



On Hölder fields clustering

Benoît Cadre, Quentin Paris

► To cite this version:

Benoît Cadre, Quentin Paris. On Hölder fields clustering. Test, 2012, 21 (2), pp.301-316. 10.1007/s11749-011-0244-4 . hal-00502677

HAL Id: hal-00502677

<https://hal.science/hal-00502677>

Submitted on 15 Jul 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

On Hölder fields clustering

Benoît CADRE, Quentin PARIS

IRMAR, ENS Cachan Bretagne, CNRS, UEB
Campus de Ker Lann

Avenue Robert Schuman, 35170 Bruz, France

cadre, paris@bretagne.ens-cachan.fr

Abstract

In this paper, we study the k -means clustering scheme based on the observations of a phenomenon modelled by a sequence of random fields X_1, \dots, X_n taking values in a Hilbert space. In the k -means algorithm, clustering is performed by computing a Voronoi partition associated with centers that minimize an empirical criterion, called distortion. The performance of the method is evaluated by comparing a theoretical distortion of empirically optimal centers to the theoretical optimal distortion. Our first result states that, provided the underlying distribution satisfies an exponential moment condition, an upper bound for the above performance criterion is $O(1/\sqrt{n})$. Then, motivated by a broad range of applications and computational matters, we use a Hölder property shared by classical random fields in stochastic modelling to construct a numerically simple algorithm that computes empirical centers based on a discretized version of the data. With a judicious choice of the discretization, we are able to recover the same performance than in the non-discretized case.

Index Terms — Random fields, Clustering, k -means, Vector quantization, Hilbert space, Empirical risk minimization.

1 Introduction

1.1 Clustering and Hölder random fields

Clustering methods aim at partitioning a complex data set into a series of piecewise groups, or *clusters*, each of which may then be regarded as a separate class of data, thus reducing overall data complexity. This unsupervised learning problem is one of the most widely used techniques in exploratory data analysis since in

many sciences, e.g. social science, biology, oceanography, meteorology, finance or computer science, practitioners try to get a first intuition about their data by identifying meaningful groups of observations. General references on the subject are to be found in Duda et al (2000), Gersho and Gray (1992), Linder (2001) among others.

Due to a great interest in stochastic modelling, the last four decades have seen the emergence of many classes of random fields, each of them corresponding to a precise phenomenon. One can quote for instance, fractional Brownian fields and their derivatives (Lindström, 1993, Mandelbrot, 1997, Mandelbrot and van Ness, 1968) that has been proved to be the key tools in the modelling of long-dependency phenomena, for instance in the analysis of river level height (Kärner, 2001) or turbulence (Frisch, 1995), to mention of few of them. In another context, we also observe that Brownian diffusion processes or Lévy fields are central objects in financial mathematics (Cont and Tankov, 2003, Lamberton and Lapeyre, 1996). In each case, clustering methods play a central role in the analysis of the data sets. One can notice that, in the above models, the common point is that the data to be clustered are random fields that share a Hölder type property. The clustering scheme studied in the paper will be based on this observation.

1.2 General clustering framework

We first recall the general clustering context, in which the observation space $(\mathcal{H}, \|\cdot\|)$ is a Hilbert space. In this setting, the data to be clustered is a sequence of independent \mathcal{H} -valued random observations X_1, \dots, X_n with the same distribution as a generic square integrable random variable X with distribution μ . We focus on the k -means clustering, which prescribes a criterion for partitioning the sample into k clusters, by minimizing the *empirical distortion*

$$W_k(\mathbf{c}, \mu_n) = \frac{1}{n} \sum_{i=1}^n \min_{j=1, \dots, k} \|X_i - c_j\|^2,$$

over all centers $\mathbf{c} = (c_1, \dots, c_k) \in \mathcal{H}^k$. Here, μ_n is the empirical measure defined by

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

for all Borel set $A \subset \mathcal{H}$. Associated with the centers c_j 's are the convex polyhedrons S_j of all points in \mathcal{H} closer to c_j than to any other center. Then, $\{S_1, \dots, S_k\}$

forms a partition of \mathcal{H} , called the Voronoi partition, and the S_j 's are the clusters of interest.

From a theoretical point of view, the performance of a clustering scheme given by the centers $\mathbf{c} = (c_1, \dots, c_k) \in \mathcal{H}^k$ is evaluated by the *theoretical distortion*

$$W_k(\mathbf{c}, \mu) = \mathbb{E} \min_{j=1, \dots, k} \|X - c_j\|^2.$$

Clustering methods aim at approximating the *clustering risk*, defined by

$$W_k^*(\mu) = \inf_{\mathbf{c} \in \mathcal{H}^k} W_k(\mathbf{c}, \mu).$$

More precisely, the performance of a clustering scheme based on the empirical centers $\mathbf{c}_n = (c_{n1}, \dots, c_{nk})$ is evaluated by

$$W_k(\mathbf{c}_n, \mu) - W_k^*(\mu).$$

Since the early work from Hartigan (1975, 1978), many authors have contributed to the study of k -means clustering based on a minimizer \mathbf{c}_n of the empirical distortion, namely

$$W_k(\mathbf{c}_n, \mu_n) = \inf_{\mathbf{c} \in \mathcal{H}^k} W_k(\mathbf{c}, \mu_n). \quad (1.1)$$

Proof of the existence of \mathbf{c}_n is to be found for instance in Graf and Luschgy (2000), Theorem 4.12. In the finite-dimensional setting, consistency properties have been studied by Pollard (1981, 1982b), Abaya and Wise (1984) among others, while rates of convergence are to be found in Pollard (1982a), Chou (1994), Linder et al (1994), Bartlett et al (1998), Linder (2000, 2001), Antos (2005) and Antos et al (2005). The infinite-dimensional setting has been considered by Biau et al (2008): it is proved in Corollary 2.1 that, provided X is bounded by R , then for any $\delta \in]0, 1[$

$$W_k(\mathbf{c}_n, \mu) - W_k^*(\mu) \leq \frac{12kR^2 + 4R\sqrt{-2\ln \delta}}{\sqrt{n}} \quad (1.2)$$

with probability at least $1 - \delta$. The proof of (??) consists in two steps: first establish the result in mean, and then conclude with the McDiarmid Inequality. In the case of non-bounded random variables, however, such a proof can not hold because McDiarmid's Inequality is based on a boundedness property of the increments.

1.3 Objectives of the paper

Despite its importance in fields clustering and due to the boundedness assumption, Inequality (??) may not be applied to the usual cases where the data are modelled by random fields like Brownian diffusion processes, fractional random fields, Lévy fields ... Many situations that naturally appear in classical stochastic modelling, as seen in subsection ??, Hence, in the non-bounded case, a natural question that arises at this step is : What are the features of the distribution μ that replace R ? Theorem ?? will be devoted to this problem.

Note also that, in view of fields clustering, the step (??), that involves a minimization in an infinite-dimensional setting, is numerically unrealistic. Based on the observation that many random fields that arise in stochastic modelling have a Hölder property, we shall construct a numerically simple algorithm involving discretized versions of the fields, i.e. step fields defined over a finite grid. More precisely, we shall study the performance of the empirically optimal centers with respect to discretized fields and we shall see in Theorem ?? that a judicious choice of the discretization level and the location of the grid points permits us to recover the same performances as in the non-discretized case.

2 Clustering in Hilbert spaces

In the whole paper, we denote by $\mathbf{c}_n = (c_{n1}, \dots, c_{nk}) \in \mathcal{H}^k$ a vector that minimizes the empirical clustering risk in \mathcal{H}^k :

$$W_k(\mathbf{c}_n, \mu_n) = \inf_{\mathbf{c} \in \mathcal{H}^k} W_k(\mathbf{c}, \mu_n). \quad (2.1)$$

Then, $\{S_{n1}, \dots, S_{nk}\}$ stands for the Voronoi partition of \mathcal{H} which is associated with the centers $c_{n1} \in S_{n1}, \dots, c_{nk} \in S_{nk}$ (for a definition and properties of the Voronoi partition, we refer the reader to Chapter 1 in Graf and Luschgy, 2000). We know from Lemma 1 in Linder (2001) that for all $j = 1, \dots, k$, the center c_{nj} has the following expression:

$$c_{nj} = \frac{\sum_{i=1}^n X_i \mathbf{1}\{X_i \in S_{nj}\}}{\sum_{i=1}^n \mathbf{1}\{X_i \in S_{nj}\}}. \quad (2.2)$$

In this section, we study the performance of the empirical centers $\mathbf{c}_n = (c_{n1}, \dots, c_{nk}) \in \mathcal{H}^k$.

We shall assume that, for some $\tau > 0$,

$$\mathbb{E} e^{\tau \|X\|} < \infty, \quad (2.3)$$

and we denote by $R(\mu)$ the quantity:

$$R(\mu) = \frac{1}{\tau} \left(1 + \omega(\mu) + \ln \mathbb{E} e^{\tau \|X\|} \right),$$

where $\omega(\mu) = \ln^- [W_{k-1}^*(\mu) - W_k^*(\mu)]$, if $\ln^- x = \max(0, -\ln x)$ stands for the negative part of $\ln x$. Recall that $W_{k-1}^*(\mu) > W_k^*(\mu)$ when the support of μ contains at least k points (e.g., see Theorem 4.12 in Graf and Luschgy, 2000).

Theorem 2.1. *Assume that (??) holds and the support of μ contains at least k points. There exists a universal constant $C > 0$ such that for all $\delta \in]0, 1[$, one has*

$$W_k(\mathbf{c}_n, \mu) - W_k^*(\mu) \leq C \frac{R(\mu)^2 k \ln(k/\delta)}{\sqrt{n}},$$

with probability $(1 - \delta) + O(e^{-rn^{1/5}})$, where $r > 0$. Moreover, for all $S > 0$, the term $O(e^{-rn^{1/5}})$ is uniform among the measures μ such that $R(\mu) \leq S$.

Due to the generality of the situation under study, the obtained value for the numerical constant C is large. However, the interest of Theorem ?? is to point out the contribution of each parameter, especially n , δ and μ .

Examples:

1. *Bounded random variable.* Though our study is not fully adapted to the bounded case, it is of importance to compare the previous result to its equivalent as given in Biau et al (2008). In the case where μ has a bounded support, i.e. $\|X\| \leq M$ for some $M > 0$, then τ can be chosen arbitrarily large, say $\tau = \infty$. Theorem ?? reveals that there exists a universal constant $C > 0$ such that for all $\delta \in]0, 1[$,

$$W_k(\mathbf{c}_n, \mu) - W_k^*(\mu) \leq C \frac{M^2 k \ln(k/\delta)}{\sqrt{n}},$$

with probability $(1 - \delta) + O(e^{-rn^{1/5}})$, where $r > 0$. In this result, the contribution of each parameter n , k , δ and M is very closed to that of Corollary 2.1 in Biau et al (2008).

2. *Diffusion process.* We let $\mathcal{H} = \mathbb{L}_2[0, 1]$ be the set of square integrable and real-valued functions on $[0, 1]$, and we assume that $X = (X(t))_{t \in [0, 1]}$ is a solution to the stochastic differential equation:

$$dX(t) = b(t, X(t))dW(t) + \sigma(t, X(t))dt, \quad (2.4)$$

where $W = (W(t))_{t \in [0, 1]}$ is a standard one-dimensional brownian motion, and b, σ are real-valued functions defined on $[0, 1] \times \mathbb{R}$. (For an overview on stochastic differential equations, we refer the reader to the book by Revuz and Yor, 1999). In a recent paper, Huang (2009) finds a wide class of diffusion processes such as (??) so that the exponential condition (??) holds. For simplicity, we shall assume for this example the stronger conditions that functions b and σ are bounded. In this case, we can prove (see the Appendix) that, for some judicious choice of $\tau > 0$,

$$R(\mu) \leq 4(\sup |b| + \sup |\sigma|)\sqrt{1 + \omega(\mu)}. \quad (2.5)$$

Then by Theorem ??, there exists a universal constant $C > 0$ such that for all $\delta \in]0, 1[$,

$$W_k(\mathbf{c}_n, \mu) - W_k^*(\mu) \leq C(\sup |b| + \sup |\sigma|)^2 (1 + \omega(\mu)) \frac{k \ln(k/\delta)}{\sqrt{n}},$$

with probability $(1 - \delta) + O(e^{-rn^{1/5}})$, where $r > 0$.

3 Hölder fields clustering

3.1 Numerical step in fields clustering

In practise, the data $x \in \mathcal{H}$ is assigned to the j -th cluster if for all $\ell = 1, \dots, k$,

$$\|x - c_{nj}\| \leq \|x - c_{n\ell}\|,$$

where $\mathbf{c}_n = (c_{n1}, \dots, c_{nk})$ is defined by (??); that is, if c_{nj} is the nearest-neighbor of x among $\{c_{n1}, \dots, c_{nk}\}$. However, when the dimension of \mathcal{H} is infinite, the step (??) is numerically unrealistic since it involves a minimization in the space \mathcal{H}^k . To circumvent this drawback, Biau et al (2008) proposed to make use of the so-called *random projections method*, leading to a minimization step in a finite dimensional space. Roughly speaking, they proved that a projected version of \mathbf{c}_n

in a $(n \ln n)$ -dimensional space has the same performance than \mathbf{c}_n (Corollary 3.1 in Biau et al, 2008).

If the previous approach has the advantage to hold whatever is the situation, it leads to many complications in term of the computational complexity (e.g., computation of the random projections, orthonormal representation of the data), and it does not fully exploit the particular form of the data under study. Indeed, as mentioned in the introduction, classical stochastic modelling often leads to consider diffusion processes, fractional brownian fields, Lévy fields, ... The common point here, is that most of these random fields satisfy a Hölder property. The aim of the next subsection is to exploit this fact in order to derive a simple and tractable method for clustering, in which the numerical step (??) is computationally feasible.

3.2 Discretized fields

For the rest of the section, we assume that $\mathcal{H} = \mathbb{L}_2([0, 1]^s)$ is the set of square integrable and real-valued functions defined on $[0, 1]^s$. (Here, the choice of $[0, 1]^s$ instead of a general space is for convenience of the reader and simplicity of the statements; it turns out that the case of a compact s -dimensional space could be considered as well.) In this setting, the random variable $X = (X(t))_{t \in [0, 1]^s}$ is a random field taking values in \mathcal{H} .

Our goal is to evaluate the performance of the natural method that simply consists to create the empirical quantizer, based on the partial informations carried by the data that are evaluated over a common finite grid $\mathcal{S} = \{t_1, \dots, t_d\}$ of $[0, 1]^s$. With this respect, the natural questions that arise are: Where must be located the points of \mathcal{S} and what must be the size of \mathcal{S} , so that the performance of an optimal quantizer associated with the discretized data is comparable to that of \mathbf{c}_n ?

We fix a partition $\pi = \{V_1, \dots, V_d\}$ of $[0, 1]^s$ such that $t_i \in V_i$ for all $i = 1, \dots, d$. In the sequel, the number d is referred to as the *discretization level*. To any function $x = (x(t))_{t \in [0, 1]^s} \in \mathbb{L}_2([0, 1]^s)$, we associate the *discretized function* $x^\pi = (x^\pi(t))_{t \in [0, 1]^s}$ defined as follows:

$$x^\pi(t) = x(t_i) \quad \text{if } t \in V_i,$$

for all $i = 1, \dots, d$. In the sequel, $\mathbb{L}_2^\pi([0, 1]^s)$ is the Hilbert space defined by

$$\mathbb{L}_2^\pi([0, 1]^s) = \{x^\pi, x \in \mathbb{L}_2([0, 1]^s)\}.$$

Let now μ_n^π be the empirical measure associated with the transformed data X_1^π, \dots, X_n^π , i.e. for any borel set $A \subset \mathbb{L}_2^\pi([0, 1]^s)$:

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

We define the discretized empirical centers $\mathbf{c}_n^\pi = (c_{n1}^\pi, \dots, c_{nk}^\pi)$ by

$$W_k(\mathbf{c}_n^\pi, \mu_n^\pi) = \inf_{\mathbf{c} \in (\mathbb{L}_2^\pi([0, 1]^s))^k} W_k(\mathbf{c}, \mu_n^\pi). \quad (3.1)$$

Observe that, provided the cells V_j have the same volume, say $\text{vol}(V_j) = v$, then for any $\mathbf{c} = (c_1, \dots, c_k) \in (\mathbb{L}_2^\pi([0, 1]^s))^k$:

$$W_k(\mathbf{c}, \mu_n^\pi) = \frac{v}{n} \sum_{i=1}^n \min_{j=1, \dots, k} \sum_{\ell=1}^d (X_i(t_\ell) - c_j(t_\ell))^2.$$

Consequently, in comparison with the minimization step (??), step (??) only requires to minimize the function

$$(\mathbb{R}^d)^k \ni (c_1, \dots, c_k) \mapsto \sum_{i=1}^n \min_{j=1, \dots, k} |\hat{X}_i - c_j|^2,$$

hence a minimization over a $d \times k$ -dimensional space. In the previous formula, $|\cdot|$ stands for the euclidean norm on \mathbb{R}^d and \hat{X}_i is the random vector with coordinates $(X_i(t_1), \dots, X_i(t_d))$. The aim now is to find the value of d , and the location of the t_i 's so that the center \mathbf{c}_n^π has the same performance than the center \mathbf{c}_n .

3.3 Result

In this subsection, we shall make use of a mean Hölder condition on the random field X . As mentioned in the introduction, this is a mild assumption which is satisfied by most of the relevant random fields that arise in stochastic modelling, such as fractional Brownian fields or diffusion processes for instance. We assume that X satisfies the inequality:

$$\mathbb{E}|X(s) - X(t)|^2 \leq L|s - t|^h, \quad \forall s, t \in [0, 1]^s, \quad (3.2)$$

for some $h > 0$ and $L > 0$. Here, $|\cdot|$ stands for the euclidean norm in $[0, 1]^s$. Observe that assumption (??) does not mean that X has Hölder paths, as illustrated

by the case of the Poisson process (for which $h = 1$).

For simplicity, we assume that $X(0) = 0$ and a stronger property than (??), namely that for some $\tau > 0$,

$$\mathbb{E} e^{\tau \|X\|_\infty} < \infty, \quad (3.3)$$

if $\|\cdot\|_\infty$ stands for the supremum norm of $X = (X(t))_{t \in [0,1]^s}$. Furthermore, we let

$$R_\infty(\mu) = \frac{1}{\tau} \left(1 + \omega(\mu) + \ln \mathbb{E} e^{\tau \|X\|_\infty} \right).$$

The next result shows how to choose the partition π and the discretization level d so that the empirical center \mathbf{c}_n^π have the same performance than \mathbf{c}_n .

Theorem 3.1. *Assume that (??) and (??) hold, and that the support of μ contains at least k points. If $d = \lceil n^{s/h} \rceil$ and for all $i = 1, \dots, d$, V_i is an hypercube with edge length $1/d^{1/s}$, then there exists a universal constant $C > 0$ such that for all $\delta \in]0, 1[$, one has*

$$W_k(\mathbf{c}_n^\pi, \mu) - W_k^*(\mu) \leq C \frac{Ls^{h/4} + kR_\infty(\mu)^2 \ln(k/\delta)}{\sqrt{n}},$$

with probability $(1 - \delta) + O(e^{-rn^{1/5}})$, for some $r > 0$.

Except for some specific cases, e.g. the diffusion process (??) in which $h = 1$, the exact value of h is usually unknown. Estimation of h has paid much attention in the litterature, for instance in the gaussian or Lévy field cases. With this respect, we refer the reader to the recent papers by Brouste et al (2007), Breton et al (2009), Coeurjolly (2008), Lacaux and Loubès (2007) and the references therein.

When $s = h$ (e.g. the Brownian diffusion process), step (??) leads to a minimization in a $k \times n$ -dimensional space, a closed result to that of Corollary 3.1 in Biau et al (2008). Nevertheless, we recall that our approach is much simpler from a computational point of view. Note also that in the case where $s < h$ (e.g. a wide part of the class of fractional Brownian fields) however, the complexity of the optimisation algorithm may be considerably improved.

Examples:

1. *Diffusion process.* Suppose that X is the process defined by (??). Burholder's inequality (see Revuz and Yor, 1999) and (??) of the Appendix ensures that assumptions (??) and (??) are satisfied for any $\tau > 0$, with $h = 1$ and $L = 4(\sup b^2 +$

$\sup \sigma^2$). By Theorems ?? and ??, up to some constant terms, the performances of the discretized centers \mathbf{c}_n^π are the same as those of \mathbf{c}_n , provided the discretization level d is n .

2. *Fractional Brownian field.* Assume that X is the fractional Brownian field of index $H \in (0, 2)$ (for an overview on fractional Brownian motion/field, we refer the reader to Pipiras and Taqqu, 2003 and Lindstrøm, 1993). In this case, both conditions (??) and (??) hold, with $L = 1$ and $h = H$. Hence, we deduce from Theorems ?? and ?? that, up to some constant terms, the performances of the empirical quantizers \mathbf{c}_n^π and \mathbf{c}_n are the same, provided the discretization level d is chosen so that $d = n^{s/H}$. In the motion case (i.e. $s = 1$) of positive correlation for instance (i.e. $H > 1$), the process has an aggregation behavior, and the numerical step (??) is reduced to a minimization in a $n^{1/H} \times k$ -dimensional space. From a computational point of view, we considerably improve the complexity of the minimization procedure, in comparison with the diffusion process above for instance.

4 Proofs

4.1 Proof of Theorem ??

It is proved in the Appendix that under the exponential moment condition (??), one has for all $p \geq 2$:

$$\mathbb{E}\|X\|^p \leq p! C^p \kappa^p, \quad (4.1)$$

where $\kappa = (1 + \ln \mathbb{E} \exp[\tau \|X\|]) / \tau$ and $C > 0$ is a universal constant. For simplicity of the proofs, we shall assume in the sequel that the universal constant $C = 1$, i.e. we have:

$$\mathbb{E}\|X\|^p \leq p! \kappa^p, \quad (4.2)$$

Furthermore, we let for any measure ν on \mathcal{H} and $\mathbf{c} = (c_1, \dots, c_k) \in \mathcal{H}^k$:

$$\bar{W}_k(\mathbf{c}, \nu) = \int_{\mathcal{H}} \min_{j=1, \dots, k} [-2 \langle x, c_j \rangle + \|c_j\|^2] \nu(dx),$$

where $\langle \cdot, \cdot \rangle$ stands for the scalar product in \mathcal{H} . Finally, we denote by B_ρ the centered ball in \mathcal{H} with radius ρ .

The proof of the next lemma will borrow some arguments in Biau et al (2008).

Lemma 4.1. *Let $\rho, t > 0$ such that $4(\kappa^2 + t\kappa/(32k\rho)) \leq \rho^2$. Then,*

$$\mathbb{P}\left(\sup_{\mathbf{c} \in B_\rho^k} |\bar{W}_k(\mathbf{c}, \mu_n) - \bar{W}_k(\mathbf{c}, \mu)| \geq t\right) \leq 32k^3 \exp\left(-\frac{nt^2}{512k^2\rho^4}\right).$$

PROOF. The classical symmetrisation argument (see Devroye et al, 1996, pp. 193-195) reveals that

$$\mathbb{P}\left(\sup_{\mathbf{c} \in B_\rho^k} |\bar{W}_k(\mathbf{c}, \mu_n) - \bar{W}_k(\mathbf{c}, \mu)| \geq t\right) \leq 4\mathbb{P}\left(S_k \geq \frac{t}{4}\right), \quad (4.3)$$

where for all $m = 1, \dots, k$:

$$S_m = \sup_{\mathbf{c} \in B_\rho^m} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i \min_{j=1, \dots, m} \ell_{c_j}(X_i) \right|,$$

with $\ell_c(x) = -2 < x, c > + \|c\|^2$ for $x, c \in \mathcal{H}$, and $\sigma_1, \dots, \sigma_n$ are i.i.d. Rademacher random variables, independent of the data X_1, \dots, X_n . Here and in the sequel, the components of $\mathbf{c} \in \mathcal{H}^m$ are denoted by (c_1, \dots, c_m) .

We now proceed to bound the rightmost term in (??). Using the equality

$$\min(a, b) = \frac{1}{2}(a + b - |a - b|), \quad a, b \in \mathbb{R},$$

we deduce that for all $u > 0$ and $m = 0, \dots, k$:

$$\begin{aligned} \mathbb{P}(S_k \geq u) &\leq \mathbb{P}\left(S_m \geq \frac{u}{2}\right) + \mathbb{P}\left(S_{k-m} \geq \frac{u}{2}\right) \\ &\quad + \mathbb{P}\left(\sup_{\mathbf{c} \in B_\rho^k} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i \left(\min_{j=1, \dots, m} \ell_{c_j}(X_i) - \min_{j=m+1, \dots, k} \ell_{c_j}(X_i) \right) \right| \geq u \right). \end{aligned}$$

According to the contraction principle (see Chapter 4 in Ledoux and Talagrand, 1991), the rightmost term is bounded by

$$\begin{aligned} &2\mathbb{P}\left(\sup_{\mathbf{c} \in B_\rho^k} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i \left(\min_{j=1, \dots, m} \ell_{c_j}(X_i) - \min_{j=m+1, \dots, k} \ell_{c_j}(X_i) \right) \right| \geq u \right) \\ &\leq 2\mathbb{P}\left(S_m \geq \frac{u}{2}\right) + 2\mathbb{P}\left(S_{k-m} \geq \frac{u}{2}\right). \end{aligned}$$

Hence, we have for all $m = 1, \dots, k$:

$$\mathbb{P}(S_k \geq u) \leq 3\mathbb{P}\left(S_m \geq \frac{u}{2}\right) + 3\mathbb{P}\left(S_{k-m} \geq \frac{u}{2}\right),$$

and we easily deduce that

$$\mathbb{P}(S_k \geq u) \leq 2k^3 \mathbb{P}\left(S_1 \geq \frac{u}{2k}\right). \quad (4.4)$$

Observe now that

$$\begin{aligned} \mathbb{P}\left(S_1 \geq \frac{u}{2k}\right) &= \mathbb{P}\left(\sup_{c \in B_\rho} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i \ell_c(X_i) \right| \geq \frac{u}{2k}\right) \\ &\leq \mathbb{P}\left(\sup_{c \in B_\rho} \frac{2}{n} \left| \sum_{i=1}^n \sigma_i \langle X_i, c \rangle \right| \geq \frac{u}{4k}\right) + \mathbb{P}\left(\frac{\rho^2}{n} \left| \sum_{i=1}^n \sigma_i \right| \geq \frac{u}{4k}\right) \\ &\leq \mathbb{P}\left(\left\| \sum_{i=1}^n \sigma_i X_i \right\| \geq \frac{un}{8k\rho}\right) + \mathbb{P}\left(\left| \sum_{i=1}^n \sigma_i \right| \geq \frac{un}{4k\rho^2}\right). \end{aligned}$$

According to Hoeffding inequality (see Bosq, 2000, Chapter 2) for the rightmost term, and Bernstein inequality for Banach-valued random variables (see Bosq, 2000, Chapter 2) for the former term, we deduce from (??) that

$$\mathbb{P}\left(S_1 \geq \frac{u}{2k}\right) \leq 2 \exp\left(-\frac{u^2 n}{128k^2 \rho^2 (\kappa^2 + u\kappa/(8k\rho))}\right) + 2 \exp\left(-\frac{u^2 n}{32k^2 \rho^4}\right).$$

We can now conclude from (??) and (??) that

$$\begin{aligned} &\mathbb{P}\left(\sup_{\mathbf{c} \in B_\rho^k} |\bar{W}_k(\mathbf{c}, \mu_n) - \bar{W}_k(\mathbf{c}, \mu)| \geq t\right) \\ &\leq 16k^3 \exp\left(-\frac{t^2 n}{2048k^2 \rho^2 (\kappa^2 + t\kappa/(32k\rho))}\right) + 16k^3 \exp\left(-\frac{t^2 n}{512k^2 \rho^4}\right). \end{aligned}$$

Therefore, provided the inequality

$$4\left(\kappa^2 + \frac{t\kappa}{32k\rho}\right) \leq \rho^2$$

is satisfied, we have

$$\mathbb{P}\left(\sup_{\mathbf{c} \in B_\rho^k} |\bar{W}_k(\mathbf{c}, \mu_n) - \bar{W}_k(\mathbf{c}, \mu)| \geq t\right) \leq 32k^3 \exp\left(-\frac{t^2 n}{512k^2 \rho^4}\right),$$

hence the lemma. \square

We now fix $\rho > 0$ such that:

$$\rho^2 \mu(B_{\rho/10}) \geq 100 \mathbb{E} \|X\|^2 \text{ and } 4 \mathbb{E} \|X\|^2 \mathbf{1}\{\|X\| \geq 2\rho/5\} < \alpha/2, \quad (4.5)$$

where $2\alpha = W_{k-1}^*(\mu) - W_k^*(\mu) > 0$. Then, one can prove that (e.g. see the proof of Theorem 4.12 in Graff and Lushgy, 2000):

$$\mathbf{c} \notin B_\rho^k \Rightarrow W_k(\mathbf{c}, \mu) \geq W_k^*(\mu) + \alpha. \quad (4.6)$$

Lemma 4.2. *Assume that ρ satisfies (??) and fix $S > 0$. Then, there exists $r > 0$ such that*

$$\mathbb{P}(\mathbf{c} \notin B_\rho^k) = O(e^{-rn^{1/5}}),$$

uniformly over all measures μ such that $R(\mu) \leq S$.

PROOF. Since ρ satisfies (??), we know from (??) that

$$\mathbb{P}(\mathbf{c}_n \notin B_\rho^k) \leq \mathbb{P}(W_k(\mathbf{c}_n, \mu) - W_k^*(\mu) \geq \alpha)$$

Observe that, since for all $\mathbf{c} = (c_1, \dots, c_k) \in \mathcal{H}^k$,

$$W_k(\mathbf{c}, \mu) = \mathbb{E} \|X\|^2 + \mathbb{E} \min_{j=1, \dots, k} [-2 \langle X, c_j \rangle + \|c_j\|^2],$$

we also have

$$\mathbb{P}(\mathbf{c}_n \notin B_\rho^k) \leq \mathbb{P}(\bar{W}_k(\mathbf{c}_n, \mu) - \bar{W}_k^*(\mu) \geq \alpha).$$

According to Theorem 4.12 in Graff and Lushgy (2000), there exists $\eta_0 > 0$ such that $\bar{W}_k^*(\mu) = \inf_{\mathbf{c} \in B_{\eta_0}^k} \bar{W}_k(\mathbf{c}, \mu)$. Then, for all $\eta \geq \eta_0$,

$$\begin{aligned} \mathbb{P}(\mathbf{c}_n \notin B_\rho^k) &\leq \mathbb{P}\left(\bar{W}_k(\mathbf{c}_n, \mu) - \bar{W}_k(\mathbf{c}_n, \mu_n) \geq \frac{\alpha}{2}, \mathbf{c}_n \in B_\eta^k\right) \\ &\quad + \mathbb{P}\left(\bar{W}_k(\mathbf{c}_n, \mu_n) - \inf_{\mathbf{c} \in B_\eta^k} \bar{W}_k(\mathbf{c}, \mu) \geq \frac{\alpha}{2}, \mathbf{c}_n \in B_\eta^k\right) + \mathbb{P}(\mathbf{c}_n \notin B_\eta^k) \\ &\leq 2\mathbb{P}\left(\sup_{\mathbf{c} \in B_\eta^k} |\bar{W}_k(\mathbf{c}, \mu_n) - \bar{W}_k(\mathbf{c}, \mu)| \geq \frac{\alpha}{2}\right) + \mathbb{P}(\mathbf{c}_n \notin B_\eta^k), \end{aligned} \quad (4.7)$$

because $\bar{W}_k(\mathbf{c}_n, \mu_n) = \inf_{\mathbf{c} \in B_\eta^k} \bar{W}_k(\mathbf{c}, \mu_n)$ when $\mathbf{c}_n \in B_\eta^k$. We know from Lemma ?? that provided η is large enough,

$$\mathbb{P}\left(\sup_{\mathbf{c} \in B_\eta^k} |\bar{W}_k(\mathbf{c}, \mu_n) - \bar{W}_k(\mathbf{c}, \mu)| \geq \frac{\alpha}{2}\right) \leq 32k^3 \exp\left(-\frac{\alpha^2 n}{1024k^2 \eta^4}\right). \quad (4.8)$$

Moreover, according to (??) we get $\|c_{nj}\| \leq \max_{i=1, \dots, n} \|X_i\|$ for all $j = 1, \dots, k$ so that

$$\begin{aligned}
\mathbb{P}(\mathbf{c}_n \notin B_\eta^k) &\leq k \max_{j=1, \dots, k} \mathbb{P}(c_{nj} \notin B_\eta) \\
&\leq k \mathbb{P}\left(\max_{i=1, \dots, n} \|X_i\| \geq \eta\right) \\
&\leq kn \mathbb{P}(\|X\| \geq \eta) \\
&\leq kn e^{-\tau\eta} \mathbb{E} e^{\tau\|X\|}.
\end{aligned} \tag{4.9}$$

Letting $\eta = n^{1/5} + \ln n / \tau$, we deduce from above and (??), (??) that

$$\mathbb{P}(\mathbf{c}_n \notin B_\rho^k) \leq \left(32k^3 + k \mathbb{E} e^{\tau\|X\|}\right) e^{-rn^{1/5}},$$

for some constant $r > 0$, hence the lemma. The uniformity result derives from (??) and (??). \square

We are now in position to prove Theorem ??.

PROOF OF THEOREM ??. Let us fix ρ that satisfies (??). For all $t > 0$,

$$\begin{aligned}
\mathbb{P}(W_k(\mathbf{c}_n, \mu) - W_k^*(\mu) \geq t) &= \mathbb{P}(\bar{W}_k(\mathbf{c}_n, \mu) - \bar{W}_k^*(\mu) \geq t) \\
&\leq \mathbb{P}\left(\bar{W}_k(\mathbf{c}_n, \mu) - \bar{W}_k^*(\mu) \geq t, \mathbf{c}_n \in B_\rho^k\right) \\
&\quad + \mathbb{P}(\mathbf{c}_n \notin B_\rho^k) \\
&\leq 2\mathbb{P}\left(\sup_{\mathbf{c} \in B_\rho^k} |\bar{W}_k(\mathbf{c}, \mu) - \bar{W}_k(\mathbf{c}, \mu_n)| \geq \frac{t}{2}\right) \\
&\quad + \mathbb{P}(\mathbf{c}_n \notin B_\rho^k).
\end{aligned}$$

By Lemmas ?? and ??, we deduce that

$$\mathbb{P}(W_k(\mathbf{c}_n, \mu) - W_k^*(\mu) \geq t) \leq 64k^3 \exp\left(-\frac{t^2 n}{2048k^2 \rho^4}\right) + O(e^{-rn^{1/5}}),$$

for some $r > 0$, provided t and ρ satisfy the inequality

$$4\left(\kappa^2 + \frac{t\kappa}{64k\rho}\right) \leq \rho^2. \tag{4.10}$$

Note that by Lemma ??, the term $O(e^{-m^{1/5}})$ is uniform over all measures μ such that $R(\mu) \leq S$, for any $S > 0$. Then, for a choice of t such that

$$t^2 \geq \frac{2048k^2\rho^4}{n} \ln \frac{64k^3}{\delta}, \quad (4.11)$$

we have

$$\mathbb{P}(W_k(\mathbf{c}_n, \mu) - W_k^*(\mu) \geq t) \leq \delta + O(e^{-m^{1/5}}). \quad (4.12)$$

It is an easy exercise to prove that when t is subject to (??), then each choice of ρ such that

$$\rho \geq 60 \left(\kappa + \frac{\ln^- \alpha}{\tau} \right),$$

satisfies both conditions (??) and (??). Consequently, for a sufficiently large numerical constant C , one has

$$W_k(\mathbf{c}_n, \mu) - W_k^*(\mu) \leq \frac{Ck}{\sqrt{n}} \left(\kappa + \frac{\ln^- \alpha}{\tau} \right)^2 \ln \frac{k}{\delta},$$

with probability $(1 - \delta) + O(e^{-m^{1/5}})$. Since $R(\mu) = \kappa + \ln^- \alpha / \tau$, the theorem is proved. \square

4.2 Proof of Theorem ??

In the sequel, μ^π stands for the law of X^π .

Lemma 4.3. (i) For all $\mathbf{c} \in \mathcal{H}^k$, $|W_k(\mathbf{c}, \mu^\pi) - W_k(\mathbf{c}, \mu)| \leq 4Ls^{h/4}d^{-h/(2s)}$.

(ii) We have $|W_k^*(\mu^\pi) - W_k^*(\mu)| \leq 4Ls^{h/4}d^{-h/(2s)}$.

(iii) If d satisfies $W_{k-1}^*(\mu) - W_k^*(\mu) \geq 8Ls^{h/4}d^{-h/(2s)}$, then $R(\mu^\pi) \leq 2R_\infty(\mu)$.

PROOF. We only prove (ii) and (iii).

(ii) According to Lemma 3 in Linder (2001) and (??),

$$\begin{aligned}
\left| W_k^*(\mu^\pi)^{1/2} - W_k^*(\mu)^{1/2} \right|^2 &\leq \mathbb{E} \|X^\pi - X\|^2 \\
&= \sum_{p=1}^d \int_{V_p} \mathbb{E} |X(t_p) - X(t)|^2 dt \\
&\leq L \sum_{p=1}^d \int_{V_p} |t_p - t|^h dt \\
&\leq \frac{L s^{h/2}}{d^{h/s}},
\end{aligned}$$

because $\text{vol}(V_p) = 1/d$ and $\text{diam}(V_p) = \sqrt{s}/d^{1/s}$ for all $p = 1, \dots, d$. Consequently,

$$\begin{aligned}
|W_k^*(\mu^\pi) - W_k^*(\mu)| &\leq 2 \max \left(W_k^*(\mu^\pi)^{1/2}, W_k^*(\mu)^{1/2} \right) \frac{\sqrt{L s^{h/2}}}{d^{h/(2s)}} \\
&\leq 2 \left(W_k^*(\mu)^{1/2} + \sqrt{L} \right) \frac{\sqrt{L s^{h/2}}}{d^{h/(2s)}} \\
&\leq \frac{4L s^{h/4}}{d^{h/(2s)}},
\end{aligned}$$

since $W_k^*(\mu) \leq \mathbb{E} \|X\|^2 \leq L$, hence (ii).

(iii) Observing that inequality (ii) holds with $k-1$ instead of k , we can deduce that, provided d is large enough so that $W_{k-1}^*(\mu) - W_k^*(\mu) \geq 8L s^{h/4} d^{-h/(2s)}$, then

$$W_{k-1}^*(\mu^\pi) - W_k^*(\mu^\pi) \geq \frac{1}{2} (W_{k-1}^*(\mu) - W_k^*(\mu)).$$

Therefore,

$$\omega(\mu^\pi) \leq \omega(\mu) + \ln 2.$$

The conclusion is now straightforward, since $\|X\| \leq \|X\|_\infty$. \square

We are now in position to prove Theorem ??.

PROOF OF THEOREM ??. Recall that the term $O(e^{-rn^{1/5}})$ in Theorem ?? holds for all measure μ such that $R(\mu) \leq S$, with a fixed $S > 0$. Therefore, by Lemma

?? and Theorem ?? applied with the measure μ^π , we have:

$$\begin{aligned} W_k(\mathbf{c}_n^\pi, \mu) - W_k^*(\mu) &= [W_k(\mathbf{c}_n^\pi, \mu) - W_k(\mathbf{c}_n^\pi, \mu^\pi)] + [W_k(\mathbf{c}_n^\pi, \mu^\pi) - W_k^*(\mu^\pi)] \\ &\quad + [W_k^*(\mu^\pi) - W_k^*(\mu)] \\ &\leq \frac{8Ls^{h/4}}{d^{h/(2s)}} + C \frac{R(\mu)^2 k \ln(k/\delta)}{\sqrt{n}}, \end{aligned}$$

with probability $(1 - \delta) + O(e^{-rn^{1/5}})$, hence the result. \square

5 Appendix

PROOF OF (??) Since the function $x \mapsto (\ln x)^p$ defined on $] \exp(p-1), \infty[$ is concave, one has according to Jensen's Inequality:

$$\begin{aligned} \mathbb{E}\|X\|^p &\leq \left(\frac{p}{\tau}\right)^p + \frac{1}{\tau^p} \mathbb{E} \left(\ln e^{\tau\|X\|} \right)^p \mathbf{1}_{\left\{ \|X\| \geq \frac{p-1}{\tau} \right\}} \\ &\leq \left(\frac{p}{\tau}\right)^p + \frac{1}{\tau^p} \left(\ln \mathbb{E} e^{\tau\|X\|} \right)^p \\ &\leq \left(\frac{p}{\tau}\right)^p \left(1 + \ln \mathbb{E} e^{\tau\|X\|} \right)^p. \end{aligned}$$

There exists a universal constant $C > 0$ such that $p^p \leq C^p p!$. Therefore,

$$\mathbb{E}\|X\|^p \leq p! \left(\frac{C}{\tau} \right)^p \left(1 + \ln \mathbb{E} e^{\tau\|X\|} \right)^p,$$

hence (??).

PROOF OF (??). We first proceed to bound $\mathbb{E} \exp(\tau\|X\|) \leq \mathbb{E} \exp(\tau\|X\|_\infty)$, where $\|\cdot\|_\infty$ stands for the supremum norm. Denote by $(Z(t))_{t \in [0,1]}$ the continuous-time martingale defined for all $t \in [0, 1]$ by

$$Z(t) = \int_0^t b(s, X(s)) dW(s).$$

Since σ is bounded, say $\bar{\sigma} = \sup |\sigma|$, we have for any $\tau > 0$:

$$\mathbb{E} e^{\tau\|X\|_\infty} \leq e^{\tau\bar{\sigma}} \mathbb{E} e^{\tau\|Z\|_\infty}. \quad (5.1)$$

Hence one only needs to bound the rightmost term. Observe that the quadratic variation $\langle Z \rangle$ of Z satisfies

$$\langle Z \rangle_t = \int_0^t b^2(s, X(s)) ds \leq \bar{b}^2,$$

for all $t \in [0, 1]$, where \bar{b} stands for the supremum of $|b|$, i.e. $\bar{b} = \sup |b|$. Therefore, by Doob's exponential inequality for continuous martingales (see Revuz-Yor, 1999):

$$\mathbb{E} e^{\tau \|Z\|_\infty} = 1 + \int_0^\infty \mathbb{P} \left(\|Z\|_\infty \geq \frac{v}{\tau} \right) e^v dv \leq 1 + 2 \int_0^\infty e^{-v^2/(2\tau^2\bar{b}^2)} e^v dv.$$

According to (??), we then have for all $\tau > 0$:

$$\mathbb{E} e^{\tau \|X\|_\infty} \leq e^{\tau \bar{\sigma}} \left(1 + \sqrt{2\pi} \tau \bar{b} e^{\tau^2 \bar{b}^2/2} \right). \quad (5.2)$$

It is an easy exercise to deduce that

$$\inf_{\tau > 0} R(\mu) = \inf_{\tau > 0} \frac{1}{\tau} \left(1 + \omega(\mu) + \ln \mathbb{E} e^{\tau \|X\|_\infty} \right) \leq 4(\sup |b| + \sup |\sigma|) \sqrt{1 + \omega(\mu)},$$

hence the result. \square

References

- Abaya, E.A. and Wise, G.L. (1984). Convergence of vector quantizers with applications to optimal quantization, *SIAM J. Appl. Math.*, pp. 183–189
- Antos, A. (2005). Improved minimax bounds on the test and training distortion of empirically designed vector quantizers, *IEEE Trans. Inform. Theory*, pp. 4022–4032.
- Antos, A., Györfy, L. and György, A. (2005). Improved convergence rates in empirical vector quantizer design, *IEEE Trans. Inform. Theory*, pp. 4013–4022.
- Bartlett, P.L. (2003). Prediction algorithms: complexity, concentration and convexity, in *Proceedings of the 13th IFAC Symposium on System Identification*, pp. 1507–1517.
- Bartlett, P.L., Boucheron, S. and Lugosi, G. (2001). Model selection and error estimation, *Machine Learning*, pp. 85–113.

- Bartlett, P.L., Linder, T. and Lugosi, G. (1998). The minimax distortion redundancy in empirical quantizer design, *IEEE Trans. Inform. Theory*, pp. 1802–1813.
- Biau, G., Devroye, L. and Lugosi, G. (2008). On the performance of clustering in Hilbert spaces, *IEEE Trans. Inform. Theory*, pp. 781–790.
- Bosq, D. (2000). *Linear processes in function spaces. Theory and applications*, Lecture Notes in Statistics 149, Springer-Verlag, New-York.
- Breton, J.C., Nourdin, I. and Peccati, G. (2009). Exact confidence intervals for the Hurst parameter of a fractional Brownian motion, *Electron. J. Stat.*, pp. 416–425.
- Chou, P.A. (1994). The distortion of vector quantizers trained on n vectors decreases to the optimum at $O_P(1/n)$, *IEEE Trans. Inform. Theory*, pp. 457–457.
- Coeurjolly, J.-F. (2008). Hurst exponent estimation of locally self-similar Gaussian processes using sample quantiles, *Ann. Statist.*, pp. 1404–1434.
- Cont, R. and Tankov, P. (2003). *Financial Modelling with Jump Processes*, Second Edition, Chapman and Hall, CRC Press, London.
- Devroye, L., Györfi, L. and Lugosi, G. (1996). *A probabilistic theory of pattern recognition*, Springer-Verlag, New-York.
- Duda, R.O., Hart, P.E. and Stork, D.G. (2000). *Pattern Classification*, Second Edition, Wiley-Interscience, New York.
- Frisch, U. (1995). *Turbulences*, Cambridge University Press, Cambridge.
- Gersho, A. and Gray, R.M. (1992). *Vector Quantization and Signal Compression*, Kluwer Academic Press, Boston.
- Graf, S. and Luschgy, H. (2000). *Foundations of quantization for probability distributions*, Lectures Notes in Mathematics 1730, Springer-Verlag, New-York.
- Hartigan, J.A. (1975). *Clustering Algorithms*, Wiley, New-York.
- Hartigan, J.A. (1978). Asymptotic distributions for clustering criteria, *Ann. Statist.*, pp. 117–131.
- Huang, W. (2009). Exponential integrability of Itô's processes, *J. Math. Anal. Appl.*, pp. 427–433.
- Kärner, O. (2001). Comments on Hurst exponent, *Geophysical research Letters*, pp. 3825–3826.
- Lacaux, C. and Loubès, J.-M. (2007). Hurst exponent estimation of fractional Lévy motions, *Alea*, pp. 143–164.
- Lamberton, D. and Lapeyre, B. (1996). *Introduction to Stochastic Calculus applied to Finance*, Chapman and Hall, CRC Press, London.

- Ledoux, M. and Talagrand, M. (1991). *Probability in Banach spaces*, Springer-Verlag, New-York.
- Linder, T. (2000). On the training distortion of vector quantizers, *IEEE Trans. Inform. Theory*, pp. 1617–1623.
- Linder, T. (2001). *Learning-theoretic methods in vector quantization*, Lecture Notes for the Advanced School on the Principle of Nonparametric Learning, Udine, Italy, July 9-13.
- Linder, T., Lugosi, G. and Zeger, K. (1994). Rates of convergence in the source coding theorem, in empirical quantizer design, and in universal lossy source coding, *IEEE Trans. Inform. Theory*, pp. 1728–1740.
- Lindstrøm, T. (1993). Fractional Brownian fields as integrals of white noise, *Bull. London Math. Soc.*, pp. 83-88.
- Mandelbrot, B. (1997). *Fractals and scaling in finance*. Selected Works of Benoit B. Mandelbrot. Discontinuity, concentration, risk. Selecta Vol. E, with a forward by R. E. Gomory. Springer-Verlag, New York.
- Mandelbrot, B. and van Ness, J. (1968). Fractional brownian motion, fractional noises and applications. *SIAM Rev.*, pp. 422–437.
- Pipiras, V. and Taqqu, M.S. (2003). Fractional calculus and its connection to fractional Brownian motion, in: *Long Range Dependence*, pp. 166-201, Birkhäuser, Basel.
- Pollard, D. (1981). Strong consistency of k -means clustering, *Ann. Statist.*, pp. 135–140.
- Pollard, D. (1982). A central limit theorem for k -means clustering, *Ann. Probab.*, pp. 199–205.
- Pollard, D. (1982). Quantization and the method of k -means, *IEEE Trans. Inform. Theory*, pp. 1728–1740.
- Revuz, D. and Yor, M. (1999). *Continuous martingales and Brownian motion*, Springer-Verlag, New-York.