Quantization and Clustering with Bregman Divergences

Aurélie Fischer

Laboratoire de Statistique Théorique et Appliquée

Université Pierre et Marie Curie – Paris VI

Boîte 158

175 rue du Chevaleret

75013 Paris, France

aurelie.fischer@upmc.fr

Abstract – This paper deals with the quantization problem of a random variable X taking values in a separable and reflexive Banach space, and with the related question of clustering independent random observations distributed as X. To this aim, we use a quantization scheme with a class of distortion measures called Bregman divergences, and provide conditions ensuring the existence of an optimal quantizer and an empirically optimal quantizer. Rates of convergence are also discussed.

Index Terms – Bregman divergences, Quantization, k-means clustering, Banach spaces, Rates of convergence.

1 Introduction

Bregman divergences are a broad class of dissimilarity measures indexed by strictly convex functions. Introduced in 1967 by Bregman [9], these proximity functions are useful in a wide range of areas, among which statistical learning and data mining (Banerjee, Merugu, Dhillon and Ghosh [4], Cesa-Bianchi and Lugosi [11]), computational geometry (Nielsen, Boissonnat and Nock [27]), natural sciences, speech processing and information theory (Gray, Buzo, Gray and Matsuyama [19]). A lot of well-known proximity measures such as squared Euclidean, Mahalanobis, Kullback-Leibler and L^2 distances are particular cases of Bregman divergences. In $\mathbb{R}^d$, a Bregman divergence d_ϕ has the form

$$d_\phi(x,y) = \phi(x) - \phi(y) - \langle x - y, \nabla\phi(y)\rangle,$$

where $\langle \cdot, \cdot \rangle$ denotes the standard inner product, and $\nabla\phi(y)$ the gradient of ϕ at y. For example, taking $\phi(x) = \|x\|_2^2$ gives back the squared Euclidean distance. The same definition is valid in Hilbert spaces, and it even generalizes to Banach spaces by setting

$$d_\phi(x,y) = \phi(x) - \phi(y) - D_y\phi(x - y),$$

with $D_y\phi$ the Fréchet derivative of ϕ at y (Alber and Butnariu [1], Frigyik, Srivastava and Gupta [16]; see also Jones and Byrne [20] and Csiszár [12]). Note that a Bregman divergence is not necessary a true metric, since it may be asymmetric or fail to satisfy the triangle inequality. However, Bregman divergences fulfill an interesting projection property which generalizes the Hilbert projection on a closed convex set, as shown in Bregman [9] for the finite-dimensional setting and Alber and Butnariu [1] for the functional case. Recently, Banerjee, Merugu, Dhillon and Ghosh [4] have established a bijection between finite-dimensional Bregman divergences and exponential families, and shown that the standard k-means clustering algorithm (Lloyd [25]) generalizes to these divergences.

Following the approach of Banerjee et al. [4], we propose in the present paper to use this class of proximity measures for quantization and clustering purposes. Quantization, also called lossy data compression in information theory, is the problem of replacing data by an efficient and compact representation which allows one to reconstruct the original observations with a certain accuracy. More formally, for a fixed integer $k \geq 1$, a random variable X with distribution μ, taking values in a set $\mathcal{X}$, will be represented by a so-called k-quantizer $q(X)$. Here q is a Borel measurable mapping from $\mathcal{X}$ to a finite subset of $\mathcal{X}$ with at most k elements. The error committed when representing X by $q(X)$ is given by the distortion

$$W(\mu, q) = \mathbb{E}d(X, q(X)),$$

where $\mathbb{E}$ denotes expectation with respect to the distribution μ and $d(\cdot,\cdot)$ is called the distortion measure. For more information on quantization, we refer the reader to Gersho and Gray [17], Graf and Luschgy [18] and Linder [24]. In practice, the distribution μ is unknown, and $W(\mu, q)$ is replaced by the empirical criterion

$$W(\mu_n, q) = \frac{1}{n}\sum_{i=1}^{n} d(X_i, q(X_i)),$$

where $X_1, \dots, X_n$ are independent random observations with distribution μ, and μ_n denotes the empirical measure associated with $X_1, \dots, X_n$, i.e.,

$$\mu_n(A) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{X_i \in A\}}$$

for any Borel subset A of $\mathcal{X}$. In this context, the problem is called clustering and it consists in grouping data items in meaningful classes by minimizing $W(\mu_n, q)$ over all possible k-quantizers. In short, the goal is to find a data-based quantizer q_n such that the clustering risk $W(\mu, q_n)$ is "close" to the optimal risk $\inf_q W(\mu, q)$ as the size of the data set grows.

To date, most of the results pertaining to the clustering problem have been reported in the finite dimensional case, that is when $\mathcal{X} = \mathbb{R}^d$ ($d \geq 1$) endowed with the Euclidean metric. However, in many applied problems, the data items are in the form of random functions rather than standard vectors, and this casts the problem into the general class of functional data clustering. Besides, Bregman divergences represent a natural tool to measure proximity between infinite-dimensional objects, such as curves or even probability measures. For a comprehensive introduction to the topic of functional data analysis, see the book of Ramsay and Silverman [29]. In this functional statistics context, Biau, Devroye and Lugosi [7] investigate clustering with Hilbert norms and Laloë explores in [22] quantization and clustering with L^1 norms in Banach spaces.

In the present contribution, we go one step further and consider the problem of quantization and clustering when $d(\cdot, \cdot)$ is a general Bregman divergence $d_\phi(\cdot, \cdot)$ defined on a reflexive and separable Banach space E. Our approach extends and completes the results presented in [4], which focuses on the finite-dimensional setting and adopts a more algorithmic-oriented point of view. The paper is organized as follows. In Section 2, we set up notation and assumptions, and recall the relevant definitions. In Section 3, we provide conditions ensuring the existence of a minimizer q^* of the distortion $W(\mu, q)$ and its empirical counterpart q_n^*. Then, in Section 4, we focus on the convergence of the distortion and prove almost sure and L^1 convergence of $W(\mu, q_n^*)$ towards $W(\mu, q^*)$. Rates of convergence which do not depend on the dimension of E are also obtained, using Rademacher averages as complexity measures. For the sake of clarity, proofs are postponed to Section 5.

2 Context and assumptions

In this section, we formally define Bregman divergences, quantization and k-means clustering. We first need some notation and assumptions. Throughout the paper, $(E, \|\cdot\|)$ will denote a separable and reflexive Banach space, and $\mathcal{C}$ will be a measurable convex subset of E. Whenever E is a Hilbert space, $\langle \cdot, \cdot \rangle$ will stand for its inner product. Recall that the relative interior of a convex set $\mathcal{C}$, denoted hereafter by $ri(\mathcal{C})$, is its interior with respect to the affine hull. Finally, we will write $\partial\mathcal{C}$ for the complement of $ri(\mathcal{C})$ in its closure $\overline{\mathcal{C}}$.

We are now in a position to state the general definition of a Bregman divergence in E (Alber and Butnariu [1], Frigyik, Srivastava and Gupta [16]).

Definition 2.1. *Let $\mathcal{C}$ be a convex subset of E, and let $\phi : \mathcal{C} \to \mathbb{R}$ be strictly convex and twice continuously differentiable on $ri(\mathcal{C})$. The Bregman divergence associated with ϕ is defined by*

$$d_\phi(x, y) = \phi(x) - \phi(y) - D_y\phi(x - y),$$

where $D_y\phi$ denotes the Fréchet derivative of ϕ at y.

In particular, when E is a Hilbert space, it reduces to

$$d_\phi(x, y) = \phi(x) - \phi(y) - \langle x - y, \nabla\phi(y)\rangle.$$

Although Bregman divergences are not true metrics, they satisfy some interesting properties, such as non-negativity and separation, convexity in the first argument and linearity. For a complete description and proofs of these basic properties, the reader is referred to Bregman [9], Nielsen, Boissonnat and Nock [27] and Frigyik, Srivastava and Gupta [15]. Table 1 collects the most common examples of Bregman divergences.

Now, let X be a random variable with distribution μ, taking values in $\mathcal{C}$. Throughout the paper, we make the following assumptions:

1. $\mathbb{E}\|X\| < +\infty$.
2. $\mathbb{E}X \in ri(\mathcal{C})$.
3. $\mathbb{E}|\phi(X)| < +\infty$ and, for all $c \in ri(\mathcal{C})$, $\mathbb{E}|D_c\phi(X)| < +\infty$. This implies in particular that $\mathbb{E}d_\phi(X, c) < +\infty$ for all c.

Let $k \geq 1$. As already mentioned in the introduction, a k-quantizer is a Borel measurable mapping $q : \mathcal{C} \subset E \to \mathbf{c}$, where $\mathbf{c} = \{c_1, \ldots, c_\ell\}$, $\ell \leq k$, is a subset of $ri(\mathcal{C})$ called its codebook. In the sequel, the elements of $\mathbf{c}$ will also be named the centers associated to q. Every $x \in \mathcal{C}$ is represented by a unique $\hat{x} = q(x) \in \mathbf{c}$ and q induces a partition of $\mathcal{C}$ in cells $S_1, \ldots, S_\ell$. Each cell S_j is made of the elements of $\mathcal{C}$ whose image by q is c_j. Every k-quantizer is characterized by its codebook $\mathbf{c} = \{c_1, \ldots, c_\ell\}$ and its partition cells $S_1, \ldots, S_\ell$.

The error committed when representing X by $q(X)$ is assessed by the distortion

$$W(\mu, q) = \mathbb{E}d_\phi(X, q(X)) = \int_{\mathcal{C}} d_\phi(x, q(x))d\mu(x). \tag{1}$$

Bregman divergence	E	$\mathcal{C}$
Squared loss	$\mathbb{R}$	$\mathbb{R}$
Exponential loss	$\mathbb{R}$	$\mathbb{R}$
Norm-like	$\mathbb{R}$	$\mathbb{R}^+$
I-divergence (dim 1)	$\mathbb{R}$	$\mathbb{R}^+$
Logistic loss	$\mathbb{R}$	$[0,1]$
Itakura-Saito (dim 1)	$\mathbb{R}$	$(0,+\infty)$
Squared Euclidean distance	$\mathbb{R}^d$	$\mathbb{R}^d$
Mahalanobis distance	$\mathbb{R}^d$	$\mathbb{R}^d$
Kullback-Leibler (discrete)	$\mathbb{R}^d$	$(d-1)-$simplex
I-divergence (discrete)	$\mathbb{R}^d$	$(\mathbb{R}^+)^d$
Squared L^2 norm	$L^2(I,m)$	$L^2(I,m)$
Kullback-Leibler (continuous)	$L^2([0,1],dt)$	$\{x \in C^0([0,1]), \int_0^1 x(t)dt = 1\ \}$
I-divergence (continuous)	$L^2([0,1],dt)$	$\{x \in C^0([0,1]), x \geq 0\}$
Itakura-Saito (continuous)	$L^2_{2\pi}(dt)$	$\{x \in C^0_{2\pi}, x > 0\}$

Bregman divergence	$\phi(x)$	$d_\phi(x,y)$
Squared loss	x^2	$(x-y)^2$
Exponential loss	e^x	$e^x - e^y - (x-y)e^y$
Norm-like	x^α	$x^\alpha + (\alpha-1)y^\alpha - \alpha x y^{\alpha-1}$
I-divergence (dim 1)	$x\ln x$	$x\ln\frac{x}{y} - (x-y)$
Logistic loss	$x\ln x + (1-x)\ln(1-x)$	$x\ln\frac{x}{y} + (1-x)\ln\left(\frac{1-x}{1-y}\right)$
Itakura-Saito (dim 1)	$-\ln x$	$\frac{x}{y} - \ln\frac{x}{y} - 1$
Squared Euclidean distance	$\\|x\\|_2^2$	$\\|x-y\\|_2^2$
Mahalanobis distance	${}^t xAx$	${}^t(x-y)A(x-y)$
Kullback-Leibler (discrete)	$\sum_{\ell=1}^d x_\ell \ln x_\ell$	$\sum_{\ell=1}^d x_\ell \ln\frac{x_\ell}{y_\ell}$
I-divergence (discrete)	$\sum_{\ell=1}^d x_\ell \ln x_\ell$	$\sum_{\ell=1}^d x_\ell \ln\frac{x_\ell}{y_\ell} - \sum_{\ell=1}^d (x_\ell - y_\ell)$
Squared L^2 norm	$\int_I x^2(t)dm(t)$	$\\|x-y\\|_{L^2}^2$
Kullback-Leibler (continuous)	$\int_0^1 x(t)\ln x(t)dt$	$\int_0^1 x(t)\ln\frac{x(t)}{y(t)}dt$
I-divergence (continuous)	$\int_0^1 x(t)\ln x(t)dt$	$\int_0^1 x(t)\ln\frac{x(t)}{y(t)} + y(t) - x(t)dt$
Itakura-Saito (continuous)	$-\frac{1}{2\pi}\int_{-\pi}^{\pi}\ln(x(\theta))d\theta$	$-\frac{1}{2\pi}\int_{-\pi}^{\pi}\big(\ln\frac{x(\theta)}{y(\theta)} - \frac{x(\theta)}{y(\theta)} + 1\big)d\theta$

Table 1: Some examples of Bregman divergences. The matrix A is supposed to be positive definite. The notation $L^2(I,m)$ stands for the set of square integrable functions on an interval $I \subset \mathbb{R}$, with respect to the positive measure m, $L^2_{2\pi}(dt)$ for the set of 2π-periodic square integrable functions, $C^0([0,1])$ denotes the set of continuous functions on $[0,1]$, and $C^0_{2\pi}$ the set of 2π-periodic continuous functions.

Let

$$W^*(\mu) = \inf_{q \in \mathcal{Q}_k} W(\mu, q),$$

where $\mathcal{Q}_k$ is the set of all k-quantizers. To get a representation that is as accurate as possible, we look for an optimal quantizer, i.e., a quantizer q^* satisfying

$$W(\mu, q^*) = W^*(\mu).$$

In a statistical context, we only have at hand independent random observations $X_1, \ldots, X_n$ with distribution μ. The empirical distortion associated with $X_1, \cdots, X_n$ is given by

$$W(\mu_n, q) = \frac{1}{n} \sum_{i=1}^{n} d_\phi(X_i, q(X_i)), \tag{2}$$

where μ_n is the empirical measure. Observe that this is just the distortion (1) calculated with μ_n instead of μ. Clustering data into k groups means looking for an optimal quantizer q_n^* with respect to the empirical distortion (2).

Codebook and partition characterize a quantizer. As in the Euclidean case, it is easy to show that among all quantizers with same codebook, the best one (with respect to the distortion) is the nearest neighbor quantizer, whose partition $S_1, \ldots, S_\ell$ is the Voronoi partition, i.e.,

$$S_1 = \{x \in \mathcal{C}, d_\phi(x, c_1) \leq d_\phi(x, c_p), p = 1, \ldots, \ell\}$$

and for $j = 2, \ldots, \ell$,

$$S_j = \{x \in \mathcal{C}, d_\phi(x, c_j) \leq d_\phi(x, c_p), p = 1, \ldots, \ell\} \backslash \bigcup_{m=1}^{j-1} S_m$$

(see Linder [24]). If an optimal quantizer exists, it is necessarily a nearest neighbor quantizer. Hence, in the sequel, we will always consider nearest neighbor quantizers. Conversely, given a partition $\{S_j\}_{j=1}^{\ell}$, with $\mu(S_j) > 0$ and $\mathbb{E}[X|X \in S_j] \in ri(\mathcal{C})$ for $j = 1, \ldots, \ell$, the best quantizer is obtained by setting

$$c_j \in \arg\min_{c \in ri(\mathcal{C})} \mathbb{E}[d_\phi(X, c)|X \in S_j] \quad \text{for } j = 1, \ldots, \ell.$$

The next proposition, proved in Section 5, extends a result of Banerjee, Guo and Wang [3] to the case of functional Bregman divergences.

Proposition 2.1. *Let d_ϕ be a Bregman divergence. If S is a Borel subset of $\mathcal{C}$ with $\mu(S) > 0$ and $\mathbb{E}[X|X \in S] \in ri(\mathcal{C})$, the function*

$$c \mapsto \mathbb{E}[d_\phi(X, c)|X \in S]$$

reaches its infimum at a unique element of $ri(\mathcal{C})$, namely $\mathbb{E}[X|X \in S]$.

Thus, for every Bregman divergence, the minimizer is the conditional expectation, just like for the squared Euclidean distance. Observe that it is the median instead of the expectation when the distortion measure is an L^1 norm.

Observe that the combination of Proposition 2.1 and the optimality of the Voronoi partition is of computational interest. Indeed, even for squared Euclidean distance, minimizing the empirical distortion is generally a computationally hard problem, the complexity of an exact algorithm being exponential in the dimension of the space. In practice, a k-means type algorithm converging to local minima yields approximate solutions, and this adapts to general Bregman divergences. More precisely, given an initial codebook, which is made for instance of data items chosen at random, the algorithm proceeds by alternating between two steps. The first one consists in computing the Voronoi partition corresponding to the current centers. Then, during the second step, the new codebook is obtained by computing the mean of the data points falling in each cluster, according to Proposition 2.1. For further information on k-means algorithms with Bregman divergences, see Banerjee et al. [4].

3 Existence of an optimal quantizer

In this section, we look for conditions ensuring the existence of an optimal quantizer q^*, i.e., a q^* such that $W(\mu, q^*) = W^*(\mu)$. Since a nearest neighbor quantizer is characterized by its codebook $\mathbf{c} = (c_1, \ldots, c_k)$, we may rewrite the distortion

$$W(\mu, \mathbf{c}) = \mathbb{E} \min_{j=1,\ldots,k} d_\phi(X, c_j)$$

and look for an optimal codebook $\mathbf{c}^*$.

The existence of a minimum rests upon a compactness argument. We distinguish the finite dimensional case (Theorem 3.1) from the general case (Theorem 3.2). In finite dimension, we prove the result by exploiting an idea of Sabin and Gray [31] based on Alexandroff one-point-compactification (see, e.g., Dudley [14]).

Theorem 3.1 (Finite-dimensional case). *Assume that the convex set $\mathcal{C}$ lies in a finite-dimensional affine space and that the following statements hold:*

1. *For all $x \in \mathcal{C}$, the function $y \mapsto d_\phi(x, y)$ is lower semi-continuous on $ri(\mathcal{C})$.*

2. *For all $(x, y) \in \mathcal{C} \times ri(\mathcal{C})$, $d_\phi(x, y) \leq \liminf_{z \to \tilde{z} \in \partial\mathcal{C}} d_\phi(x, z)$ for all $\tilde{z} \in \partial\mathcal{C}$.*

3. *For all $(x, y) \in \mathcal{C} \times ri(\mathcal{C})$, $d_\phi(x, y) \leq \liminf_{\|z\| \to +\infty} d_\phi(x, z)$.*

Then, there exists an optimal codebook $\mathbf{c}^*$, i.e.,

$$W(\mu, \mathbf{c}^*) = W^*(\mu).$$

Requirement 1 is not restrictive since $y \mapsto d_\phi(x, y)$ is continuous for most well-known Bregman divergences. Observe that ϕ and $y \mapsto D_y\phi$ are continuous on $ri(\mathcal{C})$, so that condition 1 could be replaced by lower semi-continuity of $y \mapsto D_y\phi(y)$. Roughly speaking, requirements 2 and 3 prevent a possible minimizer from running to infinity. Note that condition 3 is void whenever $\mathcal{C}$ is bounded. In this case, $\overline{\mathcal{C}}$ is compact and the existence of an optimal codebook can easily be shown without resorting to Alexandroff compactification.

When E is potentially infinite-dimensional and $\mathcal{C}$ is any convex subset of E, things are not so simple, since Alexandroff compactification only applies to locally compact spaces. As we know that E is locally compact if and only if it is finite-dimensional (see for instance Dudley [14]), this tool is not suited to the infinite-dimensional case. However, since E is reflexive, a closed and bounded convex subset of E is compact for the weak topology $\sigma(E, E')$, that is the coarsest topology on E making all continuous linear forms on E continuous. Moreover, every weakly lower semi-continuous function reaches its minimum on a weakly compact set. Thus, if we know in advance that $\mathbf{c}^*$ is to be searched for in a closed and bounded convex set, an argument of continuity suffices to show the existence of $\mathbf{c}^*$. In the sequel, $\mathcal{C}_R \subset ri(\mathcal{C})$ will denote a closed and bounded convex set of diameter $2R$. For example, $\mathcal{C}_R = B(0, R) = \{x \in E, \|x\| \leq R\}$ the closed ball of center 0 and radius R. A key fact is that $X \in \mathcal{C}_R$ implies that $\mathbf{c}^* \in \mathcal{C}_R$ if it exists, by Bregman projection (Alber and Butnariu [1]).

For further details about weak convergence and lower semi-continuous and convex functions, the reader is referred to Brezis [10] and Rockafellar [30].

Theorem 3.2 (**General case**). *Suppose that there exists $R > 0$ such that $\mathbb{P}\{X \in \mathcal{C}_R\} = 1$, and that for all $x \in \mathcal{C}$, $y \mapsto d_\phi(x, y)$ is weakly lower semi-continuous on $\mathcal{C}_R$. Then, there exists an optimal quantizer.*

Example 3.1 Convex functions which are lower semi-continuous for the norm are examples of weakly lower semi-continuous functions (see, e.g., [10]).

Observe that since the weak topology coincides with the norm topology in finite dimension, the term “weakly” in Theorem 3.2 can be dropped whenever E is finite-dimensional.

In fact, if we only have $\mathcal{C}_R \cap ri(\mathcal{C}) \neq \emptyset$ instead of $\mathcal{C}_R \subset ri(\mathcal{C})$, but ϕ is of Legendre type (see Rockafellar [30], and for the infinite-dimensional definition, Bauschke, Borwein and Combettes [6]), it remains possible to use Bregman projection to obtain the same result.

In the particular case where $d_\phi(\cdot, \cdot)$ is the squared distance induced by the inner product of a Hilbert space, it can be shown (see Section 5) that it is sufficient to look for an optimal quantizer on a ball. Hence, Theorem 3.2 admits the following corollary.

Corollary 3.1. *Let E be a Hilbert space. If $\phi(\cdot) = \|\cdot\|^2$, there exists an optimal quantizer corresponding to the Bregman divergence $d_\phi(\cdot,\cdot)$.*

In the last part of this section, we turn to the existence of an empirically optimal quantizer. In other words, we will look for a minimizer $\mathbf{c}_n^*$ of the empirical distortion

$$W(\mu_n, \mathbf{c}) = \frac{1}{n}\sum_{i=1}^{n} \min_{j=1,\ldots,k} d_\phi(X_i, c_j).$$

Since the support of the empirical measure μ_n contains at most n points, it is included in a closed ball B_R. Thus, Theorem 3.2 implies the following result.

Corollary 3.2. *Assume that for all $x \in \mathcal{C}$, $y \mapsto d_\phi(x, y)$ is weakly lower semi-continuous. Then, there exists an empirically optimal quantizer.*

As above, the term "weakly" may be omitted when E is finite-dimensional.

4 Convergence

4.1 Convergence of the distortion

Suppose that there exists an optimal codebook $\mathbf{c}_n^*$ that achieves the minimum of the empirical distortion $W(\mu_n, \mathbf{c})$. We turn our attention to the "true" distortion $W(\mu, \mathbf{c})$ for $\mathbf{c} = \mathbf{c}_n^*$ and would like to know if this quantity gets close to the minimal distortion $W^*(\mu)$ as the number n of observations becomes large.

Assuming that $\mathbf{c}^*$ exists,

$$\begin{aligned} W(\mu, \mathbf{c}_n^*) - W^*(\mu) &= W(\mu, \mathbf{c}_n^*) - W(\mu, \mathbf{c}^*) \\ &= W(\mu, \mathbf{c}_n^*) - W(\mu_n, \mathbf{c}_n^*) + W(\mu_n, \mathbf{c}_n^*) - W(\mu, \mathbf{c}^*) \\ &\leq W(\mu, \mathbf{c}_n^*) - W(\mu_n, \mathbf{c}_n^*) + W(\mu_n, \mathbf{c}^*) - W(\mu, \mathbf{c}^*) \\ &\leq 2 \sup_{\mathbf{c} \in ri(\mathcal{C})^k} |W(\mu_n, \mathbf{c}) - W(\mu, \mathbf{c})|. \end{aligned}$$

Thus, if we intend to show that $W(\mu, \mathbf{c}_n^*)$ converges to $W^*(\mu)$ as n tends to infinity, it will be enough to prove that $\sup_{\mathbf{c} \in ri(\mathcal{C})^k} |W(\mu_n, \mathbf{c}) - W(\mu, \mathbf{c})|$ vanishes as n tends to infinity.

As in the previous section, we distinguish the finite-dimensional case (Theorem 4.1) from the general case (Theorem 4.2).

Theorem 4.1 (**Finite-dimensional case**). *Assume that $\mathcal{C}$ lies in a finite-dimensional affine space and that the following statements hold:*

1. *The Bregman divergence $d_\phi(\cdot,\cdot)$ is continuous.*
2. *For all $x \in \mathcal{C}$, $\tilde{z} \in \partial\mathcal{C}$, $\lim_{z\to\tilde{z}\in\partial\mathcal{C}} d_\phi(x,z) = +\infty$.*
3. *For all $x \in \mathcal{C}$, $\lim_{\|z\|\to+\infty} d_\phi(x,z) = +\infty$.*
4. *For all $x \in \mathcal{C}$, the function $y \mapsto d_\phi(x,y)$ is convex on $ri(\mathcal{C})$.*

Then, if $\mathbf{c}_n^*$ is a minimizer of the empirical distortion,

$$\lim_{n\to+\infty} W(\mu,\mathbf{c}_n^*) = W^*(\mu) \quad a.s.$$

Note that the existence of $\mathbf{c}_n^*$ (and $\mathbf{c}^*$) is guaranteed under these assumptions.

In view of the definition of ϕ, the requirement 1 could be replaced by the continuity of $(x,y) \mapsto D_y\phi(x-y)$. Condition 4 is not necessarily satisfied for each Bregman divergence. For instance, the Itakura-Saito divergence $d_\phi(x,y) = \frac{x}{y} - \ln\frac{x}{y} - 1$ is not convex in the second argument.

As for the existence of an optimal quantizer, the infinite-dimensional setting requires further hypotheses, as expressed in the following theorem:

Theorem 4.2 (**General case**). *Assume that for all $x \in \mathcal{C}$, $y \mapsto d_\phi(x,y)$ is weakly lower semi-continuous, so that there exists a minimizer $\mathbf{c}_n^*$ of the empirical distortion. If there exists $R > 0$ such that $\mathbb{P}\{X \in \mathcal{C}_R\} = 1$, and $M = M(\phi,R) \geq 0$ such that, for all $c \in \mathcal{C}_R$, $\|D_c\phi\| \leq M$, then*

$$\lim_{n\to+\infty} W(\mu,\mathbf{c}_n^*) = W^*(\mu) \quad a.s.$$

and

$$\lim_{n\to+\infty} \mathbb{E}W(\mu,\mathbf{c}_n^*) = W^*(\mu).$$

Let us point out that the convergence results $\lim_{n\to+\infty} W(\mu,\mathbf{c}_n^*) = W^*(\mu)$ *a.s.* and $\lim_{n\to+\infty} \mathbb{E}W(\mu,\mathbf{c}_n^*) = W^*(\mu)$ always hold when $\phi(\cdot) = \|\cdot\|^2$ (Biau, Devroye and Lugosi [7]).

Let us now discuss some examples.

Example 4.1 (I-divergence in dimension 1). Here $E = \mathbb{R}$, $\mathcal{C} = \mathbb{R}^+$ and $d_\phi(x,y) = x\ln\frac{x}{y} - (x-y)$. Let $x \in \mathcal{C}$. The map $y \mapsto x\ln\frac{x}{y} - (x-y)$ is continuous and convex on $ri(\mathcal{C}) = (0,+\infty)$ (its second derivative is $\frac{x}{y^2} \geq 0$) and tends to $+\infty$ as y tends to 0 or $+\infty$. Thus there exists a quantizer whose codebook achieves the minimum of the distortion $W(\mu,\mathbf{c})$ (Theorem 3.1) as well as an empirically optimal quantizer (Corollary 3.2). Moreover, if $\mathbf{c}_n^*$ is a minimizer of the empirical distortion, almost sure convergence of $W(\mu,\mathbf{c}_n^*)$ to $W^*(\mu)$ is ensured (Theorem 4.1).

Example 4.2 (Exponential loss). Let $E = \mathcal{C} = \mathbb{R}$ and $\phi(x) = e^x$, which yields $d_\phi(x, y) = e^x - e^y - (x-y)e^y$. The function $y \mapsto e^x - e^y - (x-y)e^y$ is continuous on $\mathbb{R}$. If $\mathbb{P}\{|X| \leq R\} = 1$, there exists an optimal quantizer (Theorem 3.2), and since $\phi'(x) = e^x \leq e^R$ on $[-R, R]$, $W(\mu, \mathbf{c}_n^*)$ converges almost surely and in L^1 to $W^*(\mu)$ (Theorem 4.2).

Example 4.3 (Squared Euclidean distance). When $d_\phi(\cdot, \cdot)$ is the squared Euclidean distance, existence of an optimal quantizer, almost sure and L^1 convergence of the distortion are guaranteed.

Example 4.4 (Kullback-Leibler divergence between discrete probability measures). Here, $E = \mathbb{R}^d$, $\mathcal{C}$ is the $(d-1)$-simplex and $d_\phi(p, q) = \sum_{\ell=1}^d p_\ell \ln \frac{p_\ell}{q_\ell}$. The function $q = (q_1, \dots, q_d) \mapsto \sum_{\ell=1}^d p_\ell \ln \frac{p_\ell}{q_\ell}$ is continuous and convex on $ri(\mathcal{C}) = \{(p_1, \dots, p_d) \in (0, +\infty)^d, \sum_{\ell=1}^d p_\ell = 1\}$ and tends to $+\infty$ as one of the q_ℓ's tends to 0. Thus, there exists an optimal quantizer and we have almost sure convergence of the distortion.

Example 4.5 (Squared L^2 distance). Let $E = \mathcal{C} = L^2([0, 1], dt)$, and $d_\phi(x, y) = \int_0^1 (x(t) - y(t))^2 dt$. This is a Hilbert norm, thus the existence of a minimizer of the distortion as well as convergence are guaranteed.

Example 4.6 (I-divergence). Let $E = L^2([0, 1], dt)$ and let $\mathcal{C}$ be the set of all continuous non-negative elements of E. Here $d_\phi(p, q) = \int_0^1 [p(t) \ln \frac{p(t)}{q(t)} + q(t) - p(t)]dt$. The map $q \mapsto d_\phi(p, q)$ is continuous and convex and therefore weakly semi-continuous. Assume that $\mathbb{P}\{r \leq \|X\| \leq R\} = 1$ $(r > 0)$. Then, there exists an optimal quantizer. Moreover, we have almost sure and L^1 convergence of the distortion.

4.2 Rates of convergence

The previous section indicates that $W(\mu, \mathbf{c}_n^*)$ gets close to the minimal distortion when the sample size grows. However, it gives no information about the rates of convergence. To address this question, let us first observe that minimizing

$$W(\mu, \mathbf{c}) = \mathbb{E} \min_{j=1,\dots,k} d_\phi(X, c_j) = \mathbb{E} \min_{j=1,\dots,k} \big(\phi(X) - \phi(c_j) - D_{c_j}\phi(X - c_j)\big)$$

is equivalent to minimizing the quantity

$$\overline{W}(\mu, \mathbf{c}) = \mathbb{E} \min_{j=1,\dots,k} \big(-\phi(c_j) - D_{c_j}\phi(X - c_j)\big).$$

Similarly, to $W(\mu_n, \mathbf{c})$, we associate

$$\overline{W}(\mu_n, \mathbf{c}) = \frac{1}{n} \sum_{i=1}^n \min_{j=1,\dots,k} \big(-\phi(c_j) - D_{c_j}\phi(X_i - c_j)\big).$$

Since

$$W(\mu, \mathbf{c}_n^*) - \inf_{\mathbf{c} \in ri(\mathcal{C})^k} W(\mu, \mathbf{c}) = \overline{W}(\mu, \mathbf{c}_n^*) - \inf_{\mathbf{c} \in ri(\mathcal{C})^k} \overline{W}(\mu, \mathbf{c})$$

and

$$\begin{aligned}&\mathbb{E}\overline{W}(\mu, \mathbf{c}_n^*) - \inf_{\mathbf{c} \in ri(\mathcal{C})^k} \overline{W}(\mu, \mathbf{c}) \\ &\quad \leq \mathbb{E} \sup_{\mathbf{c} \in ri(\mathcal{C})^k} \left(\overline{W}(\mu_n, \mathbf{c}) - \overline{W}(\mu, \mathbf{c}) \right) + \mathbb{E} \sup_{\mathbf{c} \in ri(\mathcal{C})^k} \left(\overline{W}(\mu, \mathbf{c}) - \overline{W}(\mu_n, \mathbf{c}) \right) \qquad (3)\end{aligned}$$

(see Lemma 8.2 in Devroye, Györfi and Lugosi [13]), we are done if we can find upper bounds for the uniform deviation

$$\mathbb{E} \sup_{\mathbf{c} \in ri(\mathcal{C})^k} \left(\overline{W}(\mu_n, \mathbf{c}) - \overline{W}(\mu, \mathbf{c}) \right).$$

(The second term of the right-hand side of (3) can indeed be bounded by an upper bound of the first term.) The next theorem may be proved by resorting to the Rademacher averages as a complexity measure for a function class (see, e.g., Bartlett, Boucheron, and Lugosi [5]).

Theorem 4.3. *For $\mathcal{C}_R \subset ri(\mathcal{C})$, the following inequality holds:*

$$\begin{aligned}&\mathbb{E} \sup_{\mathbf{c} \in \mathcal{C}_R^k} \left(\overline{W}(\mu_n, \mathbf{c}) - \overline{W}(\mu, \mathbf{c}) \right) \\ &\qquad \leq \frac{2k}{\sqrt{n}} \Big(\sup_{c \in \mathcal{C}_R} |-\phi(c) + D_c\phi(c)| + \sup_{c \in \mathcal{C}_R} \|D_c\phi\| (\mathbb{E}\|X\|^2)^{1/2} \Big).\end{aligned}$$

Corollary 4.1. *Suppose that for all $x \in \mathcal{C}$, $y \mapsto d_\phi(x, y)$ is weakly lower semi-continuous, which ensures the existence of an optimal codebook $\mathbf{c}_n^*$. Assume that there exists $R > 0$ such that $\mathbb{P}\{X \in \mathcal{C}_R\} = 1$. If $|-\phi(c) + D_c\phi(c)|$ and $\|D_c\phi\|$ are uniformly bounded on $\mathcal{C}_R$ by $M_1 = M_1(\phi, R) \geq 0$ and $M_2 = M_2(\phi, R) \geq 0$ respectively, then*

$$\mathbb{E}W(\mu, \mathbf{c}_n^*) - W^*(\mu) \leq \frac{4k}{\sqrt{n}} \left(M_1 + M_2 (\mathbb{E}\|X\|^2)^{1/2} \right),$$

and thus

$$\mathbb{E}W(\mu, \mathbf{c}_n^*) - W^*(\mu) \leq \frac{4k}{\sqrt{n}} \left(M_1 + M_2 R \right).$$

Note that Corollary 4.1 yields dimension-free upper bounds. This is worth pointing out since E is allowed to be high (or even infinite)-dimensional.

Example 4.1 In this example, we give the bounds obtained for some usual Bregman divergences. We assume throughout that there exists $R > 0$ such that $\mathbb{P}\{\|X\| \leq R\} = 1$.

1. Squared loss. For $\phi(x) = x^2$,

$$\mathbb{E}W(\mu, \mathbf{c}_n^*) - W^*(\mu) \le \frac{4k}{\sqrt{n}} \left(R^2 + 2R(\mathbb{E}|X|^2)^{1/2}\right),$$

   and then

$$\mathbb{E}W(\mu, \mathbf{c}_n^*) - W^*(\mu) \le \frac{12kR^2}{\sqrt{n}}.$$

2. Exponential loss. For $\phi(x) = e^x$,

$$\mathbb{E}W(\mu, \mathbf{c}_n^*) - W^*(\mu) \le \frac{4k(2R-1)e^R}{\sqrt{n}}.$$

3. Squared Euclidean distance. For the squared Euclidean norm $\phi(x) = \|x\|^2$,

$$\mathbb{E}W(\mu, \mathbf{c}_n^*) - W^*(\mu) \le \frac{12kR^2}{\sqrt{n}}.$$

4. Mahalanobis distance. For $\phi(x) = {}^t xAx$ with A positive definite,

$$\mathbb{E}W(\mu, \mathbf{c}_n^*) - W^*(\mu) \le \frac{12k\|A\|R^2}{\sqrt{n}}.$$

5. Squared L^2 distance. When ϕ is a squared L^2 norm,

$$\mathbb{E}W(\mu, \mathbf{c}_n^*) - W^*(\mu) \le \frac{12kR^2}{\sqrt{n}}.$$

Remark 4.1 Some Bregman divergences, typically Kullback-Leibler, involve a logarithm, which prevents $||D_c\phi||$ from being uniformly bounded on a ball B_R. In order to circumvent this difficulty, a possible solution is to consider a class of elements of E satisfying the following assumption:

- In dimension 1, $0 < r \le X \le R < +\infty \quad a.s.$
- In dimension d $(2 \le d \le +\infty)$, when the logarithm appears in a sum or an integral, $\sum_{\ell=1}^d \ln^2(x_\ell) \le M(R)$ or $\int \ln^2(x(t))dt \le M(R)$.

Several conditions of this type can be found in the literature on Kullback-Leibler divergence. For instance, Jordan, Nguyen and Wainwright [21], who develop an estimation method for the Kullback-Leibler divergence, require an envelope condition or boundedness from above and below.

As an illustration, let $d_\phi(x, y) = \int_0^1 x(t) \ln \frac{x(t)}{y(t)} dt$. Suppose that $\mathbb{P}\{\|X\| \le R\} = 1$ for some $R > 0$ and that $\int_0^1 \ln^2(X(t))dt \le R^2$. Assuming that the codebooks belong to the same function class as X, we obtain

$$\mathbb{E}W(\mu, \mathbf{c}_n^*) - W^*(\mu) \le \frac{2kR}{\sqrt{n}}(1 + R).$$

5 Proofs

5.1 Proof of Proposition 2.1

The proof of Banerjee, Guo and Wang [3] may be adapted to the infinite-dimensional case. We will check that $\mathbb{E}[X|X \in S]$ minimizes $\mathbb{E}[d_\phi(X,c)|X \in S]$ and that it is the only element of $ri(\mathcal{C})$ with this property. For $c \in ri(\mathcal{C})$,

$$\begin{aligned}
&\mathbb{E}[d_\phi(X,c)|X \in S] - \mathbb{E}[d_\phi(X,\mathbb{E}[X|X \in S])|X \in S] \\
&\quad = \mathbb{E}[\phi(X) - \phi(c) - D_c\phi(X-c) - \phi(X) + \phi(\mathbb{E}[X|X \in S]) \\
&\qquad + D_{\mathbb{E}[X|X\in S]}\phi(X - \mathbb{E}[X|X \in S])|X \in S] \\
&\quad = \phi(\mathbb{E}[X|X \in S]) - \phi(c) - D_c\phi(\mathbb{E}[X|X \in S] - c) \\
&\quad = d_\phi(\mathbb{E}[X|X \in S], c).
\end{aligned}$$

Indeed, expectation and derivation intertwine, since the derivative is a continuous linear form (see, e.g., Proposition 1.1.6 in Arendt, Batty, Hieber and Neubrander [2]). However,

$$d_\phi(\mathbb{E}[X|X \in S], c) \geq 0$$

and $d_\phi(\mathbb{E}[X|X \in S], c) = 0$ if and only if $c = \mathbb{E}[X|X \in S]$. Hence,

$$\mathbb{E}[d_\phi(X,c)|X \in S] \geq \mathbb{E}[d_\phi(X,\mathbb{E}[X|X \in S])|X \in S],$$

and equality holds if and only if $c = \mathbb{E}[X|X \in S]$. Thus, $\mathbb{E}[X|X \in S]$ is the unique minimizer of the function $c \mapsto \mathbb{E}[d_\phi(X,c)|X \in S]$ on $ri(\mathcal{C})$.

5.2 Proof of Theorem 3.1

Setting $d_\phi(x,\tilde{z}) = \liminf_{z\to\tilde{z}\in\partial\mathcal{C}} d_\phi(x,z)$ for all $x \in \mathcal{C}$ and $\tilde{z} \in \partial\mathcal{C}$, $y \mapsto d_\phi(x,y)$ extends to a lower semi-continuous function $\overline{\mathcal{C}} \to [0,+\infty]$. We compactify $\overline{\mathcal{C}}$ by adding a point at infinity ω. Let $\tilde{\mathcal{C}} = \overline{\mathcal{C}} \cup \{\omega\}$ denote the Alexandroff compactification of $\overline{\mathcal{C}}$ (for details about the Alexandroff one-point compactification, see for instance Dudley [14]). By Tychonoff's theorem, (see, e.g., Dudley [14]) the product $\tilde{\mathcal{C}}^k$ is also compact. We set for all $x \in \mathcal{C}$, $d_\phi(x,\omega) = \liminf_{\|z\|\to+\infty} d_\phi(x,z)$. According to the assumptions, the function $y \mapsto d_\phi(x,y)$ from $\tilde{\mathcal{C}}$ to $[0,+\infty]$ is lower semi-continuous, that is the level set $\{c \in \tilde{\mathcal{C}}, d_\phi(x,c) \leq \lambda\}$ is closed for all $\lambda \in \mathbb{R}$. Since $\{\mathbf{c} \in \tilde{\mathcal{C}}^k, \min_{j=1,\ldots,k} d_\phi(x,c_j) \leq t\} = \bigcup_{j=1}^k \{\mathbf{c} \in \tilde{\mathcal{C}}^k, d_\phi(x,c_j) \leq t\}$, the level sets of $\mathbf{c} \mapsto \min_{j=1,\ldots,k} d_\phi(x,c_j)$ are also closed, i.e., this function is lower semi-continuous. Hence, for $\mathbf{c} \in \tilde{\mathcal{C}}^k$,

$$\begin{aligned}
\liminf_{\mathbf{c}' \to \mathbf{c}} W(\mu, \mathbf{c}') &= \liminf_{\mathbf{c}' \to \mathbf{c}} \int_{\mathcal{C}} \min_{j=1,\ldots,k} d_\phi(x, c'_j) d\mu(x) \\
&\geq \int_{\mathcal{C}} \liminf_{\mathbf{c}' \to \mathbf{c}} \min_{j=1,\ldots,k} d_\phi(x, c'_j) d\mu(x) \\
&\quad \text{(by Fatou's Lemma)} \\
&\geq \int_{\mathcal{C}} \min_{j=1,\ldots,k} d_\phi(x, c_j) d\mu(x) \\
&= W(\mu, \mathbf{c}).
\end{aligned}$$

Thus, $\mathbf{c} \mapsto W(\mu, \mathbf{c})$ is lower semi-continuous on the compact $\tilde{\mathcal{C}}^k$, and it reaches its minimum at some codebook $\mathbf{c}^*$. By conditions 2 and 3, we can assume that $\mathbf{c}^* \in ri(\mathcal{C})^k$: if not, we replace the coordinates which belong to $\partial\mathcal{C}$ or equal ω by elements of $ri(\mathcal{C})$. Therefore, there exists an optimal codebook $\mathbf{c}^*$, and the result is proved.

5.3 Proof of Theorem 3.2

Since $\mathbb{P}\{X \in \mathcal{C}_R\} = 1$, it suffices to look for a minimizer $\mathbf{c}^*$ of the distortion over $\mathcal{C}_R^k$. Indeed,

$$\forall c \in ri(\mathcal{C}), d_\phi(X, c) \geq d_\phi(X, \bar{c})$$

with $\bar{c}$ the Bregman projection [1] of c on $\mathcal{C}_R$. Thus, for any codebook $\mathbf{c} = (c_1, \ldots, c_k)$, if $\bar{\mathbf{c}} = (\bar{c}_1, \ldots, \bar{c}_k)$, $\mathbb{E} \min_{j=1,\ldots,k} d_\phi(X, c_j) \geq \mathbb{E} \min_{j=1,\ldots,k} d_\phi(X, \bar{c}_j)$, i.e., $W(\mu, \mathbf{c}) \geq W(\mu, \bar{\mathbf{c}})$. This shows that projecting any center on the closed and bounded convex set $\mathcal{C}_R$ can only reduce the distortion. Since E is reflexive, $\mathcal{C}_R$ is weakly compact by Kakutani's Theorem (see [10] for instance), and so is $\mathcal{C}_R^k$. Let us show that $W(\mu, \cdot)$ is weakly lower semi-continuous. The function $y \mapsto d_\phi(x, y)$ is weakly lower semi-continuous for all $x \in \mathcal{C}$. This means that for all $\lambda \in \mathbb{R}$, the level sets $\{c \in \mathcal{C}_R, d_\phi(x, c) \leq \lambda\}$ are weakly closed. As $\{\mathbf{c} \in \mathcal{C}_R^k, \min_{j=1,\ldots,k} d_\phi(x, c_j) \leq t\} = \bigcup_{j=1}^k \{\mathbf{c} \in \mathcal{C}_R^k, d_\phi(x, c_j) \leq t\}$, the level sets of $\mathbf{c} \mapsto \min_{j=1,\ldots,k} d_\phi(x, c_j)$ are weakly closed as well, and this function is weakly lower semi-continuous. If $\mathbf{c}'$ converges weakly to $\mathbf{c}$,

$$\begin{aligned}
\liminf_{\mathbf{c}' \to \mathbf{c}} W(\mu, \mathbf{c}') &= \liminf_{\mathbf{c}' \to \mathbf{c}} \int \min_{j=1,\ldots,k} d_\phi(x, c'_j) d\mu(x) \\
&\geq \int \liminf_{\mathbf{c}' \to \mathbf{c}} \min_{j=1,\ldots,k} d_\phi(x, c'_j) d\mu(x) \\
&\quad \text{(Fatou's Lemma)} \\
&\geq \int \min_{j=1,\ldots,k} d_\phi(x, c_j) d\mu(x) = W(\mu, \mathbf{c})
\end{aligned}$$

since $\mathbf{c} \mapsto \min_{j=1,\ldots,k} d_\phi(x, c_j)$ is weakly lower semi-continuous. Thus $W(\mu, \cdot)$ is weakly lower semi-continuous on a weakly compact set, which implies that it reaches its minimum, i.e., there exists $\mathbf{c}^* \in \mathcal{C}_R^k$, such that $W(\mu, \mathbf{c}^*) = W^*(\mu)$.

5.4 Proof of Corollary 3.1

The result follows from Theorem 3.2 and from the following lemma whose proof is close to the first part of the proof of Theorem 1 in Linder [24] (see also Pollard [28]).

Lemma 5.1 Let d_ϕ be a Bregman divergence. Assume that the second derivative of $\phi : E \to \mathbb{R}$ is uniformly strongly positive, i.e., there exists $m = m(\phi) > 0$ such that for all c, $D_c^2\phi(x,x) \geq m\|x\|^2$, and that there exists $M = M(\phi)$ such that for all c, $\|D_c^2\phi\| \leq M$. Then,

$$\inf_{\mathbf{c}\in E^k} W(\mu,\mathbf{c}) = \inf_{\mathbf{c}\in B_R^k} W(\mu,\mathbf{c})$$

for some $R > 0$.

Proof of Lemma 5.1. By Taylor's formula, there exists z belonging to the open segment xy, such that

$$\phi(x) = \phi(y) + D_y\phi(x-y) + \frac{1}{2}D_z^2\phi(x-y,x-y).$$

Thus,

$$d_\phi(x,y) = \frac{1}{2}D_z^2\phi(x-y,x-y),$$

which implies, by assumption,

$$\frac{m}{2}\|x-y\|^2 \leq d_\phi(x,y) \leq \frac{M}{2}\|x-y\|^2.$$

For an integer $\ell \geq 1$ and $\mathbf{c}^\ell = (c_1,\ldots,c_\ell)$, let

$$w_\ell(\mathbf{c}^\ell) = \mathbb{E}\min_{j=1,\ldots,\ell} d_\phi(x,c_j).$$

Let $W_\ell^*(\mu)$ denote the optimal distortion with respect to ℓ-quantizers. Since Corollary 3.1 corresponds to Proposition 2.1 when $k=1$, we suppose $k \geq 2$. Moreover, we can assume that the support of μ contains at least k points (otherwise, we would not look for a k-quantizer), so that $W_k^*(\mu) < W_{k-1}^*(\mu)$. Let $\varepsilon > 0$ such that

$$\varepsilon < \frac{1}{2}(W_{k-1}^*(\mu) - W_k^*(\mu)) \tag{4}$$

and let $0 < r_1 < r_2$ such that

$$\frac{m}{2}(r_2-r_1)^2\mu(B_{r_1}) > W_k^*(\mu) + \varepsilon \tag{5}$$

and

$$2M\int_{B_{2r_2}^c} \|x\|^2 d\mu(x) < \varepsilon. \tag{6}$$

We choose a codebook $\mathbf{c}^k = (c_1, \dots, c_k)$ such that

$$w_k(\mathbf{c}^k) < W_k^*(\mu) + \varepsilon.$$

This implies

$$w_k(\mathbf{c}^k) < W_{k-1}^*(\mu) - \varepsilon,$$

which ensures that $c_1, \dots, c_k$ are distinct. Assume that these elements are sorted in increasing order, that is $\|c_1\| \leq \cdots \leq \|c_k\|$. Then, $\|c_1\| \leq r_2$. To see this, suppose that $\|c_1\| > r_2$. This means that $\|c_j\| > r_2$ for all j. Thus, for $x \in B_{r_1}$,

$$\begin{aligned}\min_{j=1,\dots,k} d_\phi(x, c_j) &\geq \frac{m}{2} \min_{j=1,\dots,k} \|x - c_j\|^2 \\ &\geq \frac{m}{2} \min_{j=1,\dots,k} (\|c_j\| - \|x\|)^2 \\ &\geq \frac{m}{2}(r_2 - r_1)^2.\end{aligned}$$

Hence, $W_k^*(\mu) + \varepsilon > \frac{m}{2}(r_2 - r_1)^2 \mu(B_{r_1})$, contradicting inequality (5). We will now show that for all j, $\|c_j\| \leq C r_2$ where $C = 2 + 3\sqrt{\frac{M}{m}} > 0$. Suppose that $\|c_k\| > C r_2$. On the event $\{x \in B_{2r_2}\}$,

$$d_\phi(x, c_1) \leq \frac{M}{2}(\|x\| + \|c_1\|)^2 \leq \frac{9}{2} M r_2^2$$

and

$$d_\phi(x, c_k) \geq \frac{m}{2}(\|c_k\| - \|x\|)^2 > \frac{m}{2}(C r_2 - 2 r_2)^2 = \frac{9}{2} M r_2^2,$$

thus

$$d_\phi(x, c_1) \leq d_\phi(x, c_k).$$

On $\{x \in B_{2r_2}^c\}$,

$$d_\phi(x, c_1) \leq \frac{M}{2}(\|x\| + \|c_1\|)^2 \leq 2M\|x\|^2.$$

Then

$$d_\phi(x, c_1) \leq d_\phi(x, c_k) + 2M\|x\|^2 \mathbf{1}_{\{x \in B_{2r_2}^c\}}. \tag{7}$$

Let $\mathbf{c}^{k-1} = (c_1, \dots, c_{k-1})$ and let $\{S_j\}_{j=1}^k$ denote the Voronoi partition associated with the components of $\mathbf{c}^k$. We obtain

$$\begin{aligned}w_{k-1}(\mathbf{c}^{k-1}) &= \sum_{j=1}^{k} \int_{S_j} \min_{j=1,\dots,k-1} d_\phi(x, c_j) d\mu(x) \\ &\leq \sum_{j=1}^{k-1} \int_{S_j} d_\phi(x, c_j) d\mu(x) + \int_{S_k} d_\phi(x, c_1) d\mu(x) \\ &\leq \sum_{j=1}^{k} \int_{S_j} d_\phi(x, c_j) d\mu(x) + 2M \int_{B_{2r_2}^c} \|x\|^2 d\mu(x).\end{aligned}$$

The last statement follows from inequality (7). Then, by inequalities (6) and (4),

$$\begin{aligned} w_{k-1}(\mathbf{c}^{k-1}) &\leq w_k(\mathbf{c}^k) + \varepsilon \\ &\leq W_k^*(\mu) + 2\varepsilon \\ &< W_{k-1}^*(\mu). \end{aligned}$$

This contradicts the definition of $W_{k-1}^*(\mu)$. Thus, $w_k(\mathbf{c}^k) < W_k^*(\mu) + \varepsilon$ implies $(c_1, \ldots, c_k) \in (B_{Cr_2})^k$, and finally, setting $R = Cr_2$,

$$W_k^*(\mu) = \inf_{\mathbf{c}^k \in B_R^k} w_k(\mathbf{c}^k).$$

□

5.5 Proof of Theorem 4.1

As mentioned earlier, to prove Theorem 4.1, it is enough to show that $W(\mu_n, \cdot)$ converges uniformly to $W(\mu, \cdot)$ almost surely. The method is inspired from Sabin and Gray [31] again. As in the proof of Theorem 3.1, we define the Bregman divergence $d_\phi(\cdot, \cdot)$ on $\mathcal{C} \times \tilde{\mathcal{C}}$ with $\tilde{\mathcal{C}}$ the Alexandroff compactification of $\overline{\mathcal{C}}$. The assumptions imply that the extended function $d_\phi(\cdot, \cdot)$ is continuous. Continuous convergence on a compact set is equivalent to uniform convergence (see, e.g., Theorem 3.1.9 in Lojasiewicz [26]). Hence, as $\tilde{\mathcal{C}}^k$ is compact, it suffices to show that if $(\mathbf{c}_n)_{n \in \mathbb{N}}$ is a sequence of points in $\tilde{\mathcal{C}}^k$ converging to $\mathbf{c}$, then

$$\lim_{n \to +\infty} W(\mu_n, \mathbf{c}_n) = W(\mu, \mathbf{c}) \quad a.s.$$

By a theorem of Varadarajan (Theorem 11.4.1 in Dudley [14] for example), almost surely, the empirical measure μ_n converges weakly to μ. Since E is a separable Banach space, by Skorohod's Representation Theorem (see, e.g., Theorem 11.7.2 in [14]), there exist random variables Y and Y_n, defined on the same probability space, such that Y has distribution μ, Y_n has distribution μ_n, and Y_n converges to Y almost surely. Since the extended function $d_\phi(\cdot, \cdot)$ is continuous, $\min_{j=1,\ldots,k} d_\phi(x_n, c_{nj})$ converges to $\min_{j=1,\ldots,k} d_\phi(x, c_j)$ as $(x_n, \mathbf{c}_n)$ converges to $(x, \mathbf{c})$. Therefore, as $\mathbf{c}_n$ converges to $\mathbf{c}$, $\min_{j=1,\ldots,k} d_\phi(Y_n, c_{nj})$ converges almost surely (and thus in distribution) to $\min_{j=1,\ldots,k} d_\phi(Y, c_j)$. Moreover, for all c, $d_\phi(Y_n, c)$ converges to $d_\phi(Y, c)$ almost surely, thus also in distribution.

If for all $j = 1, \ldots, k$, $c_j = \omega$ or $c_j \in \partial\mathcal{C}$, then $W(\mu, \mathbf{c}) = +\infty$. Besides, by Fatou's Lemma,

$$\liminf_{n \to +\infty} W(\mu_n, \mathbf{c}_n) = \liminf_{n \to +\infty} \mathbb{E} \min_{j=1,\ldots,k} d_\phi(Y_n, c_{nj}) \geq \mathbb{E} \min_{j=1,\ldots,k} d_\phi(Y, c_j) = W(\mu, \mathbf{c}).$$

Thus, $\lim_{n \to +\infty} W(\mu_n, \mathbf{c}_n) = +\infty = W(\mu, \mathbf{c})$.

Otherwise, let c_m be an element of $\mathbf{c}$ belonging to $ri(\mathcal{C})$. There exists in $ri(\mathcal{C})$ a regular convex polyhedron centered at c_m, containing the c_{nm}'s for large enough n (for example, an s-dimensional hypercube centered at c_m, where s denotes the dimension of the affine subspace spanned by $ri(\mathcal{C})$). Let $\mathcal{V}$ denote the finite set of its vertices. As the function $y \mapsto d_\phi(x,y)$ is assumed to be convex, for large enough n,

$$\min_{j=1,\ldots,k} d_\phi(x, c_{nj}) \leq d_\phi(x, c_{nm}) \leq \sum_{v\in\mathcal{V}} d_\phi(x,v). \tag{8}$$

By the strong law of large numbers, almost surely, for all $v \in \mathcal{V}$,

$$\mathbb{E}d_\phi(Y_n, v) = \int d_\phi(x,v)d\mu_n(x) = \frac{1}{n}\sum_{i=1}^{n} d_\phi(X_i, v)$$

tends, as $n \to +\infty$, to

$$\mathbb{E}d_\phi(X,v) = \mathbb{E}d_\phi(Y,v).$$

According to Theorem 3.6 in [8], for any $v \in \mathcal{V}$, $d_\phi(Y_n, v)$ is uniformly integrable. This implies by (8) that the variables $\min_{j=1,\ldots,k} d_\phi(Y_n, c_{nj})$ are uniformly integrable as well. Therefore, by Theorem 3.5 in [8], $W(\mu_n, \mathbf{c}_n) = \mathbb{E}\min_{j=1,\ldots,k} d_\phi(Y_n, c_{nj})$ tends almost surely to $\mathbb{E}\min_{j=1,\ldots,k} d_\phi(Y, c_j) = W(\mu, \mathbf{c})$, as desired.

5.6 Proof of Theorem 4.2

Since $\mathbb{P}\{X \in \mathcal{C}_R\} = 1$, the centers stay in the closed and bounded convex set $\mathcal{C}_R$ as the proof of Theorem 3.2 shows. Let Y and Y_n be the random variables with distribution μ and μ_n respectively, given by Skorohod's Theorem. Then, for all $\mathbf{c}$,

$$\begin{aligned} W(\mu_n, \mathbf{c}) - W(\mu, \mathbf{c}) &= \mathbb{E}\min_{j=1,\ldots,k} d_\phi(Y_n, c_j) - \mathbb{E}\min_{j=1,\ldots,k} d_\phi(Y, c_j) \\ &= \mathbb{E}\min_{j=1,\ldots,k} (\phi(Y_n) - \phi(c_j) - D_{c_j}\phi(Y_n - c_j)) \\ &\qquad - \mathbb{E}\min_{j=1,\ldots,k} (\phi(Y) - \phi(c_j) - D_{c_j}\phi(Y - c_j)) \\ &\leq \mathbb{E}\phi(Y_n) - \mathbb{E}\phi(Y) + \mathbb{E}(-\min_{j=1,\ldots,k} D_{c_j}\phi(Y_n - Y)) \\ &\leq \mathbb{E}\phi(Y_n) - \mathbb{E}\phi(Y) + M\mathbb{E}\|Y_n - Y\|. \end{aligned}$$

Yet, one has

$$\mathbb{E}\phi(Y_n) = \int \phi(x)\mu_n(dx) = \frac{1}{n}\sum_{i=1}^{n}\phi(X_i).$$

Thus, by the strong law of large numbers, $\mathbb{E}\phi(Y_n)$ converges to $\mathbb{E}\phi(X) = \mathbb{E}\phi(Y)$ almost surely. By the triangle inequality, $\|Y\| + \|Y_n\| - \|Y_n - Y\| \geq 0$. By Fatou's Lemma,

$$\liminf_{n\to+\infty} \mathbb{E}(\|Y\| + \|Y_n\| - \|Y_n - Y\|) \geq \mathbb{E}\lim_{n\to+\infty}(\|Y\| + \|Y_n\| - \|Y_n - Y\|) = 2\mathbb{E}\|Y\|.$$

Moreover, the law of large numbers implies that $\mathbb{E}\|Y_n\|$ converges to $\mathbb{E}\|Y\|$ almost surely. Thus,

$$\mathbb{E}\|Y_n - Y\| \underset{n\to+\infty}{\longrightarrow} 0 \quad a.s.$$

Hence, almost surely,

$$\sup_{\mathbf{c}\in\mathcal{C}_R^k} |W(\mu_n,\mathbf{c}) - W(\mu,\mathbf{c})| \underset{n\to+\infty}{\longrightarrow} 0$$

This completes the proof of the first statement.

We now turn to the second assertion. The following inequality (see [13]) shows that it suffices to prove that $\mathbb{E}\sup_{\mathbf{c}\in\mathcal{C}_R^k}(W(\mu_n,\mathbf{c}) - W(\mu,\mathbf{c}))$ vanishes as n tends to infinity:

$$\begin{aligned}&\mathbb{E}W(\mu,\mathbf{c}_n^*) - \inf_{\mathbf{c}\in\mathcal{C}_R^k} W(\mu,\mathbf{c})\\ &\quad\le \mathbb{E}\sup_{\mathbf{c}\in\mathcal{C}_R^k}(W(\mu_n,\mathbf{c}) - W(\mu,\mathbf{c})) + \mathbb{E}\sup_{\mathbf{c}\in\mathcal{C}_R^k}(W(\mu,\mathbf{c}) - W(\mu_n,\mathbf{c})).\end{aligned}$$

As stated above,

$$W(\mu_n,\mathbf{c}) - W(\mu,\mathbf{c}) \le \mathbb{E}\phi(Y_n) - \mathbb{E}\phi(Y) + M\mathbb{E}\|Y_n - Y\|,$$

for all $\mathbf{c}\in\mathcal{C}_R^k$. Moreover,

$$\mathbb{E}\phi(Y_n) - \mathbb{E}\phi(Y) = \frac{1}{n}\sum_{i=1}^n \phi(X_i) - \mathbb{E}\phi(X),$$

and taking expectation with respect to the X_i's, we have

$$\mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^n \phi(X_i) - \mathbb{E}\phi(X)\Big) = 0.$$

It remains to show that $\mathbb{E}\sup_{\mathbf{c}\in\mathcal{C}_R^k}\mathbb{E}\|Y_n - Y\|$ tends to 0 as n tends to infinity. This can be done by a slight adaptation of the proof of Lemma 4.2. in Biau, Devroye and Lugosi [7].

5.7 Proof of Theorem 4.3

We first recall the definition and some useful properties of the Rademacher averages. Let $\varepsilon_1,\dots,\varepsilon_n$ be independent Rademacher random variables, that is independent random variables taking values in $\{-1,1\}$ such that $\mathbb{P}\{\varepsilon_i = -1\} = \mathbb{P}\{\varepsilon_i = 1\} = \frac{1}{2}$, independent of $X_1,\dots,X_n$. For a class $\mathcal{G}$ of functions from E to $\mathbb{R}$, the Rademacher averages of $\mathcal{G}$ are defined by

$$R_n(\mathcal{G}) = \mathbb{E}\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^n \varepsilon_i g(X_i).$$

We will use the following properties:

1. For $a \in \mathbb{R}$, $R_n(a\mathcal{G}) = |a| R_n(\mathcal{G})$, where $a\mathcal{G} = \{ag, g \in \mathcal{G}\}$.
2. $R_n(|\mathcal{G}|) \leq R_n(\mathcal{G})$, where $|\mathcal{G}| = \{|g|, g \in \mathcal{G}\}$. (This property follows from the contraction principle of Ledoux and Talagrand [23].)
3. $R_n(\mathcal{G}_1 + \mathcal{G}_2) \leq R_n(\mathcal{G}_1) + R_n(\mathcal{G}_2)$, where $\mathcal{G}_1 + \mathcal{G}_2 = \{g_1 + g_2, (g_1, g_2) \in \mathcal{G}_1 \times \mathcal{G}_2)\}$.

Theorem 4.3 is a consequence of the following lemma.

Lemma 5.2 For $c \in \mathcal{C}_R$, let ℓ_c denote the real-valued function defined by

$$\ell_c(x) = -\phi(c) - D_c\phi(x - c), \quad x \in \mathcal{C}.$$

Then,

(i)

$$\mathbb{E} \sup_{\mathbf{c} \in \mathcal{C}_R^k} \left(\overline{W}(\mu_n, \mathbf{c}) - \overline{W}(\mu, \mathbf{c}) \right) \leq 2\mathbb{E} \sup_{\mathbf{c} \in \mathcal{C}_R^k} \frac{1}{n} \sum_{i=1}^n \varepsilon_i \min_{j=1,\ldots,k} \ell_{c_j}(X_i).$$

(ii)

$$\begin{aligned} &\mathbb{E} \sup_{\mathbf{c} \in \mathcal{C}_R^k} \frac{1}{n} \sum_{i=1}^n \varepsilon_i \min_{j=1,\ldots,k} \ell_{c_j}(X_i) \\ &\leq k\Big(\mathbb{E} \sup_{c \in \mathcal{C}_R} \frac{1}{n} \sum_{i=1}^n \varepsilon_i D_c\phi(X_i) + \frac{1}{\sqrt{n}} \sup_{c \in \mathcal{C}_R} |-\phi(c) + D_c\phi(c)| \Big). \end{aligned}$$

(iii)

$$\mathbb{E} \sup_{c \in \mathcal{C}_R} \frac{1}{n} \sum_{i=1}^n \varepsilon_i D_c\phi(X_i) \leq \frac{1}{\sqrt{n}} \sup_{c \in \mathcal{C}_R} \|D_c\phi\| (\mathbb{E}\|X\|^2)^{1/2}.$$

Proof of Lemma 5.2. (i) Let $X_1', \ldots, X_n'$ be an independent copy of $X_1, \ldots, X_n$, independent of the Rademacher variables $\varepsilon_1, \ldots, \varepsilon_n$. We have

$$\begin{aligned} &\mathbb{E} \sup_{\mathbf{c} \in \mathcal{C}_R^k} \Big(\overline{W}(\mu_n, \mathbf{c}) - \overline{W}(\mu, \mathbf{c}) \Big) \\ &= \mathbb{E} \sup_{\mathbf{c} \in \mathcal{C}_R^k} \Big(\frac{1}{n} \sum_{i=1}^n \min_{j=1,\ldots,k} \ell_{c_j}(X_i) - \mathbb{E} \min_{j=1,\ldots,k} \ell_{c_j}(X) \Big) \\ &= \mathbb{E} \sup_{\mathbf{c} \in \mathcal{C}_R^k} \Big(\frac{1}{n} \sum_{i=1}^n \min_{j=1,\ldots,k} \ell_{c_j}(X_i) - \mathbb{E} \frac{1}{n} \sum_{i=1}^n \min_{j=1,\ldots,k} \ell_{c_j}(X_i') \Big) \\ &= \mathbb{E} \sup_{\mathbf{c} \in \mathcal{C}_R^k} \mathbb{E} \Big(\frac{1}{n} \sum_{i=1}^n \min_{j=1,\ldots,k} \ell_{c_j}(X_i) - \frac{1}{n} \sum_{i=1}^n \min_{j=1,\ldots,k} \ell_{c_j}(X_i') \Big| X_1, \ldots, X_n \Big). \end{aligned}$$

By Jensen's inequality,

$$\begin{aligned}
&\mathbb{E} \sup_{\mathbf{c}\in\mathcal{C}_R^k} \big(\overline{W}(\mu_n,\mathbf{c}) - \overline{W}(\mu,\mathbf{c})\big)\\
&\quad\le \mathbb{E} \sup_{\mathbf{c}\in\mathcal{C}_R^k} \frac{1}{n}\sum_{i=1}^n \Big(\min_{j=1,\ldots,k} \ell_{c_j}(X_i) - \min_{j=1,\ldots,k} \ell_{c_j}(X_i')\Big)\\
&\quad= \mathbb{E} \sup_{\mathbf{c}\in\mathcal{C}_R^k} \frac{1}{n}\sum_{i=1}^n \varepsilon_i\Big(\min_{j=1,\ldots,k} \ell_{c_j}(X_i) - \min_{j=1,\ldots,k} \ell_{c_j}(X_i')\Big)\\
&\quad\le \mathbb{E} \sup_{\mathbf{c}\in\mathcal{C}_R^k} \frac{1}{n}\sum_{i=1}^n \varepsilon_i \min_{j=1,\ldots,k} \ell_{c_j}(X_i) + \mathbb{E} \sup_{\mathbf{c}\in\mathcal{C}_R^k} \frac{1}{n}\sum_{i=1}^n (-\varepsilon_i) \min_{j=1,\ldots,k} \ell_{c_j}(X_i')\\
&\quad= 2\mathbb{E} \sup_{\mathbf{c}\in\mathcal{C}_R^k} \frac{1}{n}\sum_{i=1}^n \varepsilon_i \min_{j=1,\ldots,k} \ell_{c_j}(X_i).
\end{aligned}$$

(ii) To obtain the inequality (ii), we argue by induction on k. For $k=1$,

$$\begin{aligned}
&\mathbb{E} \sup_{c\in\mathcal{C}_R} \frac{1}{n}\sum_{i=1}^n \varepsilon_i \ell_c(X_i)\\
&\quad= \mathbb{E} \sup_{c\in\mathcal{C}_R} \frac{1}{n}\sum_{i=1}^n \varepsilon_i(-\phi(c) - D_c\phi(X_i - c))\\
&\quad\le \mathbb{E} \sup_{c\in\mathcal{C}_R} \frac{1}{n}\sum_{i=1}^n \varepsilon_i D_c\phi(X_i) + \mathbb{E} \sup_{c\in\mathcal{C}_R} \frac{|-\phi(c) + D_c\phi(c)|}{n}\Big|\sum_{i=1}^n \varepsilon_i\Big|\\
&\quad\le \mathbb{E} \sup_{c\in\mathcal{C}_R} \frac{1}{n}\sum_{i=1}^n \varepsilon_i D_c\phi(X_i) + \frac{1}{n}\sup_{c\in\mathcal{C}_R} |-\phi(c) + D_c\phi(c)|\bigg(\mathbb{E}\Big(\sum_{i=1}^n \varepsilon_i\Big)^2\bigg)^{1/2}\\
&\quad= \mathbb{E} \sup_{c\in\mathcal{C}_R} \frac{1}{n}\sum_{i=1}^n \varepsilon_i D_c\phi(X_i) + \frac{1}{\sqrt{n}}\sup_{c\in\mathcal{C}_R} |-\phi(c) + D_c\phi(c)|,
\end{aligned}$$

using the fact that the ε_i's are independent. Assume that statement (ii) is true for $k-1$, and let us show that it is true for k. Let $\mathbf{c}^{k-1} = (c_1,\ldots,c_{k-1})$.

$$\begin{aligned}
&\mathbb{E} \sup_{\mathbf{c}\in\mathcal{C}_R^k} \frac{1}{n}\sum_{i=1}^n \varepsilon_i \min_{j=1,\ldots,k} \ell_{c_j}(X_i)\\
&\quad= \mathbb{E} \sup_{\mathbf{c}\in\mathcal{C}_R^k} \frac{1}{n}\sum_{i=1}^n \varepsilon_i \min\Big(\ell_{c_k}(X_i), \min_{j=1,\ldots,k-1} \ell_{c_j}(X_i)\Big)\\
&\quad= \mathbb{E} \sup_{\mathbf{c}\in\mathcal{C}_R^k} \frac{1}{2n}\sum_{i=1}^n \varepsilon_i\Big(\ell_{c_k}(X_i) + \min_{j=1,\ldots,k-1} \ell_{c_j}(X_i) - |\ell_{c_k}(X_i) - \min_{j=1,\ldots,k-1} \ell_{c_j}(X_i)|\Big),
\end{aligned}$$

since $\min(a,b) = (a+b)/2 - |a-b|/2$. By properties of Rademacher averages,

$$\begin{aligned}
&\mathbb{E} \sup_{\mathbf{c} \in \mathcal{C}_R^k} \frac{1}{n} \sum_{i=1}^n \varepsilon_i \min_{j=1,\ldots,k} \ell_{c_j}(X_i) \\
&\leq \mathbb{E} \sup_{c \in \mathcal{C}_R} \frac{1}{n} \sum_{i=1}^n \varepsilon_i \ell_c(X_i) + \mathbb{E} \sup_{\mathbf{c}^{k-1} \in (\mathcal{C}_R)^{k-1}} \frac{1}{n} \sum_{i=1}^n \varepsilon_i \min_{j=1,\ldots,k-1} \ell_{c_j}(X_i) \\
&= \mathbb{E} \sup_{c \in \mathcal{C}_R} \frac{1}{n} \sum_{i=1}^n \varepsilon_i D_c\phi(X_i) + \frac{1}{\sqrt{n}} \sup_{c \in \mathcal{C}_R} |-\phi(c) + D_c\phi(c)| + \\
&\qquad (k-1)\Big(\mathbb{E} \sup_{c \in \mathcal{C}_R} \frac{1}{n} \sum_{i=1}^n \varepsilon_i D_c\phi(X_i) + \frac{1}{\sqrt{n}} \sup_{c \in \mathcal{C}_R} |-\phi(c) + D_c\phi(c)|\Big) \\
&= k\Big(\mathbb{E} \sup_{c \in \mathcal{C}_R} \frac{1}{n} \sum_{i=1}^n \varepsilon_i D_c\phi(X_i) + \frac{1}{\sqrt{n}} \sup_{c \in \mathcal{C}_R} |-\phi(c) + D_c\phi(c)|\Big),
\end{aligned}$$

which is the desired bound for k.

(iii) We have

$$\begin{aligned}
\mathbb{E} \sup_{c \in \mathcal{C}_R} \frac{1}{n} \sum_{i=1}^n \varepsilon_i D_c\phi(X_i) &= \mathbb{E} \sup_{c \in \mathcal{C}_R} \frac{1}{n} D_c\phi\Big(\sum_{i=1}^n \varepsilon_i X_i\Big) \\
&\leq \frac{1}{n} \sup_{c \in \mathcal{C}_R} \|D_c\phi\| \mathbb{E} \Big\|\sum_{i=1}^n \varepsilon_i X_i\Big\| \\
&\leq \frac{1}{n} \sup_{c \in \mathcal{C}_R} \|D_c\phi\| \Big(\mathbb{E} \Big\|\sum_{i=1}^n \varepsilon_i X_i\Big\|^2\Big)^{1/2} \\
&\leq \frac{1}{n} \sup_{c \in \mathcal{C}_R} \|D_c\phi\| \Big(\mathbb{E} \Big(\sum_{i=1}^n \|X_i\|\Big)^2\Big)^{1/2}.
\end{aligned}$$

Using the fact that the X_i's are independent and identically distributed,

$$\begin{aligned}
\mathbb{E} \sup_{c \in \mathcal{C}_R} \frac{1}{n} \sum_{i=1}^n \varepsilon_i D_c\phi(X_i) &\leq \frac{1}{n} \sup_{c \in \mathcal{C}_R} \|D_c\phi\| \Big(\mathbb{E} \sum_{i=1}^n \|X_i\|^2\Big)^{1/2} \\
&= \frac{1}{n} \sup_{c \in \mathcal{C}_R} \|D_c\phi\| (n\mathbb{E}\|X\|^2)^{1/2} \\
&= \frac{1}{\sqrt{n}} \sup_{c \in \mathcal{C}_R} \|D_c\phi\| (\mathbb{E}\|X\|^2)^{1/2}.
\end{aligned}$$

□

References

[1] Y. Alber and D. Butnariu. Convergence of Bregman projection methods for solving consistent convex feasibility problems in reflexive Banach spaces. *Journal of optimization theory and applications*, 92(1):33–61, January 1997.

[2] W. Arendt, J. K. Batty, M. Hieber, and F. Neubrander. *Vector-valued Laplace Transforms and Cauchy Problems*. Monographs in Mathematics. Birkhäuser, 2001.

[3] A. Banerjee, X. Guo, and H. Wang. On the Optimality of Conditional Expectation as a Bregman Predictor. *IEEE Transactions on Information Theory*, 51(7), 2005.

[4] A. Banerjee, S. Merugu, I.S. Dhillon, and J. Ghosh. Clustering with Bregman Divergences. *Journal of Machine Learning Research*, 6:1705–1749, 2005.

[5] P. L. Bartlett, S. Boucheron, and G. Lugosi. Model selection and error estimation. *Machine Learning*, 48:85–113, 2001.

[6] H. H. Bauschke, J. M. Borwein, and P. L. Combettes. Essential smoothness, essential strict convexity, and Legendre Functions in Banach Spaces. *Communications in Contemporary Mathematics*, 3(4):615–647, 2001.

[7] G. Biau, L. Devroye, and G. Lugosi. On the performance of clustering in Hilbert spaces. *IEEE Transactions on Information Theory*, 54(2), February 2008.

[8] P. Billingsley. *Convergence of Probability Measures*. Wiley Series in Probability and Statistics. John Wiley and Sons, second edition, 1999.

[9] L. M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7:200–217, 1967.

[10] H. Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Universitext. Springer, 2010.

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

[12] I. Csiszár. Generalized Projections for Non-negative Functions. *Acta Mathematica Hungarica*, 68(1-2):161–185, 1995.

[13] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Applications of Mathematics. Springer, 1996.

[14] R. M. Dudley. *Real Analysis and Probability.* Cambridge studies in advanced mathematics. Cambridge University Press, 2002.

[15] B. A. Frigyik, S. Srivastava, and M. R. Gupta. An introduction to functional derivatives. Technical Report UWEETR-2008-0001, Department of Electrical Engineering, University of Washington, Seattle, July 2008.

[16] B. A. Frigyik, S. Srivastava, and M. R. Gupta. Functional Bregman Divergence and Bayesian Estimation of Distributions. *IEEE Transactions on Information Theory*, 54(11):5130–5139, November 2008.

[17] A. Gersho and R. M. Gray. *Vector Quantization and Signal Compression.* Kluwer Academic Publishers, 1992.

[18] S Graf and H Luschgy. *Foundations of quantization for probability distributions.* Lecture Notes in Mathematics. Springer-Verlag, Berlin, 2000.

[19] R. M. Gray, A. Buzo, A. H. Gray, and Y. Matsuyama. Distortion measures for speech processing. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28(4):367–376, 1980.

[20] L. Jones and C. Byrne. General Entropy Criteria for Inverse Problems, with Applications to Data Compression, Pattern Classification, and Cluster Analysis. *IEEE Transactions on Information Theory*, 36(1), 1990.

[21] M. I. Jordan, X. Nguyen, and M. J. Wainwright. Estimating divergence functionals and the likelihood ratio by convex risk minimization. Technical Report 764, Department of Statistics, University of California, Berkeley, 2009.

[22] T. Laloë. L_1-quantization and clustering in Banach spaces. Université Montpellier 2, preprint, 2009.

[23] M. Ledoux and M. Talagrand. *Probability in Banach Spaces.* Ergebnisse der Mathematik und ihrer Grenzgebiete. Springer, 1991.

[24] T. Linder. Learning-Theoretic Methods in Vector Quantization. In L. Györfi, editor, *Principles of Nonparametric Learning*, CISM lecture Notes. Springer, 2002.

[25] S. P. Lloyd. Least Squares Quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, March 1982.

[26] S. Lojasiewicz. *An introduction to the theory of real functions.* John Wiley and Sons, 1988.

[27] F. Nielsen, J.D. Boissonnat, and R Nock. Bregman Voronoi Diagrams: Properties, Algorithms and Applications. Technical Report 6154, INRIA, March 2007.

[28] D. Pollard. Quantization and the Method of k-Means. *IEEE Transactions on Information Theory*, 28(2), March 1982.

[29] J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer, second edition, 2006.

[30] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, New Jersey, 1970.

[31] M. J. Sabin and R. M. Gray. Global convergence and empirical consistency of the Generalized Lloyd Algorithm. *IEEE Transactions on Information Theory*, 32(2):148–155, 1986.

Erratum

Theorem 4.3 and Corollary 4.1 must be replaced by the statements below.

Theorem 4.3. *Suppose that E is a type 2 Banach space with constant T_2. For $\mathcal{C}_R \subset ri(\mathcal{C})$, the following inequality holds:*

$$\mathbb{E} \sup_{\mathbf{c}\in\mathcal{C}_R^k} \left(\overline{W}(\mu_n, \mathbf{c}) - \overline{W}(\mu, \mathbf{c})\right) \leq \frac{2k}{\sqrt{n}}\Big(\sup_{c\in\mathcal{C}_R} |-\phi(c) + D_c\phi(c)| + T_2 \sup_{c\in\mathcal{C}_R} \|D_c\phi\|(\mathbb{E}\|X\|^2)^{1/2}\Big).$$

Corollary 4.1. *Suppose that E is a type 2 Banach space with constant T_2, and that, for all $x \in \mathcal{C}$, $y \mapsto d_\phi(x,y)$ is weakly lower semi-continuous, which ensures the existence of an optimal codebook $\mathbf{c}_n^*$. Assume that there exists $R > 0$ such that $\mathbb{P}\{X \in \mathcal{C}_R\} = 1$. If $|-\phi(c) + D_c\phi(c)|$ and $\|D_c\phi\|$ are uniformly bounded on $\mathcal{C}_R$ by $M_1 = M_1(\phi, R) \geq 0$ and $M_2 = M_2(\phi, R) \geq 0$ respectively, then*

$$\mathbb{E}W(\mu, \mathbf{c}_n^*) - W^*(\mu) \leq \frac{4k}{\sqrt{n}}\left(M_1 + T_2M_2(\mathbb{E}\|X\|^2)^{1/2}\right),$$

and thus

$$\mathbb{E}W(\mu, \mathbf{c}_n^*) - W^*(\mu) \leq \frac{4k}{\sqrt{n}}\left(M_1 + T_2M_2R\right).$$

Lemma 5.2 (iii) and its proof are to be modified as follows:

Lemma 5.2 (iii) If E is of type 2, with a constant T_2,

$$\mathbb{E}\Big[\sup_{c\in\mathcal{C}_R} \frac{1}{n}\sum_{i=1}^n \varepsilon_i D_c\phi(X_i)\Big] \leq \frac{T_2}{\sqrt{n}} \sup_{c\in\mathcal{C}_R} \|D_c\phi\|(\mathbb{E}\|X\|^2)^{1/2}.$$

Proof. (iii) We have

$$\begin{aligned}\mathbb{E}\Big[\sup_{c\in\mathcal{C}_R} \frac{1}{n}\sum_{i=1}^n \varepsilon_i D_c\phi(X_i)\Big] &= \mathbb{E}\Big[\sup_{c\in\mathcal{C}_R} \frac{1}{n}D_c\phi\Big(\sum_{i=1}^n \varepsilon_i X_i\Big)\Big] \\ &\leq \frac{1}{n}\sup_{c\in\mathcal{C}_R}\|D_c\phi\|\mathbb{E}\|\sum_{i=1}^n \varepsilon_i X_i\| \\ &\leq \frac{1}{n}\sup_{c\in\mathcal{C}_R}\|D_c\phi\|\Big(\mathbb{E}\|\sum_{i=1}^n \varepsilon_i X_i\|^2\Big)^{1/2}\end{aligned}$$

As E is of type 2, and since the X_i are identically distributed,

$$\begin{aligned}\mathbb{E}\Big[\sup_{c\in\mathcal{C}_R} \frac{1}{n}\sum_{i=1}^n \varepsilon_i D_c\phi(X_i)\Big] &\leq \frac{1}{n}\sup_{c\in\mathcal{C}_R}\|D_c\phi\|T_2\Big[\sum_{i=1}^n \mathbb{E}\|X_i\|^2\Big]^{1/2} \\ &= \frac{1}{n}\sup_{c\in\mathcal{C}_R}\|D_c\phi\|T_2(n\mathbb{E}\|X\|^2)^{1/2} \\ &= \frac{T_2}{\sqrt{n}}\sup_{c\in\mathcal{C}_R}\|D_c\phi\|(\mathbb{E}\|X\|^2)^{1/2}.\end{aligned}$$

□