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

HERVÉ CARDOT*, PEGGY CÉNAC** and PIERRE-ANDRÉ ZITT[†]

*Institut de Mathématiques de Bourgogne, Université de Bourgogne, 9 Avenue Alain Savary, 21078 Dijon, France. E-mail: *Herve.Cardot@u-bourgogne.fr; **Peggy.Cenac@u-bourgogne.fr; [†]Pierre-Andre.Zitt@u-bourgogne.fr*

With the progress of measurement apparatus and the development of automatic sensors, it is not unusual anymore to get large 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 in metric spaces and has nice robustness properties. It is possible to estimate the geometric median, being defined as the minimizer of a simple convex functional that is differentiable everywhere when the distribution has no atom, 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 than 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

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 large samples of functional observations. For example, Cardot *et al.* [5] analyze a sample of more than 18 000 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 $d = 86\,400$, 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 [14]). Detecting these atypical curves automatically is not straightforward 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 [33]), 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 [13] 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.* [10] have proposed random projection techniques in the context of maximal depth estimators and studied their properties via simulation studies. Note that subsampling approaches based on survey sampling with unequal probability sampling designs have also been proposed in the literature in order to reduce the computational time (Chaouch and Goga [7]).

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 [17], and whose properties have been studied in detail by Kemperman [20]. It can be defined even if the random variable does not have a finite first order moment, and it has nice robustness properties since its breakdown point is equal to 0.5. As noted in Small [33], 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 [20], Cadre [4] and Gervini [14]. Iterative estimation algorithms have been developed by Gower [15], Vardi and Zhang [34], in the multivariate setting, and by Gervini [14], for functional data. This latter algorithm requires one to invert at each step matrices whose dimension is equal to the dimension d of the data and thus requires important computational efforts. The algorithm proposed by Vardi and Zhang [34] 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, defined as the minimizer of a simple functional that is differentiable everywhere when the distribution has no atom, 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 finite dimensional situations (see Kushner and Clark [22], Ruppert [31], Benveniste *et al.* [2], Ljung *et al.* [25], Duflo [12], Kushner and Yin [23], Bottou [3], in the multivariate case and Arnaudon *et al.* [1], on manifolds). 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 [35], Dippon and Walk [11], Smale and Yao [32]).

It is also 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 [29], Pelletier [28]). 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 [11]). Nevertheless, it has been noted in an empirical study whose aim was to estimate the geometric median with functional data (Cardot *et al.* [5]) 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 [9]. 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 [34]. 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 such as $\mathbb{R}^d$ or $L^2(I)$, for some closed interval $I \subset \mathbb{R}$. We denote by $\langle \cdot, \cdot \rangle$ 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 [20])

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

When $H = \mathbb{R}$, the definition (1) of the median is equivalent to usual definition of the median based on the c.d.f., provided the median is unique. In higher-dimension spaces, definition (1) is a natural extension of the median. Note that it 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

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

(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 $\{\alpha \in H, \|\alpha\| \le A\}$ and Y is a random variable with law μ_c,

$$\forall A, \exists C_A \in [0, \infty), \forall \alpha \in \mathcal{B}(0, A), \qquad \mathbb{E}[\|Y - \alpha\|^{-1}] \le C_A.$$

- μ_d is a discrete measure, $\mu_d = \sum_i p_i \delta_{\alpha_i}$. We denote by D the support of μ_d and assume that $m \notin D$.

As shown in Kemperman [20], 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. As noted in Chaudhuri [8], the conditions on μ_c are satisfied when $H = \mathbb{R}^d$, with $d \geq 2$, whenever μ_c has a bounded density on every compact subset of $\mathbb{R}^d$. More precisely, this property is closely related to small ball probabilities, since

$$\mathbb{E}[\|Y - \alpha\|^{-1}] = \int_0^\infty \mathbb{P}[\|Y - \alpha\| \leq t^{-1}]\,dt.$$

If $\mathbb{P}[\|Y - \alpha\| \leq \varepsilon] \leq C\varepsilon^d$, for some small ε and some positive constant C, it is easy to check that

$$\mathbb{E}[\|Y - \alpha\|^{-\beta}] < \infty,$$

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 [27] 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, C_4 which depend on the process under study,

$$\mathbb{P}[\|Y - m\| \leq \varepsilon] \leq C_1 \varepsilon^{C_4} \exp(-C_2 \varepsilon^{-C_3}), \tag{2}$$

so that $\mathbb{E}[\|Y - m\|^{-\beta}] < \infty$, for all positive β. Similar properties of shifted small balls, for α close to m, can be found in Li and Shao [24].

2.2. Some convexity and robustness properties of the median

In this section, we derive quantitative convexity bounds which will be useful in the proofs. As a consequence, we are also able to bound for the gross sensitivity error, which is a classical robustness indicator (see Huber and Ronchetti [18]).

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

$$G(\alpha) := \mathbb{E}[\|X - \alpha\| - \|X\|]. \tag{3}$$

This function is convex since it is a convex combination of convex functions. However, convexity is not sufficient to get the convergence of the algorithm. 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 the discrete part $G_d(\alpha) = \sum_i p_i(\|x_i - \alpha\| - \|x_i\|)$ has been isolated. The first part is Fréchet differentiable everywhere (Kemperman [20]), 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,

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

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 the vocabulary of convex analysis, we have just chosen a particular subgradient of G on the set D of non-differentiability. It is easily seen that

$$\forall x, y, \qquad G(y) - G(x) \geq \langle \Phi(x), y - x \rangle, \tag{4}$$

which asserts that Φ is a subgradient. A short proof of inequality (4) is given in Section 6.1.

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

$$\Phi(\alpha) = 0. \tag{5}$$

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 [21] for the multivariate case and Gervini [14] for the functional one) that

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

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. The operator Γ_α is not compact, but it is bounded when $\mathbb{E}[\|X - \alpha\|^{-1}] < \infty$.

If we define $\bar{h} = h / \|h\|$, and P_h the projection onto the orthogonal complement of h,

$$\begin{aligned}
\langle h, \Gamma_\alpha h \rangle &= \|h\|^2 \mathbb{E}\left[\frac{1}{\|\alpha - X\|}\left(1 - \frac{\langle \bar{h}, \alpha - X \rangle^2}{\|\alpha - X\|^2}\right)\right] \\
&= \|h\|^2 \mathbb{E}\left[\frac{1}{\|\alpha - X\|} \frac{\|P_{\bar{h}}(\alpha - X)\|^2}{\|\alpha - X\|^2}\right].
\end{aligned} \tag{6}$$

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 [21].

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*

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

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 exist two strictly positive constants $0 < c_m \leq C_m \leq \mathbb{E}[\|X - m\|^{-1}] < \infty$, such that

$$c_m \|h\|^2 \leq \langle h, \Gamma_m h \rangle \leq C_m \|h\|^2. \tag{7}$$

As noted in Kemperman [20], the geometric median has a 50% breakdown point. Furthermore, an immediate consequence of (7) is that operator Γ_m has a bounded inverse. Thus, the gross error sensitivity, which is also a classical indicator of robustness, is bounded for the median in a separable Hilbert space. Indeed, thanks to the expression derived in Gervini [14], it is bounded as follows:

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

2.3. The algorithms

Given $X_1, X_2, \ldots, X_n$, n independent copies of X, a natural estimator of m is the solution $\widehat{m}_n$ of the empirical version of (5),

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

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 [31]; Duflo [12]) and is defined as follows:

$$\begin{aligned} Z_{n+1} &= Z_n + \gamma_n \frac{X_{n+1} - Z_n}{\|X_{n+1} - Z_n\|} \\ &= Z_n - \gamma_n U_{n+1}, \end{aligned} \tag{8}$$

with a starting point that can be random and bounded, for example, $Z_0 = X_0 \mathbf{1}_{\{\|X_0\| \leq M\}}$ 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. 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

$$\mathbb{E}[U_{n+1}|\mathcal{F}_n] = \Phi(Z_n). \tag{9}$$

Note that our particular choice of subgradient Φ ensures that this equality always holds.

Defining now by ξ the sequence of "errors" in the following estimates:

$$\xi_{n+1} = \Phi(Z_n) - U_{n+1}; \tag{10}$$

algorithm (8) can also be seen as a nonlinear Robbins–Monro algorithm,

$$Z_{n+1} = Z_n + \gamma_n\big(-\Phi(Z_n) + \xi_{n+1}\big). \tag{11}$$

Thanks to (9) and (10), the sequence (ξ_n) is a sequence of martingale differences. Let us note that the bracket of the associated martingale satisfies

$$\begin{aligned}
\mathbb{E}[\|\xi_{n+1}\|^2 | \mathcal{F}_n] &= \mathbb{E}[\|U_{n+1}\|^2 | \mathcal{F}_n] + \|\Phi(Z_n)\|^2 - 2\langle \Phi(Z_n), \mathbb{E}[U_{n+1}|\mathcal{F}_n]\rangle \\
&= \mathbb{E}[\|U_{n+1}\|^2 | \mathcal{F}_n] - \|\Phi(Z_n)\|^2 \\
&\le 1 - \|\Phi(Z_n)\|^2 \le 1. \tag{12}
\end{aligned}$$

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

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

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 spaces has been proposed by Chaudhuri [9]. In such spaces, quantiles are associated to a direction and a magnitude specified by a vector $v \in H$, such that $\|v\| < 1$. The geometric quantile of X, say m^v, corresponding to direction v and magnitude $\|v\|$ is defined, uniquely under previous assumptions, by

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

If $v = 0$, one recovers the geometric median. When $\|v\|$ is close to one, m^v is a (directed) extreme quantile. In any case, m^v is characterized by

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

so that it can be naturally estimated with the following stochastic algorithm:

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

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.

Theorem 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

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

3.2. Rates of convergence and asymptotic normality

We present now the rates of convergence of the stochastic gradient algorithm as well as the asymptotic distribution of its averaged version. The proofs are given in Section 6. More specific sequences (γ_n) are considered, and we suppose from now on that $\gamma_n = c_\gamma n^{-\alpha}$, where c_γ is a positive constant and $\alpha \in (\frac{1}{2}, 1)$. We need one additional assumption to get these rates of convergence:

(A3) There is a positive constant A such that

$$\exists C_A \in [0, \infty), \forall h \in \mathcal{B}(0, A), \qquad \mathbb{E}[\|X - (m + h)\|^{-2}] \leq C_A. \tag{13}$$

This assumption is not restrictive when the dimension is strictly larger than two, as discussed in Section 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 [12], Theorem 2.2.12, for the multivariate case) up to a logarithmic factor.

Proposition 3.2. *Assume* (A1), (A2) *and* (A3). *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*

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

Remark 3. An immediate consequence of Proposition 3.2 is that

$$\|Z_n - m\|^2 = O_P\left(\frac{\ln(n)}{n^\alpha}\right).$$

Assumption (A3) 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. *Assume* (A3) *is in force. Then,*

$$\forall h \in \mathcal{B}(0, A), \qquad \Phi(m + h) = \Gamma_m(h) + \mathrm{O}(\|h\|^2).$$

Finally, Theorem 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 [16], 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 (Gervini [14], Theorem 6).

Theorem 3.4. *Assume* (A1), (A2) *and* (A3). *Then*

$$\sqrt{n}(\overline{Z}_n - m) \underset{n \to \infty}{\overset{\mathcal{L}}{\to}} \mathcal{N}(0, \Gamma_m^{-1} \Sigma \Gamma_m^{-1}),$$

with

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

Note that with (7), operator Γ_m^{-1} is well defined; it is bounded and positive.

4. Illustrations on simulated and real data

4.1. A simulation study

A simple simulation study is performed to check the good behavior of the averaged stochastic estimator and to make a comparison with the static estimator developed by Vardi and Zhang [34]. Two points of view are considered. The first classic one consists of 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 of comparing the accuracy of both approaches when the allocated computation time is fixed in advance. We use ℝ (R Development Core Team [30]) and the function `spatial.median` from the library `ICSNP` to estimate the median with the algorithm developed by Vardi and Zhang [34].

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 accuracy of the different algorithms, we compute the following estimation error:

$$R(\widehat{m}) = \|\widehat{m} - m\|, \tag{14}$$

where $\widehat{m}$ is an estimator of m.

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 is usually done for such procedures, we fix $\alpha = 3/4$ and focus on the choice of c_γ. We also 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

$$\alpha \mapsto \frac{1}{n} \sum_{i=1}^{n} (\|X_i - \alpha\| - \|X_i\|),$$

which is the empirical version of (3).

4.1.1. *Fixed sample sizes*

We perform 1000 simulations for different sample sizes, $n = 250$, $n = 500$ and $n = 2000$. Table 1 presents basic statistics for the estimation errors (first quartile Q_1, median and third quartile Q_3), according to criterion (14), for the algorithm by Vardi and Zhang [34] and our averaged procedure considering different values for $c_\gamma \in \{0.2, 0.6, 1, 2, 5, 10, 15, 25, 50, 75\}$.

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 gives slightly better estimations. We can also remark that the averaged stochastic estimator is not very sensitive to the tuning parameter c_γ which can take values in the interval $[2, 75]$ without modifying the performances

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

Estimator	$n = 250$			$n = 500$			$n = 2000$		
	$[Q_1$	median	$Q_3]$	$[Q_1$	median	$Q_3]$	$[Q_1$	median	$Q_3]$
$c_\gamma = 0.2$	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 and Zhang	0.12	0.18	0.25	0.09	0.12	0.17	0.04	0.06	0.08

of the estimator. As a matter of fact, we noted on simulations that interesting values for c_γ are around or above $\mathbb{E}[\|X - m\|]$, 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) (for n observations in dimension d), 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 the difference is even more impressive, as we will see in the next section.

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 [34] can deal with $n = 150$ observations, whereas our recursive algorithm, coded in the ℝ language, can take into account $n = 4500$ observations so that it will give 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 $n = 90\,000$ 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 used to understand the consumers' habits regarding television viewing. The French company Médiamétrie provides official television audience rates in France. Médiamétrie works with a panel of about 9000 individuals, and the television sets of these individuals are equipped with sensors that measure the audience of the different channels at a second scale.

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 sequentially to Médiamétrie during the night. Survey sampling techniques with unequal probability sampling designs are used by Médiamétrie to select the sample, and thus the i.i.d. assumption is clearly not satisfied. Nevertheless, our aim is just to give an illustration of the ability of our averaged stochastic algorithm to deal with a large sample of very high-dimensional 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 on the 6th of September, 2010. After removing from the sample people that did not watch 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 $X_i \in \{0, 1\}^{86\,400}$, where $86\,400$ 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.
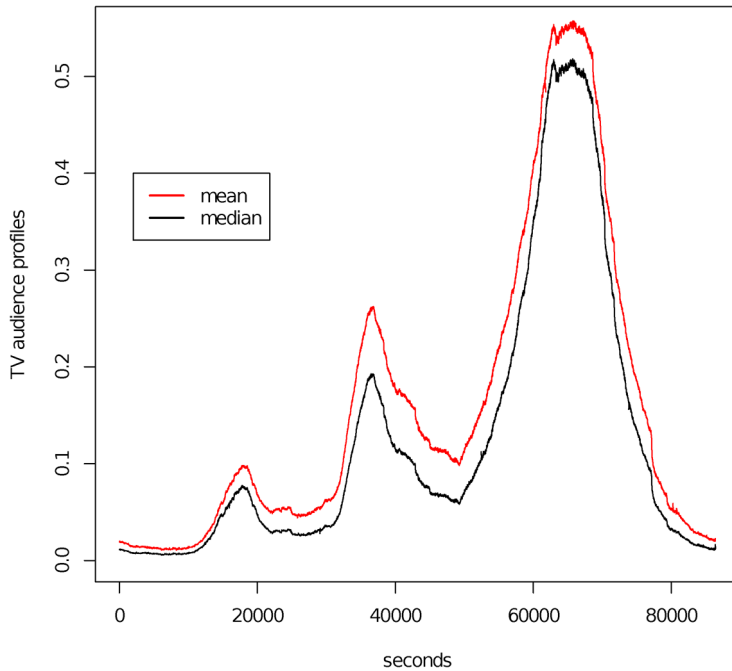
Figure 1. Estimations of the mean and the geometric median audiences, at a second scale, on the 6th september 2010.

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, even if the database is huge, it takes less than one minute for our algorithm to converge whereas we were not able to perform the estimation with the static estimator developed by Vardi and Zhang [34] because of memory constraints. 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 in the simulation experiment that they are not very sensitive to the value of the tuning parameter c_γ.

One could imagine many directions for future research that certainly deserve further attention. Taking advantage of the rapidity of our estimation procedure, one could use resampling techniques, similar to the bootstrap, in order to approximate the asymptotic distribution of the estimator given in Theorem 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 [26]) 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.* [6]) 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 individuals 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.

Finally, as noted in Section 4.2 the independence condition is rarely satisfied for real studies. A direction that deserves further investigation is to determine under which dependence conditions our results, such as Theorems 3.1 and 3.4, remain true.

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 $x \notin D$, this is clear since G is Fréchet differentiable.

Pick a point x_0 in D and recall that $\Phi_d(x_0) = \sum_{i,x_i \neq x_0} p_i \frac{x_0 - x_i}{\|x_0 - x_i\|}$. We have

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

so that Φ_d is a subgradient of G_d.

The upper bound in Proposition 2.1 follows immediately from (6) and the assumption (A2).

For the lower bound, thanks to (6), we only need to prove

$$
\forall \alpha \in \mathcal{B}(0, A), \forall u, \|u\| = 1, \qquad \langle u, \Gamma_\alpha u \rangle = \mathbb{E}\left[\frac{\|P_u(X - \alpha)\|^2}{\|X - \alpha\|^3} \right] \geq c_A, \tag{15}
$$

where P_u is the projection on the orthogonal of u. This quantity is small when $X - \alpha$ is in 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 (a 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 $y \in H$, $\|P_u y\|^2 \geq \langle y, v \rangle^2$. In particular, $\|P_u(X - \alpha)\|^2 \geq \langle X - \alpha, v \rangle^2$.

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

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

It is easily seen that the last term is bounded below by $\mathbf{Var}(\langle v, X \rangle) \geq c$, and (15) 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 Proposition 2.1 and of 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 $\Phi(\alpha_2) - \Phi(\alpha_1) = \int_0^1 f'(t)\,dt$ where $f(t) = \Phi(\alpha_t)$. One can then apply (15), t by t, with $\alpha = \alpha_t$ and $u = \frac{\alpha_2 - \alpha_1}{\|\alpha_2 - \alpha_1\|}$.

Moreover,

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

where the second line follows from (4). Since $\Phi = \lambda \Phi_c + (1 - \lambda)\Phi_d$, Corollary 2.2 is proved.

6.2. Proof of Theorem 3.1

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

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

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

Proof. 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. (11)). 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 becomes

$$\begin{aligned} \mathbb{E}[V_{n+1}|\mathcal{F}_n] &= \|Z_n - m - \gamma_n \Phi(Z_n)\|^2 + \gamma_n^2 \mathbb{E}[\|\xi_{n+1}\|^2|\mathcal{F}_n] \\ &= \|Z_n - m\|^2 - 2\gamma_n \langle Z_n - m, \Phi(Z_n) \rangle + \gamma_n^2 \big(\|\Phi(Z_n)\|^2 + \mathbb{E}[\|\xi_{n+1}\|^2|\mathcal{F}_n] \big) \quad (16) \\ &= V_n - 2\gamma_n \langle Z_n - m, \Phi(Z_n) \rangle + \gamma_n^2, \end{aligned}$$

where we used the definition of V_n and (12) 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, $\mathbb{E}[V_{n+1}|\mathcal{F}_n] \leq V_n + \gamma_n^2$. From the Robbins Siegmund theorem (see, for instance, Duflo [12], page 18), we deduce that (V_n) converges almost surely to V. Moreover, we note that $Z_n - m$ is bounded in L^2,

$$\forall n, \qquad V_n = \mathbb{E}[\|Z_n - m\|^2] \leq \mathbb{E}[\|Z_0 - m\|^2] + \sum_{k=1}^{\infty} \gamma_k^2 < \infty, \qquad (17)$$

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

We can now give the proof of the theorem.

Proof of Theorem 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 (16).

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

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

$$\sum_{n=1}^{\infty} \gamma_n \langle \Phi(Z_n), Z_n - m \rangle < +\infty \qquad \text{a.s.} \tag{18}$$

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

$$\Omega_\varepsilon = \{\omega \in \Omega \colon \exists n_\varepsilon(\omega) \geq 1, \forall n \geq n_\varepsilon(\omega), \varepsilon^2 < V_n(\omega) < \varepsilon^{-2}\}.$$

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

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

which contradicts (18) unless $\mathbb{P}(\Omega_\varepsilon) = 0$. 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.

$$\lim \|Z_n - m\| = 0 \qquad \text{a.s.} \qquad \square$$

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 [8], 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 $\|\cdot\|_L$ 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. $\square$

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

Step 1 – A spectral decomposition. Recall that Γ_m is

$$
\begin{aligned}
\Gamma_m &= \mathbb{E}\left[\frac{1}{\|X-m\|}\left(\mathbf{I}_H - \frac{(X-m)\otimes(X-m)}{\|X-m\|^2}\right)\right] \\
&= \mathbb{E}[\|X-m\|^{-1}]\mathbf{I}_H - \mathbb{E}\left[\frac{1}{\|X-m\|}\left(\frac{(X-m)\otimes(X-m)}{\|X-m\|^2}\right)\right] \qquad (19) \\
&= \mathbb{E}[\|X-m\|^{-1}]\mathbf{I}_H - \Delta_m.
\end{aligned}
$$

Since Γ_m is bounded and symmetric, it is self-adjoint. Moreover, the operator Δ_m defined by (19) is trace class: it is self-adjoint, non-negative and, if (e_j) is an orthonormal basis,

$$
\sum_j \langle e_j, \Delta_m e_j\rangle = \sum_j \mathbb{E}\left[\frac{\langle X-m, e_j\rangle^2}{\|X-m\|^3}\right]
$$

$$
\leq \mathbb{E}\left[\frac{1}{\|X-m\|}\right] < \infty.
$$

Therefore Δ_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

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

Moreover, thanks to (7), the smallest eigenvalue $\lambda_{\min}$ of Γ_m is strictly positive. For simplicity of notation, we rewrite this decomposition in the following way:

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

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

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

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 $\gamma_n \mathbb{E}[\|X-m\|^{-1}] < 1$, which we can assume without loss of generality. Elementary analysis shows that there exist constants

c_1, C_2, C_3 such that

$$\forall x \in \sigma(\Gamma_m), \qquad c_1 \exp(-s_n x) \le f_n(x) \le C_2 \exp(-s_n x),$$

$$\left| s_n - \frac{c_\gamma}{1-\alpha} n^{1-\alpha} \right| \le C_3, \tag{20}$$

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:

$$\beta_n x = \sum_{\lambda \in \Lambda} f_n(\lambda) \langle e_\lambda, x \rangle e_\lambda, \qquad x \in H;$$

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:

$$Z_{n+1} = Z_n + \gamma_n \xi_{n+1} - \gamma_n \Phi(Z_n)$$
$$= Z_n + \gamma_n \xi_{n+1} - \gamma_n \big(\Gamma_m (Z_n - m) + \delta_n \big),$$

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,

$$\forall k, \qquad Z_{k+1} - m = \alpha_k(Z_k - m) + \gamma_k \xi_{k+1} - \gamma_k \delta_k. \tag{21}$$

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

$$Z_n - m = \beta_{n-1}(Z_1 - m) + \beta_{n-1}M_n - \beta_{n-1}R_{n-1}, \tag{22}$$

where

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

The first two terms of (22) 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 behavior of f_n in equation (20) implies that

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

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

$$\mathbb{E}[\|\beta_{n-1}(Z_1 - m)\|^2] \le C \exp(-2n^{1-\alpha}) \mathbb{E}[\|Z_1 - m\|^2]. \tag{23}$$

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

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

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

$$\mathbb{E}[\|\beta_{n-1} M_n\|^2] = \sum_\lambda f_{n-1}^2(\lambda) \mathbb{E}[(M_n^\lambda)^2]$$

$$\leq \sum_\lambda \sum_{k \leq n-1} \gamma_k^2 \left(\frac{f_{n-1}(\lambda)}{f_k(\lambda)} \right)^2 \mathbb{E}[\langle \xi_{k+1}, e_\lambda \rangle^2].$$

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

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

This uniformity in λ allows us to reconstruct $\mathbb{E}[\|\xi_{k+1}\|^2]$, which is bounded by 1, thanks to (12). We obtain

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

$$\leq \sum_{k \leq n-1} \gamma_k^2 \left(\frac{f_{n-1}(\lambda_{\min})}{f_k(\lambda_{\min})} \right)^2 \mathbb{E}[\|\xi_{k+1}\|^2]$$

$$\leq \sum_{k \leq n-1} \gamma_k^2 \left(\frac{f_{n-1}(\lambda_{\min})}{f_k(\lambda_{\min})} \right)^2.$$

Now we use bounds (20) on β_n,

$$\mathbb{E}[\|\beta_{n-1} M_n\|^2] \leq \frac{C_2^2}{c_1^2} \sum_{k \leq n-1} \gamma_k^2 \exp\left(-\sum_{j=k+1}^n \gamma_j \right)$$

$$\leq C \sum_{k \leq n-1} \gamma_k^2 \exp\left(-\frac{1}{1-\alpha} (n^{1-\alpha} - k^{1-\alpha}) \right). \tag{24}$$

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

$$l(n)^{1-\alpha} = n^{1-\alpha} - c_\alpha \ln(n), \tag{25}$$

with c_α to be chosen later. The first part of the sum (24) (for $k \leq l(n)$) gives us

$$\sum_{k \leq l(n)} \gamma_k^2 \exp\left(-\frac{1}{1-\alpha}(n^{1-\alpha} - k^{1-\alpha})\right) \leq \sum_{k \leq l(n)} \gamma_k^2 \exp\left(-\frac{c_\alpha}{1-\alpha} \ln(n)\right)$$
$$\leq c_\gamma^2 \sum_{k \leq n} k^{-2\alpha - c_\alpha/(1-\alpha)}. \tag{26}$$

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 (24), for $k > l(n)$, we bound the exponential by 1 and γ_k by $\gamma_{l(n)}$.

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

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 (26). Finally we get

$$\mathbb{E}[\|\beta_{n-1} M_n\|^2] \leq C \frac{\ln(n)}{n^\alpha}. \tag{27}$$

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

$$\exists r, C_r \forall k, \qquad \|Z_k - m\| \leq r \implies \|\delta_k\| \leq C_r \|Z_k - m\|^2. \tag{28}$$

Since Z_n converges a.s. to m, we deduce two things about δ_k: it is almost surely bounded, and (28) becomes a.s. eventually true. To use these facts we introduce the following sequence of events:

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

for a value of K to be chosen later, and $l(n)$ defined by (25). 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 reuse the definition of $l(n)$ and cut the sum in two parts. For $\omega \in \Omega_N$, and $n \geq N$,

$$\|\beta_{n-1} R_n\|^2 \leq \left(\sum_{k=1}^n \gamma_k \|\beta_{n-1}\beta_k^{-1}\| \|\delta_k\|\right)^2$$
$$\leq 2N^2 \left(\sum_{k=1}^{l(n)} \gamma_k \|\beta_{n-1}\beta_k^{-1}\|\right)^2 + 2C_r^2 \left(\sum_{k=l(n)+1}^n \gamma_k \|Z_k - m\|^2\right)^2$$

$$\leq 2N^2 \left(\sum_{k=1}^{l(n)} \gamma_k \|\beta_{n-1}\beta_k^{-1}\| \right)^2 + 2\frac{C_r^2}{K^2}(n-l(n))\gamma_{l(n)} \sum_{k=l(n)+1}^{n} \gamma_k \|Z_k - m\|^2,$$

where we used the crude bound $\|\delta_k\| \leq N$ in the first part, and for the second part, $\|\beta_{n-1}\beta_k^{-1}\| \leq 1$ 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 $n \geq N$,

$$\|\beta_{n-1}R_n\|^2 \leq \frac{CN^2}{n^{42}} + \frac{C}{K^2} \sum_{k=l(n)+1}^{n} \gamma_k \|Z_k - m\|^2. \tag{29}$$

Combining now (22), (23), (27) and (29), we get, for $n \geq N$ and some new constant C

$$\mathbb{E}[\mathbf{1}_{\Omega_N}\|Z_n - m\|^2] \leq \frac{C\ln(n)}{n^{\alpha}} + \frac{C}{K^2} \sum_{k=l(n)+1}^{n} \gamma_k \mathbb{E}[\mathbf{1}_{\Omega_N}\|Z_k - m\|^2]$$

$$\leq \frac{C\ln(n)}{n^{\alpha}} + \frac{C'}{K^2} \sup_{l(n)<k\leq n} \mathbb{E}[\mathbf{1}_{\Omega_N}\|Z_k - m\|^2].$$

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

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

Defining $u_n = \mathbb{E}[\mathbf{1}_{\Omega_N}\|Z_n - m\|^2]$, this reads,

$$\forall n \geq N, \qquad u_n \leq \frac{C\ln(n)}{n^{\alpha}} + \frac{1}{2} \max_{l(n)<k\leq n} u_k. \tag{30}$$

Let us prove by induction that, for some N' large enough, and for $C'' = 4C$,

$$\forall n \geq N', \qquad u_n \leq \frac{C''\ln(n)}{n^{\alpha}}.$$

Suppose that $u_k \leq \frac{C''\ln(k)}{k^{\alpha}}$ for all $k \leq n$, and let us prove that $u_{n+1} \leq \frac{C''\ln(n+1)}{(n+1)^{\alpha}}$. Using (30), we know that

$$u_{n+1} \leq \frac{C\ln(n+1)}{(n+1)^{\alpha}} + \frac{1}{2} \max_{l(n+1)<k\leq n+1} u_k$$

$$\leq \frac{C\ln(n+1)}{(n+1)^{\alpha}} + \frac{1}{2} \max\left(u_{n+1}, \frac{C''\ln(l(n+1))}{l(n+1)^{\alpha}} \right).$$

If the max on the right hand side is u_{n+1}, we get

$$\frac{1}{2}u_{n+1} \le \frac{C\ln(n+1)}{(n+1)^\alpha}$$

which is the desired result since $2C \le 4C = C''$. If this is not the case, then the max is $\frac{C''\ln(l(n+1))}{l(n+1)^\alpha}$. However, $l(n+1) \sim (n+1)$ so for n larger than some N', $\frac{\ln(l(n+1))}{l(n+1)^\alpha} \le \frac{3}{2}\frac{\ln(n+1)}{(n+1)^\alpha}$. Hence

$$u_{n+1} \le \frac{C\ln(n+1)}{(n+1)^\alpha} + \frac{3}{4}\frac{C''\ln(n+1)}{(n+1)^\alpha}$$

$$\le 4C\frac{\ln(n+1)}{(n+1)^\alpha} = C''\frac{\ln(n+1)}{(n+1)^\alpha}.$$

This concludes the induction step and the proof of Proposition 3.2. □

6.4. Proof of Theorem 3.4

We use the same decomposition as in Pelletier [28]. It consists in linearizing the target function Φ around the true value m. Recall the following decomposition of the error (21):

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

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,

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

and rearranging the previous expression, we obtain

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

Summing these equalities, it becomes

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

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

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

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

$$\frac{\widetilde{M}_n}{\sqrt{n}} \underset{n \to \infty}{\overset{\mathcal{L}}{\to}} \mathcal{N}(0, \Sigma),$$

we need to check that the assumptions of Theorem 5.1 in Jakubowski [19] are fulfilled. We first have that the martingale difference sequence is *a.s.* bounded, $\forall n \|\xi_n\| \le 2$. Let us define

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

which can also be decomposed as follows:

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

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

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

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

$$\|\Phi(Z_n) \otimes \Phi(Z_n)\|_L \to 0 \qquad \text{a.s.}$$

With similar arguments, it is easy to show that

$$\left\| \Sigma - \mathbb{E}\left[\frac{(X - Z_n)}{\|X - Z_n\|} \otimes \frac{(X - Z_n)}{\|X - Z_n\|} \Big| \mathcal{F}_n \right] \right\|_L \le 2\mathbb{E}\left[\left\| \frac{(X - Z_n)}{\|X - Z_n\|} - \frac{(X - m)}{\|X - m\|} \right\| \Big| \mathcal{F}_n \right]$$

$$\le 4\mathbb{E}\left[\frac{1}{\|X - m\|} \right] \|Z_n - m\|,$$

so that $\|\Sigma_n - \Sigma\|_L \to 0$ a.s., when n tends to infinity. Then condition 5.2 in Jakubowski [19] is satisfied and is a consequence of a direct application of Chow's Lemma; see, for instance, Duflo [12], page 22.

Now, it remains to prove that

$$\frac{1}{\sqrt{n}} \left(\frac{T_{n+1}}{\gamma_n} - \sum_{k=2}^{n} T_k \left[\frac{1}{\gamma_k} - \frac{1}{\gamma_{k-1}} \right] + \sum_{k=1}^{n} \delta_k \right) \underset{n \to \infty}{\overset{P}{\to}} 0. \tag{31}$$

Let us denote by A_n, A'_n and A''_n the three terms.

Recall that $\mathbb{E}[\mathbf{1}_{\Omega_N} \|T_n\|^2] \le C_N \frac{\ln(n)}{n^\alpha}$, thanks to Proposition 3.2. For the first term $A_n = \frac{T_{n+1}}{\sqrt{n}\gamma_n}$, we have

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

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

Let us turn to the second term A'_n. Since $\gamma_k^{-1} - \gamma_{k-1}^{-1} \le 2\alpha c_\gamma^{-1} k^{\alpha-1}$, we have, for two positive constants C_0, C_1,

$$\mathbb{E}[\|A'_n\|\mathbf{1}_{\Omega_N}] \le \frac{2\alpha c_\gamma^{-1}}{\sqrt{n}} \sum_{k\le n} \mathbb{E}[\mathbf{1}_{\Omega_N}\|T_k\|]k^{\alpha-1}$$

$$\le \frac{C_0}{\sqrt{n}} \sum_{k\le n} \sqrt{\ln(k)}k^{\alpha/2-1}$$

$$\le C_1\sqrt{\ln(n)}n^{\alpha/2-1/2},$$

which goes to zero since $\alpha < 1$. Therefore $A'_n \xrightarrow[n\to\infty]{P} 0$.

Finally, for the last term A''_n, since there exists a positive constant C_2 such that $\|\delta_k\| \le C_2\|Z_k - m\|^2$, we have

$$\mathbb{E}[\mathbf{1}_{\Omega_N}\|A''_n\|] \le \frac{1}{\sqrt{n}} \sum_{k\le n} \mathbb{E}[\mathbf{1}_{\Omega_N}\|T_k\|^2]$$

$$\le \frac{C_N}{\sqrt{n}} \sum_{k\le n} \ln(k)k^{-\alpha}.$$

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 (31) holds, and Theorem 3.4 is finally proved.

Acknowledgements

References

[1] Arnaudon, M., Dombry, C., Phan, A. and Yang, L. (2010). Stochastic algorithms for computing means of probability measures. Preprint. Available at http://hal.archives-ouvertes.fr/hal-00540623/PDF/algo_means4.pdf.

[2] Benveniste, A., Métivier, M. and Priouret, P. (1990). *Adaptive Algorithms and Stochastic Approximations. Applications of Mathematics* (*New York*) **22**. Berlin: Springer. MR1082341

[3] Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In *Compstat 2010* (Y. Lechevallier and G. Saporta, eds.) 177–186. Heidelberg: Springer.

[4] Cadre, B. (2001). Convergent estimators for the L_1-median of a Banach valued random variable. *Statistics* **35** 509–521. MR1880179

[5] Cardot, H., Cénac, P. and Chaouch, M. (2010). Stochastic approximation to the multivariate and the functional median. In *Compstat 2010* (Y. Lechevallier and G. Saporta, eds.) 421–428. Heidelberg: Springer.

[6] Cardot, H., Cénac, P. and Monnez, J.M. (2012). A fast and recursive algorithm for clustering large datasets with k-medians. *Comput*. *Statist*. *Data Anal*. **56** 1434–1449.

[7] Chaouch, M. and Goga, C. (2010). Design-based estimation for geometric quantiles with application to outlier detection. *Comput*. *Statist*. *Data Anal*. **54** 2214–2229. MR2720483

[8] Chaudhuri, P. (1992). Multivariate location estimation using extension of R-estimates through U-statistics type approach. *Ann*. *Statist*. **20** 897–916. MR1165598

[9] Chaudhuri, P. (1996). On a geometric notion of quantiles for multivariate data. *J*. *Amer*. *Statist*. *Assoc*. **91** 862–872. MR1395753

[10] Cuevas, A., Febrero, M. and Fraiman, R. (2007). Robust estimation and classification for functional data via projection-based depth notions. *Comput*. *Statist*. **22** 481–496. MR2336349

[11] Dippon, J. and Walk, H. (2006). The averaged Robbins–Monro method for linear problems in a Banach space. *J*. *Theoret*. *Probab*. **19** 166–189. MR2256485

[12] Duflo, M. (1997). *Random Iterative Models*. *Applications of Mathematics* (*New York*) **34**. Berlin: Springer. Translated from the 1990 French original by Stephen S. Wilson and revised by the author. MR1485774

[13] Fraiman, R. and Muniz, G. (2001). Trimmed means for functional data. *Test* **10** 419–440. MR1881149

[14] Gervini, D. (2008). Robust functional estimation using the median and spherical principal components. *Biometrika* **95** 587–600. MR2443177

[15] Gower, J.C. (1974). Algorithm as 78: The mediancentre. *J*. *R*. *Stat*. *Soc*. *Ser*. *C Appl*. *Stat*. **23** 466–470.

[16] Haberman, S.J. (1989). Concavity and estimation. *Ann*. *Statist*. **17** 1631–1661. MR1026303

[17] Haldane, J.B.S. (1948). Note on the median of a multivariate distribution. *Biometrika* **35** 414–417.

[18] Huber, P.J. and Ronchetti, E.M. (2009). *Robust Statistics*, 2nd ed. *Wiley Series in Probability and Statistics*. Hoboken, NJ: Wiley. MR2488795

[19] Jakubowski, A. (1988). Tightness criteria for random measures with application to the principle of conditioning in Hilbert spaces. *Probab*. *Math*. *Statist*. **9** 95–114. MR0945679

[20] 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) 217–230. Amsterdam: North-Holland. MR0949228

[21] Koltchinskii, V.I. (1997). M-estimation, convexity and quantiles. *Ann*. *Statist*. **25** 435–477. MR1439309

[22] Kushner, H.J. and Clark, D.S. (1978). *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. *Applied Mathematical Sciences* **26**. New York: Springer. MR0499560

[23] Kushner, H.J. and Yin, G.G. (2003). *Stochastic Approximation and Recursive Algorithms and Applications*, 2nd ed. *Applications of Mathematics* (*New York*) **35**. New York: Springer. MR1993642

[24] Li, W.V. and Shao, Q.M. (2001). Gaussian processes: Inequalities, small ball probabilities and applications. In *Stochastic Processes*: *Theory and Methods*. *Handbook of Statist*. **19** 533–597. Amsterdam: North-Holland. MR1861734

[25] Ljung, L., Pflug, G. and Walk, H. (1992). *Stochastic Approximation and Optimization of Random Systems*. *DMV Seminar* **17**. Basel: Birkhäuser. MR1162311

[26] 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), *Vol*. *I*: *Statistics* 281–297. Berkeley, CA: Univ. California Press. MR0214227

[27] Nazarov, A.I. (2009). Exact L_2-small ball asymptotics of Gaussian processes and the spectrum of boundary-value problems. *J*. *Theoret*. *Probab*. **22** 640–665. MR2530107

[28] Pelletier, M. (2000). Asymptotic almost sure efficiency of averaged stochastic algorithms. *SIAM J. Control Optim.* **39** 49–72 (electronic). MR1780908

[29] Polyak, B.T. and Juditsky, A.B. (1992). Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim.* **30** 838–855. MR1167814

[30] R Development Core Team (2010). A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

[31] Ruppert, D. (1985). A Newton–Raphson version of the multivariate Robbins–Monro procedure. *Ann. Statist.* **13** 236–245. MR0773164

[32] Smale, S. and Yao, Y. (2006). Online learning algorithms. *Found. Comput. Math.* **6** 145–170. MR2228737

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

[34] Vardi, Y. and Zhang, C.H. (2000). The multivariate L_1-median and associated data depth. *Proc. Natl. Acad. Sci. USA* **97** 1423–1426 (electronic). MR1740461

[35] Walk, H. (1977). An invariance principle for the Robbins–Monro process in a Hilbert space. *Z. Wahrsch. Verw. Gebiete* **39** 135–150. MR0494742