Efficient and fast estimation of the geometric median in Hilbert spaces with an averaged stochastic gradient algorithm.

Hervé Cardot, Peggy Cénac, Pierre-André Zitt

Efficient and fast estimation of the geometric median in Hilbert spaces with an averaged stochastic gradient algorithm.

Hervé CARDOT, Peggy CÉNAC, Pierre-André ZITT
Institut de Mathématiques de Bourgogne, Université de Bourgogne,
9 Rue Alain Savary, 21078 Dijon, France
email: {Herve.Cardot, Peggy.Cenac, Pierre-Andre.Zitt}@u-bourgogne.fr

January 21, 2011

Abstract

With the progress of measurement apparatus and the development of automatic sensors it is not unusual anymore to get thousands of samples of observations taking values in high dimension spaces such as functional spaces. In such large samples of high dimensional data, outlying curves may not be uncommon and even a few individuals may corrupt simple statistical indicators such as the mean trajectory. We focus here on the estimation of the geometric median which is a direct generalization of the real median and has nice robustness properties. The geometric median being defined as the minimizer of a simple convex functional that is differentiable everywhere when the distribution has no atoms, it is possible to estimate it with online gradient algorithms. Such algorithms are very fast and can deal with large samples. Furthermore they also can be simply updated when the data arrive sequentially. We state the almost sure consistency and the L^2 rates of convergence of the stochastic gradient estimator as well as the asymptotic normality of its averaged version. We get that the asymptotic distribution of the averaged version of the algorithm is the same as the classic estimators which are based on the minimization of the empirical loss function. The performances of our averaged sequential estimator, both in terms of computation speed and accuracy of the estimations, are evaluated with a small simulation study. Our approach is also illustrated on a sample of more 5000 individual television audiences measured every second over a period of 24 hours.

Keywords. CLT, functional data, geometric quantiles, high dimension, L^1-median, online algorithms, recursive estimation, Robbins-Monro algorithm, spatial median.
MSC2010 classification. 62L20, 62G05, 62G35, 60B12, 68W27

1 Introduction

With the progress of measurement apparatus, the development of automatic sensors and the increasing storage performances of computers it is not unusual anymore to get thousands of samples of functional observations. For example Cardot et al. (2010a) analyse more than 18000 electricity consumption curves measured every half hour over a period of two weeks. Our study is motivated by the estimation of the central point of a sample of $n = 5423$ vectors of $\mathbb{R}^d$, with $d = 86400$, which correspond to individual television audiences measured every second over a period of 24 hours.

In such large samples of high dimensional data, outlying curves may not be uncommon and a even few individuals may corrupt simple statistical indicators such as the mean trajectory or the principal components (Gervini (2008)). Detecting these atypical curves automatically is not straightforwards in such a high dimensional and large sample context and considering directly robust techniques is an interesting alternative. There are many robust location indicators in the multivariate setting (Small (1990)) but most of them require high computational efforts to be estimated, even for small sample sizes, when the dimension is relatively large. For example, Fraiman and Muniz (2001) have extended the notion of trimmed means to a functional context in order to get robust estimators of the mean profile. In order to deal with the dimensionality issue and to reduce the computation time, Cuevas et al. (2007) have proposed random projection approaches in the context of maximal depth estimators and studied their properties via simulation studies. Note that sub-sampling approaches based on survey sampling techniques have also been proposed in the literature in order to reduce the sample size with a controlled probabilistic procedure and thus reduce the computational time (Chaouch and Goga (2010)).

We focus here on the geometric median, also called L^1-median or spatial median, which is a direct generalization of the real median proposed by Haldane (1948) and whose properties have been studied in details by Kemperman (1987). As noted in Small (1990), one drawback of the geometric median is that it is not affine equivariant. Nevertheless, it is invariant to translation and scale changes and thus is well adapted to functional data which are observed with the same units at each instant of time. In a functional context, consistent estimators of the L^1-median have been proposed by Kemperman (1987), Cadre (2001) and Gervini (2008). Iterative estimation algorithms have been developed by Gower (1974), Vardi and Zhang (2000) in the multivariate setting and by Gervini (2008) for functional data. This latter algorithm requires to invert at each step matrices whose dimension is equal to the dimension d of the data and thus need important computational efforts. The algorithm proposed by Vardi and Zhang (2000) is much faster and only requires $O(nd)$ operations at each iteration, where n is the sample size. Nevertheless, these estimation procedures are not adapted when the data arrive sequentially, they need to store all the data and they cannot be simply updated.

In this paper, we explore another direction. The geometric median being defined as the minimizer of a simple functional that is differentiable everywhere when the distribution has no atoms, it is possible to estimate it with online gradient algorithms. Such algorithms are very fast and can be simply updated when the data arrive sequentially. There is a vast literature on stochastic gradient algorithms which mainly focus on the multivariate case (Kushner and Clark (1978), Ruppert (1985), Benveniste et al. (1990), Ljung et al. (1992), Duflo (1997), Kushner and Yin (2003), Bottou (2010)). The literature is much less abundant when one has to consider online observations taking values in a functional space (usually an infinite dimensional Banach or Hilbert space) and most works focus on linear algorithms (Walk (1977), Dippon and Walk (2006), Smale and Yao (2006)).

It is known in the multivariate setting that averaging procedures can lead to efficient estimation procedure under additional assumptions on the noise and when the target is defined as the minimizer of a strictly convex function (Polyak and Juditsky (1992), Pelletier (2000)). There is little work on averaging when considering random variables taking values in Hilbert spaces and, as far as we know, they only deal with linear algorithms (Dippon and Walk (2006)). Nevertheless, it has been noted in an empirical study whose aim was to estimate the geometric median with functional data (Cardot et al. (2010a)) that averaging could improve in an important way the accuracy of the estimators.

The paper is organized as follows. We first fix notations, give some properties of the

geometric median and present our stochastic gradient algorithm as well as its averaged version. We also note that our study extends directly to the estimation of geometric quantiles defined by Chaudhuri (1996). In a third section we state the almost sure consistency and the L^2 rates of convergence of the stochastic gradient estimators as well as the asymptotic normality of its averaged version. We get that the asymptotic distribution of the averaged version of the algorithm is the same as the classic estimators. A fourth section is devoted to a small simulation study which aims at comparing the performances of our estimator with the static algorithm developed by Vardi and Zhang (2000). The comparison is performed according to two points of view, for the same sample size and for the same computation time. We also analyze a real example with a large sample of individual television audiences measured every second over a period of 24 hours. The proofs are gathered in Section 6.

2 The algorithms and some properties of the geometric median

2.1 Definitions and assumptions

Let H be a separable Hilbert space (think $H = \mathbb{R}^d$ or $H = L^2(I)$, for some closed interval $I \subset \mathbb{R}$). We denote by $\langle .,. \rangle$ its inner product and by $\|\cdot\|$ the associated norm.

The geometric median m of a random variable X taking values in H is defined by (see Kemperman (1987)),

$$m := \arg\min_{u \in H} \mathbb{E}\left[\|X-u\| - \|X\|\right]. \tag{1}$$

Note that the general definition (1) does not assume the existence of the first order moment of $\|X\|$. We suppose from now on that the following assumptions are fulfilled.

A1. The random variable X is not concentrated on a straight line: for all $v \in H$, there is $w \in H$ such that $\langle v, w \rangle = 0$ and

$$\mathbf{Var}(\langle w, X \rangle) > 0. \tag{2}$$

A2. The law of X is a mixing of two "nice" distributions : $\mu_X = \lambda\mu_c + (1-\lambda)\mu_d$, where

- μ_c is not strongly concentrated around single points: if $\mathcal{B}(0,A)$ is the ball $\{\alpha \in H, \|\alpha\| \leq A\}$,

$$\forall A, \exists C_A \in [0,\infty), \forall \alpha \in \mathcal{B}(0,A), \quad \mathbb{E}\left[\|X-\alpha\|^{-1}\right] \leq C_A. \tag{3}$$

- μ_d is a discrete measure, $\mu_d = \sum_i p_i \delta_{\alpha_i}$, which does not charge the median m. We denote by D the support of μ_d.

As shown in Kemperman (1987), assumption (A1) ensures that the median m is uniquely defined.

The second assumption could probably be relaxed, but it is general enough for most natural examples. The conditions on μ_c, for example, are satisfied when $H = \mathbb{R}^d$, with $d \geq 2$, whenever μ_c has a bounded density on every compact subset of $\mathbb{R}^d$ (as noted in Chaudhuri (1992)). More precisely, this property is closely related to small ball probabilities since

$$\mathbb{E}\left[\|X-m\|^{-1}\right] = \int_0^\infty \mathbb{P}\left[\|X-m\| \leq t^{-1}\right] dt.$$

If $\mathbb{P}\left[\|X-m\| \leq \epsilon\right] \leq C\epsilon^d$, for small ϵ and some positive constant C, it is easy to check that

$$\mathbb{E}\left[\|X-m\|^{-\beta}\right] < \infty,$$

whenever $0 \leq \beta < d$.

When $H = L^2(I)$, the dimension is not finite and small ball probabilities have been derived for some particular classes of Gaussian processes (see Nazarov (2009) for a recent reference). In this case, by symmetry of the distribution, the median m is equal to the mean, and many processes satisfy, for positive constants C_1, C_2, C_3 and a constant C_4 which depend on the process under study,

$$\mathbb{P}\left[\|X - m\| \leq \epsilon\right] \leq C_1 \epsilon^{C_4} \exp(-C_2 \epsilon^{-C_3}), \tag{4}$$

so that $\mathbb{E}\left[\|X - m\|^{-\beta}\right] < \infty$, for all positive β.

2.2 Some convexity and robustness properties of the median

Recalling the definition of the median (eq. (1)), let us denote by $G : H \mapsto \mathbb{R}$ the function we would like to minimize:

$$G(\alpha) := \mathbb{E}\left[\|X - \alpha\| - \|X\|\right]. \tag{5}$$

This function is convex since it is a convex combination of convex functions. To ensure the convergence of the algorithm, we will need quantitative bounds on its convexity.

Under assumptions (A1) and (A2) this function can be decomposed in two parts:

$$G(\alpha) = \lambda G_c(\alpha) + (1 - \lambda) G_d(\alpha),$$

where we isolate the discrete part $G_d(\alpha) = \sum_i p_i(\|x_i - \alpha\| - \|x_i\|)$. The first part is Fréchet differentiable everywhere (Kemperman (1987)), so G is differentiable except on D, the support of the discrete part μ_d . We denote by $\Phi = \lambda \Phi_c + (1 - \lambda)\Phi_d$ its Fréchet derivative,

$$\Phi(\alpha) := \nabla_\alpha G = -\mathbb{E}\left[\frac{X - \alpha}{\|X - \alpha\|}\right]. \tag{6}$$

Remark 1. *It will be useful to define Φ on the set D. If $x \in D$, we define G_x by "forgetting" x,*

$$G_x(y) = \sum_{i, x_i \neq x} p_i(\|x_i - y\| - \|x_i\|).$$

This function is Fréchet differentiable in x, and we let

$$\Phi_d(x) = \sum_{i, x_i \neq x} p_i \frac{x - x_i}{\|x - x_i\|}.$$

In other words, up to a multiplicative constant (accounting for the loss of the mass p_x), $\Phi(\alpha)$ is just Φ computed for a different law of X, the one where we delete the Dirac mass $p_\alpha \delta_\alpha$.

In the vocabulary of convex analysis, what we have done is just to choose one particular subgradient of G on the set D of non-differentiability. In particular, it is easily seen (we give a short proof in the annex) that:

$$\forall x, y, \quad G(y) - G(x) \geq \langle \Phi(x), y - x \rangle$$

(which asserts that Φ is a subgradient).

The median m is then the unique solution of the nonlinear equation,

$$\Phi(\alpha) = 0. \tag{7}$$

To exhibit some useful strong convexity and robustness properties of the median we need to introduce the Hessian of functional G, for $\alpha \in H \setminus D$. It is denoted by Γ_α, maps H

to H and it is easy to check (see Koltchinskii (1997) for the multivariate case and Gervini (2008) for the functional one) that

$$\Gamma_\alpha = \mathbb{E}\left[\frac{1}{\|X-\alpha\|}\left(\mathbf{I}_H - \frac{(X-\alpha)\otimes(X-\alpha)}{\|X-\alpha\|^2}\right)\right], \tag{8}$$

where $\mathbf{I}_H$ is the identity operator in H and $u \otimes v(h) = \langle u, h\rangle\, v$, for u, v and h belonging to H. Operator Γ_α is not compact but it is bounded when $\mathbb{E}\left[\|X-\alpha\|^{-1}\right] < \infty$.

Let us look at $\langle h, \Gamma_\alpha h\rangle$: if we define $\bar{h} = h/\|h\|$, and P_h the projection onto the orthogonal complement of h,

$$\begin{aligned}\langle h, \Gamma_\alpha h\rangle &= \|h\|^2 \mathbb{E}\left[\frac{1}{\|\alpha - X\|}\left(1 - \frac{\langle \bar{h}, \alpha - X\rangle^2}{\|\alpha - X\|^2}\right)\right] \\ &= \|h\|^2 \mathbb{E}\left[\frac{1}{\|\alpha - X\|}\frac{\|P_{\bar{h}}(\alpha - X)\|^2}{\|\alpha - X\|^2}\right].\end{aligned} \tag{9}$$

We can now state a strong convexity property of functional G which can be seen as an extension to an infinite dimensional setting of Proposition 4.1 in Koltchinskii (1997).

Proposition 2.1. *Recall that $\mathcal{B}(0, A)$ is the ball of radius A in H. Under assumptions A1 and A2, there is a strictly positive constant c_A, such that:*

$$\forall \alpha \in \mathcal{B}(0, A) \setminus D, \forall h \in H, \quad c_A \|h\|^2 \leq \langle h, \Gamma_\alpha h\rangle \leq C_A \|h\|^2.$$

In other words, G is strictly convex in H and it is strongly convex on any bounded set, as shown in the following corollary.

Corollary 2.2. *Assume hypotheses of Proposition 2.1 are fulfilled. For any strictly positive A, there is a strictly positive constant c_A such that:*

$$\forall \alpha_1, \alpha_2 \in \mathcal{B}(0, A)^2, \qquad \langle \Phi(\alpha_2) - \Phi(\alpha_1), \alpha_2 - \alpha_1\rangle \geq c_A \|\alpha_2 - \alpha_1\|^2.$$

As a particular case of Proposition 2.1, we get that there are two strictly positive constants $0 < c_m \leq C_m \leq \mathbb{E}\left[\|X-m\|^{-1}\right] < \infty$, such that

$$c_m \|h\|^2 \leq \langle h, \Gamma_m h\rangle \leq C_m \|h\|^2. \tag{10}$$

As noted in Kemperman (1987), the geometric median has a 50 % breakdown point. Furthermore, an immediate consequence of (10) is that operator Γ_m has a bounded inverse. Thus, the gross error sensitivity, which is also a classical indicator of robustness (see Huber and Ronchetti (2009)), is bounded for the median in a separable Hilbert space. Indeed, thanks to the expression derived in Gervini (2008), it is bounded as follows,

$$\sup_{z\in H}\left\|\Gamma_m^{-1}\left(\frac{z-m}{\|z-m\|}\right)\right\| \quad \leq \quad \frac{1}{c_m}. \tag{11}$$

2.3 The algorithms

When observing a sample $X_1, X_2, \ldots, X_n$ of n independent realizations of X, a natural estimator of m is the solution $\widehat{m}_n$ of the empirical version of (7),

$$\sum_{i=1}^{n} \frac{X_i - \widehat{m}_n}{\|X_i - \widehat{m}_n\|} = 0. \tag{12}$$

The solution $\widehat{m}_n$ is defined implicitly and is found by iterative algorithms.

We propose now an alternative and simple estimation algorithm which can be seen as a stochastic gradient algorithm (Ruppert (1985); Duflo (1997)) and is defined as follows

$$Z_{n+1} = Z_n + \gamma_n \frac{X_{n+1} - Z_n}{\|X_{n+1} - Z_n\|} \tag{13}$$

$$= Z_n - \gamma_n U_{n+1}, \tag{14}$$

with a starting point that can be random and bounded, *e.g.* $Z_0 = X_0 \mathbf{1}_{\{\|X_0\| \leq M\}}$ for some positive constant M fixed in advance, or deterministic. If $X_{n+1} = Z_n$, we set $U_{n+1} = 0$ and $Z_{n+1} = Z_n$ so the algorithm does not move Z_{n+1}. The sequence of descent steps γ_n controls the convergence of the algorithm. The direction U_{n+1} is an "estimate" of the gradient Φ of G at Z_n since the conditional expectation given the sequence of σ-algebra $\mathcal{F}_n = \sigma(Z_1, \ldots, Z_n) = \sigma(X_1, \ldots, X_n)$ satisfies

$$\mathbb{E}\left[U_{n+1} | \mathcal{F}_n\right] = \Phi(Z_n). \tag{15}$$

Note that our particular choice of a subgradient Φ on the points where there is a mass was done to ensure that this equality always holds.

Defining now by ξ the sequence of "errors" in these estimates,

$$\xi_{n+1} = \Phi(Z_n) - U_{n+1}, \tag{16}$$

algorithm (13) can also be seen as a non linear Robbins-Monro algorithm,

$$Z_{n+1} = Z_n + \gamma_n \left(-\Phi(Z_n) + \xi_{n+1}\right). \tag{17}$$

Thanks to (15) and (16), the sequence (ξ_n) is a sequence of martingale differences. Let us note that the bracket of the associated martingale satisfies,

$$\mathbb{E}\left[\|\xi_{n+1}\|^2 \Big| \mathcal{F}_n\right] = \mathbb{E}\left[\|U_{n+1}\|^2 \Big| \mathcal{F}_n\right] - \|\Phi(Z_n)\|^2 \tag{18}$$

$$= 1 - \|\Phi(Z_n)\|^2 \leq 1. \tag{19}$$

Our second algorithm consists in averaging all the estimated past values,

$$\overline{Z}_{n+1} = \overline{Z}_n + \frac{1}{n+1}\left(Z_{n+1} - \overline{Z}_n\right)$$

with $\overline{Z}_0 = 0$, so that $\overline{Z}_n = \frac{1}{n}\sum_{i=1}^n Z_i$.

Remark 2. *An extension of the notion of quantiles in Euclidean and Hilbert space has been proposed by Chaudhuri (1996), the geometric median being a particular case. Consider a vector $v \in H$, such that $\|v\| < 1$. The geometric quantile of X, say m^v, corresponding to direction v, is defined, uniquely under previous assumptions, by*

$$m^v = \arg\min_{u \in H} \mathbb{E}\left[\|X - u\| + \langle X - u, v\rangle\right].$$

It is characterized by

$$\Phi_u(m^v) = \Phi(m^v) - v = 0,$$

so that it can be estimated with the following stochastic algorithm

$$\widehat{m}^v_{n+1} = \widehat{m}^v_n + \gamma_n \left(\frac{X_{n+1} - \widehat{m}^v_n}{\|X_{n+1} - \widehat{m}^v_n\|} + v\right),$$

as well as with its averaged version.

3 Convergence results

3.1 Almost sure convergence of the stochastic gradient algorithm

We first state the almost sure consistency of our sequence of estimators Z_n under classical and general assumptions on the descent steps γ_n.

Proposition 3.1. *If (A1) and (A2) hold, and if $(\gamma_n)_{n\in\mathbb{N}}$ satisfies the usual conditions:*

$$\sum_{n\geq 1}\gamma_n = \infty, \qquad \sum_{n\geq 1}\gamma_n^2 < \infty,$$

then

$$\lim_{n\to\infty}\|Z_n - m\| = 0, \quad a.s.$$

3.2 Rates of convergence and asymptotic normality

We study now the rates of convergence of the stochastic gradient algorithm as well as the asymptotic distribution of its averaged version. For these results, we restrict ourselves to more specific sequences (γ_n) given by $\gamma_n = c_\gamma n^{-\alpha}$, where c_γ is a positive constant and $\alpha \in (\frac{1}{2}, 1)$.

The following proposition states that, on events of arbitrarily high probability, the functional estimator Z_n attains the classical rates of convergence in quadratic mean (see (Duflo, 1997, theorem 2.2.12) for the multivariate case) up to a logarithmic factor.

Proposition 3.2. *Assume (A1)-(A2) and suppose that there is a positive constant A such that*

$$\exists C_A \in [0,\infty), \forall h \in \mathcal{B}(0,A), \quad \mathbb{E}\left[\|X-(m+h)\|^{-2}\right] \leq C_A. \tag{20}$$

Then, there exist an increasing sequence of events $(\Omega_N)_{N\in\mathbb{N}}$, and constants C_N, such that $\Omega = \bigcup_{N\in\mathbb{N}}\Omega_N$, and

$$\forall N, \quad \mathbb{E}\left[\mathbf{1}_{\Omega_N}\|Z_n - m\|^2\right] \leq C_N\gamma_n \ln\left(\sum_{k=1}^{n}\gamma_k\right) \leq C_N\frac{\ln(n)}{n^\alpha}.$$

The additional assumption, $\sup_{h\in\mathcal{B}(0,A)}\mathbb{E}\left[\|X-(m+h)\|^{-2}\right] < \infty$, is not restrictive when the dimension is strictly larger than two as discussed in (4). Hypothesis (20) is needed to bound the difference between G and its quadratic approximation, in a neighborhood of m as stated in the following Lemma.

Lemma 3.3. *Suppose there is a positive constant A such that*

$$\exists C_A \in [0,\infty), \forall h \in \mathcal{B}(0,A), \quad \mathbb{E}\left[\|X-(m+h)\|^{-2}\right] \leq C_A.$$

Then,

$$\Phi(m+h) = \Gamma_m(h) + O\left(\|h\|^2\right).$$

Finally, Proposition 3.4 stated below probably gives the most important result of this work. It is shown that the averaged estimator $\overline{Z}_n$ and the classic static estimator $\widehat{m}_n$ have the same asymptotic distribution. Consequently, for large sample sizes, it is possible to get, very quickly, estimators which are as efficient, at first order, as the slower static one $\widehat{m}_n$. Note that the asymptotic distribution of $\widehat{m}_n$ has been derived in the multivariate case by Haberman (1989), Theorem 6.1. For variables taking values in a Hilbert space, such asymptotic distribution has only been proved for a particular case, when the support of X is a finite dimensional space (Theorem 6 in Gervini (2008)).

Proposition 3.4. *Assume (A1)-(A2) and suppose that for some positive constant A,*

$$\sup_{h\in\mathcal{B}(0,A)} \mathbb{E}\left[\|X-(m+h)\|^{-2}\right] < \infty.$$

Then,

$$\sqrt{n}\left(\overline{Z}_n - m\right) \xrightarrow[\mathrm{n}\to\infty]{\mathcal{L}} \mathcal{N}\left(0, \Gamma_m^{-1}\Sigma\Gamma_m^{-1}\right),$$

with,

$$\Sigma = \mathbb{E}\left[\frac{(X-m)}{\|X-m\|}\otimes\frac{(X-m)}{\|X-m\|}\right].$$

Note that with (10), operator Γ_m^{-1} is well defined, it is bounded and positive. The proofs of Lemma 3.3 and Propositions 3.2 and 3.4 are given in Section 6.

4 An illustration on simulated and real data

4.1 A simulation study

A simple simulation study is performed to check the good behavior of the averaging estimator and we make a comparison of our approach with the static estimator developed by Vardi and Zhang (2000) considering two points of view. The first classic one consists in evaluating the performances of these two different approaches for different sample sizes. The second one, which is the point of view that should be adopted when computation time matters, consists in comparing the accuracy of both approaches when the allocated computation time is fixed in advance. The Vardi & Zhang estimator is computed thanks to the function `spatial.median` from the R library `ICSNP`.

For simplicity, we consider random variables taking values in $\mathbb{R}^3$ and make simulations of Gaussian random vectors with median $m = (0,0,0)$ and covariance matrix:

$$\Gamma = \begin{pmatrix} 3 & 2 & 1 \\ 2 & 4 & -0.5 \\ 1 & -0.5 & 2 \end{pmatrix}.$$

In order to compare the algorithms, we evaluate the estimation error of an estimator $\widehat{m}$, with the following criterion:

$$R(\widehat{m}) = \|\widehat{m} - m\|. \tag{21}$$

Our averaged estimator depends on the tuning parameters α and c_γ which control the descent steps $\gamma_k = c_\gamma k^{-\alpha}$. It is well known that for the particular case $\alpha = 1$, the choice of parameter c_γ is crucial for the convergence and depends on the second derivative of G in m which is unknown in practice. As usually done for such procedures, we fix $\alpha = 3/4$ and focus on the choice of c_γ. Taking advantage of the rapidity of our recursive algorithm,

we decide to adopt the following strategy; run in parallel the algorithm for 10 initial points chosen randomly in the sample and then select the best estimate $\widehat{m}$ which corresponds to the minimum value of the empirical version of (1), that is the minimum value of

$$\frac{1}{n}\sum_{i=1}^{n}\left(\|X_i - \hat{m}\| - \|X_i\|\right).$$

4.1.1 Fixed sample sizes

We perform 1000 simulations for different sample sizes, $n = 250$, $n = 500$ and $n = 2000$. Table 1 presents the estimation error (first quartile Q1, median and third quartile Q3), according to criterion (21), for the algorithm by Vardi and Zhang (2000) and our averaged procedure considering different values for $c_\gamma \in \{0.2, 0.6, 1, 2, 5, 10, 15, 25, 50, 75\}$.

Table 1: Comparison of the estimation errors for different sample sizes

	n=250			n=500			n=2000		
Estimator	[Q1	median	Q3]	[Q1	median	Q3]	[Q1	median	Q3]
$c_\gamma = 0.2$	0.45	0.60	0.80	0.38	0.53	0.69	0.25	0.35	0.47
$c_\gamma = 0.6$	0.21	0.29	0.40	0.15	0.21	0.29	0.06	0.09	0.12
$c_\gamma = 1$	0.15	0.22	0.31	0.11	0.16	0.21	0.05	0.08	0.10
$c_\gamma = 2$	0.15	0.21	0.30	0.09	0.15	0.20	0.05	0.07	0.10
$c_\gamma = 5$	0.13	0.19	0.25	0.09	0.13	0.18	0.04	0.06	0.09
$c_\gamma = 10$	0.13	0.18	0.25	0.09	0.13	0.18	0.04	0.06	0.09
$c_\gamma = 15$	0.12	0.18	0.25	0.09	0.13	0.18	0.04	0.06	0.08
$c_\gamma = 25$	0.13	0.19	0.26	0.09	0.13	0.18	0.04	0.06	0.09
$c_\gamma = 50$	0.13	0.19	0.26	0.09	0.13	0.18	0.04	0.06	0.09
$c_\gamma = 75$	0.14	0.20	0.27	0.09	0.14	0.19	0.05	0.07	0.09
Vardi & Zhang	0.12	0.18	0.25	0.09	0.12	0.17	0.04	0.06	0.08

At first, we can note that even for moderate sample sizes the averaged procedure performs well in comparison with the Vardi and Zhang estimator which only performs slightly better. We can also remark that the averaged stochastic estimator is not much sensitive to the value of the tuning parameter c_γ which can take values in the interval $[2, 75]$ without modifying the performances of the estimator. As a matter of fact, we noted on simulations that interesting values for c_γ are around or above $\mathbb{E}\left[\|X - m\|\right]$, which is about 2.7 for this particular simulation study.

4.1.2 Fixed computation time

Even if both algorithms require computation times which are $O(nd)$, the averaged stochastic gradient approach is much faster (on the same computer, with procedures coded in the same R language). For example, in previous simulations, if the sample size is $n = 1000$, the averaged estimator is about 30 times faster. When the dimension gets larger, as seen in the real data example, the difference is even more impressive. If the audience is measured every minute over a period of 24 hours, we get vectors of dimension $d = 1440$ and then, for a sample size $n = 6000$, the averaged estimator takes about 1 second, whereas it takes 66 seconds for the estimator proposed by Vardi & Zhang. When one has measurements every second, so that $d = 86400$, it takes about 60 seconds for our procedure to estimate the median.

Let us suppose the allocated time for computation is limited and fixed in advance, say 1 second, and compare the sample sizes that can be handled by the different algorithms. The static estimator by Vardi and Zhang (2000) can deal with $n = 150$ observations, whereas our recursive algorithm, coded in the R language, can take account of $n = 4500$, so that it gives much better estimates of the median, as seen in Table 1. Finally, if the algorithm is coded in **C** and called from R, then it is at least 20 times faster than its R analogue, so that it can deal with at least $n = 90000$ observations, during the same second.

4.2 Estimation of the median television audience profile

The analysis of audience profiles for different channels, or different days of the year, is an essential tool to understand the consumers' habits as regards television. The French society Médiamétrie provides official television audience rates in France. Médiamétrie works with a panel of about 9000 individuals; the television sets of these individuals are equipped with sensors that measure the audience of the different channels.

A sample of around 7000 people is drawn every day in this panel and the television consumption of the people belonging to this sample is recorded every second. The data are then sent to Médiamétrie during the night. Thus recursive techniques are well adapted to deal with such kind of data. Moreover, Médiamétrie has noted in these samples the presence of some atypical behaviors so that robust techniques may be helpful.

We focus our study on the estimation of the television audience profile during the 6th september 2010. After removing people from the sample that have not watched television at all on that day, we finally get a sample of size $n = 5423$. For each element i of the sample, we have a vector $X_i \in \{0, 1\}^{86400}$, where 86400 is the number of seconds within a day, and zero values correspond to seconds during the day where i is not watching television.

A classical audience indicator is given by the mean profile, drawn in Figure 1, which is simply the proportion of people watching television at every second over the considered period of time. We compare this classical indicator with the geometric median, whose estimation is drawn in black in Figure 1. We can first note that both estimators have the same shape along time, showing three peaks of audience during the day with higher audience rates between 8 and 10 PM. Estimated values are smaller for the geometric median which is less sensitive to small perturbations and outliers. This also indicates that the distribution of the individual audience curves is not symmetric around the mean profile.

From a computational point of view, it takes less than one minute for our algorithm to converge. The value of the tuning parameter was chosen to be $c_\gamma = 400$, it leads to a value of about 92 for the empirical loss criterion.

5 Concluding remarks

The experimental results confirm that averaged recursive estimators of the geometric median relying on stochastic gradient approaches are of particular interest when one has to deal with large samples of data and potential outliers. Furthermore, when the allocated computation time is limited and fixed in advance and the data arrive online these techniques can deal, in a recursive way, with larger sample sizes and finally provide estimations that are much more accurate than static estimation procedures. We have also noted that they are not very sensitive to the values of the tuning parameters c_γ that control the gain.

One could imagine many directions for future research that certainly deserve further attention. Taking advantage of the rapidity of our estimation procedures, one could use resampling techniques, similar to the bootstrap, in order to approximate the asymptotic

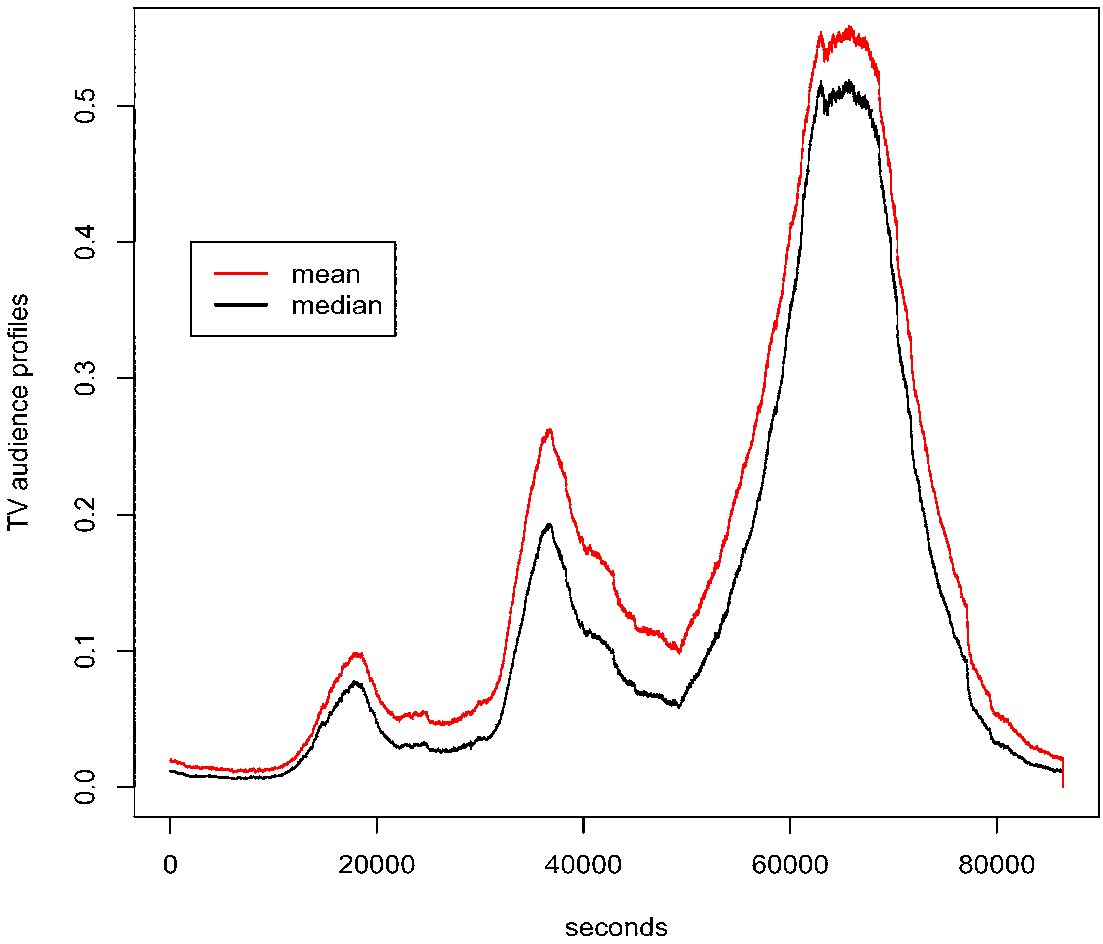

Figure 1: Estimations of the mean and and the geometric median audiences, at a second scale, during the 6th september 2010.

distribution of the estimator given in Proposition 3.4 and then build pointwise confidence intervals. Proving rigorously the validity of such techniques is far beyond the scope of this paper.

Our procedure can also be extended readily for online clustering, adapting the well known MacQueen algorithm (MacQueen (1967)) to the L_1 context. Even if the criterion to be optimized is not convex anymore, it can be proved that stochastic gradient approaches converge almost surely to the set of stationary points (Cardot et al. (2010b)) and thus are interesting candidates for online clustering.

Another direction of interest is online estimation of the conditional geometric median when real covariates are available. For instance, the age or the size of the city where individual live are known by Médiamétrie and it can be possible to take such information into account in order to get varying time regression models that can also be estimated in a very fast way thanks to sequential approaches.

6 Proofs

6.1 Convexity — Proofs of Proposition 2.1 and corollary 2.2

We first show that Φ is a subgradient of G. For points $x \notin D$, it is clear (since G is Fréchet differentiable).

Pick a point x in D (say it is x_0). We will show that Φ_d is a subgradient of G_d, where we have defined $\Phi_d(x_0) = \sum_i p_i \frac{x - x_i}{\|x - x_i\|}$. This follows from a simple computation:

$$\begin{aligned}
\langle \Phi_d(x_0), y - x_0 \rangle &= \sum_{i \neq 0} p_i \frac{\langle x_0 - x_i, y - x_0 \rangle}{\|x_0 - x_i\|} \\
&= \sum_{i \neq 0} p_i \frac{\langle x_0 - x_i, y - x_i \rangle}{\|x_0 - x_i\|} - \sum_{i \neq 0} p_i \|x_0 - x_i\| \\
&\leq \sum_{i \neq 0} p_i \|y - x_i\| - \sum_{i \neq 0} p_i \|x_0 - x_i\| \\
&\leq G_d(y) - G_d(x_0).
\end{aligned}$$

The upper bound in proposition 2.1 follows immediately from (9) and the assumption (A2).

For the lower bound, thanks to (9), we only need to prove:

$$\forall \alpha \in \mathcal{B}(0, A), \forall u, \|u\| = 1, \qquad \langle u, \Gamma_\alpha u \rangle = \mathbb{E}\left[\frac{\|P_u(X - \alpha)\|^2}{\|X - \alpha\|^3} \right] \geq c_A, \tag{22}$$

where P_u is the projection on the orthogonal of u. This quantity is small when $X - \alpha$ is in $\text{span}(u)$.

Recall that (by (A1)), X is not supported on a line. Consider the set of subspaces $K \subset H$ satisfying: $\forall x \in K, \mathbf{Var}(\langle x, X \rangle) = 0$. Suppose that this set is non-empty, and let H' be a maximal element in it (this exists by Zorn's lemma). The orthogonal of H' has at least dimension 2 (otherwise, we get a contradiction to A1). Let v_1, v_2 be two orthogonal vectors in $H' \perp$. Let $v_t = \cos(t)v_1 + \sin(t)v_2$. The map

$$t \mapsto \mathbf{Var}(\langle v_t, X \rangle)$$

is continuous on a compact set. Its minimum cannot be zero (since this would contradict the maximality of H'). Therefore there exists a c such that, for all unit v in the plane spanned by (v_1, v_2), $\mathbf{Var}(\langle X, v \rangle) \geq c$.

The orthogonal of u (an hyperplane) and the (2-dimensional) plane spanned by v_1 and v_2 necessarily intersect: there exists a unit vector $v \in \text{span}(v_1, v_2)$ such that $\langle u, v \rangle = 0$. Therefore, for all $x \in H$, $\|P_u(x - \alpha)\|^2 \geq \langle x, v \rangle^2$.

Suppose first that X is a.s. bounded by K. Then

$$\mathbb{E}\left[\frac{\langle v, X - \alpha \rangle^2}{\|X - \alpha\|^3} \right] \geq \frac{1}{(A + K)^3} \mathbb{E}\left[\langle v, X - \alpha \rangle^2 \right].$$

It is easily seen that the last term is bounded below by $\mathbf{Var}(\langle v, X \rangle) \geq c$ and (22) holds with

$$c_A = \frac{1}{(K + A)^3} c.$$

To get rid of the boundedness assumption on X, we can just choose K large enough so that $\mathbf{Var}(\langle v, X \mathbf{1}_{\|X\| \leq K} \rangle)$ is strictly positive for $v = v_1, v_2$.

The corollary is a consequence of the proposition 2.1 and the fact that Φ is a subgradient. Indeed, the inequality holds for Φ_c by interpolation: for an elementary proof, define $\alpha_t = (1-t)\alpha_1 + t\alpha_2$, and write $\Phi(\alpha_2) - \Phi(\alpha_1) = \int_0^t f'(t)dt$ where $f(t) = \Phi(\alpha_t)$. One can then apply (22), t by t, with $\alpha = \alpha_t$ and $u = \frac{\alpha_2 - \alpha_1}{\|\alpha_2 - \alpha_1\|}$.

Moreover,

$$\begin{aligned}\langle \Phi_d(\alpha_2) - \Phi_d(\alpha_1), \alpha_2 - \alpha_1 \rangle &= \langle \Phi_d(\alpha_2), \alpha_2 - \alpha_1 \rangle + \langle \Phi_d(\alpha_1), \alpha_1 - \alpha_2 \rangle \\ &\geq G(\alpha_2) - G(\alpha_1) + G(\alpha_1) - G(\alpha_2) \\ &= 0.\end{aligned}$$

Since $\Phi = \lambda\Phi_c + (1-\lambda)\Phi_d$, the corollary 2.2 is proved.

6.2 Proof of Proposition 3.1.

The proof of Proposition 3.1 follows a classical strategy and consists of two steps.

Lemma 6.1. *Under the hypotheses of Proposition 3.1, there is a random variable V such that, $\mathbb{E}\left[|V|^2\right] < \infty$, and*

$$\lim_{n\to\infty} \| Z_n - m \|^2 = V, \quad a.s.$$

Proof of Lemma 6.1. Let us consider $V_n := \|Z_n - m\|^2$. Recall that $Z_{n+1} = Z_n - \gamma_n \Phi(Z_n) + \gamma_n \xi_{n+1}$ (cf. (17)). Therefore

$$V_{n+1} = \|Z_n - m - \gamma_n \Phi(Z_n)\|^2 + \gamma_n^2 \|\xi_{n+1}\|^2 + 2\gamma_n \langle \xi_{n+1}, Z_n - \gamma_n \Phi(Z_n) \rangle .$$

If we condition with respect to $\mathcal{F}_n$, the last term disappears since (ξ_n) is a martingale difference sequence and it comes:

$$\begin{aligned}\mathbb{E}\left[V_{n+1}|\mathcal{F}_n\right] &= \|Z_n - m - \gamma_n \Phi(Z_n)\|^2 + \gamma_n^2 \mathbb{E}\left[\|\xi_{n+1}\|^2 \Big| \mathcal{F}_n\right] \\ &= \|Z_n - m\|^2 - 2\gamma_n \langle Z_n - m, \Phi(Z_n)\rangle + \gamma_n^2 \left(\|\Phi(Z_n)\|^2 + \mathbb{E}\left[\|\xi_{n+1}\|^2 \Big| \mathcal{F}_n\right]\right) \\ &= V_n - 2\gamma_n \langle Z_n - m, \Phi(Z_n)\rangle + \gamma_n^2, \end{aligned} \tag{23}$$

where we used the definition of V_n and (19) for the last term. Since G is convex, using Corollary 2.2, we get:

$$\langle Z_n - m, \Phi(Z_n)\rangle = \langle Z_n - m, \Phi(Z_n) - \Phi(m)\rangle \geq 0.$$

Therefore, for all n, $\mathbb{E}\left[V_{n+1}|\mathcal{F}_n\right] \leq V_n + \gamma_n^2$. From the Robbins Siegmund theorem (see for instance (Duflo, 1997, page 18)), we deduce that (V_n) converges almost surely to V. Moreover, we note that $Z_n - m$ is bounded in L^2,

$$\forall n, \quad V_n = \mathbb{E}\left[\|Z_n - m\|^2\right] \leq \mathbb{E}\left[\|Z_0 - m\|^2\right] + \sum_{k=1}^{\infty} \gamma_k^2 < \infty, \tag{24}$$

□

whenever $\mathbb{E}\left[\|Z_0 - m\|^2\right] < \infty$, which is satisfied for example if $Z_0 = X_0 \mathbf{1}_{\{\|X_0\| \leq M\}}$, with $M < \infty$.

We can now give the proof the proposition.

Proof of proposition 3.1. Lemma 6.1 shows that the sequence V_n converges almost surely. Let us check now that its limit is zero. Let us take expectations in equation (23):

$$\begin{aligned}\mathbb{E}\left[V_{n+1}\right] &= \mathbb{E}\left[V_n\right] + \gamma_n^2 - 2\gamma_n \mathbb{E}\left[\langle \Phi(Z_n), Z_n - m\rangle\right] \\ &= \mathbb{E}\left[V_0\right] + \sum_{k=1}^{n} \gamma_k^2 - 2\sum_{k=1}^{n} \gamma_k \mathbb{E}\left[\langle \Phi(Z_k), Z_k - m\rangle\right]. \end{aligned} \tag{25}$$

The sequence $\sum_{k=1}^{n} \gamma_k \mathbb{E}\left[\langle \Phi(Z_k), Z_k - m\rangle\right]$ has positive terms, and is bounded above by $\mathbb{E}\left[V_0\right] + \sum_{k=1}^{\infty} \gamma_k^2$, therefore it converges. This implies in particular that

$$\sum_{n=1}^{\infty} \gamma_n \left\langle \Phi(Z_n), Z_n - m\right\rangle < +\infty \quad \text{a.s.} \tag{26}$$

This convergence cannot happen unless Z_n converges to m. Indeed, for each $\epsilon \in]0,1[$, let us introduce the set

$$\Omega_\epsilon = \left\{\omega \in \Omega : \ \exists n_\epsilon(\omega) \geq 1, \forall n \geq n_\epsilon(\omega), \quad \epsilon^2 < V_n(\omega) < \epsilon^{-2}\right\}.$$

For $\omega \in \Omega_\epsilon$, we have with Corollary 2.2,

$$\sum_{n\geq 1} \gamma_n \left\langle \Phi(Z_n(\omega)), Z_n(\omega) - m\right\rangle \geq \left(\sum_{n \geq n_\epsilon(\omega)} \gamma_n\right) \inf_{\epsilon < \|\alpha - m\| < \epsilon^{-1}} \left\langle \Phi(\alpha), \alpha - m\right\rangle = \infty,$$

which contradicts (26) unless $\mathbb{P}(\Omega_\epsilon) = 0$. Since V_n converges a.s. to a finite limit, and $\{\lim V_n \in [c, c^{-1}]\} \subset \Omega_{c/2}$, the only possible limit is zero:

$$\lim \|Z_n - m\| = 0, \quad \text{a.s.}$$ □

6.3 Proof of Lemma 3.3 and Proposition 3.2

Proof of Lemma 3.3. Consider, for $h \in \mathcal{B}(0, A)$, the function $f_h(t) = \Phi(m + th)$, defined for $t \in [0,1]$. We have $f_h(0) = \Phi(m) = 0$ and $f_h(1) = \Phi(m + h)$. It is also clear that the first order derivative $f_h'(t)$ of function f_h satisfies $f_h'(t) = \Gamma_{m+th}$. Consequently, a Taylor expansion with integral remainder of f_h about $t = 0$ gives us

$$\Phi(m + h) = \Phi(m) + \int_0^1 \Gamma_{m+th}(h)\, dt.$$

By Lemma 5.7 in Chaudhuri (1992), there is a constant M_A such that for all $t \in [0,1]$,

$$\|\Gamma_{m+th} - \Gamma_m\|_L \leq M_A \|h\|$$

where $\|.\|_L$ is the usual norm for bounded linear operators. Since $\Phi(m) = 0$, one gets

$$\|\Phi(m + h) - \Gamma_m(h)\| \leq \sup_{t\in[0,1]} \|\Gamma_{m+th} - \Gamma_m\|_L \|h\| \leq M_A \|h\|^2,$$

and this concludes the proof. □

Proof of Proposition 3.2. The proof is composed of 5 steps.

Step 1 — a spectral decomposition. Recall that Γ_m is:

$$\begin{aligned}\Gamma_m &= \mathbb{E}\left[\frac{1}{\|X-m\|}\left(\mathbf{I}_H - \frac{(X-m)\otimes(X-m)}{\|X-m\|^2}\right)\right] \\ &= \mathbb{E}\left[\|X-m\|^{-1}\right]\mathbf{I}_H - \mathbb{E}\left[\frac{1}{\|X-m\|}\left(\frac{(X-m)\otimes(X-m)}{\|X-m\|^2}\right)\right] \\ &= \mathbb{E}\left[\|X-m\|^{-1}\right]\mathbf{I}_H - T_m. \end{aligned} \tag{27}$$

Since Γ_m is bounded and symmetric, it is self-adjoint. Moreover, the operator T_m defined by (27) is trace class: it is self-adjoint, non negative, and if (e_j) is an orthonormal basis,

$$\begin{aligned}\sum_j \langle e_j, T_m e_j\rangle &= \sum_j \mathbb{E}\left[\frac{\langle X-m, e_j\rangle^2}{\|X-m\|^3}\right] \\ &\leq \mathbb{E}\left[\frac{1}{\|X-m\|}\right] < \infty.\end{aligned}$$

Therefore T_m is compact, and there is an increasing sequence of eigenvalues (λ_j), with possible repetitions, and an orthonormal basis (v_j) of eigenvectors in H such that:

$$\begin{aligned}\forall j \in \mathbb{N}, \quad \Gamma_m v_j &= \lambda_j v_j, \\ \lambda_j &\xrightarrow{j\to\infty} \mathbb{E}\left[\|X-m\|^{-1}\right], \\ \sigma(\Gamma_m) &= \{\lambda_j, j\in\mathbb{N}\} \cup \left\{\mathbb{E}\left[\|X-m\|^{-1}\right]\right\}.\end{aligned}$$

Moreover, thanks to (10), the smallest eigenvalue λ_{min} of Γ_m is strictly positive. For simplicity of notation, we rewrite this decomposition as follows,

$$\Gamma_m x = \sum_{\lambda\in\Lambda} \lambda \langle e_\lambda, x\rangle e_\lambda, \quad x \in H,$$

where Λ is the multiset $\{\lambda_j, j\in\mathbb{N}\}$, that can account for eigenspaces of dimension larger than 1.

In the following, we will need the operators:

$$\alpha_k = \mathbf{I}_H - \gamma_k \Gamma_m, \qquad \beta_n = \alpha_n \alpha_{n-1}\cdots\alpha_1. \tag{28}$$

Since Γ_m is bounded, these operators are well defined. Introducing the sequence of real functions, for $n\in\mathbb{N}$,

$$f_n(x) = \prod_{k=1}^n (1-\gamma_k x),$$

we see that $f_n(\cdot)$ and $f_n^{-1}(\cdot)$ are well defined on $\sigma(\Gamma_m)$, provided $\gamma_n \mathbb{E}\left[\|X-m\|^{-1}\right] < 1$, which we can assume without loss of generality. Elementary analysis shows that there exist constants c_1, C_2, C_3 such that:

$$\forall x \in \sigma(\Gamma_m), \quad c_1 \exp\left(-s_n x\right) \leq f_n(x) \leq C_2 \exp\left(-s_n x\right), \tag{29}$$

$$\left| s_n - \frac{c_\gamma}{1-\alpha} n^{1-\alpha}\right| \leq C_3, \tag{30}$$

where we recall that $s_n = \sum_{k=1}^n \gamma_k$, and $\gamma_k = c_\gamma k^{-\alpha}$. Then each operator β_n can be also expressed as follows:

$$\beta_n x = \sum_{\lambda\in\Lambda} f_n(\lambda) \langle e_\lambda, x\rangle e_\lambda, \quad x\in H, \tag{31}$$

their inverses are bounded operators, and satisfy: $\beta_n^{-1} x = \sum_{\lambda\in\Lambda} f_n^{-1}(\lambda) \langle e_\lambda, x\rangle e_\lambda$.

Step 2 — Decomposition of the algorithm. Let us rewrite the algorithm in the following way

$$\begin{aligned} Z_{n+1} &= Z_n + \gamma_n \xi_{n+1} - \gamma_n \Phi(Z_n) \\ &= Z_n + \gamma_n \xi_{n+1} - \gamma_n (\Gamma_m (Z_n - m) + \delta_n) \end{aligned}$$

where $\delta_n = \Phi(Z_n) - \Gamma_m(Z_n - m)$ is the difference between the gradient of G and the gradient of its quadratic approximation. Therefore:

$$\forall k, \quad Z_{k+1} - m = \alpha_k (Z_k - m) + \gamma_k \xi_{k+1} - \gamma_k \delta_k \tag{32}$$

Rewriting $\alpha_{n-1}\alpha_{n-2}\cdots\alpha_{k+1}$ as $\beta_{n-1}\beta_k^{-1}$, we get by induction,

$$Z_n - m = \beta_{n-1}(Z_1 - m) + \beta_{n-1} M_n - \beta_{n-1} R_{n-1}, \tag{33}$$

where

$$R_n = \sum_{k=1}^{n-1} \gamma_k \beta_k^{-1} \delta_k \qquad\qquad M_n = \sum_{k=1}^{n-1} \gamma_k \beta_k^{-1} \xi_{k+1}.$$

The first two terms of (33) are what we would get if G was exactly quadratic: a deterministic gradient part going to m, and a noise part; R_n is the error term. We will look at each of these terms in turn.

Step 3 — The deterministic term. We want to bound $\beta_{n-1}(Z_1 - m)$. The asymptotic behaviour of f_n in eq. (29) implies that

$$\|\beta_{n-1}\| \leq C_2 \exp\left(-s_n \lambda_{min}\right),$$

where $\lambda_{min} > 0$ is the smallest eigenvalue of Γ_m. Therefore

$$\mathbb{E}\left[\|\beta_{n-1}(Z_1 - m)\|^2\right] \leq C \exp\left(-2n^{1-\alpha}\right) \mathbb{E}\left[\|Z_1 - m\|^2\right]. \tag{34}$$

Step 4 — The martingale. The fact that the β_k are operators (instead of real numbers) makes matters more complicated. To deal with this problem, we use the spectral decomposition of the sequence of self-adjoint operators (β_k).

More precisely, we decompose $M_n = \sum_{\lambda\in\Lambda} \langle e_\lambda, M_n\rangle e_\lambda = \sum_\lambda M_n^\lambda e_\lambda$. For each $\lambda \in \Lambda$, M_n^λ is a martingale, and

$$\mathbb{E}[(M_n^\lambda)^2] = \sum_{k\leq n-1} \gamma_k^2 f_k^{-2}(\lambda) \mathbb{E}\left[\langle \xi_{k+1}, e_\lambda\rangle^2 \middle| \mathcal{F}_k\right],$$

since $\mathbb{E}\left[\langle \xi_{k'}, e_\lambda\rangle \langle \xi_{k+1}, e_\lambda\rangle | \mathcal{F}_k\right] = 0$ when $k' < k+1$. Summing now over $\lambda \in \Lambda$, we get:

$$\begin{aligned} \mathbb{E}\left[\|\beta_{n-1} M_n\|^2\right] &= \sum_\lambda f_{n-1}^2(\lambda) \mathbb{E}\left[(M_n^\lambda)^2\right], \\ &\leq \sum_\lambda \sum_{k\leq n-1} \gamma_k^2 \left(\frac{f_{n-1}(\lambda)}{f_k(\lambda)}\right)^2 \mathbb{E}\left[\langle \xi_{k+1}, e_\lambda\rangle^2\right]. \end{aligned} \tag{35}$$

However, for any k, n, and any $\lambda \in \Lambda$,

$$\frac{f_{n-1}(\lambda)}{f_k(\lambda)} \leq \prod_{j=k+1}^{n-1} (1 - \lambda\gamma_j) \leq \frac{f_{n-1}(\lambda_{min})}{f_k(\lambda_{min})}.$$

This uniformity in λ allows us to reconstruct $\mathbb{E}\left[\|\xi_{k+1}\|^2\right]$, which is bounded by 1, thanks to (19). We obtain:

$$\begin{aligned}\mathbb{E}\left[\|\beta_{n-1}M_n\|^2\right] &\leq \sum_{k\leq n-1}\gamma_k^2\left(\frac{f_{n-1}(\lambda_{min})}{f_k(\lambda_{min})}\right)^2\sum_{\lambda}\mathbb{E}\left[\langle\xi_{k+1},e_\lambda\rangle^2\right],\\ &\leq \sum_{k\leq n-1}\gamma_k^2\left(\frac{f_{n-1}(\lambda_{min})}{f_k(\lambda_{min})}\right)^2\mathbb{E}\left[\|\xi_{k+1}\|^2\right],\\ &\leq \sum_{k\leq n-1}\gamma_k^2\left(\frac{f_{n-1}(\lambda_{min})}{f_k(\lambda_{min})}\right)^2.\end{aligned}$$

Now we use the bounds (29) on β_n:

$$\begin{aligned}\mathbb{E}\left[\|\beta_{n-1}M_n\|^2\right] &\leq \frac{C_2^2}{c_1^2}\sum_{k\leq n-1}\gamma_k^2\exp\left(-\sum_{j=k+1}^{n}\gamma_j\right)\\ &\leq C\sum_{k\leq n-1}\gamma_k^2\exp\left(-\frac{1}{1-\alpha}\left(n^{1-\alpha}-k^{1-\alpha}\right)\right).\end{aligned}\tag{36}$$

The exponential terms are very small when k is much smaller than n, therefore we isolate the last terms. To do that, we choose $l(n)$ such that,

$$l(n)^{1-\alpha} = n^{1-\alpha} - c_\alpha\ln(n)\,,\tag{37}$$

with c_α to be chosen later. The first part of the sum (36) (for $k \leq l(n)$) gives us:

$$\begin{aligned}\sum_{k\leq l(n)}\gamma_k^2\exp\left(-\frac{1}{1-\alpha}\left(n^{1-\alpha}-k^{1-\alpha}\right)\right) &\leq \sum_{k\leq l(n)}\gamma_k^2\exp\left(-\frac{c_\alpha}{1-\alpha}\ln(n)\right)\\ &\leq c_\gamma^2\sum_{k\leq n}k^{-2\alpha-\frac{c_\alpha}{1-\alpha}}.\end{aligned}\tag{38}$$

This can be made smaller than any prescribed inverse power of n, if we choose c_α large enough. In the second part of the sum (36), for $k > l(n)$, we bound the exponential by 1 and γ_k by $\gamma_{l(n)}$:

$$\sum_{k>l(n)}\gamma_k^2\exp\left(-\frac{1}{1-\alpha}\left(n^{1-\alpha}-k^{1-\alpha}\right)\right) \leq (n-l(n))\gamma_{l(n)}^2.$$

The number of terms $n - l(n)$ is equivalent to $\frac{c_\alpha}{1-\alpha}\ln(n)n^\alpha$, and $\gamma_{l(n)} \sim c_\gamma n^{-\alpha}$. Therefore, the whole second term is equivalent to $c\ln(n)n^{-\alpha}$, where c depends on c_α and c_γ. For c_α large enough, this dominates the first term (38). Finally we get:

$$\mathbb{E}\left[\|\beta_{n-1}M_n\|^2\right] \;\leq\; C\frac{\ln(n)}{n^\alpha}.\tag{39}$$

Step 5 — the error term and the conclusion. The error term is $\beta_{n-1}\sum_{k=1}^{n}\gamma_k\beta_k^{-1}\delta_k$, where $\delta_k = \Phi(Z_k) - \Gamma_m(Z_k - m)$. With Lemma 3.3, we get that

$$\exists r, C_r \quad \forall k, \|Z_k - m\| \leq r \implies \|\delta_k\| \leq C_r\|Z_k - m\|^2\,.\tag{40}$$

Since Z_n converges a.s. to m, we deduce two things about δ_k: it is almost surely bounded, and (40) becomes a.s. eventually true. To use these facts we introduce the following sequence of events:

$$\Omega_N = \left\{ \omega, \begin{array}{l} \forall n \geq N, \forall k \geq n - l(n), \quad \|Z_k - m\| \leq 1/K \text{ and } \|\delta_k\| \leq C_r \|Z_k - m\|^2 \\ \forall k, \|\delta_k\| \leq N. \end{array} \right\}.$$

for a value of K to be chosen later, and $l(n)$ defined by (37). This sequence is increasing and $\bigcup \Omega_N = \Omega$; from now on we work on Ω_N.

Once more, since $\beta_{n-1}\beta_k^{-1}$ is very small when k is much smaller than n, only the last terms in the sum defining R_n matter. This is why we re-use the definition of $l(n)$ and cut the sum in two parts. For $\omega \in \Omega_N$, and $n \geq N$,

$$\begin{aligned} \|\beta_{n-1}R_n\|^2 &\leq \left(\sum_{k=1}^{n} \gamma_k \left\| \beta_{n-1}\beta_k^{-1} \right\| \|\delta_k\| \right)^2 \\ &\leq 2N^2 \left(\sum_{k=1}^{l(n)} \gamma_k \left\| \beta_{n-1}\beta_k^{-1} \right\| \right)^2 + 2C_r^2 \left(\sum_{k=l(n)+1}^{n} \gamma_k \|Z_k - m\|^2 \right)^2 \\ &\leq 2N^2 \left(\sum_{k=1}^{l(n)} \gamma_k \left\| \beta_{n-1}\beta_k^{-1} \right\| \right)^2 + 2\frac{C_r^2}{K^2}(n - l(n))\gamma_{l(n)} \sum_{k=l(n)+1}^{n} \gamma_k \|Z_k - m\|^2 . \end{aligned}$$

where we used the crude bound $\|\delta_k\| \leq N$ in the first part, and for the second part, $\left\| \beta_{n-1}\beta_k^{-1} \right\| \leq 1$ and the definition of Ω_N.

As before, it is easy to see that the first term is bounded by any prescribed inverse power of n, say n^{-42}. For the second term, we already know that $(n - l(n))\gamma_{l(n)}$ is bounded. Therefore, on Ω_N and for $n \geq N$,

$$\|\beta_{n-1}R_n\|^2 \leq \frac{CN^2}{n^{42}} + \frac{C}{K^2} \sum_{k=l(n)+1}^{n} \gamma_k \|Z_k - m\|^2 . \tag{41}$$

Combining now (33), (34), (39) and (41), we get, for $n \geq N$ and some new constant C

$$\begin{aligned} \mathbb{E}\left[\mathbf{1}_{\Omega_N} \|Z_n - m\|^2 \right] &\leq \frac{C\ln(n)}{n^\alpha} + \frac{C}{K^2} \sum_{k=l(n)+1}^{n} \gamma_k \mathbb{E}\left[\mathbf{1}_{\Omega_N} \|Z_k - m\|^2 \right] \\ &\leq \frac{C\ln(n)}{n^\alpha} + \frac{C'}{K^2} \sup_{l(n)<k\leq n} \mathbb{E}\left[\mathbf{1}_{\Omega_N} \|Z_k - m\|^2 \right]. \end{aligned}$$

Let us choose K such that $K^2 \geq 2C'$. Then

$$\forall n \geq N, \quad \mathbb{E}\left[\mathbf{1}_{\Omega_N} \|Z_n - m\|^2 \right] \leq \frac{C\ln(n)}{n^\alpha} + \frac{1}{2} \max_{l(n)<k\leq n} \mathbb{E}\left[\mathbf{1}_{\Omega_N} \|Z_k - m\|^2 \right].$$

The fact that $l(n)$ is close enough to n allows us to prove by induction that, for some constant C',

$$\forall n \geq N', \quad \mathbb{E}\left[\mathbf{1}_{\Omega_N} \|Z_n - m\|^2 \right] \leq \frac{C'\ln(n)}{n^\alpha}.$$

This concludes the proof of Proposition 3.2. □

6.4 Proof of Proposition 3.4

We use the same decomposition as in Pelletier (2000). It consists in linearizing the target function Φ around the true value m. Recall the following decomposition of the error (32),

$$\forall k, \quad Z_{k+1} - m = (\mathbf{I}_H - \gamma_k \Gamma_m)(Z_k - m) + \gamma_k \xi_{k+1} - \gamma_k \delta_k,$$

where ξ_k is a martingale difference sequence and δ_k are error terms, $\delta_k := \Phi(Z_k) - \Gamma_m(Z_k - m)$. Defining now,

$$T_n := Z_n - m, \quad \overline{T}_n := \overline{Z}_n - m \quad \text{and} \quad M_{n+1} := \sum_{k=1}^{n} \xi_{k+1},$$

and rearranging the previous expression, we obtain:

$$\Gamma_m T_k = \xi_{k+1} - \delta_k + \frac{1}{\gamma_k}\left(T_k - T_{k+1}\right).$$

Summing these equalities, it comes,

$$n\Gamma_m \overline{T}_n = \sum_{k=1}^{n} \frac{1}{\gamma_k}\left(T_k - T_{k+1}\right) - \sum_{k=1}^{n} \delta_k + M_{n+1}.$$

Applying Abel's transform, and dividing by $\sqrt{n}$ yields:

$$\sqrt{n}\Gamma_m \overline{T}_n = \frac{1}{\sqrt{n}}\left(\frac{T_1}{\gamma_1} - \frac{T_{n+1}}{\gamma_n} + \sum_{k=2}^{n} T_k\left[\frac{1}{\gamma_k} - \frac{1}{\gamma_{k-1}}\right] - \sum_{k=1}^{n} \delta_k\right) + \frac{1}{\sqrt{n}} M_{n+1}.$$

To prove that last term is a martingale for which the CLT holds,

$$\frac{M_n}{\sqrt{n}} \xrightarrow[\mathrm{n}\to\infty]{\mathcal{L}} \mathcal{N}\left(0, \Sigma\right),$$

we need to check that the assumptions of Theorem 5.1 in (Jakubowski, 1988) are fulfilled. We first have that the martingale difference sequence is *a.s.* bounded, $\forall n \left\|\xi_n\right\| \leq 2$. Let us define

$$\Sigma_n = \mathbb{E}\left[\xi_{n+1} \otimes \xi_{n+1} | \mathcal{F}_n\right], \tag{42}$$

which can also be decomposed as follows

$$\Sigma_n = \mathbb{E}\left[\frac{(X - Z_n)}{\|X - Z_n\|} \otimes \frac{(X - Z_n)}{\|X - Z_n\|} \middle| \mathcal{F}_n\right] - \Phi(Z_n) \otimes \Phi(Z_n). \tag{43}$$

Since $\Phi(m) = 0$, we have by a direct computation,

$$\left\|\Phi(Z_n)\right\| \leq \mathbb{E}\left[\frac{2}{\|X - m\|}\right] \left\|Z_n - m\right\|.$$

Using now, for $(a, b) \in H \times H$, the inequality $\left\|a \otimes b\right\|_L \leq \left\|a\right\| \left\|b\right\|$, where $\left\|a \otimes b\right\|_L$ is the usual the norm for linear operators, we directly get, with Proposition 3.1,

$$\left\|\Phi(Z_n) \otimes \Phi(Z_n)\right\|_L \to 0, \quad a.s.$$

With similar arguments, it is easy to show that

$$\begin{aligned}\left\|\Sigma - \mathbb{E}\left[\frac{(X - Z_n)}{\|X - Z_n\|} \otimes \frac{(X - Z_n)}{\|X - Z_n\|} \middle| \mathcal{F}_n\right]\right\|_L &\leq 2\mathbb{E}\left[\left\|\frac{(X - Z_n)}{\|X - Z_n\|} - \frac{(X - m)}{\|X - m\|}\right\| \middle| \mathcal{F}_n\right] \\ &\leq 4\mathbb{E}\left[\frac{1}{\|X - m\|}\right] \left\|Z_n - m\right\|,\end{aligned}$$

so that $\|\Sigma_n - \Sigma\|_L \to 0$ *a.s.*, when n tends to infinity. Then condition 5.2 in (Jakubowski, 1988) is satisfied and is a consequence of a direct application of Chow's Lemma, see for instance (Duflo, 1997, page 22).

Now, it remains to prove that

$$\frac{1}{\sqrt{n}}\left(\frac{T_{n+1}}{\gamma_n} - \sum_{k=2}^{n} T_k\left[\frac{1}{\gamma_k} - \frac{1}{\gamma_{k-1}}\right] + \sum_{k=1}^{n}\delta_k\right) \xrightarrow[\mathrm{n}\to\infty]{\mathrm{P}} 0. \tag{44}$$

Let us denote by A_n, B_n and C_n the three terms. We will use on each term the following lemma:

Lemma 6.2. *Let Ω_N be such that $\Omega_N \uparrow \Omega$, and let X_n be a sequence of random variables such that, for all N,*

$$\mathbb{E}\left[\mathbf{1}_{\Omega_N} \cdot \|X_n\|\right] \xrightarrow[n\to\infty]{} 0.$$

Then $X_n \xrightarrow[\mathrm{n}\to\infty]{\mathrm{P}} 0$.

Proof of lemma. The convergence in L^1 implies that, for all N, $\mathbf{1}_{\Omega_N} X_n$ converges in probability to zero. Then

$$\limsup_n \mathbb{P}\left[\|X_n\| > \epsilon\right] \leq \mathbb{P}\left[\Omega_N^c\right] + \limsup_n \mathbb{P}\left[\mathbf{1}_{\Omega_N}\|X_n\| > \epsilon\right] \leq 1 - \mathbb{P}\left[\Omega_N\right].$$

Since this holds for all N, the lemma is proved. □

Recall that $\mathbb{E}\left[\mathbf{1}_{\Omega_N}\|T_n\|^2\right] \leq C_N\frac{\ln(n)}{n^\alpha}$, thanks to Proposition 3.2.

For the first term $A_n = \frac{T_{n+1}}{\sqrt{n}\gamma_n}$, we have:

$$\mathbb{E}\left[\mathbf{1}_{\Omega_N}\|A_n\|^2\right] \leq C'_N n^{2\alpha-1}\frac{\ln(n)^2}{n^{2\alpha}} = \frac{C'_N\ln(n)^2}{n}$$

so $A_n \xrightarrow[\mathrm{n}\to\infty]{\mathrm{P}} 0$.

Let us turn to the second term B_n. Since $\gamma_k^{-1} - \gamma_{k-1}^{-1} \leq 2\alpha c_\gamma^{-1} k^{\alpha-1}$,

$$\begin{aligned}
\mathbb{E}\left[\|B_n\|\,\mathbf{1}_{\Omega_N}\right] &\leq \frac{2\alpha c_\gamma^{-1}}{\sqrt{n}}\sum_{k\leq n}\mathbb{E}\left[\mathbf{1}_{\Omega_N}\|T_k\|\right]k^{\alpha-1} \\
&\leq \frac{C_0}{\sqrt{n}}\sum_{k\leq n}\sqrt{\ln(k)}k^{\alpha/2-1} \\
&\leq \kappa\sqrt{\ln(n)}n^{\alpha/2-1/2},
\end{aligned}$$

which goes to zero since $\alpha < 1$ (C_0 and κ stand for two positive constants). Therefore $B_n \xrightarrow[\mathrm{n}\to\infty]{\mathrm{P}} 0$.

Finally, for the last term C_n, since there exists a positive constant C_1 such that $\|\delta_k\| \leq C_1\|Z_k - m\|^2$, we have:

$$\begin{aligned}
\mathbb{E}\left[\mathbf{1}_{\Omega_N}\|C_n\|\right] &\leq \frac{1}{\sqrt{n}}\sum_{k\leq n}\mathbb{E}\left[\mathbf{1}_{\Omega_N}\|T_k\|^2\right] \\
&\leq \frac{C_N}{\sqrt{n}}\sum_{k\leq n}\ln(k)k^{-\alpha}.
\end{aligned}$$

Since the right hand side term converges to zero (as can be seen e.g. by Kronecker's lemma, using the fact that $\alpha > 1/2$), $C_n \xrightarrow[n\to\infty]{P} 0$, therefore (44) holds, and proposition 3.4 is finally proved.

Acknowledgements. We thank the company Médiamétrie for allowing us to illustrate our methodologies with their data.

References

Benveniste, A., Métivier, M., and Priouret, P. (1990). *Adaptive Algorithms and Stochastic Approximations*, volume 22 of *Applications of Mathematics*. Springer-Verlag, New York.

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In Lechevallier, Y. and Saporta, G., editors, *Compstat 2010*, pages 177–186. Physica Verlag, Springer.

Cadre, B. (2001). Convergent estimators for the L_1-median of a Banach valued random variable. *Statistics*, 35(4):509–521.

Cardot, H., Cénac, P., and Chaouch, M. (2010a). Stochastic approximation to the multivariate and the functional median. In Lechevallier, Y. and Saporta, G., editors, *Compstat 2010*, pages 421–428. Physica Verlag, Springer.

Cardot, H., Cénac, P., and Monnez, J.-M. (2010b). Fast clustering of large datasets with sequential k-medians : a stochastic gradient approach. Technical report, Institut de Mathématiques de Bourgogne.

Chaouch, M. and Goga, C. (2010). Design-based estimation for geometric quantiles with application to outliers detection. *Computational Statistics and Data Analysis*, 54:2214–2229.

Chaudhuri, P. (1992). Multivariate location estimation using extension of R-estimates through U-statistics type approach. *Ann. Statist.*, 20:897–916.

Chaudhuri, P. (1996). On a geometric notion of quantiles for multivariate data. *J. Amer. Statist. Assoc.*, 91(434):862–872.

Cuevas, A., Febrero, M., and Fraiman, R. (2007). Robust estimation and classification for functional data via projection-based depth notions. *Computational Statistics*, 22:481–496.

Dippon, J. and Walk, H. (2006). The averaged Robbins-Monro method for linear problems in a Banach space. *J. Theoret. Probab.*, 19(1):166–189.

Duflo, M. (1997). *Random iterative models*, volume 34 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin. Translated from the 1990 French original by Stephen S. Wilson and revised by the author.

Fraiman, R. and Muniz, G. (2001). Trimmed means for functional data. *TEST*, 10:419–440.

Gervini, D. (2008). Robust functional estimation using the median and spherical principal components. *Biometrika*, 95(3):587–600.

Gower, J. C. (1974). Algorithm as 78: The mediancentre. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 23(3):466–470.

Haberman, J. (1989). Concavity and estimation. *Ann. Statist.*, 17:1631–1661.

Haldane, J. B. S. (1948). Note on the median of a multivariate distribution. *Biometrika*, 35(3-4):414–417.

Huber, P. and Ronchetti, E. (2009). *Robust Statistics*. John Wiley and Sons, second edition.

Jakubowski, A. (1988). Tightness criteria for random measures with application to the principle of conditioning in Hilbert spaces. *Probab. Math. Statist.*, 9(1):95–114.

Kemperman, J. H. B. (1987). The median of a finite measure on a Banach space. In *Statistical data analysis based on the L_1-norm and related methods (Neuchâtel, 1987)*, pages 217–230. North-Holland, Amsterdam.

Koltchinskii, V. I. (1997). *M*-estimation, convexity and quantiles. *Ann. Statist.*, 25(2):435–477.

Kushner, H. J. and Clark, D. S. (1978). *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. Springer-Verlag, Berlin.

Kushner, H. J. and Yin, G. G. (2003). *Stochastic approximation and recursive algorithms and applications*, volume 35 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, second edition. Stochastic Modelling and Applied Probability.

Ljung, L., Pflug, G., and Walk, H. (1992). *Stochastic Approximation and Optimization of Random Systems*. Birkhäuser, Boston.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability (Berkeley, Calif., 1965/66)*, pages Vol. I: Statistics, pp. 281–297. Univ. California Press, Berkeley, Calif.

Nazarov, A. (2009). Exact l_2-small ball asymptotics of gaussian processes and the spectrum of boundary-value problems. *J. Theoret. Probab.*, 22:640–665.

Pelletier, M. (2000). Asymptotic almost sure efficiency of averaged stochastic algorithms. *SIAM J. Control Optim.*, 39(1):49–72 (electronic).

Polyak, B. and Juditsky, A. (1992). Acceleration of stochastic approximation. *SIAM J. Control and Optimization*, 30:838–855.

Ruppert, D. (1985). A Newton-Raphson version of the multivariate Robbins-Monro procedure. *Ann. Statist.*, 13(1):236–245.

Smale, S. and Yao, Y. (2006). Online learning algorithms. *Found. Comput. Math.*, 6(2):145–170.

Small, C. G. (1990). A survey of multidimensional medians. *International Statistical Review / Revue Internationale de Statistique*, 58(3):263–277.

Vardi, Y. and Zhang, C.-H. (2000). The multivariate L_1-median and associated data depth. *Proc. Natl. Acad. Sci. USA*, 97(4):1423–1426 (electronic).

Walk, H. (1977). An invariance principle for the Robbins-Monro process in a Hilbert space. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, 39(2):135–150.