have

$$\begin{aligned}
&\Bigl|\mathrm{cost}^{\mathrm{D}_A}(P, C) \ - \ \mathrm{cost}_w^{\mathrm{D}_A}(S, C)\Bigr| \\
&\qquad\qquad = \left|\sum_{p \in P} \mathrm{D}_A(p, C) - \sum_{s \in S} w(s) \, \mathrm{D}_A(s, C)\right| \qquad (6.27)\\
&\qquad\qquad \leq \sum_{p \in P} \bigl|\mathrm{D}_A(p, C) - \mathrm{D}_A(s_p, C)\bigr| \ . \qquad (6.28)
\end{aligned}$$

Now, consider a partition $P = P_{\mathrm{near}} \cup P_{\mathrm{far}}$, where

$$P_{\mathrm{near}} = \bigl\{p \in P \bigm| \mathrm{D}_A(p, s_p) \leq \varepsilon^2 \, \mathrm{D}_A(p, C)\bigr\} \tag{6.29}$$

consists of all $p \in P$ that are very close to their representative s_p, while

$$P_{\mathrm{far}} = \bigl\{p \in P \bigm| \mathrm{D}_A(p, s_p) > \varepsilon^2 \, \mathrm{D}_A(p, C)\bigr\} \tag{6.30}$$

consists of the points where the distance between p and s_p is at least a constant fraction of $\mathrm{D}_A(p, C)$. We derive bounds on $\bigl|\mathrm{D}_A(p, C) - \mathrm{D}_A(s_p, C)\bigr|$ for $p \in P_{\mathrm{near}}$ and $p \in P_{\mathrm{far}}$ individually. These bounds are given in Lemma

6.9 and Lemma 6.10 below. Using Lemma 6.9 and 6.10 we obtain

$$\begin{aligned}
&\left|\operatorname{cost}^{\mathrm{D}_A}(P,C) - \operatorname{cost}_w^{\mathrm{D}_A}(S,C)\right| \\
&\quad\leq \sum_{p\in P_{\mathrm{near}}} \left|\mathrm{D}_A(p,C) - \mathrm{D}_A(s_p,C)\right| + \sum_{p\in P_{\mathrm{far}}} \left|\mathrm{D}_A(p,C) - \mathrm{D}_A(s_p,C)\right| && (6.31)\\
&\quad\leq 3\varepsilon \sum_{p\in P_{\mathrm{near}}} \mathrm{D}_A(p,C) + \frac{3}{\varepsilon}\sum_{p\in P_{\mathrm{far}}} \mathrm{D}_A(p,s_p) && (6.32)\\
&\quad\leq 3\varepsilon \operatorname{cost}^{\mathrm{D}_A}(P,C) + \frac{3}{\varepsilon}\operatorname{cost}^{\mathrm{D}_A}(P,S)\ . && (6.33)
\end{aligned}$$

Hence, we just have to give a bound on $\frac{3}{\varepsilon}\operatorname{cost}^{\mathrm{D}_A}(P,S)$, that is, a bound on the clustering cost of P using the coreset points from S as centers. This bound, of course, is provided by inequality (6.26) (and proven in detail in Lemma 6.11 below). We obtain

$$\left|\operatorname{cost}^{\mathrm{D}_A}(P,C) - \operatorname{cost}_w^{\mathrm{D}_A}(S,C)\right| \leq 6\varepsilon \operatorname{cost}^{\mathrm{D}_A}(P,C)\ . \qquad (6.34)$$

In the remainder of this section, we give proof to the lemmas from the argumentation above. We start by proving Lemma 6.9 and Lemma 6.10 which give bounds on $\left|\mathrm{D}_A(p,C) - \mathrm{D}_A(s_p,C)\right|$ for $p \in P_{\mathrm{near}}$ and $p \in P_{\mathrm{far}}$ individually. To prove both lemmas, we make use of the following observation.

Claim 6.8. *For all $p \in P$ we have*

$$\left|\mathrm{D}_A(p,C) - \mathrm{D}_A(s_p,C)\right| \leq \mathrm{D}_A(p,s_p) + 2\sqrt{\mathrm{D}_A(p,s_p)\cdot \mathrm{D}_A(p,C)}\ . \qquad (6.35)$$

Proof. Let c_p be the center from C closest to p, and let c_s be the center from C closest to s_p, i.e., $\mathrm{D}_A(p,C) = \mathrm{D}_A(p,c_p)$ and $\mathrm{D}_A(s_p,C) = \mathrm{D}_A(s_p,c_s)$. Note that we have

$$\begin{aligned}
&\left|\,\mathrm{D}_A(p,c_p) - \mathrm{D}_A(s_p,c_s)\right| \\
&\quad= \left|\sqrt{\mathrm{D}_A(p,c_p)} - \sqrt{\mathrm{D}_A(s_p,c_s)}\right| \cdot \left(\sqrt{\mathrm{D}_A(p,c_p)} + \sqrt{\mathrm{D}_A(s_p,c_s)}\right) && (6.36)
\end{aligned}$$

From Lemma 2.14 we know that the (positive) square root $\sqrt{\mathrm{D}_A(\cdot,\cdot)}$ of a Mahalanobis distance D_A is a metric, i.e, $\sqrt{\mathrm{D}_A(\cdot,\cdot)}$ is symmetric and does obey the triangle inequality. Hence, considering the first factor of (6.36), on one hand we have

$$\sqrt{\mathrm{D}_A(p,c_p)} \leq \sqrt{\mathrm{D}_A(p,c_s)} \leq \sqrt{\mathrm{D}_A(p,s_p)} + \sqrt{\mathrm{D}_A(s_p,c_s)}\ . \qquad (6.37)$$

This leads to

$$\sqrt{\mathrm{D}_A(p, c_p)} - \sqrt{\mathrm{D}_A(s_p, c_s)} \leq \sqrt{\mathrm{D}_A(p, s_p)}\ . \tag{6.38}$$

On the other hand, we have

$$\sqrt{\mathrm{D}_A(s_p, c_s)} \leq \sqrt{\mathrm{D}_A(s_p, c_p)} \leq \sqrt{\mathrm{D}_A(p, s_p)} + \sqrt{\mathrm{D}_A(p, c_p)}\ . \tag{6.39}$$

This leads to

$$\sqrt{\mathrm{D}_A(s_p, c_s)} - \sqrt{\mathrm{D}_A(p, c_p)} \leq \sqrt{\mathrm{D}_A(p, s_p)}\ . \tag{6.40}$$

Hence, using inequalities (6.38) and (6.40), we obtain as bound of the first factor of equation (6.36)

$$\left|\sqrt{\mathrm{D}_A(p, c_p)} - \sqrt{\mathrm{D}_A(s_p, c_s)}\right| \leq \sqrt{\mathrm{D}_A(p, s_p)}\ . \tag{6.41}$$

For the second factor of (6.36), using inequality (6.39) we find

$$\sqrt{\mathrm{D}_A(p, c_p)} + \sqrt{\mathrm{D}_A(s_p, c_s)} \leq 2\sqrt{\mathrm{D}_A(p, c_p)} + \sqrt{\mathrm{D}_A(p, s_p)}\ . \tag{6.42}$$

Using equation (6.36) in combination with the bounds from (6.41) and (6.42), we conclude

$$\begin{aligned} &|\,\mathrm{D}_A(p, C) - \mathrm{D}_A(s_p, C)| \\ &\qquad \leq \sqrt{\mathrm{D}_A(p, s_p)} \cdot \left(2\sqrt{\mathrm{D}_A(p, C)} + \sqrt{\mathrm{D}_A(p, s_p)}\right) && (6.43) \\ &\qquad = \mathrm{D}_A(p, s_p) + 2\sqrt{\mathrm{D}_A(p, s_p) \cdot \mathrm{D}_A(p, C)}\ . && (6.44) \end{aligned}$$

□

Lemma 6.9. *If* $\mathrm{D}_A(p, s_p) \leq \varepsilon^2\, \mathrm{D}_A(p, C)$ *then*

$$|\mathrm{D}(p, C) - \mathrm{D}(s_p, C)| \leq 3\varepsilon\, \mathrm{D}(p, C)\ . \tag{6.45}$$

Proof. Using Claim 6.8 and $\mathrm{D}_A(p, s_p) \leq \varepsilon^2\, \mathrm{D}_A(p, C)$ we have

$$\begin{aligned} |\mathrm{D}_A(p, C) - \mathrm{D}_A(s_p, C)| &\leq \mathrm{D}_A(p, s_p) + 2\sqrt{\mathrm{D}_A(p, s_p) \cdot \mathrm{D}_A(p, C)} && (6.46) \\ &\leq (\varepsilon^2 + 2\varepsilon)\, \mathrm{D}_A(p, C) && (6.47) \\ &\leq 3\varepsilon\, \mathrm{D}_A(p, C) && (6.48) \end{aligned}$$

for $\varepsilon \leq 1$. □

Lemma 6.10. *If* $\mathrm{D}_A(p, s_p) > \varepsilon^2\,\mathrm{D}_A(p, C)$ *then*

$$|\mathrm{D}(p, C) - \mathrm{D}(s_p, C)| \leq \frac{3}{\varepsilon}\,\mathrm{D}(p, s_p)\ . \tag{6.49}$$

Proof. Using Claim 6.8 and $\mathrm{D}_A(p, C) < \frac{1}{\varepsilon^2}\,\mathrm{D}_A(p, s_p)$ we have

$$\begin{aligned} |\mathrm{D}_A(p, C) - \mathrm{D}_A(s_p, C)| &\leq \mathrm{D}_A(p, s_p) + 2\sqrt{\mathrm{D}_A(p, s_p) \cdot \mathrm{D}_A(p, C)} && (6.50)\\ &< \left(1 + \frac{2}{\varepsilon}\right)\mathrm{D}_A(p, s_p) && (6.51)\\ &\leq \frac{3}{\varepsilon}\,\mathrm{D}_A(p, s_p) && (6.52)\end{aligned}$$

for $\varepsilon \leq 1$. □

Finally, we give proof to Lemma 6.11 that guarantees that by construction of Har-Peled-Mazumdar coresets, the k-median cost of P using S as center points is small when compared to the optimal k-median cost of P.

Lemma 6.11. *For any* $C \subseteq \mathbb{R}^d$ *of size* $|C| = k$ *we have*

$$\mathrm{cost}^{\mathrm{D}_A}(P, S) \leq \varepsilon^2\, opt_k^{\mathrm{D}_A}(P)\ . \tag{6.53}$$

Proof. As in the coreset construction from Section 6.1.1, for $i = 1, 2, \ldots, \kappa$ and $j = 1, 2, \ldots, \nu$ let V'_{ij} denote the partition of a region of the $\mathbb{R}^d$ that contains all points from P. Recall that V'_{ij} is contained in an axis-parallel cube of side length $r_j = \frac{\varepsilon}{3}\sqrt{\frac{1}{\alpha d}2^j R}$ centered at point $q_i \in Q$. Furthermore, let V_{ij} be the preimages of V'_{ij} under the linear transformation B. We have

$$\begin{aligned} \mathrm{cost}^{\mathrm{D}_A}(P, S) &\leq \sum_{p \in P} \mathrm{D}_A(p, s_p) && (6.54)\\ &\leq \sum_{i=1}^{\kappa} \sum_{p \in P \cap V_{i0}} \mathrm{D}_A(p, s_p) + \sum_{i=1}^{\kappa}\sum_{j=1}^{\nu} \sum_{p \in P \cap V_{ij}} \mathrm{D}_A(p, s_p) && (6.55)\end{aligned}$$

Using Claim 6.12 and Claim 6.13 stated below we obtain

$$\mathrm{cost}^{\mathrm{D}_A}(P, S) \leq \frac{\varepsilon^2}{9} nR + \frac{8\varepsilon^2}{9\alpha}\,\mathrm{cost}^{\mathrm{D}_A}(P, Q) \tag{6.56}$$

where

$$R = \frac{1}{\alpha n}\,\mathrm{cost}^{\ell_2^2}(P', Q') = \frac{1}{\alpha n}\,\mathrm{cost}^{\mathrm{D}_A}(P, Q)\ . \tag{6.57}$$

Since $\mathrm{cost}^{\mathrm{D}_A}(P,Q) \leq \alpha\, opt_k^{\mathrm{D}_A}(P)$ and $R \leq \frac{1}{n}\, opt_k^{\mathrm{D}_A}(P)$ we conclude

$$\mathrm{cost}^{\mathrm{D}_A}(P,S) \leq \frac{\varepsilon^2}{9}\, opt_k^{\mathrm{D}_A}(P) + \frac{8\varepsilon^2}{9}\, opt_k^{\mathrm{D}_A}(P) = \varepsilon^2\, opt_k^{\mathrm{D}_A}(P)\ . \tag{6.58}$$

□

Claim 6.12. *If $p \in P \cap V_{i0}$ for any $i \in \{1,2,\ldots,\kappa\}$ then*

$$\mathrm{D}_A(p, s_p) \leq \frac{\varepsilon^2}{9} R\ . \tag{6.59}$$

Proof. Since $p', s'_p \in V'_{i0}$ both lie inside a cube of side length $r_0 = \frac{\varepsilon}{3}\sqrt{\frac{1}{\alpha d} R}$ we have

$$\|p' - s'_p\| \leq \sqrt{d}\, r_0 = \frac{\varepsilon}{3}\sqrt{\frac{1}{\alpha} R}\ . \tag{6.60}$$

Using $\alpha \geq 1$ we conclude

$$\mathrm{D}_A(p, s_p) = \|p' - s'_p\|^2 \leq \frac{\varepsilon^2}{9} R\ . \tag{6.61}$$

□

Claim 6.13. *If $p \in P \cap V_{ij}$ for any $i \in \{1,2,\ldots,\kappa\}$ and $j \in \{1,2,\ldots,\nu\}$ then*

$$\mathrm{D}_A(p, s_p) \leq \frac{8\varepsilon^2}{9\alpha}\, \mathrm{D}_A(p, q_i)\ . \tag{6.62}$$

Proof. Since $p' \in V'_{ij}$ with $j \geq 1$ we have $p' \notin V'_{i,t}$ for all $t < j$ and, hence,

$$\|p' - q'_i\| \geq \frac{1}{2}\sqrt{2^{j-1} R}\ . \tag{6.63}$$

Furthermore, since $p', s'_p \in V'_{ij}$ both lie inside a cube of side length $r_j = \frac{\varepsilon}{3}\sqrt{\frac{1}{\alpha d} 2^j R}$ we have

$$\|p' - s'_p\| \leq \sqrt{d}\, r_j = \frac{\varepsilon}{3}\sqrt{\frac{1}{\alpha} 2^j R} \leq \frac{2\varepsilon}{3}\sqrt{\frac{2}{\alpha}} \|p' - q'_i\|\ . \tag{6.64}$$

Hence, we conclude

$$\mathrm{D}_A(p, s_p) = \|p' - s'_p\|^2 \leq \frac{8\varepsilon^2}{9\alpha}\|p' - q'_i\|^2 = \frac{8\varepsilon^2}{9\alpha}\, \mathrm{D}_A(p, q_i)\ . \tag{6.65}$$

□

6.2 A new coreset construction based on non-uniform sampling

In this section we give a new, randomized coreset construction for the Mahalanobis k-median problem based on the non-uniform sampling approach from Chapter 5. In this construction, the first coreset point is chosen uniformly at random among the input points. After that, iteratively, any further coreset point is obtained by choosing an input point non-uniformly at random with probability proportional to the distance towards the already chosen coreset points. This construction has been first proposed in [Ackermann et al., 2010b] for the Euclidean k-means problem. Here, we give a generalization of this result which is applicable to the Mahalanobis k-median problem.

In a way our new approach can be seen as a randomized version of the deterministic coreset construction from [Har-Peled and Mazumdar, 2004]. In Section 6.1, we have shown how a coreset can be constructed by carefully placing points in a grid-like fashion to cover all input points. During this process, the regions close to the initial approximate centers which are expected to contain a larger number of input points are provided with a larger number of coreset points. On the other hand, the outer regions of the input set, that contain only a small number of input points, are provided with fewer coreset points, but are not ignored. A quite similar behavior can be observed when the coreset points are chosen iteratively at random with probability proportional to the distance towards the already chosen coreset points: Regions that are crowded with input points are preferred by the non-uniform sampling due to the combined probability mass of this region. On the other hand, outliers are not ignored since the non-uniform sampling will prefer these points due to their single, large contribution to the total cost. Hence, intuitively, a very similar covering to the deterministic case is obtained. In fact, we show that this intuition is true and that our new construction yields a coreset of approximately the same size as the coresets from [Har-Peled and Mazumdar, 2004], up to a logarithmic factor.

This new approach has several, practical advantages. First, our new construction is easy to implement. Second, in contrast to many other coreset construction, the non-uniform sampling scheme does not need an initial constant factor approximation to build the coreset. Rather, in the light of the result from Chapter 5, such an approximation is computed on-the-fly while building the coreset. Third, another practical advantage is that the size of the coreset has not to be fixed prior to the coreset construction. Instead, the size can be adjusted on demand while building the coreset. At any point during the construction, the distances of the input points toward the coreset points (which have to be stored anyway to enable the non-uniform sampling) can be taken as indicator whether additional coreset points yield any significant benefit. Furthermore, it should be noted that for any fixed coreset size m the running time of all operations necessary to construct the coreset does only have a low polynomial dependency on the dimension d. Hence, if we choose to construct a coreset of a size merely polynomial in d, we should still obtain a fair coreset (albeit not a (k, ε)-coreset) in time polynomial in d which does not seem to be possible using grid-based constructions like the one from [Har-Peled and Mazumdar, 2004].

A practical implementation of coresets based on the result of this section has been given in [Ackermann et al., 2010b]. There, a data structure called the *coreset tree* has been used to allow non-uniform sampling according to approximately the same distribution as the distribution considered in this section, but at a considerably improved running time. In doing so, an efficient clustering algorithm for huge data sets in the data streaming model has been obtained that performs quite well in practice. The empirical evidence given in [Ackermann et al., 2010b] suggests that this algorithm is on a par, if not superior, in terms of quality and running time when compared to other widely used clustering algorithms for data streams, such as the local improvement algorithm given in [Guha et al., 2000] and [O'Callaghan et al., 2002], or algorithm BIRCH from [Zhang et al., 1996].

6.2.1 Coreset Construction

Let $P \subseteq \mathbb{R}^d$ be of size n. We show how to construct a (k, ε)-coreset for the k-median problem with respect to a Mahalanobis distance D_A by using the non-uniform sampling approach from Chapter 5. Recall that this approach is an iterative process as follows:

1. Choose an initial point s_1 uniformly at random from P.

2. Let S be the set of already chosen points from P. Then element $p \in P$ is chosen with probability $\frac{\mathrm{D}_A(p,S)}{\mathrm{cost}(P,S)}$ as next element of S.

3. Repeat step 2 until S contains the desired number of points.

As in Chapter 5, we say S is chosen *at random according to* D_A.

For our coreset construction, let $S = \{s_1, s_2, \ldots, s_m\}$ be a set of m points chosen at random according to D_A. Let $S_1, S_2, \ldots, S_m$ be the partion of P induced by assigning each $p \in P$ to their closest $s_i \in S$. That is, $p \in S_i$ if and only if $s_i = \arg\min_{s \in S} \mathrm{D}_A(p, s)$, breaking ties arbitrarily. Furthermore, for each point $s_i \in S$ we assign weight $w(s_i) = |S_i|$. In Section 6.2.2 below, we show the following theorem.

Theorem 6.14. *Let $0 < \varepsilon, \delta < 1$ be arbitrary constants. There exists an $m = \Theta\big(d^{d/2}\delta^{-d/2}\varepsilon^{-d}k\log(n)\log^{d/2}(d^{d/2}\delta^{-d/2}\varepsilon^{-d}k\log n)\big)$ such that with probability at least $1-\delta$ the weighted multiset S is a $(k, 6\varepsilon)$-coreset of P.*

From Chapter 5 we know that a set of size m can be sampled according to D_A using at most $\mathcal{O}(mn)$ arithmetic operations, including evaluations of Mahalanobis distance D_A. Here, Mahalanobis distance D_A can be evaluated in time $\mathcal{O}(d^2)$. Hence, with high probability coreset S of size

$$|S| = m = 2^{\mathcal{O}(d\log d)}\varepsilon^{-d}k\log(n)\log^{d/2}\big(\varepsilon^{-1}k\log(n)\big) \qquad (6.66)$$

can be computed in time $\mathcal{O}(d^2n|S|)$. Theorem 6.3 is an immediate consequence of Theorem 6.14.

6.2.2 Proof of Theorem 6.14

Now we give a proof of Theorem 6.14. Let $C \subseteq \mathbb{R}^d$ be an arbitrary set of k centers. For all $p \in P$ let s_p denote the closest coreset point from S. We say s_p is the representative of p in S. Analogousely to inequalities (6.27) to (6.33) from Section 6.1.3 we obtain

$$\Big|\mathrm{cost}(P, C) - \mathrm{cost}_w(S, C)\Big| \leq 3\varepsilon\,\mathrm{cost}(P, C) + \frac{3}{\varepsilon}\,\mathrm{cost}(P, S)\ . \qquad (6.67)$$

Hence, we still need to give a bound on $\frac{3}{\varepsilon}\mathrm{cost}(P, S)$, that is, a bound on the clustering cost of P using the sampled coreset points as centers. Furthermore, this bound has to be given in terms of $\mathcal{O}(\varepsilon^2)\,\mathrm{cost}(P, C)$.

Intuitively, it is clear that such a bound exists. From Theorem 5.3 we know that, with high probability, non-uniform sampling of m points according to D_A yields a factor $\mathcal{O}(\log m)$-approximation of the optimal m-medians of P, i.e.,

$$\mathrm{cost}(P, S) \leq \mathcal{O}(\log m)\, opt_m(P)\ . \qquad (6.68)$$

The following lemma assures that if we use a large enough number of clusters m, we have

$$opt_m(P) \leq \gamma \, opt_k(P) \tag{6.69}$$

with an arbitrarily small $\gamma > 0$. Hence,

$$\mathrm{cost}(P, S) \leq \mathcal{O}(\gamma \log m) \, \mathrm{cost}(P, C) \ . \tag{6.70}$$

By simultaneously choosing the right parameters γ and m, we make the factor $\mathcal{O}(\gamma \log m)$ as small as desired.

We start by giving a prove to the following lemma.

Lemma 6.15. *Let $0 < \gamma \leq 1$. If $m \geq (9d/\gamma)^{d/2} k \lceil \log(n) + 2 \rceil$ then*

$$opt_m(P) \leq \gamma \, opt_k(P) \ . \tag{6.71}$$

Proof. We show that there exists a set $G \subseteq \mathbb{R}^d$ of size at most m with $\mathrm{cost}(P, G) \leq \gamma \, opt_k(P)$. Thus, the lemma follows.

To this end, let C_P be an optimal solution to the Mahalanobis k-median problem of P, i.e., $\mathrm{cost}(P, C_P) = opt_k(P)$. Using C_P as initial $[1, 1]$-bicriteria approximation, we consider the construction of a $(k, \sqrt{\gamma})$-coreset G according to the construction of Har-Peled and Mazumdar from Section 6.1. Hence, we have

$$|G| \leq (9d/\gamma)^{d/2} k \lceil \log(n) + 2 \rceil \leq m \ . \tag{6.72}$$

Thus, using Lemma 6.11 we conclude

$$opt_m(P) \leq \mathrm{cost}(P, G) \leq \gamma \, opt_k(P) \ . \tag{6.73}$$

□

Now, recall that a Mahalanobis distance D_A is a 1-similar Bregman divergence. Thus, from Theorem 5.3 and Markov's inequality we know that with probability at least $1 - \delta$ we have

$$\mathrm{cost}(P, S) \leq \frac{8}{\delta}(2 + \ln m) \, opt_m(P) \ . \tag{6.74}$$

Lemma 6.16 below guarantees that by using Lemma 6.15 with the right choice of parameters γ and m we have

$$\begin{aligned} \mathrm{cost}(P, S) &\leq \frac{8\gamma}{\delta}(2 + \ln m) \, opt_k(P) && (6.75) \\ &\leq \varepsilon^2 \, opt_k(P) && (6.76) \\ &\leq \varepsilon^2 \, \mathrm{cost}(P, C) \ . && (6.77) \end{aligned}$$

Hence, with probability at least $1-\delta$ we obtain from inequality (6.67) and (6.77)

$$|\mathrm{cost}(P,C) - \mathrm{cost}_w(S,C)| \leq 6\varepsilon \, \mathrm{cost}(P,C) \ . \tag{6.78}$$

This proves Theorem 6.14.

Lemma 6.16. *There exist m, γ with*

$$m = \Theta\left(d^{d/2}\delta^{-d/2}\varepsilon^{-d}k\log(n)\log(d^{d/2}\delta^{-d/2}\varepsilon^{-d}k\log n)\right) \ , \tag{6.79}$$

$$\gamma \leq \frac{\varepsilon^2\delta}{8(2+\ln m)} \tag{6.80}$$

such that we have

$$m \geq (9d/\gamma)^{d/2}k\lceil\log(n)+2\rceil \ . \tag{6.81}$$

Proof. Let $e \approx 2.718\ldots$ denote the base of the natural logarithm. For simplicity of notation, we define

$$L = (171d)^{d/2}\delta^{-d/2}\varepsilon^{-d}k\lceil\log(n)+2\rceil \ . \tag{6.82}$$

As choice of m and γ, we consider

$$m = e^{d/2}L\ln^{d/2}(L) \ , \tag{6.83}$$

$$\gamma = \frac{\varepsilon^2\delta}{16\ln m} \ . \tag{6.84}$$

Hence, we find

$$m = \Theta\left(d^{d/2}\delta^{-d/2}\varepsilon^{-d}k\log(n)\log(d^{d/2}\delta^{-d/2}\varepsilon^{-d}k\log n)\right) \ , \tag{6.85}$$

$$\gamma \leq \frac{\varepsilon^2\delta}{8(2+\ln m)} \tag{6.86}$$

for $m \geq 8$. Observe that by choice of γ we have

$$(9d/\gamma)^{d/2}k\lceil\log n+2\rceil = (171d)^{d/2}\delta^{-d/2}\varepsilon^{-d}k\lceil\log(n)+2\rceil \ \ln^{d/2}(m) \tag{6.87}$$

$$= L\ln^{d/2}(m) \ . \tag{6.88}$$

Hence, to show that inequality (6.81) is satisfied, it is sufficient to show that $m \geq L\ln^{d/2}(m)$.

To this end, note that by definition of L we have

$$d = o(\ln L) \ . \tag{6.89}$$

Thus, we know that for large enough L we have

$$\ln(L) + \frac{d}{2}\ln\ln(L) + \frac{d}{2} \leq e \ln L \ . \tag{6.90}$$

Using our choice of m and inequality (6.90), we conclude

$$\begin{aligned} L \ln^{d/2}(m) &= L\Big(\ln\big(e^{d/2} L \ln^{d/2}(L)\big)\Big)^{\frac{d}{2}} && (6.91)\\ &= L\left(\ln(L) + \frac{d}{2}\ln\ln(L) + \frac{d}{2}\right)^{\frac{d}{2}} && (6.92)\\ &\leq e^{d/2} L \ln^{d/2}(L) && (6.93)\\ &= m \ . && (6.94)\end{aligned}$$

□

6.3 Discussion

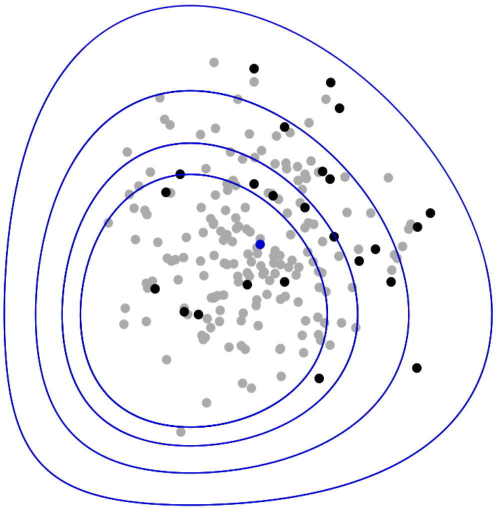
In this chapter we have presented two (strong) coreset constructions applicable to the Mahalanobis k-median problem. The first construction is deterministic and a direct generalization of the coreset construction given in [Har-Peled and Mazumdar, 2004]. The size of these coresets is only linear in the number of clusters k and only logarithmic in the number of points n, but exponential in the dimension d. Our second construction is a new, randomized coreset construction based on the non-uniform sampling scheme from Chapter 5. This new construction yields coresets of approximately the same size as the deterministic construction from [Har-Peled and Mazumdar, 2004], up to a logarithmic factor.

It is noteworthy that while both constructions are well suited for the case of Mahalanobis distances, they do not generalize easily to more general k-median clustering problems, even if the dissimilarity measure features some approximate metric properties such as a μ-similar Bregman divergence D_φ. Subtle technical difficulties arise if we want to show the existence of a strong (k, ε)-coreset for arbitrarily small ε. The main problem we encounter is that the approximate metric properties do not provide much help when our goal is to give bounds on the difference of two dissimilarities, e.g., a bound on $|\mathrm{D}_\varphi(p,C)-\mathrm{D}_\varphi(s,C)|$ in terms of $\varepsilon\,\mathrm{D}_\varphi(p,C)$ for arbitrary small ε. Instead, the proofs of this chapter rely on the fact that a Mahalanobis distance in general (such as the squared Euclidean distance in particular) is the square of a metric. Unfortunately, the Mahalanobis distances are the only Bregman divergences that are the square of a metric.

To overcome these problems, in the next chapter, we introduce a relaxed notion of coresets which we call weak coresets. Not only do we show that there is an efficient construction of weak coresets for the μ-similar Bregman k-median problem, but we also show that there are weak coresets of size merely polynomial in k, polylogarithmic in the number of points n, and independent of the dimension d.

However, it remains an open question whether there exist strong coresets of small size for the k-median problem with respect to a general Bregman divergence, or even just in the case when we are restricted to μ-similar Bregman divergences.

7 Weak coresets for μ-similar Bregman divergences



Several constructions for strong coresets like the coresets given in Chapter 6 have been proposed for the k-median problem in metric and Euclidean spaces. Usually, the analysis of these coreset constructions relies heavily on the fact that the underlying distance measure is a metric (such as the Euclidean distance), or at least the square of a metric (such as Mahalanobis distances). No strong coreset constructions are known for the general Bregman k-median problem using Bregman divergences other than Mahalanobis distances.

From Section 2.2 we know that μ-similar Bregman divergences feature at least some quasi-metric properties, such as triangle inequality and symmetry within a constant factor of $\mathcal{O}(1/\mu)$. Unfortunately, a straightforward adaptation of existing coreset analyses using these quasi-metric properties only leads to $(k, \Theta(1))$-coresets at best. Subtle technical difficulties arise from the asymmetry and the lack of triangle inequality if we want to show the existence of a strong (k, ε)-coreset for arbitrarily small ε.

However, when using coresets to design faster approximation algorithms, it turns out that the classical definition of strong coresets seems to be unnecessary strict. Recall that we demand in the definition of strong (k, ε)-coresets that the clustering cost of the original set P and the weighted coreset S have to be approximately the same with respect to *any* set of k centers from domain $\mathbb{X} \subseteq \mathbb{R}^d$. This means that the number of centers that have to be considered may be infinite or even uncountable. Now, assume that we want to construct a coreset S by preprocessing the input data P for a certain, fixed approximation algorithm. In this case, we are not interested in the cost of P and S towards an arbitrary set of centers. We are merely interested in comparing the cost of P and S with respect to the output centers of this particular approximation algorithm, and with respect to the to optimal k-medians of the given input instance. However, for each finite input set there are (usually) only a finite number of possible approximate medians computed by a fixed approximation algorithm, as well as only finitely many optimal medians. So, it turns out that we only need the clustering cost of P and S to be approximately the same with respect to this finite number of relevant center points. This observation leads to a relaxed notion of coresets.

In our relaxed notion of coresets, only center points from a finite but significant set $\Gamma \subseteq \mathbb{X}$ are considered. We call this a Γ-*weak coreset*. Weak coresets have been introduced in [Feldman et al., 2007] to construct coresets for the Euclidean k-means problem with a size that is independent of the number of input points n and dimension d. However, our notion of weak coresets for the generalized k-median problem differs slightly from the previous definition.

Definition 7.1. *Let* D *be a dissimilarity measure on domain* $\mathbb{X} \subseteq \mathbb{R}^d$*, let* $P \subseteq \mathbb{X}$ *be finite, and let* $\Gamma \subseteq \mathbb{X}$ *be arbitrary. A weighted multiset* $S \subseteq \mathbb{X}$ *with weight function* $w : S \to \mathbb{R}_{\geq 0}$ *such that* $\sum_{s \in S} w(s) = |P|$ *is called a* Γ-weak (k, ε)-coreset *of* P *for the* k*-median problem with respect to* D *if for all* $C \subseteq \Gamma$ *of size* $|C| = k$ *we have*

$$\left|\mathrm{cost}^{\mathrm{D}}(P, C) - \mathrm{cost}^{\mathrm{D}}_w(S, C)\right| \leq \varepsilon \,\mathrm{cost}^{\mathrm{D}}(P, C) \ . \tag{7.1}$$

In this chapter, we give a randomized construction of weak coresets for the μ-similar Bregman k-median problem. This randomized construction is based on uniform sampling. In doing so, we present the first coreset construction applicable to the Bregman k-median clustering problem. This construction is an adaptation of an earlier construction of strong coresets from [Chen, 2006] and [Chen, 2009], originally proposed in the context of the k-median and the k-means problem in metric and Euclidean spaces. We prove that Chen's construction yields at least weak coresets for the Bregman k-median problem. The main result of Section 7.1 is summarized in the following theorem.

Theorem 7.2. *Let* D_φ *be a* μ*-similar Bregman divergence on* $\mathbb{X} \subseteq \mathbb{R}^d$ *and let* $P \subseteq \mathbb{X}$ *be of size* n*. For any finite* $\Gamma \subseteq \mathbb{X}$ *there exists a* Γ*-weak* (k,ε)*-coreset of* P *of size* $\mathcal{O}\left(\varepsilon^{-2} k \log(n) \log(k|\Gamma|^k \log n)\right)$ *for the* k*-median problem with respect to* D_φ*. Furthermore, given a set of medians of a constant factor approximate* k*-median clustering of* P*, such a weak coreset can be constructed with high probability using uniform sampling. This construction requires at most* $\mathcal{O}\left(kn + \varepsilon^{-2} k \log(n) \log(k|\Gamma|^k \log n)\right)$ *arithmetic operations, including evaluations of* D_φ*.*

Note that the size of the Γ-weak coreset depends only logarithmically on the size of Γ. Hence, if the set of relevant center points for an input set P and a given approximation algorithm is only polynomial in n and independent of d, we find that the size of our Γ-weak coreset will be at most polylogarithmic in n, and independent of d.

Furthermore, in Section 7.2, we show how to apply these weak coresets to the $(1+\varepsilon)$-approximation algorithm Cluster from Chapter 4. To this end, we give a combinatorial analysis of the number of relevant center points, that is, the number of potential output points of algorithm Cluster. We show that for each input set P of size n the set Γ of relevant center points is indeed of size at most polynomial in n. In doing so, we prove that we can speed up the running time of algorithm Cluster by building a Γ-weak coreset and using this coreset as a smaller input set, as is stated in the following theorem.

Theorem 7.3. *Let* $P \subseteq \mathbb{X}$ *be of size* n*. There exists an algorithm that with constant probability computes a* $(1+\varepsilon)$*-approximate solution of the* k*-median problem with respect to* μ*-similar Bregman divergence* D_φ *for input instance* P *using at most* $\mathcal{O}(kn) + 2^{\mathcal{O}(k/\varepsilon \log(k/\varepsilon))} \log^{k+2}(n)$ *arithmetic operations, including evaluations of* D_φ*.*

For a number of Bregman divergences, such as the Kullback-Leibler divergence or the Itakura-Saito divergence, this result improves significantly over the previousely asymptotically fastest known $(1+\varepsilon)$-approximation algorithm from Chapter 4. The results from this Chapter have been published earlier in [Ackermann and Blömer, 2009].

7.1 Construction of Γ-weak coresets

In this section we present a weak coreset construction applicable to a μ-similar Bregman divergences D_φ on domain $\mathbb{X} \subseteq \mathbb{R}^d$. In particular, we show how to construct a Γ-weak coreset for an arbitrary but fixed and finite Γ.

7.1.1 Chen's coreset construction for Bregman divergences

We give an adaptation of Chen's coreset construction from [Chen, 2006] and [Chen, 2009]. Chen's construction has been originally proposed as a construction of strong coresets in the context of the Euclidean k-means and k-median problem, as well as in the context of the discrete version of metric k-means and k-median problems. We show that an adaptation of this construction to μ-similar Bregman divergences yields at least weak (k, ε)-coresets.

In a nutshell, Chen's construction is as follows. Given the centers of a constant factor approximate clustering, the input set is partitioned by assigning each point to their closest center. Each set of this partition is further divided into ring sets centered around their common center point. This division is obtained in a way such that all points in a common ring set differ in their distance to their center by no more than a factor of 2. Finally, a prespecified number of m points is sampled uniformly at random from each ring set. The union of these sample points forms the coreset. Each coreset point is assigned with a weight proportional to the number of points from their ring set. An illustration of this coreset construction can be found in Figure 7.1.

We now give the coreset construction in detail. In the following, let $P \subseteq \mathbb{X}$ with $|P| = n$, and let $A = \{a_1, a_2, \dots, a_\kappa\}$ be the medians of an $[\alpha, \beta]$-bicriteria approximation for the optimal k-median clustering of P

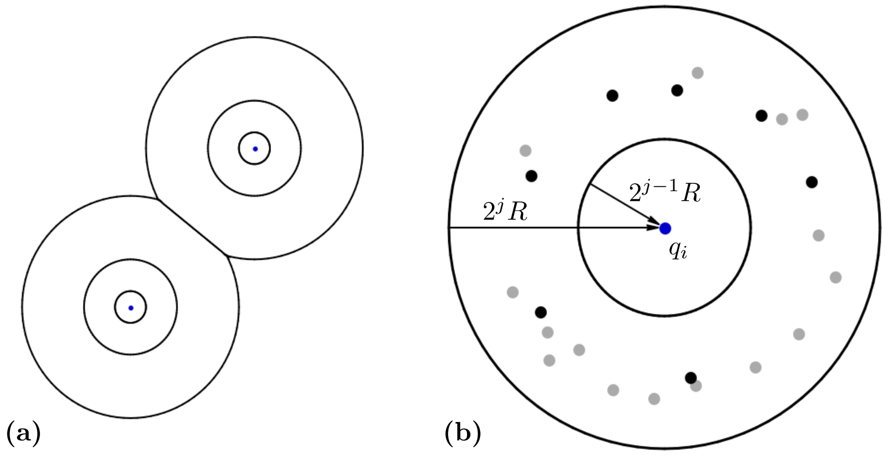


Figure 7.1: An illustration of Chen's coreset construction. **(a)** For each approximate center q_i, the input set P is partitioned by Bregman balls of radius 2^jR centered at q_i. **(b)** From each ring set of the partion, a fixed number of m representative points is chosen uniformly at random and is added to the coreset. Each coreset point is assigned a weight proportional to the number of input points from its ring set.

with respect to μ-similar Bregman divergence D_φ, i.e.,

$$\mathrm{cost}(P, A) \le \alpha\, opt_k(P) \tag{7.2}$$

and

$$|A| = \kappa \le \beta k \ . \tag{7.3}$$

A simple and fast algorithm to obtain an $[\mathcal{O}(\mu^{-2} \log k), 1]$-bicriteria approximation for μ-similar Bregman divergences is given in Chapter 5.

Let $A_1, A_2, \ldots, A_\kappa$ be the partition of P induced by assigning each $p \in P$ to their closest $a_i \in A$, i.e. $p \in A_i$ if and only if $a_i = \arg\min_{a \in A} \mathrm{D}_\varphi(p, a)$, breaking ties arbitrarily. Furthermore, let

$$R = \frac{1}{\alpha n} \mathrm{cost}(P, A) \ . \tag{7.4}$$

Note that $R \le \frac{1}{n}\, opt_k(P)$. Furthermore, let

$$U_\varphi(a_i, r) = \{x \in \mathbb{X} \,|\, \mathrm{D}_\varphi(x, a_i) \le r\} \tag{7.5}$$

denote the D_φ-ball of radius r centered at a_i. Using A, we define a partition $\{P_{ij}\}_{i,j}$ of P by

$$P_{i0} = P_i \cap U_\varphi(a_i, R) \tag{7.6}$$

for $i = 1, 2, \ldots, \kappa$ and

$$P_{ij} = P_i \cap \big(U_\varphi(a_i, 2^j R) \setminus U_\varphi(a_i, 2^{j-1} R)\big) \tag{7.7}$$

for $i = 1, 2, \ldots, \kappa$ and $j = 1, 2, \ldots, \nu$, where $\nu = \lceil \log(\alpha n) \rceil$. Note that $\{P_{ij}\}_{i,j}$ is indeed a partition of P since the existence of a $p \in P$ with $\mathrm{D}_\varphi(p, A) > 2^\nu R$ leads to

$$\mathrm{cost}(P, A) \geq \mathrm{D}_\varphi(p, A) > 2^\nu R \geq \alpha n R = \mathrm{cost}(P, A) \tag{7.8}$$

which is a contradiction.

Assume we have fixed a number $m \in \mathbb{N}$. For each i, j let S_{ij} be a uniform sample multiset from P_{ij} of size $|S_{ij}| = m$. Let $w(s) = \frac{1}{m}|P_{ij}|$ be the weight associated with $s \in S_{ij}$. We define $S = \bigcup_{i,j} S_{ij}$ of size

$$|S| = \kappa \nu m = \beta k m \lceil \log(\alpha n) \rceil \tag{7.9}$$

as our weak coreset. In Section 7.1.2 we prove the following theorem.

Theorem 7.4. *Let $\Gamma \subseteq \mathbb{X}$ be an arbitrary finite set. Then there exists an $m = \Theta\left(\alpha^2 \varepsilon^{-2} \mu^{-2} \log\big(\beta \delta^{-1} k |\Gamma|^k \log(\alpha n)\big)\right)$ such that with probability at least $1 - \delta$ the weighted multiset S is a Γ-weak $(k, 7\varepsilon)$-coreset of P.*

The interesting aspect of this result is that the size of the weak coreset depends only logarithmically on the size of Γ. Hence, if we know that the set of relevant center points for an input set P is, say, only polynomial in n, we find that the size of our Γ-weak coreset will be only polylogarithmic in n. We will make use of this observation in Section 7.2. Also note that we do not have to know the exact content of set Γ to construct the Γ-weak coreset. We only need a size bound on Γ to know how many points should be sampled from each set P_{ij}.

The partion $\{P_{ij}\}_{i,j}$ can be computed using at most $\mathcal{O}(kn)$ arithmetic operations, including evaluations of D_φ. The Γ-weak coreset S can be sampled from $\{P_{ij}\}_{i,j}$ in time $\mathcal{O}(|S|)$. Hence, Theorem 7.2 is an immediate consequence of Theorem 7.4.

7.1.2 Proof of Theorem 7.4

To prove Theorem 7.4 we will make use of the following probabilistic concentration bound, given in [Haussler, 1992]. This bound states that the average of a finite number of values from a fixed region can be well approximated by the average value of a constant sized sample set.

Lemma 7.5 ([Haussler, 1992]). *Let $f : P \to \mathbb{R}$ and $F \in \mathbb{R}$ be such that we have $0 \leq f(p) \leq F$ for all $p \in P$. Let $S \subseteq P$ be a uniform sample multiset of size $|S| \geq \frac{1}{2}\varepsilon^{-2}\ln(2\delta^{-1})$ for constants $\varepsilon > 0$ and $\delta > 0$. Then we have*

$$\Pr\left[\left|\frac{1}{|P|}\sum_{p\in P} f(p) - \frac{1}{|S|}\sum_{s\in S} f(s)\right| \leq \varepsilon F\right] \geq 1-\delta\ . \tag{7.10}$$

Our strategy to prove Theorem 7.4 is as follows. First, we prove inequality (7.1) with high probability for an arbitrary but fixed set C of size k. After that, we use the union bound to show that with probability at least $1-\delta$ inequality (7.1) is satisfied for all $C \subseteq \Gamma$ of size k.

Lemma 7.6. *Let $C \subseteq \mathbb{X}$ be a fixed set of size $|C| = k$. If we have*

$$m \geq 8\alpha^2\varepsilon^{-2}\mu^{-2}\ln\left(2\delta^{-1}\kappa\nu|\Gamma|^k\right) \tag{7.11}$$

then with probability $1-\delta|\Gamma|^{-k}$ we have

$$|\mathrm{cost}(P,C) - \mathrm{cost}_w(S,C)| \leq 7\varepsilon\,\mathrm{cost}(P,C)\ . \tag{7.12}$$

Proof. We prove the Lemma in two steps. First, we use the concentration bound from Lemma 7.5 to give an upper bound on the difference $|\mathrm{cost}(P_{ij},C) - \mathrm{cost}_w(S_{ij},C)|$ for fixed i,j with (very) high probability. After that, we use the union bound to show that with high probability this upper bound holds for all i,j. Summing up over all i,j concludes the proof.

To this end, initially, fix i,j. We have

$$\begin{aligned}&\left|\mathrm{cost}(P_{ij},C) - \mathrm{cost}_w(S_{ij},C)\right|\\ &\quad= |P_{ij}|\cdot\left|\frac{1}{|P_{ij}|}\sum_{p\in P_{ij}} \mathrm{D}_\varphi(p,C) - \frac{1}{|P_{ij}|}\sum_{s\in S_{ij}} w(s)\,\mathrm{D}_\varphi(s,C)\right|\ .\end{aligned} \tag{7.13}$$

For all $p \in P_{ij}$ we define a function f_{ij} by

$$f_{ij}(p) = \mathrm{D}_\varphi(p,C)\ . \tag{7.14}$$

Obviously, we have $f_{ij}(p) \geq 0$ for all $p \in P_{ij}$. Let $q^* \in P_{ij}$ denote a point that minimizes f_{ij}. Furthermore, let D_U be a Mahalanobis distance such that

$$\mu\,\mathrm{D}_U(p,q) \leq \mathrm{D}_\varphi(p,q) \leq \mathrm{D}_U(p,q) \tag{7.15}$$

for all $p, q \in \mathbb{X}$. Using the double triangle inequality of D_U (Lemma 2.15) we deduce that for all $p \in P_{ij}$ we have

$$\begin{aligned} f_{ij}(p) &\leq \mathrm{D}_\varphi(p, c^*) && (7.16)\\ &\leq \mathrm{D}_U(p, c^*) && (7.17)\\ &\leq 4\big(\mathrm{D}_U(q^*, c^*) + \mathrm{D}_U(q^*, a_i) + \mathrm{D}_U(p, a_i)\big) && (7.18)\\ &\leq \frac{4}{\mu}\big(\mathrm{D}_\varphi(q^*, c^*) + \mathrm{D}_\varphi(q^*, a_i) + \mathrm{D}_\varphi(p, a_i)\big) && (7.19)\\ &\leq \frac{4}{\mu}\big(\mathrm{D}_\varphi(q^*, C) + 2^{j+1}R\big)\ , && (7.20) \end{aligned}$$

where $c^* = \arg\min_{c\in C} \mathrm{D}_\varphi(q^*, c)$. Here, inequality (7.20) holds since by construction of P_{ij} we have $\mathrm{D}_\varphi(q^*, a_i) \leq 2^j R$ and $\mathrm{D}_\varphi(p, a_i) \leq 2^j R$. Using f_{ij} and $w(s) = |P_{ij}|/|S_{ij}|$ for all $s \in S_{ij}$, equation (7.13) can be written as

$$\big|\mathrm{cost}(P_{ij}, C) - \mathrm{cost}_w(S_{ij}, C)\big| = |P_{ij}| \cdot \left| \frac{1}{|P_{ij}|} \sum_{p\in P_{ij}} f_{ij}(p) - \frac{1}{|S_{ij}|} \sum_{s\in S_{ij}} f_{ij}(s) \right| . \tag{7.21}$$

By using Lemma 7.5 with function f_{ij}, bound $F_{ij} = \frac{4}{\mu}\big(\mathrm{D}_\varphi(q^*, C) + 2^{j+1}R\big)$, and $|S_{ij}| = m$, with probability at least $1 - \delta(\kappa\nu|\Gamma|^k)^{-1}$ we obtain

$$\big|\mathrm{cost}(P_{ij}, C) - \mathrm{cost}_w(S_{ij}, C)\big| \leq \frac{\varepsilon}{\alpha} |P_{ij}| \big(\mathrm{D}_\varphi(q^*, C) + 2^{j+1}R\big)\ . \tag{7.22}$$

Now, note that by choice of $q^* \in P_{ij}$ we have

$$|P_{ij}|\,\mathrm{D}_\varphi(q^*, C) \leq \sum_{p\in P_{ij}} \mathrm{D}_\varphi(p, C) = \mathrm{cost}(P_{ij}, C)\ . \tag{7.23}$$

Furthermore, let us derive a bound on $2^{j+1}|P_{ij}|R$ for all $j \geq 0$. In case $j = 0$ we have

$$2^{j+1}|P_{ij}|R = 2\,|P_{ij}|R\ . \tag{7.24}$$

On the other hand, if $j \geq 1$ by construction of P_{ij} for all $p \in P_{ij}$ we have

$$2^{j-1} R \leq \mathrm{D}_\varphi(p, A) \;, \tag{7.25}$$

and we find

$$2^{j+1}|P_{ij}|R \leq 4 \sum_{p \in P_{ij}} \mathrm{D}_\varphi(p, A) = 4 \operatorname{cost}(P_{ij}, A) \;. \tag{7.26}$$

Hence, in either case we obtain

$$2^{j+1}|P_{ij}|R \leq 4 \operatorname{cost}(P_{ij}, A) + 2\,|P_{ij}|R \;. \tag{7.27}$$

Therefore, using inequality (7.22) in combination with inequalities (7.23) and (7.27) we obtain

$$\begin{aligned} &\big|\operatorname{cost}(P_{ij}, C) - \operatorname{cost}_w(S_{ij}, C)\big| \\ &\qquad \leq \frac{\varepsilon}{\alpha}\big(\operatorname{cost}(P_{ij}, C) + 4 \operatorname{cost}(P_{ij}, A) + 2\,|P_{ij}|R\big) \;. \end{aligned} \tag{7.28}$$

So far we have learned that for fixed i, j, inequality (7.28) holds with probability $1 - \delta(\kappa\nu|\Gamma|^k)^{-1}$. Using the union bound, we find that with probability at least $1 - \delta|\Gamma|^{-k}$, inequality (7.28) holds for all $i = 1, \ldots, \kappa$ and $j = 1, \ldots, \nu$. Hence, using the triangle inequality of the reals and summing up over all i, j we have

$$\begin{aligned} &\big|\operatorname{cost}(P, C) - \operatorname{cost}_w(S, C)\big| \\ &\quad \leq \sum_{i,j} |\operatorname{cost}(P_{ij}, C) - \operatorname{cost}_w(S_{ij}, C)| && (7.29) \\ &\quad \leq \frac{\varepsilon}{\alpha}\Big(\sum_{i,j} \operatorname{cost}\,(P_{ij}, C) + 4 \sum_{i,j} \operatorname{cost}\,(P_{ij}, A) + 2R \sum_{i,j} |P_{ij}|\Big) && (7.30) \\ &\quad = \frac{\varepsilon}{\alpha}\big(\operatorname{cost}(P, C) + 4 \operatorname{cost}(P, A) + 2nR\big) \;. && (7.31) \end{aligned}$$

Using $R \leq \frac{1}{n}\, opt_k(P)$ and $\operatorname{cost}(P, A) \leq \alpha\, opt_k(P)$ we conclude

$$\begin{aligned} &\big|\operatorname{cost}(P, C) - \operatorname{cost}_w(S, C)\big| \\ &\qquad \leq \frac{\varepsilon}{\alpha}\big(\operatorname{cost}(P, C) + 4 \operatorname{cost}(P, A) + 2\, opt_k(P)\big) && (7.32) \\ &\qquad \leq \frac{\varepsilon}{\alpha}\big(\operatorname{cost}(P, C) + (4\alpha + 2)\, opt_k(P)\big) && (7.33) \\ &\qquad \leq 7\varepsilon \operatorname{cost}(P, C) && (7.34) \end{aligned}$$

since $\alpha > 1$. □

By Lemma 7.6, for a fixed choice of $C \subseteq \Gamma$ we have inequality (7.12) with probability $1 - \delta|\Gamma|^{-k}$. Since there are at most $\binom{|\Gamma|}{k} \leq |\Gamma|^k$ subsets $C \subseteq \Gamma$ of size k, using the union bound we obtain that with probability $1 - \delta$ the weighted multiset S is a Γ-weak $(k, 7\varepsilon)$-coreset of P. This proves Theorem 7.4.

7.2 Application to Bregman k-median clustering

In this section we show how to use Γ-weak coresets to improve the asymptotic running time of an existing $(1 + \varepsilon)$-approximation algorithm for the μ-similar Bregman k-median clustering problem, thereby giving proof to Theorem 7.3. In particular, we improve the adaptation of algorithm Cluster from Chapter 4 that works on weighted input sets. To this end, we will use the following strategy:

1. In a preprocessing step we construct a Γ-weak coreset of the input set with respect to a carefully defined, small Γ that includes all medians relevant to P when using the adaptation of algorithm Cluster.

2. After that, we run the adaptation of algorithm Cluster with the Γ-weak coreset as weighted input set. The computed approximate medians are returned.

If we have identified the right set Γ it turns out that the computed approximate medians are a $(1+\varepsilon)$-approximate solution to the Bregman k-median clustering problem. However, the definition of Γ is the crucial part of this approach.

To give a precise description of the set of relevant center points we have to clarify which points are considered to be *relevant* to us in the given approach. The idea behind using coresets in a preprocessing step is that any solution computed using the coreset as input set should be approximately as good or as bad for the coreset as the solution is for the original input set. This property is captured by inequality (7.1) from our definition of weak coresets, and we want to make use of this inequality for the output points computed by our algorithm. So, an input point is relevant to our approach if such a point turns up as an approximate center point computed by algorithm Cluster.

Of course, we do not know these optput points in advance since both our coreset construction as well as our approximation algorithm are randomized. Hence, what we really need is a guarantee that the number of all *potential* output points, computed during any run of algorithm Cluster started with any coreset from Chen's construction, is small. Fortunately, due to the nature of our coreset construction and of algorithm Cluster this is indeed the case, as is stated in the following Lemma.

Lemma 7.7. *Fix an input set P and a bicriteria approximation A as given in Chen's coreset construction. Then there exists a set $\Gamma_{P,A}$ of size*

$$|\Gamma_{P,A}| \leq n^{\mathcal{O}(\varepsilon^{-1}\mu^{-1})} \tag{7.35}$$

such that for every potential (S, w) from Chen's coreset construction applied to P and A, and for every possible output C of algorithm Cluster *started with weighted set S, we have $C \subseteq \Gamma_{P,A}$.*

Proof. The proof of this lemma is somewhat technical and requires some insight into the operation of algorithm Cluster. However, the main idea of the proof is straightforward. Let P be of size $|P| = n$. Any cluster center $c \in C$ computed by algorithm Cluster is the weighted centroid of an $\mathcal{O}(\varepsilon^{-1}\mu^{-1})$-sized subset of S. We know that there are at most $n^{\mathcal{O}(\varepsilon^{-1}\mu^{-1})}$ subsets of S of size $\mathcal{O}(\varepsilon^{-1}\mu^{-1})$. In addition, the number of different weights that can be assigned to a point during a run of the algorithm started with a coreset from Chens's construction is bounded by n. Thus, the number of all possible weighted subsets that define an approximate median can be bounded by $n^{\mathcal{O}(\varepsilon^{-1}\mu^{-1})}$.

We now give a detailed proof of the lemma. Since P and A are fixed, so is the partition $\{P_{ij}\}_{i,j}$ of P from Chen's coreset construction. Let S be a weighted multiset with weight function w obtained by the coreset construction. Without loss of generality, we may assume $|S| \leq n$. Let m be the constant number of elements sampled uniformly at random from each P_{ij} to obtain S.

First, let us ignore the weight function w. Recall that according to Corollary 4.22 each approximate median from the output of Cluster is obtained as the (weighted) centroid of a subset of size $\mathcal{O}(\varepsilon^{-1}\mu^{-1})$ from S. Hence, each potential output point corresponds to a constant sized subset of S. Let N denote the number of such subsets. Then we have

$$N \leq \binom{|S|}{\mathcal{O}(\varepsilon^{-1}\mu^{-1})} \leq n^{\mathcal{O}(\varepsilon^{-1}\mu^{-1})} \ . \tag{7.36}$$

Now let us consider the weight function w. Since $|P_{ij}|$ and m are fixed and independent of the random choices made during the construction of S, the (initial) weight of each point from input set S is also fixed. However, during the execution of algorithm Cluster, sometimes the weight of a single point is split during the pruning phase of the algorithm. Hence, the weight of some points will change. So, we have to analyze the number of different weights that may be assigned to point $s \in S$ during a run of algorithm Cluster.

To this end, observe that the behavior and output of algorithm Cluster will not change when the weight function of the input set is scaled by a constant. Therefore, let us consider set S with weight function $\hat{w}$ such that $\hat{w}(s) = m\, w(s)$. Since for all $s \in P_{ij}$ we have $w(s) = \frac{1}{m}|P_{ij}|$, it follows that function $\hat{w}$ has only integral weights $\hat{w}(s) = |P_{ij}|$. Hence, there is a one-to-one correspondence to a run of algorithm Cluster on unweighted input multiset $\hat{S}$ where each $s \in S$ is replaced by $\hat{w}(s)$ copies of s. Since splitting the weights of a point from S corresponds to the situation when some points from $\hat{S}$ are pruned and some are not, we find that the weights of $\hat{w}$ remain integral during a run of algorithm Cluster. Hence, there will be at most $|P_{ij}| \leq n$ different weights assigned to point $s \in P_{ij}$.

We conclude that the number W of different weight functions assigned to a fixed $\mathcal{O}(\varepsilon^{-1}\mu^{-1})$-sized subsets of input set S is bounded by

$$W \leq n^{\mathcal{O}(\varepsilon^{-1}\mu^{-1})} \ . \tag{7.37}$$

Thus, the number of different $\mathcal{O}(\varepsilon^{-1}\mu^{-1})$-sized subsets with different weight functions is at most

$$NW \leq \left(n^{\mathcal{O}(\varepsilon^{-1}\mu^{-1})}\right)^2 = n^{\mathcal{O}(\varepsilon^{-1}\mu^{-1})} \ . \tag{7.38}$$

Of course, the same bound applies to the number of weighted centroids of $\mathcal{O}(\varepsilon^{-1}\mu^{-1})$-sized subsets of S. Thus, inequality (7.38) provides a bound to the number of all possible output points of algorithm Cluster. □

In addition to the observations above, we also want to compare the solutions obtained by algorithm Cluster to the optimal k-medians of input set P. Thus, the optimal k-medians of P are also considered to be relevant to our approach.

Now, we can give the definition of Γ explicitly. Let $\Gamma_{P,A}$ be as given by Lemma 7.7 and let C_P denote the optimal k-medians of P. Then we define

$$\Gamma = \Gamma_{P,A} \cup C_P \ . \tag{7.39}$$

```
CoreCluster($R, k$):
  $P$  set of input points
  $k$  number of medians to be found
1: Obtain $\mathcal{O}(\mu^{-2}\log k)$-approximation $A$ using algorithm BregMeans++($P, k$) from Chapter 5
2: Build $\Gamma$-weak $(k, \varepsilon)$-coreset $S$ of $P$, using $A$ and Chen's coreset construction with $|\Gamma| \leq n^{\mathcal{O}(\varepsilon^{-1}\mu^{-1})}$.
3: Run algorithm WeightedCluster on weighted input set $S$ to obtain $(1+\varepsilon)$-approximate $k$-median set $\tilde{C}$.
4: Return $\tilde{C}$ as set of approximate medians for $P$.
```

Figure 7.2: The clustering algorithm using Γ-weak coresets

We obtain the following bound on $|\Gamma|$.

Lemma 7.8. $|\Gamma| \leq n^{\mathcal{O}(\varepsilon^{-1}\mu^{-1})}$

Proof. We know $|C_P| = k$. By Lemma 7.7 we have $|\Gamma_{P,A}| \leq n^{\mathcal{O}(\varepsilon^{-1}\mu^{-1})}$. Since $k \leq n$ we obtain

$$|\Gamma| \leq n^{\mathcal{O}(\varepsilon^{-1}\mu^{-1})} + k \leq n^{\mathcal{O}(\varepsilon^{-1}\mu^{-1})} \ . \tag{7.40}$$

□

Note that the definition of Γ depends only on P and A, and is independent of the random choices made during the construction of coreset S. Also note that we do not have to know the exact content of set Γ to construct a Γ-weak coreset using Chen's construction: We only need a size bound on Γ, and this bound is given by Lemma 7.8.

Our new approximation algorithm for the Bregman k-median clustering problem is summarized in Figure 7.2. We call this algorithm CoreCluster. Using algorithm BregMeans++ from Chapter 5 we obtain an initial $[\alpha, \beta]$-bicriteria approximation with $\alpha = \mathcal{O}(\mu^{-2}\log k)$ and $\beta = 1$. Hence, from Theorem 7.2 and Lemma 7.8 we learn that S is a Γ-weak coreset S of size

$$|S| = \mathcal{O}\left(\varepsilon^{-3}\mu^{-5}k^2\log^2(k)\log^2(n)\right) \ . \tag{7.41}$$

That is, the size of S is independent of the dimension d and depends only polylogarithmically on the size n of the input set P.

In the following theorem we prove that with constant probability algorithm CORECLUSTER indeed computes a $(1+\varepsilon)$-approximate solution of the μ-similar Bregman k-median problem.

Theorem 7.9. *Let $0 < \varepsilon \leq \frac{1}{2}$. With constant probability, algorithm* CORECLUSTER *computes a solution $\tilde{C}$ of the k-median problem with respect to μ-similar Bregman divergence D_φ for input instance P satisfying*

$$\mathrm{cost}(P, \tilde{C}) \leq (1 + 7\varepsilon)\, opt_k(P) \ . \tag{7.42}$$

Proof. Since each of the steps 1–3 of algorithm CORECLUSTER succeeds at least with constant probability, we may assume that with constant probability steps 1–3 yield the desired result.

Let C_P denote the optimal k-medians for P and let C_S denote the optimal k-medians for weighted set S, i.e. $\mathrm{cost}(P, C_P) = opt_k(P)$ and $\mathrm{cost}_w(S, C_S) = opt_k(S, w)$. Using $\tilde{C} \subseteq \Gamma_{P,A}$ and the fact that S is a Γ-weak (k, ε)-coreset we obtain from inequality (7.1)

$$\mathrm{cost}(P, \tilde{C}) \leq \frac{1}{1-\varepsilon}\, \mathrm{cost}_w(S, \tilde{C}) \ . \tag{7.43}$$

Since $\tilde{C}$ is a $(1+\varepsilon)$-approximation for weighted input set S we get

$$\mathrm{cost}(P, \tilde{C}) \leq \frac{1+\varepsilon}{1-\varepsilon}\, \mathrm{cost}_w(S, C_S) \leq \frac{1+\varepsilon}{1-\varepsilon}\, \mathrm{cost}_w(S, C_P). \tag{7.44}$$

Using $C_P \subseteq \Gamma$ and inequality (7.1) we obtain

$$\mathrm{cost}(P, \tilde{C}) \leq \frac{(1+\varepsilon)^2}{1-\varepsilon}\, \mathrm{cost}(P, C_P) \leq (1 + 7\varepsilon)\, opt_k(P) \tag{7.45}$$

since $\frac{(1+\varepsilon)^2}{1-\varepsilon} \leq 1 + 7\varepsilon$ for $\varepsilon \leq \frac{1}{2}$. □

We still have to give a bound on the running time. In the following theorem we show that algorithm CORECLUSTER has indeed an improved running time compared to the running time of algorithm CLUSTER from Chapter 4.

Theorem 7.10. *Let $P \subseteq \mathbb{X}$ of size $|P| = n$. Algorithm* CORECLUSTER *started with parameters (P, k) and fixed parameter $\varepsilon > 0$ requires at most $\mathcal{O}(kn) + 2^{\mathcal{O}(k\varepsilon^{-1}\mu^{-1}\log(k\varepsilon^{-1}\mu^{-1}))} \log^{k+2}(n)$ arithmetic operations, including evaluations of D_φ.*

Proof. Let $T(n,k)$ denote the running time of algorithm CORECLUSTER started with n input points and k approximate medians to be found. From Chapter 5 we know that we can obtain approximation A using at most $\mathcal{O}(kn)$ arithmetic operations. Theorem 7.2 states that coreset S can be constructed using at most $\mathcal{O}(kn+|S|)$ operations. Furthermore, recall that a μ-similar Bregman divergence D_φ satisfies the $[\gamma,\delta]$-sampling property with $m_{\gamma,\delta} = \mathcal{O}(\varepsilon^{-1}\mu^{-1})$ for $\gamma = \varepsilon/3$ and constant δ (cf. Lemma 4.21). Thus, according to Theorem 4.15, algorithm WEIGHTEDCLUSTER started with coreset S of total weight $w(S) = n$ requires at most

$$T(n,k) = 2^{\mathcal{O}(k\varepsilon^{-1}\mu^{-1}\log(k\varepsilon^{-1}\mu^{-1}))}|S|\log^k(n) \tag{7.46}$$

arithmetic operations. In addition to that, from equation (7.41) we know that $|S| = \mathcal{O}\left(\varepsilon^{-3}\mu^{-5}k^2\log^2(k)\log^2(n)\right)$. We conclude

$$T(n,k) = 2^{\mathcal{O}(k\varepsilon^{-1}\mu^{-1}\log(k\varepsilon^{-1}\mu^{-1}))}\log^{k+2}(n)\ . \tag{7.47}$$

□

Hence, as long as the number of clusters k is small compared to the number of input points, i.e., $k = o(\log n/\operatorname{loglog} n)$, the running time of algorithm CORECLUSTER improves significantly over the running time of algorithm CLUSTER from Chapter 4. Theorem 7.3 is an immediate consequence of Theorem 7.9 and Theorem 7.10.

7.3 Discussion

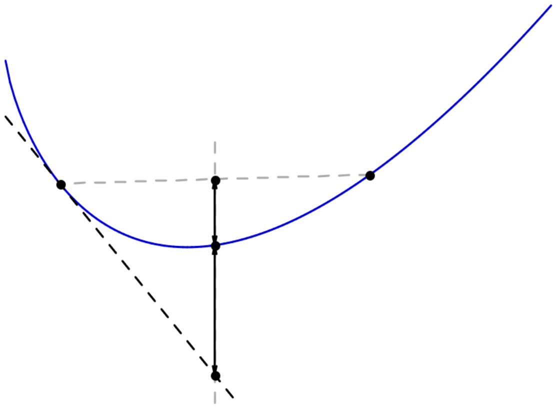
In this chapter we introduced our concept of weak coresets. We have shown that there exist small weak coresets for the μ-similar Bregman k-median problem, and we have shown how to construct such a weak coreset explicitly using uniform sampling. Furthermore, we have shown how to use weak coresets to significantly speed up algorithm CLUSTER from Chapter 4. In doing so, we presented the asymptotically fastest algorithm currently known for the k-median problem with respect to μ-similar Bregman divergences. Due to the low dependency of the running time on d this algorithm is particularly relevant for high-dimensional settings.

We should also mention that our application of weak coresets does not only generalize the result from [Chen, 2009] considering the Euclidean k-means problem (as an instance of the Mahalanobis k-median problem), but the use of weak coresets even improves Chen's result by a factor of d. In [Chen, 2009], the size of the strong coresets obtained is linear in d. This leads to a running time of $\mathcal{O}(dkn) + d^2 2^{(k/\varepsilon)^{\mathcal{O}(1)}} \log^{k+2}(n)$. Since the size of our weak coresets is independent of dimension d and the Euclidean distance can be evaluated in time $\mathcal{O}(d)$, we obtain a running time that depends only linearly on d.

It is noteworthy that our notion of weak coresets can be applied to any μ-similar Bregman k-median clustering algorithm if the combinatorial complexity of the algorithm's possible outputs is small. For instance, assume that the input set P and a constant factor bicriteria approximation A are fixed. Let $\Gamma_{P,A}$ denote the union of every possible output point of a given clustering algorithm started for every potential weighted (k, ε)-coreset S from Chen's coreset construction applied to P and A. If $f(n, k, d, \varepsilon)$ is a function such that $|\Gamma_{P,A}| \leq f(n, k, d, \varepsilon)$ for every P of size n and every A of size $\mathcal{O}(k)$, then Chen's construction yields a weak (k, ε)-coreset of size $(k/\varepsilon)^{\mathcal{O}(1)} \log(n) \log\bigl(f(n, k, d, \varepsilon)\bigr)$. Hence, weak coresets can be used to speed up any given clustering algorithm, as long as the number of potential output center points of the algorithm is small enough. Unfortunately, algorithm Cluster from Chapter 4 and CoreCluster from this chapter are the only $(1 + \varepsilon)$-approximation algorithms currently known for the Bregman k-median problem.

Of course, some open problems remain. First of all, it is still unknown whether there exist any strong coresets for the μ-similar Bregman k-median problem. Considering the result from [Feldman et al., 2007] (or even [Har-Peled and Kushal, 2005]) the question arises whether one can construct weak (or even strong) coresets for the μ-similar Bregman k-median problem that are independent of the size of the point set n. Also, in the light of the results from Chapter 6, are there strong coresets for the Mahalanobis k-median problem of size independent of d? Furthermore, it remains an open problem to construct coresets for non-similar Bregman divergences with singularities in their domain.

8 On the limits of using uniform sampling



Many algorithms, like our algorithm Cluster from Chapter 4 and algorithm BregMeans++ from Chapter 5, rely on the power of random sampling to obtain good approximate solutions to the Bregman k-median problem with high probability. In particular, we have shown in Chapter 4 that we can always give a linear time approximation scheme, provided that the 1-median problem can be approximated within a factor of $(1+\varepsilon)$ by inspection of a uniform random sample set of merely constant size. Furthermore, we have shown that for all μ-similar Bregman divergences, with high probability a sample set of size $\mathcal{O}(\varepsilon^{-1}\mu^{-1})$ is indeed sufficient to obtain a $(1+\varepsilon)$-approximate 1-median.

Naturally, the question arises whether a property like μ-similarity is actually necessary to achieve such a sampling result for a given Bregman divergence D_φ. In this chapter, we provide strong evidence that this is indeed the case. More precisely, in Section 8.1 we show that if uniform sampling provides a constant factor approximation guarantee for finding the 1-median of any input set, then the domain $\mathbb{X}$ of D_φ has to be free of any singularities (that is, there must not exist $p, q \in \mathbb{X}$ with $\mathrm{D}_\varphi(p, q) = \infty$), and it is even unlikely that the domain may come arbitrarily close towards having such singularities. As we already know from Lemma 2.17, this implies μ-similarity for some constant $\mu > 0$ (under some mild assumptions on the second derivatives of φ). Hence, this observation provides strong evidence for the conjecture that, in fact, for all Bregman divergences sampleability implies μ-similarity. One consequence of this observation is that techniques that rely on approximation by constant sized uniform sampling (like algorithm Cluster from Chapter 4) are limited to the case of μ-similar Bregman divergences.

In addition, in Section 8.2 we show that the intuition from Section 8.1 can be made explicit by taking into account the concrete analytical properties of a given generating function φ. In detail, for some Bregman divergences we show explicitly that the assumption of sampleability indeed implies μ-similarity, namely for the Kullback-Leibler divergence and the Itakura-Saito divergence.

8.1 Sampleable Bregman divergences avoid singularities

The sampling property from Chapter 4 assures that with probability at least $1 - \delta$ the centroid of a constant number of sample points chosen uniformly at random from input set P is a $(1 + \varepsilon)$-approximation of the 1-median of P. In this chapter, we use a slightly different notion of sampleability that focuses on the use of a single uniform sample point. In detail, we make use of the following formal definition.

Definition 8.1. *A dissimilarity measure* D *on domain* $\mathbb{X}$ *is called* (γ, δ)*-sampleable if for all finite* $P \subseteq \mathbb{X}$ *and a single uniform sample point* $s \in P$ *we have*

$$\Pr\left[\mathrm{cost}(P, s) \leq (1 + \gamma)\, opt_1(P)\right] \geq 1 - \delta \ . \tag{8.1}$$

Please note that in the case of μ-similar Bregman divergences this notion of sampleability differs from the sampling property of Chapter 4 only in the quantity of the constants γ and δ. That is, obviously, if D_φ is (γ,δ)-sampleable, then D_φ also satisfies the $[\gamma,\delta]$-sampling property from Chapter 4 using a sample of size $m = 1$. On the other hand, if — according to Lemma 4.20 — we have the $[\gamma,\delta]$-sampling property with $m \geq \frac{1}{\gamma\delta\mu}$, then we have that D_φ is also (γ',δ')-sampleable as long as $\gamma' \geq \frac{1}{\delta'\mu}$.

In this section, our goal is to provide evidence that the domain of a sampleable Bregman divergence necessarily avoids all singularities. To this end, let us consider a (γ,δ)-sampleable Bregman divergence D_φ on domain $\mathbb{X}$. Furthermore, let $n \in \mathbb{N}$ be such that

$$n > \frac{1}{\delta} \ . \tag{8.2}$$

For arbitrary $p, q \in \mathbb{X}$ with $p \neq q$ we consider a multiset $P \subseteq \mathbb{X}$ of size n consisting of one copy of point p and $n-1$ copies of point q. Recall that we know from Lemma 3.6 that the centroid

$$c_P = \frac{1}{n}\,p + \frac{n-1}{n}\,q \tag{8.3}$$

is the optimal 1-median of P. By construction of P, we obtain the following lemma.

Lemma 8.2. *Let D_φ be (γ,δ)-sampleable, and let p, q, c_P, P, and n be as above. Then we have*

$$\mathrm{D}_\varphi(c_P, q) \leq \frac{\gamma}{n}\, opt_1(P) \ . \tag{8.4}$$

Proof. Let

$$X = \left\{ x \in P \,\middle|\, \mathrm{D}_\varphi(c_P, x) \leq \frac{\gamma}{n}\, opt_1(P) \right\} \ . \tag{8.5}$$

From the central identity of Lemma 3.5 we know that the elements of X are exactly the points from P that are γ-approximate medians of P. Since D_φ is (γ,δ)-sampleable and $\delta > 1/n$, we know that

$$|X| \geq \delta n > 1 \ . \tag{8.6}$$

That is, X contains at least two points from P. Thus, we have $q \in X$, and the lemma follows. □

Furthermore, using the sampleability of D_φ and the result from Lemma 8.2 we observe the following property of the generating function φ.

Lemma 8.3. *Let* D_φ *be* (γ, δ)*-sampleable, and let* p*,* q*,* c_P*,* P*, and* n *be as above. Then we have*

$$\varphi(c_P) - \left(\varphi(q) + \frac{1}{n}\nabla\varphi(q)^\top(p-q)\right) \le \gamma\left(\frac{\varphi(p) + (n-1)\varphi(q)}{n} - \varphi(c_P)\right) . \tag{8.7}$$

Proof. From Lemma 8.2 we obtain

$$\varphi(c_P) - \varphi(q) - \nabla\varphi(q)^\top(c_P - q) = \mathrm{D}_\varphi(c_P, q) \le \frac{\gamma}{n}\, opt_1(P) . \tag{8.8}$$

Furthermore, we have

$$\begin{aligned} opt_1(P) &= \mathrm{D}_\varphi(p, c_P) + (n-1)\,\mathrm{D}_\varphi(q, c_P) &(8.9)\\ &= \varphi(p) - \varphi(c_P) - \nabla\varphi(c_P)^\top(p - c_P) \\ &\quad + (n-1)\big(\varphi(q) - \varphi(c_P) - \nabla\varphi(c_P)^\top(q - c_P)\big) . &(8.10)\end{aligned}$$

Since $(p - c_P) = \frac{n-1}{n}(p-q)$ and $(q - c_P) = -\frac{1}{n}(p-q)$ we obtain

$$\nabla\varphi(c_P)^\top(p - c_P) = \frac{n-1}{n}\nabla\varphi(c_P)^\top(p-q) = -(n-1)\nabla\varphi(c_P)^\top(q - c_P) \tag{8.11}$$

and

$$opt_1(P) = \varphi(p) + (n-1)\varphi(q) - n\varphi(c_P) . \tag{8.12}$$

Using (8.8) and (8.12) we conclude

$$\varphi(c_P) - \varphi(q) - \nabla\varphi(q)^\top(c_P - q) \le \frac{\gamma}{n}\big(\varphi(p) + (n-1)\varphi(q) - n\varphi(c_P)\big) \tag{8.13}$$

or, equivalently,

$$\varphi(c_P) - \left(\varphi(q) + \frac{1}{n}\nabla\varphi(q)^\top(p-q)\right) \le \gamma\left(\frac{\varphi(p) + (n-1)\varphi(q)}{n} - \varphi(c_P)\right) \tag{8.14}$$

since $c_P - q = \frac{1}{n}(p-q)$. □

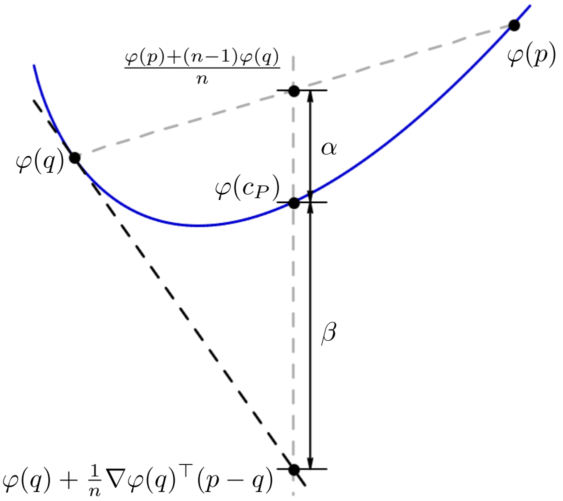

$\frac{\varphi(p)+(n-1)\varphi(q)}{n}$
$\varphi(p)$
α
$\varphi(q)$
$\varphi(c_P)$
β
$\varphi(q) + \frac{1}{n}\nabla\varphi(q)^\top(p-q)$

Figure 8.1: A geometric interpretation of Lemma 8.3. For all points p, q from domain $\mathbb{X}$, the ratio β/α has to be bounded by γ.

Lemma 8.3 has an interesting geometric interpretation considering the graph of the generating function φ. Observe that the left-hand side of inequality (8.7) gives the difference between the graph of φ and the tangent of φ through point q, evaluated at point c_P. On the other hand, the right-hand side of inequality (8.7) depends on the difference between the graph of φ and the chord connecting $\big(p, \varphi(p)\big)$ and $\big(q, \varphi(q)\big)$, again evaluated at point c_P. Thus, the ratio of both differences gives information about the curvature of the function φ. Now, recall that the result of Lemma 8.3 states that if D_φ is (γ, δ)-sampleable, then these differences are within a factor of γ from each other. Hence, a Bregman divergence on domain $\mathbb{X}$ is not sampleable unless the curvature of φ on domain $\mathbb{X}$ is bounded.

Intuitively, this observation immediately rules out the existence of singularities on the domain of a sampleable Bregman divergence. To see this connection, assume that there exist singularities on $\mathbb{X}$. That is, there exists a point $q \in \mathbb{X} \setminus \mathrm{ri}(\mathbb{X})$ such that the directional derivatives at point q tend to $\pm\infty$. Furthermore, let $\{q_\nu\}_{\nu\in\mathbb{N}}$ be an infinite sequence of points from $\mathrm{ri}(\mathbb{X})$ with $q_\nu \to q$ for $\nu \to \infty$. Hence, for $\nu \to \infty$ the difference between the graph and the tangent through q_ν must converge to ∞ and, eventually, violates inequality (8.7). We obtain that such a point q may not exist in $\mathbb{X}$. In addition to that, since the ratio of both differences is bounded by constant γ, we even find that domain $\mathbb{X}$ may not come arbitrarily close to these singularities.

This geometric interpretation of Lemma 8.3 is depicted in Figure 8.1. As we will see in the next section, the intuition given above can be made explicit when considering concrete Bregman divergences such as the Kullback-Leibler divergence and the Itakura-Saito divergence.

8.2 Explicit domain bounds for some Bregman divergences

In this section, we show that the intuition from Section 8.1 can be made explicit by taking into account the concrete analytical properties of a given generating function φ. To this end, we apply the observation from Lemma 8.3 to two concrete Bregman divergences, namely the Kullback-Leibler divergence and the Itakura-Saito divergence. Under the assumption of sampleability we derive bounds on the domain of these Bregman divergences. In doing so we show that their domain avoids all singularities, and that in the case of the Kullback-Leibler divergence and the Itakura-Saito divergence, sampleability indeed implies μ-similarity.

Example 1: Kullback-Leibler divergence. We show the following result for the Kullback-Leibler divergence

$$\mathrm{D}_{\mathrm{KL}}(p, q) = \sum_{i=1}^{d} \left(p_i \ln \frac{p_i}{q_i} - p_i + q_i \right) \ , \tag{8.15}$$

where $p = (p_1, p_2, \ldots, p_d)^\top \in \mathbb{R}^d_{\geq 0}$ and $q = (q_1, q_2, \ldots, q_d)^\top \in \mathbb{R}^d_{\geq 0}$.

Lemma 8.4. *If* D_{KL} *on domain* $[\lambda, \upsilon]^d$ *with* $0 \leq \lambda < \upsilon$ *is* (γ, δ)*-sampleable, then we have*

$$\frac{\lambda}{\upsilon} > \frac{1}{2} \left(\frac{\delta}{e(1+\delta)} \right)^{1+\gamma} \tag{8.16}$$

where $e \approx 2.718\ldots$ *denotes the base of the natural logarithm.*

Proof. Without loss of generality, we may assume $\lambda < \upsilon/2$ since otherwise the claim is trivially true. Let $n \geq 2$ be the unique integer such that $n-1 \leq 1/\delta < n$. In the following, we consider an input multiset $P \subseteq [\lambda, \upsilon]^d$

of size n consisting of one copy of point $p = (\upsilon, a, \ldots, a)^\top$ and $n-1$ copies of point $q = (\lambda, a, \ldots, a)^\top$, where a denotes an arbitrary real number with $\lambda < a < \upsilon$. From Lemma 3.6 we know that the centroid

$$c_P = \frac{1}{n}p + \frac{n-1}{n}q = \left(\frac{\upsilon + (n-1)\lambda}{n}, a, \ldots, a\right)^\top \tag{8.17}$$

is the optimal 1-median of P.

From Lemma 8.2 above we obtain

$$\mathrm{D_{KL}}(c_P, q) \leq \frac{\gamma}{n}\, opt_1(P)\ . \tag{8.18}$$

Considering the left-hand side of inequality (8.18) we find that

$$\mathrm{D_{KL}}(c_P, q) = \frac{\upsilon + (n-1)\lambda}{n} \ln\left(\frac{\upsilon + (n-1)\lambda}{n\lambda}\right) - \frac{\upsilon + (n-1)\lambda}{n} + \lambda \tag{8.19}$$

$$= \frac{\upsilon + (n-1)\lambda}{n} \ln\left(1 + \frac{\upsilon - \lambda}{n\lambda}\right) - \frac{\upsilon - \lambda}{n} \tag{8.20}$$

$$\geq \frac{\upsilon}{n} \ln\left(1 + \frac{\delta(\upsilon - \lambda)}{(1+\delta)\lambda}\right) - \frac{\upsilon - \lambda}{n} \tag{8.21}$$

since $\upsilon + (n-1)\lambda \geq \upsilon$ and $1/n \geq \delta/(1+\delta)$. On the other hand, considering the right-hand side of inequality (8.18), we derive an upper bound on $opt_1(P)$. First, we find that

$$\mathrm{D_{KL}}(p, c_P) = \upsilon \ln\left(\frac{n\upsilon}{\upsilon + (n-1)\lambda}\right) - \upsilon + \frac{\upsilon + (n-1)\lambda}{n} \tag{8.22}$$

$$< \upsilon \ln\left(\frac{n\upsilon}{\upsilon + (n-1)\lambda}\right) \tag{8.23}$$

$$\leq \upsilon \ln(n) \tag{8.24}$$

since $\frac{\upsilon+(n-1)\lambda}{n} < \upsilon$ for $\lambda < \upsilon$ and $\frac{n\upsilon}{\upsilon+(n-1)\lambda} \leq n$ for $\lambda \geq 0$. We also find

$$\mathrm{D_{KL}}(q, c_P) = \lambda \ln\left(\frac{n\lambda}{\upsilon + (n-1)\lambda}\right) - \lambda + \frac{\upsilon + (n-1)\lambda}{n} \tag{8.25}$$

$$< \frac{\upsilon - \lambda}{n} \tag{8.26}$$

since $\lambda \ln\left(\frac{n\lambda}{\upsilon+(n-1)\lambda}\right) < 0$ for $0 \leq \lambda < \upsilon$. Hence,

$$\begin{aligned} opt_1(P) &= \mathrm{D}_{\mathrm{KL}}(p, c_P) + (n-1)\,\mathrm{D}_{\mathrm{KL}}(q, c_P) && (8.27) \\ &< \upsilon \ln(n) + \frac{n-1}{n}(\upsilon - \lambda) && (8.28) \\ &< \upsilon \ln\left(\frac{1+\delta}{\delta}\right) + \upsilon - \lambda && (8.29) \end{aligned}$$

since $n \leq (1+\delta)/\delta$ and $(n-1)/n < 1$.

Using inequalities (8.18), (8.21), and (8.29) we find

$$\upsilon \ln\left(1 + \frac{\delta(\upsilon-\lambda)}{(1+\delta)\lambda}\right) - (\upsilon - \lambda) < \gamma\upsilon \ln\left(\frac{1+\delta}{\delta}\right) + \gamma(\upsilon - \lambda) \quad (8.30)$$

or, equivalently,

$$\upsilon \ln\left(1 + \frac{\delta(\upsilon-\lambda)}{(1+\delta)\lambda}\right) < \gamma\upsilon \ln\left(\frac{1+\delta}{\delta}\right) + (1+\gamma)(\upsilon - \lambda)\ . \quad (8.31)$$

Using $\upsilon/2 < \upsilon - \lambda < \upsilon$ for $0 < \lambda < \upsilon/2$ we obtain

$$\ln\left(1 + \frac{\delta\upsilon}{2(1+\delta)\lambda}\right) < \gamma \ln\left(\frac{1+\delta}{\delta}\right) + 1 + \gamma\ . \quad (8.32)$$

Now, we apply the exponential function to both sides. We conclude

$$\frac{\delta\upsilon}{2(1+\delta)\lambda} < 1 + \frac{\delta\upsilon}{2(1+\delta)\lambda} < \left(\frac{1+\delta}{\delta}\right)^{\gamma} \cdot e^{1+\gamma} \quad (8.33)$$

and, thus,

$$\frac{\lambda}{\upsilon} > \frac{1}{2}\left(\frac{\delta}{e(1+\delta)}\right)^{1+\gamma}\ . \quad (8.34)$$

□

Note that the proof of Lemma 8.4 can be easily generalized to the case of an arbitrary convex domain $\mathbb{X}$ with smallest enclosing bounding box $[\lambda, \upsilon]^d$. As we have already learned from Lemma 2.19, we know that D_{KL} is a (λ/υ)-similar Bregman divergence, provided that $\lambda/\upsilon > 0$ as is implied by Lemma 8.4. Hence, we obtain the following corollary for the Kullback-Leibler divergence.

Corollary 8.5. *If the Kullback-Leibler divergence D_{KL} is (γ, δ)-sampleable on domain $\mathbb{X}$ with constants $\gamma, \delta > 0$, then we have that D_{KL} on domain $\mathbb{X}$ is μ-similar for some constant $\mu > 0$.*

Example 2: Itakura-Saito divergence. In analogy to the result for the Kullback-Leibler divergence, we show the following result for the Itakura-Saito divergene

$$\mathrm{D}_{\mathrm{IS}}(p,q)=\sum_{i=1}^{d}\left(\frac{p_i}{q_i}-\ln\frac{p_i}{q_i}-1\right)\ , \tag{8.35}$$

where $p=(p_1,p_2,\dots,p_d)^\top\in\mathbb{R}^d_{\geq 0}$ and $q=(q_1,q_2,\dots,q_d)^\top\in\mathbb{R}^d_{\geq 0}$.

Lemma 8.6. *If* D_{IS} *on domain* $[\lambda,\upsilon]^d$ *with* $0\leq\lambda<\upsilon$ *is* (γ,δ)*-sampleable, then we have*

$$\frac{\lambda^2}{\upsilon^2}\geq\frac{\delta^8}{256\gamma^4(1+2\delta)^4}\ . \tag{8.36}$$

Proof. Without loss of generality, we may assume $\lambda<\upsilon/4$ since otherwise the claim is trivially true. Let $n\geq 2$ be the unique integer such that $n-1\leq 1/\delta<n$. In the following, we consider an input multiset $P\subseteq[\lambda,\upsilon]^d$ of size n consisting of one copy of point $p=(\upsilon,a,\dots,a)^\top$ and $n-1$ copies of point $q=(\lambda,a,\dots,a)^\top$, where a denotes an arbitrary real number with $\lambda<a<\upsilon$. From Lemma 3.6 we know that the centroid

$$c_P=\frac{1}{n}p+\frac{n-1}{n}q=\left(\frac{\upsilon+(n-1)\lambda}{n},a,\dots,a\right)^\top \tag{8.37}$$

is the optimal 1-median of P.

From Lemma 8.2 above we obtain

$$\mathrm{D}_{\mathrm{IS}}(c_P,q)\leq\frac{\gamma}{n}\,opt_1(P)\ . \tag{8.38}$$

Considering the left-hand side of inequality (8.38) we find that

$$\begin{aligned}\mathrm{D}_{\mathrm{IS}}(c_P,q)&=\frac{\upsilon+(n-1)\lambda}{n\lambda}-\ln\left(\frac{\upsilon+(n-1)\lambda}{n\lambda}\right)-1 &(8.39)\\ &=\left(1+\frac{\upsilon-\lambda}{n\lambda}\right)-\ln\left(1+\frac{\upsilon-\lambda}{n\lambda}\right)-1\ . &(8.40)\end{aligned}$$

Note that

$$1+\frac{\upsilon-\lambda}{n\lambda}>1+\frac{1}{n} \tag{8.41}$$

since $\frac{\upsilon-\lambda}{\lambda} > 1$ for $\lambda < \upsilon/4$. Using Claim 8.7 stated below with $\rho = 1 + \frac{1}{n}$ and $t = 1 + \frac{\upsilon-\lambda}{n\lambda}$ we obtain

$$\begin{aligned} \mathrm{D_{IS}}(c_P, q) &\geq (t-\rho) \cdot \frac{\rho - 1}{\rho} && (8.42)\\ &= \frac{\upsilon - 2\lambda}{n\lambda} \cdot \frac{1}{n+1} && (8.43)\\ &> \frac{1}{2n(n+1)} \cdot \frac{\upsilon}{\lambda} && (8.44) \end{aligned}$$

for $\lambda < \upsilon/4$. On the other hand, considering the right-hand side of inequality (8.38), we derive an upper bound on $opt_1(P)$. First, we find that

$$\begin{aligned} \mathrm{D_{IS}}(p, c_P) &= \frac{n\upsilon}{\upsilon + (n-1)\lambda} - \ln\left(\frac{n\upsilon}{\upsilon + (n-1)\lambda}\right) - 1 && (8.45)\\ &< n - \ln(1) - 1 && (8.46)\\ &= n - 1 && (8.47) \end{aligned}$$

since $1 < \frac{n}{\upsilon+(n-1)\lambda} < n$ for $0 < \lambda < \upsilon$. At the same time, we find

$$\begin{aligned} \mathrm{D_{IS}}(q, c_P) &= \frac{n\lambda}{\upsilon + (n-1)\lambda} - \ln\left(\frac{n\lambda}{\upsilon + (n-1)\lambda}\right) - 1 && (8.48)\\ &< 1 - \ln\left(\frac{\lambda}{\upsilon}\right) - 1 && (8.49)\\ &= \ln\left(\frac{\upsilon}{\lambda}\right) && (8.50) \end{aligned}$$

since $\frac{\lambda}{\upsilon} < \frac{n\lambda}{\upsilon+(n-1)\lambda} < 1$ for $0 < \lambda < \upsilon$. Hence,

$$\begin{aligned} opt_1(P) &= \mathrm{D_{IS}}(p, c_P) + (n-1)\,\mathrm{D_{IS}}(q, c_P) && (8.51)\\ &< n - 1 + (n-1)\ln\left(\frac{\upsilon}{\lambda}\right) && (8.52)\\ &< 2(n-1)\ln\left(\frac{\upsilon}{\lambda}\right) && (8.53) \end{aligned}$$

since $1 < \ln(\upsilon/\lambda)$ for $\lambda < \upsilon/4$. Using inequalities (8.38), (8.44), and (8.53) we find

$$\frac{1}{2(n+1)} \cdot \frac{\upsilon}{\lambda} < 2\gamma(n-1)\ln\left(\frac{\upsilon}{\lambda}\right) \qquad (8.54)$$

or, equivalently,

$$\frac{\upsilon}{\lambda} < 4\gamma(n-1)(n+1)\ln\left(\frac{\upsilon}{\lambda}\right) \tag{8.55}$$

$$\leq 4\gamma\frac{1+2\delta}{\delta^2}\ln\left(\frac{\upsilon}{\lambda}\right) \ , \tag{8.56}$$

where the last inequality is due to $n-1 \leq 1/\delta$. Using $\ln(t) \leq \sqrt{t}$ for all $t \geq 0$ we conclude

$$\frac{\upsilon}{\lambda} < 4\gamma\frac{1+2\delta}{\delta^2}\sqrt{\frac{\upsilon}{\lambda}} \tag{8.57}$$

which leads to

$$\frac{\lambda^2}{\upsilon^2} > \frac{\delta^8}{256\gamma^4(1+2\delta)^4} \ . \tag{8.58}$$

□

Claim 8.7. *Let $\rho > 1$ be arbitrary. Then for all $t \geq \rho$ we have*

$$t - \ln(t) - 1 \geq (t-\rho)\frac{\rho-1}{\rho} \ . \tag{8.59}$$

Proof. In the following, let $f : \mathbb{R}_{\geq 0} \to \mathbb{R}$ denote the function given by

$$f(t) = t - \ln(t) - 1 \tag{8.60}$$

for all $t \geq 0$, and let

$$f'(t) = 1 - \frac{1}{t} = \frac{t-1}{t} \tag{8.61}$$

denote the first-order derivative of f. Note that since f' is strictly increasing on $\mathbb{R}_{\geq 0}$ we have that f is strictly convex on $\mathbb{R}_{\geq 0}$. Furthermore, let $g : \mathbb{R}_{\geq 0} \to \mathbb{R}$ denote the tangent of f at point ρ, that is,

$$g(t) = f(\rho) + (t-\rho)f'(\rho) = \rho - \ln(\rho) - 1 + (t-\rho)\frac{\rho-1}{\rho} \ . \tag{8.62}$$

Since f is strictly convex and $\rho - \ln(\rho) - 1 \geq 0$ we know that for all $t \geq 0$ we have

$$f(t) \geq g(t) \geq (t-\rho)\frac{\rho-1}{\rho} \ . \tag{8.63}$$

□

As in the case of the Kullback-Leibler divergence, the proof of Lemma 8.6 can be easily generalized to the case of an arbitrary convex domain $\mathbb{X}$ with smallest enclosing bounding box $[\lambda, \upsilon]^d$. We already know from Lemma 2.20 that D_{IS} is a (λ^2/υ^2)-similar Bregman divergence, provided that $\lambda^2/\upsilon^2 > 0$ as is implied by Lemma 8.6. Hence, we obtain the following corollary for the Itakura-Saito divergence.

Corollary 8.8. *If the Itakura-Saito divergence D_{IS} is (γ, δ)-sampleable on domain $\mathbb{X}$ with constants $\gamma, \delta > 0$, then we have that D_{IS} on domain $\mathbb{X}$ is μ-similar for some constant $\mu > 0$.*

8.3 Discussion

In this Chapter, we studied the properties of a sampleable Bregman divergence D_φ. In particular, we have shown that if with high probability a constant factor approximate solution of the 1-median problem can be obtained by sampling a single point uniformly at random, then D_φ has no singularities on its domain. In doing so, we have given strong evidence that any sampleable Bregman divergences is also μ-similar for some constant $\mu > 0$.

We also have shown that the intuition above can be made explicit by taking into account the concrete analytical properties of a given generating function φ. Under the assumption of sampleability we provided bounds on the domain of two concrete Bregman divergences, namely the Kullback-Leibler divergence and the Itakura-Saito divergence. In doing so we showed that their domain avoids all singularities, and that in the case of the Kullback-Leibler divergence and the Itakura-Saito divergence, sampleability indeed implies μ-similarity.

However, we did not give a formal proof in the case of a general Bregman divergence with an arbitrary generating function φ. It remains an open problem to prove an analogous result in this general setting. Nevertheless, the author of this thesis is convinced that such a result can be obtained in the case of a general Bregman k-median problem under only mild assumptions on the second derivatives of φ. Hence, based on the observations made in this chapter, we state the following conjecture.

Conjecture 8.9. *Let D_φ be a Bregman divergence on domain $\mathbb{X}$ such that φ is twice differantiable, and that the Hessian $\nabla^2\varphi(t)$ is symmetric positive definite for all $t \in \mathbb{X}$. If D_φ is (γ, δ)-sampleable on domain $\mathbb{X}$ with constants*

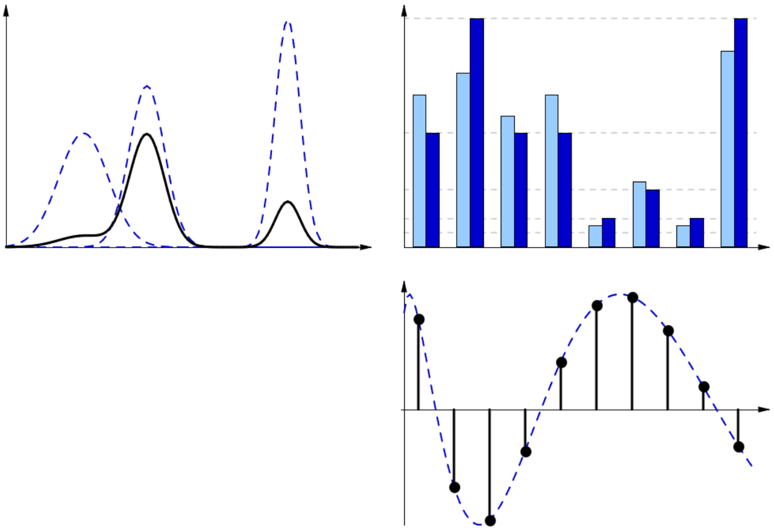
$\gamma, \delta > 0$*, then we have that* D_φ *on domain* $\mathbb{X}$ *is* μ*-similar for some constant* $\mu > 0$.

A related open problem is the question whether the observations that lead to this conjecture can be generalized to the case of approximation by sampling according to D_φ. It is remarkable that in algorithm BREG-MEANS++ from Chapter 5, as well as in all other publications addressing this sampling technique (namely [Sra et al., 2008] and [Nock et al., 2008]), μ-similarity or equivalent properties are always assumed to obtain any result. It is quite likely that the observations from this chapter can be adapted to the case of sampling according to D_φ by a more sophisticated analysis. Yet, it is still an open problem to prove this intuition.

Conjecture 8.9 implies that the restriction on μ-similar Bregman divergences is, in fact, necessary for the uniform sampling techniques employed in this thesis, and that this restriction can not be avoided. Unfortunately, non-similar instances of the Bregman k-median problem occur quite frequently in practice. For instance, in many text mining applications k-median clustering by Kullback-Leibler divergence is considered. In this context it is desireable that the underlying probability vectors are sparse (that is, many entries of the vector have 0 probability). However, these sparse vectors are singularities of the Kullback-Leibler divergence, such as $\mathrm{D}_{\mathrm{KL}}(p, q) = \infty$ for $p = (1, 0, 0, \ldots, 0)^\top$ and $q = (0, 1, 0, \ldots, 0)^\top$.

Hence, an interesting direction for further research is the development of algorithms and techniques for the Bregman k-median problem in the presence of singularities. As Conjecture 8.9 implies, there is little hope that this can be achieved by using uniform sampling alone to approximate the cluster medians. Up to date, the only algorithm known that computes an approximate solution for a non-similar Bregman k-median problem is the $\mathcal{O}(\log n)$-approximation algorithm of [Chaudhuri and McGregor, 2008] for the case of the Kullback-Leibler divergence on domain $\mathbb{R}^d_{\geq 0}$. This algorithm makes use of a deterministic polynomial time reduction from the Kullback-Leibler k-median problem to the Euclidean k-means problem and does not rely on uniform sampling. It remains an open problem to give approximation algorithms with a provable approximation guarantee for other non-similar instances, or the Bregman k-median problem in general.

A Applications



Clustering with Bregman divergences is a problem that arises in many different disciplines of computer science, such as machine learning, data compression, data mining, speech processing, image analysis, or pattern recognition. To provide a small overview, in this appendix, we give a brief introduction into three different practical applications that benefit from efficient Bregman clustering algorithms. These applications feature the Bregman k-median problem with respect to three different Bregman divergences, namely the Mahalanobis distances (Section A.1), the Kullback-Leibler divergence (Section A.2), and the Itakura-Saito divergence (Section A.3). These three dissimilarity measures are also our canonical instances of Bregman divergences throughout this thesis.

A.1 Estimating mixtures of identical Gaussian sources

One of the main tasks in statistical inference is to estimate the unknown parameters of the underlying model of a random source. Usually, such a source can only be perceived through examination of the features of a finite number of repeatable observations. The nature of these features may vary widely from application to application. For example, the height and body weight of people may be collected in a survey; in digital image processing the pixel-wise color values of digital images may be considered; or in experimental chemistry the molecular composition of synthesized chemical solutions may be studied. For the sake of this exposition, we will always assume that an observation is described as a vector of real valued features. These feature vectors are also called the *sample data*.

When the goal is to estimate the underlying regularity of a source, it is inevitable to make some assumptions on the general structure of the random process that generates the sample data (cf. [Bishop, 2008]). First of all, it is convenient to assume that the observations made are the outcome of an independent and identically distributed random experiment. Second, we have to fix the *model* of the experiment. That is, we have to select a parameterized family of probability distributions that are assumed to describe the outcome of the random experiment. The parameters of this model are to be inferred from the sample data.

One particular choice of model for random variables over the reals that is justified in many real world scenarios is the *Gaussian distribution*, also known as the normal distribution. The Gaussian distribution is a unimodal distribution which is parameterized by its mean (which also happens to be the mode of the distribution) and its standard deviation. Formally, we write $X \sim \mathcal{N}(\mu, \sigma^2)$ if random variable X takes values from $\mathbb{R}$ according to a Gaussian distribution with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 \in \mathbb{R}_+$. The probability density function $p : \mathbb{R} \to [0,1]$ of random variable $X \sim \mathcal{N}(\mu, \sigma^2)$ is given by

$$p\left(a \,\middle|\, \mu, \sigma^2\right) = \frac{1}{(2\pi)^{1/2}\sigma} \exp\left(-\frac{(a-\mu)^2}{2\sigma^2}\right) \tag{A.1}$$

for all $a \in \mathbb{R}$. When the observations yield feature vectors over $\mathbb{R}^d$, the d-dimensional Gaussion distribution is parameterized by its means $\mu \in \mathbb{R}^d$ and a symmetric positive definite covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$, formally

$X \sim \mathcal{N}^d(\mu, \Sigma)$. In this case, the probability density function $p : \mathbb{R}^d \to [0, 1]$ of an $X \sim \mathcal{N}^d(\mu, \Sigma)$ is denoted by

$$p\,(a \,|\, \mu, \Sigma) = \frac{1}{(2\pi)^{d/2} \det(\Sigma)^{1/2}} \exp\left(-\frac{1}{2}(a - \mu)^\top \Sigma^{-1} (a - \mu)\right) \qquad \text{(A.2)}$$

for all $a \in \mathbb{R}^d$.

The choice of the Gaussian distribution as model is justified since it arises quite naturally in many different contexts. Most prominently, it is known from the central limit theorem that the sum of n identical and independently distributed random variables is distributed according to a Gaussian distribution for $n \to \infty$ (see [Chung and AitSahlia, 2006]). Furthermore, the distribution that maximizes the differential entropy of a random variable is the Gaussian distribution (cf. [Bishop, 2008]). However, on the downside, the d-dimensional Gaussian distribution is always unimodal, and this might not reflect the nature of the random source we are out to model.

One particular choice of model which is frequently used under the assumption of a k-modal source distribution is the *mixture model* of Gaussian distributions. In the Gaussian mixture model, it is assumed that the source distribution is the convex combination of k Gaussian distributions. That is, probability density function $p : \mathbb{R}^d \to [0, 1]$ is given by

$$p\,(a \,|\, \pi, \mu_1, \ldots, \mu_k, \Sigma_1, \ldots, \Sigma_k) = \sum_{i=1}^{k} \pi_i \, p\,(a \,|\, \mu_i, \Sigma_i) \qquad \text{(A.3)}$$

for all $a \in \mathbb{R}^d$, where $\pi = (\pi_1, \pi_2, \ldots, \pi_k)^\top$ with $\pi_i > 0$ and $\sum_{i=1}^{d} \pi_i = 1$ describes a discrete probability distribution between the k Gaussian distributions. Hereby, each Gaussian distribution is parameterized by their own mean $\mu_i \in \mathbb{R}^d$ and covariance matrix $\Sigma_i \in \mathbb{R}^{d \times d}$. It is known that by using a sufficiently large number of Gaussians even relatively complicated source distributions can be approximated using the mixture model of Gaussian distributions (see [McLachlan and Basford, 1988]).

If all k Gaussians employ the same covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$, the model is called the *mixture model of identical Gaussian distributions*. We will concentrate on the case of identical Gaussian sources for a fixed covariance matrix Σ throughout the rest of this section. The free parameters of this model that have to be estimated are the means $\mu_1, \mu_2, \ldots, \mu_k \in \mathbb{R}^d$ as well as the mixture component distribution $\pi \in \mathbb{R}^k_{\geq 0}$.

We now show how estimating these parameters of the mixture model of identical Gaussians is related to finding the k-medians of a Bregman k-median problem. To this end, let us focus on the case $k = 1$. In the sequel, we assume covariance matrix Σ to be an arbitrary fixed constant. A particular heuristic frequently applied when estimating the parameters of a model is the *maximum likelihood principle.* This principle states that if a has been observed then among all parameters the parameters should be chosen that maximize the likelihood[1] function $p\,(a\,|\,\mu, \Sigma)$. Now, assume that a number of n feature vectors $a_1, a_2, \ldots, a_n \in \mathbb{R}^d$ have been observed. Let $p\,(a_1, a_2, \ldots, a_n\,|\,\mu, \Sigma)$ denote the probability density function of the joint distribution of these observations. Since we assume the source to be identically and independently distributed, we obtain

$$p\,(a_1, a_2, \ldots, a_n\,|\,\mu, \Sigma) = \prod_{i=1}^{n} p\,(a_i\,|\,\mu, \Sigma)\ . \tag{A.4}$$

From the fact that the natural logarithm is strictly increasing we know that maximizing equation (A.4) is equivalent to maximizing

$$\begin{aligned}
&\ln p\,(a_1, a_2, \ldots, a_n\,|\,\mu, \Sigma) \\
&\quad = \sum_{i=1}^{n} \ln p\,(a_i\,|\,\mu, \Sigma) && \text{(A.5)} \\
&\quad = \sum_{i=1}^{n} \ln \left(\frac{1}{(2\pi)^{d/2} \det(\Sigma)^{1/2}} \exp\left(-\frac{1}{2} (a_i - \mu)^\top \Sigma^{-1} (a_i - \mu) \right) \right) && \text{(A.6)} \\
&\quad = -\frac{n}{2} \ln \left((2\pi)^d \det(\Sigma) \right) - \frac{1}{2} \sum_{i=1}^{n} (a_i - \mu)^\top \Sigma^{-1} (a_i - \mu)\ . && \text{(A.7)}
\end{aligned}$$

Since n, d, and Σ are fixed we obtain the maximum likelihood if and only if the sum on the right-hand side of (A.7) is minimized. But this, of course, is equivalent to finding the 1-median of $P = \{a_1, a_2, \ldots, a_n\}$ with respect to the Mahalanobis distance D_A where $A = \Sigma^{-1}$ since

$$\operatorname{cost}\,(P, \mu) = \sum_{i=1}^{n} \mathrm{D}_A\,(a_i, \mu) = \sum_{i=1}^{n} (a_i - \mu)^\top \Sigma^{-1} (a_i - \mu)\ . \tag{A.8}$$

[1] Note that if μ and Σ are fixed and a is variable the term $p\,(a\,|\,\mu, \Sigma)$ is called the *probability density function*, while when a is fixed and the parameters μ, Σ are variable the term $p\,(a\,|\,\mu, \Sigma)$ is called the *likelihood function* of observation a.

This observation can be extended to the mixture model with $k > 1$ (see Chapter 9 of [Bishop, 2008] for an detailed description). We learn that, for fixed covariance matrix Σ, estimating the parameters of a mixture of identical Gaussian sources is equivalent to solving the Mahalanobis k-median problem. In particular, the means $\mu_1, \mu_2, \ldots, \mu_k$ are given by the k-medians of P, and the mixture component distribution π is obtained from the relative size of the optimal clustering $P_1, P_2, \ldots, P_k$, i.e. $\pi_i = |P_i|/|P|$.

Note that the approach given above assumes covariance matrix Σ to be a constant known in advance. In a more general setting, we are also interested in estimating Σ from the sample data. The task of simultaneously estimating μ and Σ is solved by the famous EM algorithm [Dempster et al., 1977]. However, in the context of the EM algorithm, the corresponding clustering computed is a so-called *soft clustering*. That is, points are not assigned to a single cluster but are rather provided with a discrete probability distribution on the k different clusters. Please note that the notion of soft clusterings is not considered in this thesis.

A.2 Model reduction for data compression

Lossless data compression is an indispensable tool of modern data transmission and modern data storage systems. Here, the objective is to find an encoding of a given piece of data — usually called the *message* — that uses a smaller representation than the original data. The size of such a representation is measured in terms of information-bearing units, such as bits or computer words. Compression is achieved by assigning shorter codewords to the more likely messages, and (necessarily) larger codewords to the less likely messages. Furthermore, the encoding has to guarantee that any message can be uniquely recovered by a matching decoding scheme.

Usually, the compression of a message takes place in two phases. The first phase is known as *modeling*. In the modeling step, the encoder seeks to obtain a simple yet accurate statistical model of the message. In particular, the message can be modeled as a random stream of symbols from a finite symbol alphabet $A = \{a_1, a_2, \ldots, a_d\}$. The j-th symbol of the stream is given by a random variable X_j distributed among the symbols from A according to a discrete probability mass function $p_j : A \to [0, 1]$, such that

$$p_j(a) = \Pr[X_j = a] \tag{A.9}$$

for all $a \in A$. This probability distribution is also called the *prediction* of the j-th symbol. Models that are used in practice include the *memoryless source model*, as well as the *Markov model*. In a memoryless source model, the same fixed probability distribution is used for each symbol of the stream, independent of the actual symbols encountered in the stream. In a Markov model of order $t \in \mathbb{N}$, the predictive distribution of the j-th symbol of the stream depends on the last t symbols of the stream. These past symbols are also called the *context* of the current symbol. Further, more sophisticated context based models are used in practice, such as *prediction by partial matching* (PPM, cf. [Cleary and Witten, 1984]).

The second phase of data compression is referred to as *coding*. In this step, a description of the model and a description of the actual message in terms of the given model is stored. For instance, assume that each symbol of the stream is stored as a uniquely decodeable bit string. For each position j of the stream, let the codeword in the case that $a \in A$ occurs as the j-th symbol be given by $C_j(a) \in \{0,1\}^*$. To achieve a good compression of the data, the length of the codeword $|C_j(a)|$ is determined by probability $p_j(a)$. The famous source coding theorem from the seminal work of [Shannon, 1948] states that an optimal compression is achieved if $C_j(a)$ has a bit length of

$$|C_j(a)| = \log \frac{1}{p_j(a)} \ . \tag{A.10}$$

Please note that this is an idealized simplification of Shannon's result since, obviously, a codeword consists of an integral number of bits. However, for simplicity of exposition, we will ignore this inaccuracy. Also note that there exist binary encoding schemes such that the codeword length of a message symbol with probability p comes asymptotically arbitrary close to $\log(1/p)$ (e.g., the arithmetic coding scheme due to the American information theorists Peter Elias, cf. [Abramson, 1963] or [Jelinek, 1968]).

Assuming that p_j is indeed an accurate model of X_j, we obtain that the expected codeword length $\mathrm{E}\big[|C_j(X_j)|\big]$ of the j-th symbol is given by

$$\mathrm{E}\big[|C_j(X_j)|\big] = \sum_{i=1}^{d} p_j(a_i) \log \frac{1}{p_j(a_i)} \ . \tag{A.11}$$

Here, the term $\mathrm{H}(X_j) = \sum_{i=1}^{d} p_j(a_i) \log \frac{1}{p_j(a_i)}$ is also known as the (binary) *entropy* of X_j. In the following, let

$$C(X_1, X_2, \dots, X_n) = C_1(X_1) \circ C_2(X_2) \circ \dots \circ C_n(X_n) \tag{A.12}$$

denote the encoding of a sequence of n symbols of the source stream, that is, the concatenation of the codewords of the symbols of the source stream. Then the expected encoded length of the n symbols using an accurate model of the stream is given by

$$\begin{aligned} \mathrm{E}\big[|C(X_1, X_2, \ldots, X_n)|\big] &= \sum_{i=1}^{n} \mathrm{E}\big[|C_j(X_j)|\big] && \text{(A.13)} \\ &= \sum_{i=1}^{n} \sum_{i=1}^{d} p_j(a_i) \log \frac{1}{p_j(a_i)} \ . && \text{(A.14)} \end{aligned}$$

However, in addition to the encoded symbols, a description of the underlying model has to be stored with the compressed data in order to let the decoder regain the compressed message. For instance, assume a stream features a source distribution according to a Markov model of a very large order t and a symbol alphabet of size d. Then the Markov model consists of d^t different contexts, each represented by a different probability distribution on the symbol alphabet. The space required to store this model alone can easily outweigh any compression achieved by the encoding of the message.

One idea to avoid this problem is to derive a new model from the source model which uses a far smaller number of probability distributions. To this end, groups of similar probability distributions are represented by a single prototypical distribution for each group. This approach is called *model reduction.* In this case, of course, encoder and decoder are using an inaccurate prediction of the j-th symbol of the stream to encode this symbol. That is, if q_j is the inaccurate model used to encode the j-th symbol by using encoding function C'_j, we obtain

$$|C'_j(a)| = \log \frac{1}{q_j(a)} \ . \qquad \text{(A.15)}$$

Hence, if p_j is the accurate model of X_j, then the expected encoded length of a sequence of n symbols of the source stream using the inaccurate model q_j to build the codewords is given by

$$\begin{aligned} \mathrm{E}\big[|C'(X_1, X_2, \ldots, X_n)|\big] &= \sum_{i=1}^{n} \mathrm{E}\big[|C'_j(X_j)|\big] && \text{(A.16)} \\ &= \sum_{i=1}^{n} \sum_{i=1}^{d} p_j(a_i) \log \frac{1}{q_j(a_i)} \ . && \text{(A.17)} \end{aligned}$$

Given the n probability distributions of the source model, the goal of the *model reduction problem* is to find a number of k prototypical probability distributions that minimize the expected loss of compression when using the inaccurate model given by the prototypes. This expected loss of compression is given by

$$\begin{aligned}
&\mathrm{E}\big[|C'(X_1, X_2, \ldots, X_n)|\big] - \mathrm{E}\big[|C(X_1, X_2, \ldots, X_n)|\big] \\
&\qquad = \sum_{i=1}^{n} \sum_{i=1}^{d} p_j(a_i) \log \frac{1}{q_j(a_i)} - \sum_{i=1}^{n} \sum_{i=1}^{d} p_j(a_i) \log \frac{1}{p_j(a_i)} \qquad \text{(A.18)} \\
&\qquad = \sum_{i=1}^{n} \sum_{i=1}^{d} p_j(a_i) \log \frac{p_j(a_i)}{q_j(a_i)} \;, \qquad \text{(A.19)}
\end{aligned}$$

where equation (A.19) is nothing else but the Kullback-Leibler divergence between probability vectors $p_j = \big(p_j(a_1), p_j(a_2), \ldots, p_j(a_d)\big)^\top$ and $q_j = \big(q_j(a_1), q_j(a_2), \ldots, q_j(a_d)\big)^\top$ with $\sum_{i=1}^{d} p_j(a_i) = 1$ and $\sum_{i=1}^{d} q_j(a_i) = 1$. Hence, computing a set of representatives for probability distributions in a statistical model immediately leads to a k-median clustering problem with respect to the Kullback-Leibler divergence on the probability simplex

$$\mathbb{X} = \left\{ (x_1, x_2, \ldots, x_d)^\top \in \mathbb{R}^d_{\geq 0} \;\middle|\; \sum_{i=1}^{d} x_i = 1 \right\} \;. \qquad \text{(A.20)}$$

A.3 Codebook generation for vector quantization in speech processing

Audio signals, such as speech or music, are analog signals. In order to let an analog signal be stored in digital form or be transmitted over a digital channel it is necessary to find a digital representation of the signal. This digitization process is also called *quantization*. Furthermore, the digital data needs to be converted back into an analog signal in order to be heard through analog devices like audio speakers.

To give a formal description of this quantization/reconstruction process, assume that the range of magnitudes of an analog signal is given by $[-a, a] \subseteq \mathbb{R}$. An analog waveform is frequently given as a sequence of real-valued waveform magnitudes which are called *samples*. For a fixed rate of b bits per sample, quantization is obtained by encoding each waveform sample as a bit string of size b, using an encoding function $f : [-a, a] \to \{0, 1\}^b$.

The digital bit strings are converted back into analog signals using a decoding function $g : \{0,1\}^b \to [-a,a]$. In this context the set $G = \bigl\{g(x) \bigm| x \in \{0,1\}^b\bigr\}$ is also called the *codebook* of the quantization process.

Obviously, quantization is a lossy encoding of the original waveform. This loss is measured by using a fitting distortion function $\delta : [-a,a]^2 \to \mathbb{R}$, where $\delta(x, \hat{x})$ denotes the distortion between an original sample x and the reconstructed value $\hat{x} = g\bigl(f(x)\bigr)$. Given a sequence of n samples $x_1, x_2, \ldots, x_n$, the *average distortion* $\Delta(x_1, x_2, \ldots, x_n)$ is given by

$$\Delta(x_1, x_2, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^{n} \delta(x_i, \hat{x}_i) \ . \tag{A.21}$$

The main objective of the *codebook generation problem* is to find encoding and decoding functions f, g that minimize the average distortion for a given set of training data.

One classical method of quantization is *pulse code modulation (PCM)* which was developed 1939 by the English inventor Alec H. Reeves (cf. [Waggener, 1994]). PCM is the direct scalar digitization of a speech waveform given as a sequence of waveform magnitudes. To this end, the range $[-a,a]$ is partitioned into 2^b intervals. The intervals are labeled using the bit strings from $\{0,1\}^b$, and f assigns each waveform sample to the bit string label of the interval that contains the sample. Furthermore, each $g(x)$ is defined as the midpoint of the interval labeled by $x \in \{0,1\}^b$. The distortion measure is assumed to be given by the squared error $\delta(x, \hat{x}) = (x - \hat{x})^2$. It is easy to see that the optimal codebook of the PCM approach is obtained by finding the 2^b-medians with respect to the squared Euclidean distance in dimension $d = 1$ [Max, 1960].

A one-dimensional approach like PCM is also referred to as *scalar quantization.* Surprisingly, it turns out that the efficiency of such a quantization scheme can be improved by a simple generalization to higher dimensions. A fundamental result of Shannon's rate distortion theory implies that, at a fixed bit rate, d analog signals can be digitized with lower average distortion if they are treated as one d-dimensional vector instead of d scalar values separately. More precisely, this observation holds no matter whether the d signals are correlated or independent. This leads to the following generalization: Instead of a single value, a sequence of d waveform samples is considered as vector from $[-a,a]^d$. For a rate of b bits per sample the encoding function is given by $f : [-a,a]^d \to \{0,1\}^{bd}$ and the decoding function is given by $g : \{0,1\}^{bd} \to [-a,a]^d$. This form of digitization is called *vector quantization.* Again, the objective of the codebook generation problem is to find f, g that minimizes average distortion for n training vectors from $[-a, a]^d$. Furthermore, a straightforward generalization of the PCM approach is applicable to vector quantization [Linde et al., 1980]. Again, it is not difficult to see that for $\delta(x, \hat{x}) = \|x - \hat{x}\|^2$ the optimal codebook is obtained by finding the 2^{bd}-medians with respect to the squared Euclidean distance in dimension d.

The fidelity of a restored signal can be further improved through preprocessing of the waveform samples, especially through *predictive filtering.* In predictive filtering, the aim is to give a description of the magnitude of the next sample through the values of the recent samples as well as a fixed number of real-valued model parameters. Instead of the waveform sample magnitudes, these model parameters are to be transmitted over the digital channel. One particular model that is frequently used is *linear predictive coding (LPC).* Here a linear predictor of order m of the next sample magnitude x_n is given by the linear combination

$$\hat{x}_n = \sum_{i=1}^{m} a_i x_{n-i} \tag{A.22}$$

of the recent sample magnitudes $x_{n-m}, x_{n-m+1}, \ldots, x_{n-1}$ and the real-valued model parameters $a_1, a_2, \ldots, a_m$. Since the model parameters are real-valued, again, this calls out for vector quantization analogously to the case when transmitting waveform samples.

However, in speech processing, careful consideration has to be given to the choice of the distortion function. To be of any value, a distortion measure has to be "analytically tractable, computable from sampled data, and, most important, subjectively meaningful" [Buzo et al., 1980]. If the analog signal is given as waveform samples, usually the squared Euclidean distance is the distortion measure of choice because of its tractability and computability, and because of its interpretation in the context of maximum likelihood estimation for waveform samples. Unfortunately, to take into account the perceivable quality of a speech signal, it is not only important how much distortion is induced on the waveform by applying vector quantization. It is also of great importance to minimize the effect of the quantization on the frequency spectrum of the signal, that is, to minimize the distortion of the spectrum when applying a quantization of the parameters. Here, the squared Euclidean distance does not necessarily provide a subjectively meaningful distortion measure if the signal is given as the parameters of a discrete all-pole model of the spectral parametres in LPC.

To overcome this problem, Itakura and Saito proposed a new distortion measure in [Itakura and Saito, 1968] that has proven to be subjectively meaningful with respect to the spectral properties of LPC parameters. This distortion measure, today known as the Itakura-Saito divergence, is given as

$$\mathrm{D_{IS}}\big(z, \hat{z}\big) = \sum_{i=1}^{d} \left(\frac{z_i}{\hat{z}_i} - \ln \frac{z_i}{\hat{z}_i} - 1 \right) \tag{A.23}$$

for $z = (z_1, z_2, \ldots, z_d)$ and $\hat{z} = (\hat{z}_1, \hat{z}_2, \ldots, \hat{z}_d)$, where z and $\hat{z}$ are the Z-transform[2] of the LPC model parameters a and $\hat{a}$, respectively, evaluated at d fixed points from the frequency spectrum. It has been shown in [Itakura and Saito, 1968] that the Itakura-Saito divergence arises as an approximation to the maximum likelihood estimator in LPC. Furthermore, the Itakura-Saito divergence asymptotically results from the *discrimination information minimization principle* [Gray et al., 1981]. If we consider the optimal codebook generation problem for vector quantization of LPC parameters minimizing the average Itakura-Saito divergence as spectral distortion measure, we obtain an instance of the Itakura-Saito k-median problem.

[2] The Z-transform is the discrete version of the Laplace transform, which transforms a discrete time series of sample points (from the so-called time domain) into a continuous spectral representation in the frequency domain.

B Mathematical fundamentals

In this appendix we give a short summary of the mathematical fundamentals assumed to be common knowledge in this thesis. Results are stated in brief to support the argumentation given in this thesis, and without any intention of completeness. The formal proofs to these fundamental results are omitted. The interested reader is directed to the provided references for rigorous proof of the statements given in this appendix.

B.1 The vectorspace $\mathbb{R}^d$

B.1.1 Vector arithmetic and inner product

For $x, y \in \mathbb{R}^d$ and $\lambda \in \mathbb{R}$ we define an *addition* $x + y$ on $\mathbb{R}^d$ as the componentwise addition of the coordinates of x and y, and a *scalar product* $\lambda \cdot x$ as the componentwise product of the coordinates of x by factor λ. That is, for $x = (x_1, x_2, \ldots, x_d)^\top$ and $y = (y_1, y_2, \ldots, y_d)^\top$ we have

$$x + y = (x_1 + y_1, x_2 + y_2, \ldots, x_d + y_d)^\top \in \mathbb{R}^d \ , \tag{B.1}$$

$$\lambda \cdot x = (\lambda x_1, \lambda x_2, \ldots, \lambda x_d)^\top \in \mathbb{R}^d \ . \tag{B.2}$$

Since $(\mathbb{R}^d, +, \cdot)$ forms an $\mathbb{R}$-vectorspace these operations are associative, commutative, and obey the distributive law

$$\lambda(x + y) = \lambda x + \lambda y \ . \tag{B.3}$$

Furthermore, for $x, y \in \mathbb{R}^d$ the *inner product* of x and y, denoted by $x^\top y$, is defined as

$$x^\top y = \sum_{i=1}^{d} x_i y_i \in \mathbb{R} \ . \tag{B.4}$$

Using this inner product, we denote the *(Euclidean) ℓ_2-norm* on $\mathbb{R}^d$ by

$$\|x\| = \sqrt{x^\top x} = \sqrt{\sum_{i=1}^{d} x_i^2} \in \mathbb{R} \ . \tag{B.5}$$

B.1.2 Distances and metrics

Let $\mathbb{X} \subseteq \mathbb{R}^d$ be arbitrary. A function $\mathrm{D} : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ is called a *distance* if the following conditions are satisfied.

a) *Non-Negativity*: For all $x, y \in \mathbb{X}$ we have $\mathrm{D}(x, y) \geq 0$.

b) *Identity of indiscernibles*: We have $\mathrm{D}(x, y) = 0$ if and only if $x = y$.

c) *Symmetry*: For all $x, y \in \mathbb{X}$ we have $\mathrm{D}(x, y) = \mathrm{D}(y, x)$.

Furthermore, a distance D is called a *metric*, if the following additional condition is satisfied.

d) *Triangle inequality*: For all $x, y, z \in \mathbb{X}$ we have

$$\mathrm{D}(x, z) \leq \mathrm{D}(x, y) + \mathrm{D}(y, z) \ . \tag{B.6}$$

Distances and metrics express the intuitive notion of dissimilarity between elements from $\mathbb{R}^d$. Many different distances and metrics are used to measure this dissimilarity, varying from application to application. Most notably, it is easy to check that the well-known Euclidean distance

$$\mathrm{D}_{\ell_2}(x, y) = \|x - y\| = \sqrt{\sum_{i=1}^{d} (x_i - y_i)^2} \tag{B.7}$$

of two points $x = (x_1, x_2, \ldots, x_d) \in \mathbb{R}^d$ and $y = (y_1, y_2, \ldots, y_d) \in \mathbb{R}^d$ is a metric on $\mathbb{R}^d$. Here, $\| \cdot \|$ denotes the ℓ_2-norm of $\mathbb{R}^d$.

B.1.3 Convex sets

Intuitively, a subset from the vectorspace $\mathbb{R}^d$ is said to be convex if for every pair of points within the subset we have that the line segment that joins them is also within the subset. More precisely, a non-empty set $\mathbb{X} \subseteq \mathbb{R}^d$ is called *convex* if for all $x, y \in \mathbb{X}$ and for all $\lambda \in [0, 1]$ we have

$$\lambda x + (1 - \lambda) y \in \mathbb{X} \ . \tag{B.8}$$

Following this definition, we find that any *convex combination*

$$y = \sum_{i=1}^{n} \lambda_i x_i \tag{B.9}$$

of points $x_1, x_2, \ldots, x_n \in \mathbb{X}$ with positive λ_i and $\sum_{i=1}^{n} \lambda_i = 1$ lies within $\mathbb{X}$. In particular, it follows that for any finite subset $P \subseteq \mathbb{X}$ the arithmetic mean of P (also known as the *centroid* or the *center of gravity*)

$$c_P = \frac{1}{|P|} \sum_{p \in P} p \tag{B.10}$$

lies within $\mathbb{X}$.

B.1.4 Relative interior

From time to time, it is convenient to distinguish between the points "on the border" and the points that lie "inside" of a convex set $\mathbb{X}$. To this end, the *relative interior* $\mathrm{ri}(\mathbb{X})$ of convex set $\mathbb{X}$ gives the interior of the affine hull of $\mathbb{X}$. Formally, we define

$$\mathrm{ri}(\mathbb{X}) = \left\{ x \in \mathbb{X} \,\middle|\, \forall y \in \mathbb{X} \; \exists z \in \mathbb{X} \; \exists 0 < \lambda < 1 : x = \lambda y + (1 - \lambda) z \right\} \ . \tag{B.11}$$

Note that if $\mathbb{X}$ is non-singleton, then $\mathrm{ri}(\mathbb{X})$ is non-empty.

B.2 Inequalities

B.2.1 Triangle inequality of the reals

The subadditivity of the absolute value of real numbers $x, y \in \mathbb{R}$ yields

$$|x + y| \leq |x| + |y| \ . \tag{B.12}$$

Inequality (B.12) is called the *triangle inequality of the reals*. Using this triangle inequality, we also find that

$$\left|\sum_{i=1}^{n} x_i\right| \leq \sum_{i=1}^{n} |x_i| \tag{B.13}$$

for any number of elements $x_1, x_2, \ldots, x_n \in \mathbb{R}$.

B.2.2 Bounds to the binomial coefficient

For $n, k \in \mathbb{N}_0$ with $n \geq k$ the *binomial coefficient*

$$\binom{n}{k} = \frac{n!}{(n-k)!\,k!} \tag{B.14}$$

denotes the coefficient of the monomial x^k in the polynomial expansion of the binomial power $(1+x)^n$. It is a well-known fact from combinatorics that $\binom{n}{k}$ equals the number of distinct subsets of size k from a superset of size n. It is easy to see (for instance, cf. [Graham et al., 1994]) that the binomial coefficient is bounded from above by

$$\binom{n}{k} \leq \frac{n^k}{k!} \leq n^k \tag{B.15}$$

and from below by

$$\binom{n}{k} \geq \frac{n^k}{k^k} \ . \tag{B.16}$$

B.2.3 Bounds on the harmonic number

The *harmonic series* is the infinite series

$$\sum_{i=1}^{\infty} \frac{1}{i} = 1 + \frac{1}{2} + \frac{1}{3} + \ldots \ . \tag{B.17}$$

It is known that the harmonic series diverges to ∞, although rather slowly. The n-th partial sum of the harmonic series

$$H_n = \sum_{i=1}^{n} \frac{1}{i} \tag{B.18}$$

is called the *n-th harmonic number*. It is a classical observation by Swiss mathematician Leonhard Euler that the growth of H_n is approximately as fast as the growth of natural logarithm $\ln(n)$. In particular, it is known (cf. [Graham et al., 1994]) that for all $n \in \mathbb{N}$ we have

$$\ln(n) < H_n \leq \ln(n) + 1 \; . \tag{B.19}$$

B.2.4 Chebyshev's sum inequality

Let $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$ be ordered sequences of non-decreasing real numbers, i.e., $a_1 \leq a_2 \leq \ldots \leq a_n$ and $b_1 \leq b_2 \leq \ldots \leq b_n$. Then we have

$$n \sum_{i=1}^{n} a_i b_i \geq \left(\sum_{i=1}^{n} a_i \right) \left(\sum_{i=1}^{n} b_i \right) \; . \tag{B.20}$$

Inequality (B.20) is called *Chebyshev's sum inequality*, named after the Russian mathematician Pafnuty Chebyshev. A formal proof of this inequality can be found in [Hardy et al., 1952]. If both sequences are identical, that is, $a_i = b_i$ for all $i = 1, 2, \ldots, n$, we obtain

$$n \sum_{i=1}^{n} a_i^2 \geq \left(\sum_{i=1}^{n} a_i \right)^2 \tag{B.21}$$

as a special case of Chebyshev's sum inequality.

B.3 Calculus

B.3.1 Partial derivatives

For a continuous function $f : \mathbb{X} \to \mathbb{R}$ with $\mathbb{X} \subseteq \mathbb{R}^d$ and any direction vector $v \in \mathbb{R}^d$ with $\|v\| = 1$ the term

$$\frac{\partial}{\partial v} f(t) = \lim_{\lambda \to 0} \frac{f(t + \lambda v) - f(t)}{\lambda} \tag{B.22}$$

is called the *partial derivative* of f in direction v, provided that the limit exists. The function f is called *differentiable* if this limit exists for all $t \in \mathbb{X}$

and for all $v \in \mathbb{R}^d$ with $\|v\| = 1$. If the partial derivative is obtained in direction of a unit vector $e_i = (0, \ldots, 1, \ldots, 0) \in \mathbb{R}^d$ we also write

$$\frac{\partial}{\partial t_i} f(t) = \frac{\partial}{\partial e_i} f(t) \tag{B.23}$$

for $t = (t_1, t_2, \ldots, t_d) \in \mathbb{X}$. The vector ∇f of the partial derivatives of f with respect to the unit vectors is called the *gradient* of f, that is, we define

$$\nabla f(t) = \begin{pmatrix} \frac{\partial}{\partial t_1} f(t) \\ \frac{\partial}{\partial t_2} f(t) \\ \vdots \\ \frac{\partial}{\partial t_d} f(t) \end{pmatrix} . \tag{B.24}$$

It is an important observation that the partial derivative in direction $v \in \mathbb{R}^d$ can be expressed in terms of the gradient, i.e.,

$$\frac{\partial}{\partial v} f(t) = \nabla f(t)^\top v . \tag{B.25}$$

Furthermore, for $v, w \in \mathbb{R}^d$ we call

$$\frac{\partial^2}{\partial w \partial v} f(t) = \lim_{\lambda \to 0} \frac{\frac{\partial}{\partial v} f(t + \lambda w) - \frac{\partial}{\partial v} f(t)}{\lambda} \tag{B.26}$$

the *second order partial derivative* of f in direction v and w. The function f is called *twice differentiable* if this limit exists for all $t \in \mathbb{X}$ and for all $v, w \in \mathbb{R}^d$ with $\|v\| = \|w\| = 1$. The second order partial derivatives in direction of the unit vectors form the so-called *Hessian matrix* of f, that is, for all $t = (t_1, t_2, \ldots, t_d)^\top \in \mathbb{X}$ we define

$$\nabla^2 f(t) = \begin{pmatrix} \frac{\partial^2}{\partial t_1^2} f(t) & \frac{\partial^2}{\partial t_1 \partial t_2} f(t) & \cdots & \frac{\partial^2}{\partial t_1 \partial t_d} f(t) \\ \frac{\partial^2}{\partial t_2 \partial t_1} f(t) & \frac{\partial^2}{\partial t_2^2} f(t) & \cdots & \frac{\partial^2}{\partial t_2 \partial t_d} f(t) \\ \vdots & & & \\ \frac{\partial^2}{\partial t_d \partial t_1} f(t) & \frac{\partial^2}{\partial t_d \partial t_2} f(t) & \cdots & \frac{\partial^2}{\partial t_d^2} f(t) \end{pmatrix} . \tag{B.27}$$

By Clairaut-Schwarz's theorem (cf. [Courant and John, 1974]), it is known that if f has continuous second order partial derivatives, then the partial derivatives of f commute, that is

$$\frac{\partial^2}{\partial t_j \partial t_i} f(t) = \frac{\partial^2}{\partial t_i \partial t_j} f(t) \tag{B.28}$$

for any $1 \leq i, j \leq d$. Thus, in this case, $\nabla^2 f(t)$ is a symmetric matrix.

B.3.2 Mean value theorem

The *mean value theorem* is one of the central theorems of calculus. Intuitively, it is equivalent to the geometric observation that for any differentiable function f and for any two distinct preimages $x, y \in \mathbb{X}$ there is an intermediate point ξ on the line segment through x and y where the slope of the graph of f (i.e., the partial derivatives in direction $x - y$) equals the slope of the chord from $\big(x, f(x)\big)$ to $\big(y, f(y)\big)$.

Theorem B.1 (mean value theorem)**.** *Let $f : \mathbb{X} \to \mathbb{R}$ be a differentiable function on domain $\mathbb{X} \subseteq \mathbb{R}^d$, and let $x, y \in \mathbb{X}$ be two distinct points. Then there exists an $\xi \in \overline{xy}$ on the line segment through x and y, i.e.,*

$$\overline{xy} = \{z \in \mathbb{X} \,|\, \exists 0 \leq \lambda \leq 1 : z = \lambda x + (1 - \lambda) y\} \ , \tag{B.29}$$

such that

$$\nabla f(\xi)^\top (x - y) = f(x) - f(y) \ . \tag{B.30}$$

A proof of this theorem can be found in [Courant and John, 1974].

B.3.3 Taylor expansion

The *Taylor series expansion* gives a description of a function in terms of evaluations of its derivatives at a fixed expansion point. In one dimension, the Taylor expansion of an infinitely often differantiable function $f : \mathbb{X} \to \mathbb{R}$ with $\mathbb{X} \subseteq \mathbb{R}$ is given by

$$f(t) = \sum_{i=0}^{\infty} \frac{1}{i!} (t - t_0)^i f^{(i)}(t_0) \tag{B.31}$$

for any $t \in \mathbb{X}$ and an arbitrary expansion point $t_0 \in X$. Here $f^{(i)}$ denotes the i-th derivative of f.

A finite partial sum of series (B.31) can be used to obtain an approximation of $f(t)$. There are several ways to express the error of this approximation. One particular description is given by the *Lagrange form* of the error term.

Theorem B.2 (Taylor expansion with Lagrange error in one dimension)**.** *Let $f : \mathbb{X} \to \mathbb{R}$ with $\mathbb{X} \subseteq \mathbb{R}$ be an $(n + 1)$-times differentiable function.*

Furthermore, let $t_0 \in \mathbb{X}$ *be arbitrary. Then for each* $t \in \mathbb{X}$ *there exists an* ξ_t *in the interval between* t *and* t_0 *such that*

$$f(t) = \sum_{i=0}^{n} \frac{1}{i!}(t - t_0)^i f^{(i)}(t_0) + R_n(\xi_t) \tag{B.32}$$

where $R_n(\xi_t)$ *is given by*

$$R_n(\xi_t) = \frac{1}{(n+1)!}(t - t_0)^{n+1} f^{(n+1)}(\xi_t) \ . \tag{B.33}$$

Equation (B.32) is called the n-th order Taylor expansion *of* $f(t)$ *at point* t_0*, and* $R_n(\xi_t)$ *is called the* Lagrange error term*.*

A proof of Theorem B.2 can be found in [Courant and John, 1974]. The Taylor expansion can be generalized to real valued functions $f : \mathbb{X} \to \mathbb{R}$ with $\mathbb{X} \subseteq \mathbb{R}^d$ in arbitrary dimension d. For the sake of brevity we omit a full description of this general case. The reader is directed to [Courant and John, 1974] for an in detail discussion of this topic. We just mention that in the case of $n = 1$, the first-order Taylor expansion of a twice differentiable function f can be given in terms of the gradient and the Hessian matrix of f.

Theorem B.3 (Taylor expansion with Lagrange error)**.** *Let* $f : \mathbb{X} \to \mathbb{R}$ *with* $\mathbb{X} \subseteq \mathbb{R}^d$ *be a twice differentiable function. Furthermore, let* $t_0 \in \mathbb{X}$ *be arbitrary. Then for each* $t \in \mathbb{X}$ *there exists an* ξ_t *on the line segment through* t *and* t_0 *such that*

$$f(t) = f(t_0) + \nabla f(t_0)^\top (t - t_0) + R_1(\xi_t) \tag{B.34}$$

where $R_1(\xi_t)$ *is given by*

$$R_1(\xi_t) = \frac{1}{2}(t - t_0)^\top \nabla^2 f(\xi_t)\, (t - t_0) \ . \tag{B.35}$$

B.3.4 Convex functions

Intuitively, a function f into the reals is called *convex* if for any line segment within its domain the value at the midpoint of the line segment does not exceed the arithmetic mean of the values at the ends of the line segment. From a geometric viewpoint, this is equivalent to the observation that for any two preimages x, y the graph of f lies completely beneath the chord from $\big(x, f(x)\big)$ to $\big(y, f(y)\big)$.

Formally, we use the following definition. A function $f : \mathbb{X} \to \mathbb{R}$ on convex domain $\mathbb{X} \subseteq \mathbb{R}^d$ is called *convex* if for all distinct points $x, y \in \mathbb{X}$ and all $0 < \lambda < 1$ we have

$$f\big(\lambda x + (1-\lambda)y\big) \leq \lambda f(x) + (1-\lambda) f(y) \ . \tag{B.36}$$

We say f is *strictly convex* if inequality (B.36) holds with strict inequality. Furthermore, f is called *(strictly) concave* if $-f$ is (strictly) convex.

It is an important property of convex functions that they are continuous on the relative interior of their domain (cf. [Boyd and Vandenberghe, 2004]).

B.3.5 Positive definiteness

A matrix $A \in \mathbb{R}^{d \times d}$ is called *positive semi-definite* if for all $x \in \mathbb{R}^d$ we have

$$x^\top A\, x \geq 0 \ . \tag{B.37}$$

Furthermore, if inequality (B.37) holds with with strict inequality for all $x \in \mathbb{R}^d \setminus \{0\}$, we say matrix A is *positive definite* on $\mathbb{X}$.

In many ways, positive semi-definite matrices are the generalization of non-negative real numbers. For instance, it is known from calculus in higher dimensions that a twice differentiable function $f : \mathbb{X} \to \mathbb{R}$ with $\mathbb{X} \subseteq \mathbb{R}^d$ is convex if and only if the Hessian matrix $\nabla^2 f(t)$ is positive semi-definite for all $t \in \mathbb{X}$ (see [Boyd and Vandenberghe, 2004]).

Unfortunately, in the case of strict convexity, we only obtain the following weaker implication: If the Hessian matrix $\nabla^2 f(t)$ is positive definite for all $t \in \mathbb{X}$, then f is strictly convex. In general, the converse of this implication is not true, since even if f is a strictly convex function there may exist $t \in \mathbb{X}$ and $x \in \mathbb{R}^d \setminus \{0\}$ with $x^\top \nabla^2 f(t)\, x = 0$. However, it is known that the set of elements $t \in \mathbb{X}$ with this property forms at most a discrete subset of $\mathbb{X}$ (again, see [Boyd and Vandenberghe, 2004]).

Furthermore, it is a well-known fact from linear algebra that for all positive definite matrices $A \in \mathbb{R}^{d \times d}$ there exists a non-singular matrix $B \in \mathbb{R}^{d \times d}$ with $A = B^\top B$. Such a matrix B can be obtained by computing the Cholesky decomposition of A (cf. [Trefethen and Bau, 1997]).

B.4 Probability theory

B.4.1 Probability, expectation, and variance

Let Ω be an arbitrary, discrete set of outcomes of a random experiment, called a *sample space*. A *probability distribution* among these outcomes is given by a real-valued function $p : \Omega \to \mathbb{R}$ with

a) *Non-negativity*: $p(\omega) \geq 0$ for all $\omega \in \Omega$,

b) *Normalization*: $\sum_{\omega \in \Omega} p(\omega) = 1$.

A subset A of sample space Ω is called an *event*. We extend the definition of probability distributions to all subsets $A \subseteq \Omega$ by

$$\Pr[A] = \sum_{\omega \in A} p(\omega) \ . \tag{B.38}$$

For any $A \subseteq \Omega$ the term $\Pr[A]$ is called the *probability* of event A. Event A is assumed to be true if the random experiment yields an outcome ω from A. Otherwise, A is assumed to be false. Hence, set A is implicitly identified with the truth assignment of a Boolean statement such as "$\omega \in A$", and we also write $\Pr[\omega \in A]$ instead. In addition, the set theoretic operations union, intersection, and complement are identified with the logic operations disjunction, conjunction, and negation, respectively. If the event A is given by a joint event $A = A_1 \wedge A_2$, we sometimes write $\Pr\left[A_1, A_2\right]$ instead of $\Pr\left[A_1 \wedge A_2\right]$ for the sake of a concise notation.

Let $X : \Omega \to \mathbb{R}$ be an arbitrary function on sample space Ω. In this context, X is called a *random variable*, and for all $a \in X(\Omega)$ we have

$$\Pr[X(\omega) = a] = \Pr[\omega \in X^{-1}(a)] \ . \tag{B.39}$$

Usually, the argument of $X(\omega)$ is omitted, and we simply write X instead. Two random variables X, Y are called *independent*, if

$$\Pr[X = a, Y = b] = \Pr[X = a] \cdot \Pr[Y = b] \tag{B.40}$$

for all $a \in X(\Omega)$ and $b \in Y(\Omega)$.

The *expectation* of a random variable X is denoted by

$$\mathrm{E}[X] = \sum_{\omega \in \Omega} \Pr\big[\omega\big] \cdot X(\omega) = \sum_{a \in X(\Omega)} \Pr\big[X = a\big] \cdot a \ . \tag{B.41}$$

It is an important property that the expectation is linear. That is, given linear combination $a_1X_1 + a_2X_2 + \ldots + a_nX_n$ of a finite number of random variables $X_1, X_2, \ldots, X_n$ and coefficients $a_1, a_2, \ldots, a_n \in \mathbb{R}$, we have

$$\mathrm{E}\left[\sum_{i=1}^{n} a_i X_i\right] = \sum_{i=1}^{n} a_i \, \mathrm{E}[X_i] \ . \tag{B.42}$$

Furthermore, the *variance* of random variable X is given by

$$\mathrm{Var}[X] = \mathrm{E}\Big[\big(X - \mathrm{E}[X]\big)^2\Big] \ . \tag{B.43}$$

B.4.2 Law of conditional probability

The probability of an event A under the assumption that event B with $\Pr[B] > 0$ is true is denoted by

$$\Pr\left[A \,|\, B\right] = \frac{\Pr[A, B]}{\Pr[B]} \ . \tag{B.44}$$

This probability is called the *conditional probability* of A given B.

Furthermore, if $A_1, A_2, \ldots, A_n$ are a finite number of arbitrary events with $\Pr[A_1, A_2, \ldots, A_n] > 0$ then we have

$$\Pr[A_1, A_2, \ldots, A_n] = \prod_{i=1}^{n} \Pr\left[A_i \,|\, A_1, A_2, \ldots, A_{i-1}\right] \ , \tag{B.45}$$

which is known as the *law of conditional probability* (for a formal proof see [Chung and AitSahlia, 2006]).

B.4.3 Law of total probability

It is easy to see (cf. [Chung and AitSahlia, 2006]) that for any partition of the sample space Ω into a finite number of disjoint events $B_1, B_2, \ldots, B_n$ and for any event $A \subseteq \Omega$ we obtain

$$\Pr[A] = \sum_{i=1}^{n} \Pr[B_i] \Pr\left[A \,|\, B_i\right] \ . \tag{B.46}$$

Equation (B.46) is called the *law of total probability.*

B.4.4 Law of total expectation

In analogy to the definition of conditional probability, the *conditional expectation* gives the expectation of a random variable X under the assumption that event B with $\Pr[B] > 0$ is true. The conditional expectation is denoted by

$$\mathrm{E}\left[X \mid B\right] = \sum_{a \in X(\Omega)} \Pr\left[X(\omega) = a \mid B\right] \cdot a \ . \tag{B.47}$$

Using the definition of the expectation and the law of total probability, for a finite number of disjoint events $B_1, B_2, \ldots, B_n$ we obtain the *law of total expectation*:

$$\mathrm{E}[X] = \sum_{i=1}^{n} \Pr[B_i]\, \mathrm{E}\left[X \mid B_i\right] \ . \tag{B.48}$$

B.4.5 Union bound

The *union bound* (which is also known as the *Boole-Bonferroni inequality*) gives an upper bound on the probability that any of two events A, B is true in terms of the sum of the individual probabilities of A and B. That is,

$$\Pr[A \vee B] \leq \Pr[A] + \Pr[B] \ . \tag{B.49}$$

Moreover, for a finite number of arbitrary events $A_1, A_2, \ldots, A_n$ we obtain

$$\Pr[A_1 \vee A_2 \vee \ldots \vee A_n] \leq \sum_{i=1}^{n} \Pr[A_i] \ . \tag{B.50}$$

These bounds are an immediate consequence of the inclusion-exclusion principle from combinatorics.

B.4.6 Markov's inequality

Markov's inequality gives a simple upper bound on the probability of the event that a positive random variable exceeds its expectation by a given factor. More precisely, let $X \geq 0$ be a positive random variable with finite expectation $\mathrm{E}[X] < \infty$. Then for any constant $c \geq 1$ we have

$$\Pr\bigl[X \geq c\, \mathrm{E}[X]\bigr] \leq \frac{1}{c} \ . \tag{B.51}$$

Inequality (B.51) is named after the Russian mathematician Andrey Markov. A formal proof can be found in [Chung and AitSahlia, 2006].

B.4.7 Chebyshev's inequality

Another important probabilistic concentration bound is known as *Chebyshev's inequality*: Let X be a random variable with finite variance, i.e., $\mathrm{Var}[X] < \infty$. Then for any constant $\rho > 0$ we have

$$\Pr\Big[\big|X - E[X]\big| \geq \rho\Big] \leq \frac{\mathrm{Var}[X]}{\rho^2} \ . \tag{B.52}$$

Chebyshev's inequality is named after the Russian mathematician Pafnuty Chebyshev. A formal proof can be found in [Chung and AitSahlia, 2006].

B.4.8 Chernoff bounds

Let $X_1, X_2, \ldots, X_n \in \{0, 1\}$ be a finite number of random variables of independent and identically distributed random experiments with probabilities $\Pr[X_i = 1] = p$ and $\Pr[X_i = 0] = 1 - p$ for all $i = 1, 2, \ldots, n$ and for some constant $0 \leq p \leq 1$. Furthermore, let $Y = X_1 + X_2 + \ldots + X_n$ be a random variable with expectation $\mathrm{E}[Y] = pn$.

The *Chernoff bounds* assure that with (very) high probability the random variable Y is close to its expectation. In detail, for any constant $\rho > 0$ we have

$$\Pr\big[Y \geq (1 + \rho)\,\mathrm{E}[Y]\big] \leq \exp\left(-\frac{\min(\rho, \rho^2)}{3}\,\mathrm{E}[Y]\right) \ , \tag{B.53}$$

$$\Pr\big[Y \leq (1 - \rho)\,\mathrm{E}[Y]\big] \leq \exp\left(-\frac{\rho^2}{2}\,\mathrm{E}[Y]\right) \ . \tag{B.54}$$

These bounds are named after the American statistician Herman Chernoff. A formal proof can be found in [McDiarmid, 1998].

List of Figures

Bibliography

[Abramson, 1963] Abramson, N. (1963). *Information Theory and Coding.* Electronic Science Series. McGraw-Hill, New York.

[Ackermann and Blömer, 2009] Ackermann, M. R. and Blömer, J. (2009). Coresets and approximate clustering for Bregman divergences. In *Proceedings of the 20th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '09)*, pages 1088–1097. Society for Industrial and Applied Mathematics.

[Ackermann and Blömer, 2010] Ackermann, M. R. and Blömer, J. (2010). Bregman clustering for separable instances. In *Proceedings of the 12th Scandinavian Symposium and Workshop on Algorithm Theory (SWAT '10).* Springer. To appear.

[Ackermann et al., 2008] Ackermann, M. R., Blömer, J., and Sohler, C. (2008). Clustering for metric and non-metric distance measures. In *Proceedings of the 19th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '08)*, pages 799–808. Society for Industrial and Applied Mathematics.

[Ackermann et al., 2010a] Ackermann, M. R., Blömer, J., and Sohler, C. (2010a). Clustering for metric and non-metric distance measures. *ACM Transactions on Algorithms.* Special issue on SODA '08. To appear.

[Ackermann et al., 2010b] Ackermann, M. R., Lammersen, C., Märtens, M., Raupach, C., Sohler, C., and Swierkot, K. (2010b). StreamKM++: A clustering algorithm for data streams. In *Proceedings of the Workshop on Algorithm Engineering and Experiments (ALENEX '10)*, pages 175–187. Society for Industrial and Applied Mathematics.

[Agarwal et al., 2005] Agarwal, P. K., Har-Peled, S., and Varadarajan, K. R. (2005). Geometric approximation via coresets. In Goodman, J. E., Pach, J., and Welzl, E., editors, *Combinatorial and Computational Geometry*, volume 52 of *Mathematical Sciences Research Institute (MSRI) publications*, pages 1–30. Cambridge University Press, New York.

[Ailon et al., 2006] Ailon, N., Chazelle, B., Comandur, S., and Liu, D. (2006). Self-improving algorithms. In *Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '06)*, pages 261–270. Society for Industrial and Applied Mathematics.

[Aloise et al., 2009] Aloise, D., Deshpande, A., Hansen, P., and Popat, P. (2009). $\mathcal{NP}$-hardness of Euclidean sum-of-squares clustering. *Machine Learning*, 75(2):245–248.

[Arora et al., 1998] Arora, S., Raghavan, P., and Rao, S. (1998). Approximation schemes for euclidean -medians and related problems. In *Proceedings of the 30th Annual ACM Symposium on Theory of Computing (STOC '98)*, pages 106–113.

[Arthur et al., 2009] Arthur, D., Manthey, B., and Röglin, H. (2009). k-means has polynomial smoothed complexity. In *Proceedings of the 50th Symposium on Foundations of Computer Science (FOCS '09)*. IEEE Computer Society. To appear.

[Arthur and Vassilvitskii, 2007] Arthur, D. and Vassilvitskii, S. (2007). `k-means++`: the advantages of careful seeding. In *Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '07)*, pages 1027–1035. Society for Industrial and Applied Mathematics.

[Assouad, 1983] Assouad, P. (1983). Plongements lipschitziens dans $\mathbb{R}^n$. *Bulletin de la Société Mathématique de France*, 111(4):429–448.

[Bădoiu et al., 2002] Bădoiu, M., Har-Peled, S., and Indyk, P. (2002). Approximate clustering via core-sets. In *Proceedings of the 34th Annual ACM Symposium on Theory of Computing (STOC '02)*, pages 250–257. Association for Computing Machinery.

[Bajaj, 1988] Bajaj, C. L. (1988). The algebraic degree of geometric optimization problems. *Discrete and Computational Geometry*, 3(1):177–191.

[Baker and McCallum, 1998] Baker, L. D. and McCallum, A. K. (1998). Distributional clustering of words for text classification. In *Proceedings* *of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '98)*, pages 96–103. Association for Computing Machinery.

[Banerjee et al., 2005a] Banerjee, A., Guo, X., and Wang, H. (2005a). On the optimality of conditional expectation as a Bregman predictor. *IEEE Transactions on Information Theory*, 51(7):2664–2669.

[Banerjee et al., 2005b] Banerjee, A., Merugu, S., Dhillon, I. S., and Ghosh, J. (2005b). Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6:1705–1749.

[Bishop, 2008] Bishop, C. M. (2008). *Pattern Recognition and Machine Learning.* Information Sience and Statistics. Springer, New York.

[Boros and Hammer, 1989] Boros, E. and Hammer, P. L. (1989). On clustering problems with connected optima in Euclidean spaces. *Discrete Mathematics*, 75(1-3):81–88.

[Boyd and Vandenberghe, 2004] Boyd, S. P. and Vandenberghe, L. (2004). *Convex Optimization.* Cambridge University Press, New York.

[Bregman, 1967] Bregman, L. M. (1967). The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7:200–217.

[Brucker, 1977] Brucker, P. (1977). On the complexity of clustering problems. In *Optimization and Operations Research: Proceedings of a Workshop Held at the University of Bonn. Lecture Notes in Economics and Mathematical Systems 157*, pages 45–54. Springer.

[Buzo et al., 1980] Buzo, A., Gray, Jr., A., Gray, R. M., and Markel, J. D. (1980). Speech coding based upon vector quantization. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(5):562–574.

[Censor and Lent, 1981] Censor, Y. and Lent, A. (1981). An iterative row-action method for interval convex programming. *Journal of Optimization Theory and Applications*, 34(3):321–353.

[Censor and Zenios, 1997] Censor, Y. and Zenios, S. A. (1997). *Parallel Optimization: Theory, Algorithms, and Applications.* Numerical Mathematics and Scientific Computation. Oxford University Press, New York.

[Chaudhuri and McGregor, 2008] Chaudhuri, K. and McGregor, A. (2008). Finding metric structure in information theoretic clustering. In *Proceedings of the 21st Annual Conference on Learning Theory (COLT '08)*, pages 391–402. Omnipress.

[Chen, 2006] Chen, K. (2006). On k-median clustering in high dimensions. In *Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '06)*, pages 1177–1185. Society for Industrial and Applied Mathematics.

[Chen, 2009] Chen, K. (2009). On coresets for k-median and k-means clustering in metric and Euclidean spaces and their applications. *SIAM Journal on Computing*, 39(3):923–947.

[Chung and AitSahlia, 2006] Chung, K. L. and AitSahlia, F. (2006). *Elementary Probability Theory: with stochastic processes and an introduction to mathematical finance.* Undergraduate Texts in Mathematics. Springer, New York, 4th edition.

[Cleary and Witten, 1984] Cleary, J. G. and Witten, I. H. (1984). Data compression using adaptive coding and partial string matching. *IEEE Transactions on Communications*, 32(4):396–402.

[Cormen et al., 2009] Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (2009). *Introduction to Algorithms.* The MIT Press, Cambridge, 3rd edition.

[Courant and John, 1974] Courant, R. and John, F. (1974). *Introduction to Calculus and Analysis*, volume 2. Wiley Interscience, New York.

[Cover and Thomas, 2006] Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory.* Wiley-Interscience, Hoboken, 2nd edition.

[Csiszár, 1991] Csiszár, I. (1991). Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems. *The Annals of Statistics*, 19(4):2032–2066.

[Czumaj and Sohler, 2004] Czumaj, A. and Sohler, C. (2004). Sublinear-time approximation for clustering via random sampling. In *Proceedings of the 31st International Colloquium on Automata, Languages and Programming (ICALP '04)*, pages 396–407. Springer.

[Dasgupta, 2007] Dasgupta, S. (2007). The hardness of k-means clustering. Technical Report CS2007-0890, University of California, San Diego.

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

[Dhillon et al., 2003] Dhillon, I. S., Mallela, S., and Kumar, R. (2003). A divisive information-theoretic feature clustering algorithm for text classifcation. *Journal of Machine Learning Research*, 3:1265–1287.

[Feldman et al., 2007] Feldman, D., Monemizadeh, M., and Sohler, C. (2007). A PTAS for k-means clustering based on weak coresets. In *Proceedings of the 23rd ACM Symposium on Computational Geometry (SCG '07)*, pages 11–18. Association for Computing Machinery.

[Fernandez de la Vega et al., 2003] Fernandez de la Vega, W., Karpinski, M., Kenyon, C., and Rabani, Y. (2003). Approximation schemes for clustering problems. In *Proceedings of the 35th Annual ACM Symposium on Theory of Computing (STOC '03)*, pages 50–58. Association for Computing Machinery.

[Frahling and Sohler, 2005] Frahling, G. and Sohler, C. (2005). Coresets in dynamic geometric data streams. In *Proceedings of the 27th Annual ACM Symposium on Theory of Computing (STOC '05)*, pages 209–217. Association for Computing Machinery.

[Graham et al., 1994] Graham, R. L., Knuth, D. E., and Patashnik, O. (1994). *Concrete Mathematics: A Foundation for Computer Science*. Addison-Wesley, Boston, 2nd edition.

[Gray et al., 1981] Gray, R. M., Gray, Jr., A., Rebolledo, G., and Shore, J. E. (1981). Rate-distortion speech coding with a minimum discrimination information distortion measure. *IEEE Transactions on Information Theory*, 27(6):708–720.

[Guha et al., 2000] Guha, S., Mishra, N., Motwani, R., and O'Callaghan, L. (2000). Clustering data streams. In *Proceedings of the 41st Symposium on Foundations of Computer Science (FOCS '00)*, pages 359–366. IEEE Computer Society.

[Gupta et al., 2003] Gupta, A., Krauthgamer, R., and Lee, J. R. (2003). Bounded geometries, fractals and low-distortion embeddings. In *Proceedings of the 44th Symposium on Foundations of Computer Science (FOCS '03)*, pages 534–543. IEEE Computer Society.

[Har-Peled and Kushal, 2005] Har-Peled, S. and Kushal, A. (2005). Smaller coresets for k-median and k-means clustering. In *Proceedings of the 21st Annual Symposium on Computational Geometry (SCG '05)*, pages 126–134. Association for Computing Machinery.

[Har-Peled and Mazumdar, 2004] Har-Peled, S. and Mazumdar, S. (2004). On coresets for k-means and k-median clustering. In *Proceedings of the 36th Annual ACM Symposium on Theory of Computing (STOC '04)*, pages 291–300. Association for Computing Machinery.

[Hardy et al., 1952] Hardy, G. H., Littlewood, J. E., and Pólya, G. (1952). *Inequalities.* Cambridge Mathematical Library. Cambridge University Press, Cambridge, 2nd edition.

[Hasegawa et al., 1993] Hasegawa, S., Imai, H., Inaba, M., Katoh, N., and Nakano, J. (1993). Efficient algorithms for variance-based k-clustering. In *Proceedings of the 1st Pacific Conference on Computer Graphics and Applications (Pacific Graphics '93)*, pages 75–89. World Scientific.

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

[Heinonen, 2001] Heinonen, J. (2001). *Lectures on analysis on metric spaces.* Universitext. Springer, New York.

[Inaba et al., 1994] Inaba, M., Katoh, N., and Imai, H. (1994). Applications of weighted Voronoi diagrams and randomization to variance-based k-clustering. In *Proceedings of the 10th ACM Symposium on Computational Geometry (SCG '94)*, pages 332–339. Association for Computing Machinery.

[Indyk and Thorup, 2000] Indyk, P. and Thorup, M. (2000). Approximate 1-medians. Unpublished manuscript.

[Itakura and Saito, 1968] Itakura, F. and Saito, S. (1968). Analysis synthesis telephony based on the maximum likelihood method. In *Reports of the 6th International Congress on Acoustics*, pages 17–20. Elsevier.

[Jain et al., 2002] Jain, K., Mahdian, M., and Saberi, A. (2002). A new greedy approach for facility location problems. In *Proceedings of the 34th Annual ACM Symposium on Theory of Computing (STOC '02)*, pages 731–740. Association for Computing Machinery.

[Jelinek, 1968] Jelinek, F. (1968). *Probabilistic Information Theory: Discrete and Memoryless Models.* McGraw-Hill Series in Systems Science. McGraw-Hill, New York.

[Kanungo et al., 2002] Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., and Wu, A. Y. (2002). An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):881–892.

[Kolliopoulos and Rao, 1999] Kolliopoulos, S. G. and Rao, S. (1999). A nearly linear-time approximation scheme for the Euclidean κ-median problem. In *Proceedings of the 7th Annual European Symposium on Algorithms (ESA '99)*, pages 378–389. Springer.

[Kullback and Leibler, 1951] Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86.

[Kumar et al., 2004] Kumar, A., Sabharwal, Y., and Sen, S. (2004). A simple linear time $(1+\varepsilon)$-approximation algorithm for k-means clustering in any dimensions. In *Proceedings of the 45th Annual IEEE Symposium on Foundations of Computer Science (FOCS '04)*, pages 454–462. IEEE Computer Society.

[Kumar et al., 2005] Kumar, A., Sabharwal, Y., and Sen, S. (2005). Linear time algorithms for clustering problems in any dimensions. In *Proceedings of the 32nd International Colloquium on Automata, Languages and Programming (ICALP '05)*, pages 1374–1385. Springer.

[Linde et al., 1980] Linde, Y., Buzo, A., and Gray, R. M. (1980). An algorithm for vector quantizer design. *IEEE Transactions on Communications*, 28(1):84–95.

[Lloyd, 1982] Lloyd, S. P. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137.

[Mahajan et al., 2009] Mahajan, M., Nimbhorkar, P., and Varadarajan, K. R. (2009). The planar k-means problem is $\mathcal{NP}$-hard. In *3rd Annual Workshop on Algorithms and Computation (WALCOM '09)*, pages 274–285. Springer.

[Mahalanobis, 1936] Mahalanobis, P. C. (1936). On the generalized distance in statistics. In *Proceedings of the National Institute of Sciences of India*, volume 2(1), pages 49–55. Indian National Science Academy.

[Manthey and Röglin, 2009] Manthey, B. and Röglin, H. (2009). Worst-case and smoothed analysis of k-means clustering with Bregman divergences. In *20th International Symposium on Algorithms and Computation (ISAAC '09)*, volume 5878 of *Lecture Notes in Computer Science*, pages 1024–1033. Springer.

[Matoušek, 2000] Matoušek, J. (2000). On approximate geometric k-clustering. *Discrete and Computational Geometry*, 24(1):61–84.

[Max, 1960] Max, J. (1960). Quantizing for minimum distortion. *IRE Transactions on Information Theory*, 6(1):7–12.

[McDiarmid, 1998] McDiarmid, C. J. H. (1998). Concentration. In Habib, M., McDiarmid, C. J. H., Ramírez Alfonsín, J. L., and Reed, B. A., editors, *Probabilistic methods for algorithmic discrete mathematics*, volume 16 of *Algorithms and Combinatorics*, pages 195–248. Springer, Berlin.

[McLachlan and Basford, 1988] McLachlan, G. J. and Basford, K. E. (1988). *Mixture models: Inference and applications to clustering.* Statistics: Textbooks and Monographs. Marcel Dekker, New York.

[Megiddo and Supowit, 1984] Megiddo, N. and Supowit, K. J. (1984). On the complexity of some common geometric location problems. *SIAM Journal on Computing*, 13(1):182–196.

[Mishra et al., 2001] Mishra, N., Oblinger, D., and Pitt, L. (2001). Sublinear time approximate clustering. In *Proceedings of the 12th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '01)*, pages 439–447. Society for Industrial and Applied Mathematics.

[Nielsen et al., 2007] Nielsen, F., Boissonnat, J.-D., and Nock, R. (2007). On Bregman Voronoi diagrams. In *Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '07)*, pages 746–755. Society for Industrial and Applied Mathematics.

[Nock et al., 2008] Nock, R., Luosto, P., and Kivinen, J. (2008). Mixed Bregman clustering with approximation guarantees. In *Proceedings of the 19th European Conference on Machine Learning (ECML '08)*, pages 154–169. Springer.

[O'Callaghan et al., 2002] O'Callaghan, L., Meyerson, A., Motwani, R., Mishra, N., and Guha, S. (2002). Streaming-data algorithms for high-quality clustering. In *18th International Conference on Data Engineering (ICDE '02)*, pages 685–696. IEEE Computer Society.

[Ostrovsky et al., 2006] Ostrovsky, R., Rabani, Y., Schulman, L. J., and Swamy, C. (2006). The effectiveness of Lloyd-type methods for the k-means problem. In *Proceedings of the 47th Annual Symposium on Foundations of Computer Science (FOCS '06)*, pages 165–176. IEEE Computer Society.

[Pereira et al., 1993] Pereira, F. C. N., Tishby, N., and Lee, L. (1993). Distributional clustering of english words. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL '93)*, pages 183–190. Association for Computational Linguistics.

[Shannon, 1948] Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656.

[Slonim and Tishby, 1999] Slonim, N. and Tishby, N. (1999). Agglomerative information bottleneck. In *Advances in Neural Information Processing Systems 12 (NIPS 12)*, pages 617–623. The MIT Press.

[Sra et al., 2008] Sra, S., Jegelka, S., and Banerjee, A. (2008). Approximation algorithms for Bregman clustering, co-clustering and tensor clustering. Technical Report MPIK-TR-177, Max Planck Institute for Biological Cybernetics.

[Thorup, 2005] Thorup, M. (2005). Quick k-median, k-center, and facility location for sparse graphs. *SIAM Journal on Computing*, 34(2):405–432.

[Trefethen and Bau, 1997] Trefethen, L. N. and Bau, III., D. (1997). *Numerical Linear Algebra.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia.

[Vattani, 2009] Vattani, A. (2009). k-means requires exponetially many iterations even in the plane. In *Proceedings of the 25th Annual Symposium on Computational Geometry (SCG '09)*, pages 324–332. Association for Computing Machinery.

[Waggener, 1994] Waggener, W. M. (1994). *Pulse Code Modulation Techniques: with applications in communications and data recording.* Solomon Press, New York.

[Weber, 1909] Weber, A. (1909). *Über den Standort der Industrien. 1.Teil: Reine Theorie des Standorts.* J.C.B. Mohr Verlag, Tübingen.

[Wesolowsky, 1993] Wesolowsky, G. O. (1993). The Weber problem: History and perspective. *Location Science*, 1(1):5–23.

[Zhang et al., 1996] Zhang, T., Ramakrishnan, R., and Livny, M. (1996). Birch: An efficient data clustering method for very large databases. In *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data(SIGMOD '96)*, pages 103–114. Association for Computing Machinery.