Online stochastic Newton methods for estimating the geometric median and applications

Antoine Godichon-Baggioni, Wei Lu

Online stochastic Newton methods for estimating the geometric median and applications

Antoine Godichon-Baggioni$^{(*)}$ and Wei Lu$^{(**)}$
$(*)$ Sorbonne Université, Laboratoire de Probabilités, Statistique et Modélisation
$(**)$ INSA Rouen Normandie, Laboratoire de Mathématiques de l'INSA
antoine.godichon_baggioni@upmc.fr, wei.lu@insa-rouen.fr

Abstract

In the context of large samples, a small number of individuals might spoil basic statistical indicators like the mean. It is difficult to detect automatically these atypical individuals, and an alternative strategy is using robust approaches. This paper focuses on estimating the geometric median of a random variable, which is a robust indicator of central tendency. In order to deal with large samples of data arriving sequentially, online stochastic Newton algorithms for estimating the geometric median are introduced and we give their rates of convergence. Since estimates of the median and those of the Hessian matrix can be recursively updated, we also determine confidences intervals of the median in any designated direction and perform online statistical tests.

Keywords: Geometric median; stochastic Newton algorithm; online estimation, stochastic optimization

1 Introduction

Large samples of observations are now commonplace due to advancements in measurement technology and improved computer storage capabilities. In such a large sample context, even a small number of individuals might spoil basic statistical indicators like the mean. Detecting automatically these atypical individuals is difficult, and adopting robust approaches is an appealing alternative. It is well known that the median is a robust indicator of central tendency, and here we concentrate on the geometric median of a random variable in $\mathbb{R}^p$. The geometric median, also called spatial median or multivariate L_1 median, is firstly introduced in [9] and [13]. It is defined as the minimizer of L_1 distances to observations of a random variable. It has nice robustness properties such as a breakdown point at 0.5 [16, 14].

Recently, the geometric median attached more attention in the field of machine learning. For example, in [17], authors proposed a L_1 median filter as a tool of their mesh denoising method, which helps their model to preserve geometric features; an optimization algorithm of section line extraction was established based on geometric median [20], the author considered the geometric median because it has the characteristics of noise immunity; an image filtering algorithm [7] has

been proposed based on a spatial median filter, which shows better performances than mean filters, since the median is more robust than the mean when noisy pixels are present in the image.

In this paper, we focus on the estimation of the geometric median. An iterative algorithm called Weiszfeld's algorithm has been developed [19, 12, 18], and the method has been improved in [1]. The algorithm is simple and fast, but the procedure is not adapted in the case where data are acquired sequentially from files too large to be loaded into memory. To overcome this, and since the geometric median is defined as the minimizer of a convex function, an averaged stochastic gradient algorithm for estimating the geometric median has been proposed in [6]. However, as a first-order algorithm, in practice it can be very sensitive to the Hessian structure of the function to minimize [2, 3].

In order to overcome this, we propose here new stochastic Newton type algorithms for estimating the geometric median. One difficulty encountered by stochastic Newton algorithms is the update of the inverse of the Hessian estimates. Our recursive estimation of the inverse of the Hessian is based on the Sherman-Morrison formula [8], which avoids an expensive inverse matrix calculation. In order to overcome possible initialization problems, we also propose a weighted averaged version [3]. Thanks to the asymptotic efficiency of the algorithms, and since one can recursively estimate the covariance matrix, we introduce online confidence intervals of the geometric median in a chosen direction and to perform online statistical hypothesis tests.

The paper is organized as follows: we describe the general framework and explain the method for estimating recursively the inverse of the Hessian in Section 2. In Section 3 we present stochastic Newton algorithms and we state their rates of convergence. A simulation study for comparing the performances of different algorithms is also given. Section 4 is devoted to establishing recursive confidence intervals and performing online statistical tests for the geometric median. The proofs are gathered in Section 5.

2 Framework

2.1 General framework

The geometric median m of a random variable X taking values in $\mathbb{R}^p$ is the minimizer of the convex function $G : \mathbb{R}^p \longrightarrow \mathbb{R}$ defined for all $h \in \mathbb{R}^p$ by [13]

$$G(h) =: \mathbb{E}\left[g(X,h)\right] = \mathbb{E}\left[\|X-h\| - \|X\|\right].$$

Note that this definition does not assume the uniqueness of the median or the existence of the first order moment of $\|X\|$. From now on we suppose that following assumptions are fulfilled.

- **Assumption 1.** The random variable X is not concentrated around single points : there exists $C_6 > 0$ such that for all $h \in \mathbb{R}^p$,

$$\mathbb{E}\left[\frac{1}{\|X-h\|^6}\right] \leq C_6.$$

- **Assumption 2.** The random variable X is not concentrated on a straight line : for all $h \in \mathbb{R}^p$, there exists $h' \in \mathbb{R}^p$ such that $\langle h, h' \rangle \neq 0$ and

$$\mathbb{V}\mathrm{ar}\left[\langle X, h' \rangle\right] > 0.$$

Note that in Assumption 1, the order is usually obtained to be 2 in the literature [6], and we increase the order to 6 for technical reasons, i.e. it is used to obtain the convergence rate of the Hessian's estimates that will be presented later. According to [14], Assumption 2 ensures that the function G is strictly convex, so that the median m is uniquely defined. As shown in [6], the function G is differentiable everywhere, and one can check that the gradient is defined for all $h \in \mathbb{R}^p$ by:

$$\nabla G(h) = -\mathbb{E}\left[\frac{X-h}{\|X-h\|}\right].$$

Moreover, the function G is twice differentiable everywhere and its Hessian is given by [15]

$$\nabla^2 G(h) = \mathbb{E}\left[\frac{1}{\|X-h\|}\left(I_p - \frac{(X-h)(X-h)^T}{\|X-h\|^2}\right)\right].$$

According to [6], $\nabla^2 G(h)$ is positive definite under Assumptions 1 and 2. This is of particular interest to stochastic Newton type algorithms for estimating the geometric median, in which the information given by the Hessian matrix of the function G will be taken into account.

2.2 Some recalls on the averaged stochastic gradient algorithm

An averaged stochastic gradient algorithm has been proposed in [6] for estimating the geometric median. Given X_1, $X_2, \ldots, X_n, X_{n+1}, \ldots,$ i.i.d copies of X, the stochastic gradient algorithm is given by

$$m_{n+1}^{(SG)} = m_n^{(SG)} + \gamma_n \frac{X_{n+1} - m_n^{(SG)}}{\left\|X_{n+1} - m_n^{(SG)}\right\|}, \tag{1}$$

where γ_n is a sequence of descent steps. Its averaged version consists of averaging all the estimated past values, which is defined recursively by

$$\overline{m}_{n+1} = \overline{m}_n + \frac{1}{n+1}\left(m_{n+2}^{(SG)} - \overline{m}_n\right) \tag{2}$$

with $m_0^{(SG)}$ bounded and $\overline{m}_0 = m_0^{(SG)}$. Thus, the estimation can be easily updated. This algorithm has been deeply studied: its asymptotic efficiency is given in [6], while the L^p rates are derived in [10]. Moreover, the non-asymptotic behavior of this algorithm has also been studied in [4] by giving non-asymptotic confidence balls based on the derivation of improved L^2 rates of convergence. However, it's a first-order algorithm and thus can be very sensitive to the structure of Hessian of the function we try to minimize [3], which means here that the random variable X is fairly concentrated towards a straight line for instance.

2.3 How to estimate the inverse of the Hessian

We will then focus on Newton type methods, which is more adapted to deal with ill-conditioned problems. A major difficulty encountered by stochastic Newton algorithm is that we must be able to update the inverse of the Hessian estimate with a cost, in terms of computation time, as low as possible. We explain now how to estimate the inverse of the Hessian $\nabla^2 G(m)$ in a recursive way

when m is known. The idea is to obtain an estimate of the form $\frac{1}{n}\sum_{k=1}^{n} a_k\phi_k\phi_k^T$ to apply Riccati's formula [8]. We have

$$\nabla^2 g(X,h) = \frac{1}{\|X-h\|}\left(I_p - \frac{(X-h)(X-h)^T}{\|X-h\|^2}\right).$$

Note that $\left(I_p - \frac{(X-h)(X-h)^T}{\|X-h\|^2}\right)^2 = I_p - \frac{(X-h)(X-h)^T}{\|X-h\|^2}$, we then have

$$\nabla^2 g(X,h) = \|X-h\|\left(\nabla^2 g(X,h)\right)^2. \tag{3}$$

In addition, according to Taylor's theorem,

$$\nabla g(X,h+\alpha Z) - \nabla g(X,h) = \int_0^1 \nabla^2 g(X,h+t\alpha Z)dt\alpha Z,$$

where $Z \sim \mathcal{N}(0,I_p)$ and $\alpha > 0$. Therefore, an estimate of $\nabla^2 G(m)$ is given by

$$\hat{H}_n = \frac{1}{n+1}\left(\sum_{k=1}^{n}\frac{\|X_k-m\|}{\alpha_k^2}\Phi_k\Phi_k^T + H_0\right), \tag{4}$$

where $\alpha_k = \frac{1}{k\ln(k+1)}$, $H_0 = I_p$ and Φ_k is defined by

$$\Phi_k := \nabla g(X_k, m+\alpha_k Z_k) - \nabla g(X_k,m) = \int_0^1 \nabla^2 g(X_k, m+t\alpha_k Z_k)dt\alpha_k Z_k,$$

where $(Z_k)_k$ are standard independent Gaussian vectors for any $k \geq 1$. Indeed, one can check that $\mathbb{E}\left[\frac{\|X_k-m\|}{\alpha_k^2}\Phi_k\Phi_k^T\right] \to \nabla^2 G(m)$. In addition, with the help of Riccati's formula [8], $H_{n+1}^{-1} = (n+1)^{-1}\hat{H}_n^{-1}$ can be easily updated as

$$H_{n+1}^{-1} = H_n^{-1} - \frac{\|X_{n+1}-m\|}{\alpha_{n+1}^2}\left(1 + \frac{\|X_{n+1}-m\|}{\alpha_{n+1}^2}\phi_{n+1}^T H_n^{-1}\phi_{n+1}\right)^{-1} H_n^{-1}\phi_{n+1}\phi_{n+1}^T H_n^{-1}.$$

Thus, knowing m, we are able to estimate recursively the inverse of the Hessian with complexity $\mathcal{O}\left(p^2\right)$ (instead of $\mathcal{O}\left(p^3\right)$) for each iteration.

3 Stochastic Newton methods

In this section we introduce two stochastic Newton methods for estimating the median m : a stochastic Newton algorithm and its weighted averaged version. We also give theoretical guarantees on their convergence. We recall that $(X_n)_{n\geq 1}$ is a sequence of independent random vectors, of same distribution as vector X and $(Z_n)_{n\geq 1}$ is a sequence of independent standard Gaussian vectors.

3.1 Stochastic Newton algorithm

3.1.1 Definition of the algorithm

We now introduce stochastic Newton estimates, defined recursively for all $n \geq 0$ by

$$m_{n+1} = m_n + \frac{1}{n+1} \widetilde{H}_n^{-1} \frac{X_{n+1} - m_n}{\|X_{n+1} - m_n\|}, \tag{5}$$

where m_0 is bounded. Let $(\tilde{\beta}_n)_{n \geq 1}$ be the strictly positive sequence of real numbers defined for any $n \geq 1$ by $\tilde{\beta}_n = \frac{c_\beta}{n^\beta}$ with $0 < \beta < \frac{1}{2}$ and $c_\beta > 0$. The matrix $\widetilde{H}_n$ is given for any $n \geq 0$ by

$$\widetilde{H}_n = \overline{H}_n + \frac{1}{n+1} \sum_{k=1}^{n} \tilde{\beta}_k Z_k Z_k^T, \tag{6}$$

where $\overline{H}_n$ is the recursive estimate of the Hessian $\nabla^2 G(m)$ defined by

$$\overline{H}_n = \frac{1}{n+1} \left(\sum_{k=1}^{n} \frac{\|X_k - m_{k-1}\|}{\alpha_k^2} \phi_k \phi_k^T + H_0 \right),$$

with for any $k \geq 1$, $\phi_k = \nabla g(X_k, m_{k-1} + \alpha_k Z_k) - \nabla g(X_k, m_{k-1})$, $\alpha_k = \frac{1}{k \ln(k+1)}$ and H_0 is symmetric positive. We add the term $\sum_{k=1}^{n} \tilde{\beta}_k Z_k Z_k^T$ in order to control the eigenvalues of the Hessian estimate (see Section 5), which is necessary to obtain the convergence of the algorithm [3]. Thanks to Riccati's formula [8] , $H_n^{-1} = (n+1)^{-1} \overline{H}_n^{-1}$ can be updated in two steps, leading to

$$H_{n+1/2}^{-1} = H_n^{-1} - \frac{\|X_{n+1} - m_n\|}{\alpha_{n+1}^2} \left(1 + \frac{\|X_{n+1} - m_n\|}{\alpha_{n+1}^2} \phi_{n+1}^T H_n^{-1} \phi_{n+1} \right)^{-1} H_n^{-1} \phi_{n+1} \phi_{n+1}^T H_n^{-1},$$

$$H_{n+1}^{-1} = H_{n+1/2}^{-1} - \tilde{\beta}_k \left(1 + \tilde{\beta}_k Z_{n+1}^T H_{n+1/2}^{-1} Z_{n+1} \right)^{-1} H_{n+1/2}^{-1} Z_{n+1} Z_{n+1}^T H_{n+1/2}^{-1}.$$

Therefore, this algorithm allows us to update the estimation of the Hessian matrix and the estimation of the geometric median in a recursive way.

3.1.2 Convergence results

The following theorem gives the almost sure rates of convergence as well as the asymptotic efficiency of the stochastic Newton estimates. Note that its asymptotic efficiency allows us to construct confidence intervals and carry out tests (discussed in Section 4).

Theorem 1 *Assume that Assumptions 1 and 2 hold, then the stochastic Newton estimate m_n defined by* (5) *converges almost surely towards m and*

$$\|m_n - m\|^2 = \mathcal{O}\left(\frac{\ln n}{n} \right) a.s.$$

Furthermore, the Hessian estimates defined in (6) *satisfy for all $\delta > 0$*

$$\left\| \widetilde{H}_n - H \right\|^2 = \mathcal{O}\left(\max\left\{ \frac{(\ln n)^{1+\delta}}{n}, \frac{c_\beta}{n^{2\beta}} \right\} \right) a.s.$$

Finally,

$$\sqrt{n}\,(m_n - m) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, H^{-1}\Sigma H^{-1}\right),$$

where $\Sigma = \mathbb{E}\left[\nabla g(X,m)\nabla g(X,m)^T\right]$.

The proof is given in Section 5. Observe that the price to pay in order to control the eigenvalues of the estimates of the Hessian is a loss in term of rate of convergence of the estimates. More precisely, it makes appear a term which converges at a rate $n^{-2\beta}$ instead of n^{-1}.

3.2 Weighted Averaged Stochastic Newton Algorithm

3.2.1 Definition of the algorithm

In order to improve in practice the behavior of the estimates in case of bad initializations, we now introduce a Weighted Averaged Stochastic Newton algorithm (WASN) [3] defined recursively for all $n \geq 0$ by:

$$\hat{m}_{n+1} = \hat{m}_n + \frac{c_\gamma}{\left(n+1+c'_\gamma\right)^\gamma} \widetilde{H}_{n,\tau}^{-1} \frac{X_{n+1} - \hat{m}_n}{\|X_{n+1} - \hat{m}_n\|} \tag{7}$$

$$m_{n+1,\tau} = (1-\tau_{n+1}) m_{n,\tau} + \tau_{n+1}\hat{m}_{n+1}, \tag{8}$$

where $c_\gamma > 0$, $c'_\gamma \geq 0$ and $\gamma \in \left(\frac{1}{2}, 1\right)$. The weighted averaging sequence $(\tau_n)_{n\geq 1}$ is chosen of the following way : $\tau_n = \frac{\ln(n+1)^\omega}{\sum_{k=0}^n \ln(k+1)^\omega}$ for any $n \geq 0$ and $\omega \geq 0$. Notice that the case where $\omega = 0$ corresponds to the averaged stochastic Newton algorithm (ASN). The recursive estimate of the Hessian is defined by :

$$\widetilde{H}_{n,\tau} = \frac{1}{n+1}\left(\sum_{k=1}^n \frac{\|X_k - m_{k-1,\tau}\|}{\alpha_k^2}\phi_{k,\tau}\phi_{k,\tau}^T + H_0\right) + \frac{1}{n+1}\sum_{k=1}^n \tilde{\beta}_k Z_k Z_k^T. \tag{9}$$

where for any $k \geq 1$, $\phi_{k,\tau} = \nabla g(X_k, m_{k-1,\tau} + \alpha_k Z_k) - \nabla g(X_k, m_{k-1,\tau})$. In order to control the eigenvalue of $\widetilde{H}_{n,\tau}$, $(\tilde{\beta}_n)_{n\geq 1}$ should be the sequence of real numbers defined by $\tilde{\beta}_n = \frac{c_\beta}{n^{1-\beta}}$ with $0 < \beta < \gamma - \frac{1}{2}$ and $c_\beta > 0$. Following the same procedure as for the stochastic Newton algorithm, we can always update $H_{n,\tau}^{-1} = (n+1)^{-1}\overline{H}_{n,\tau}^{-1}$ with Riccati's formula [8].

3.2.2 Convergence results

The following theorem shows that under identical assumptions, the WASN estimates are still asymptotically efficient.

Theorem 2 *Suppose Assumptions 1 and 2 hold, then the Weighted Averaged Stochastic Newton estimates* $\hat{m}_n$ *and* m_n, τ *converge almost surely towards* m*. In addition,*

$$\|\hat{m}_n - m\|^2 = \mathcal{O}\left(\frac{\ln n}{n^\gamma}\right) a.s. \qquad \textit{and} \qquad \|m_{n,\tau} - m\|^2 = \mathcal{O}\left(\frac{\ln n}{n}\right) a.s.$$

Furthermore, the Hessian estimate defined by (9) *satisfies for all* $\delta > 0$

$$\left\| \widetilde{H}_{n,\tau} - H \right\|^2 = \mathcal{O}\left(\max\left\{ \frac{(\ln n)^{1+\delta}}{n}, \frac{c_\beta}{n^{2\beta}} \right\} \right) a.s.$$

Finally,

$$\sqrt{n}\left(m_{n,\tau} - m\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, H^{-1}\Sigma H^{-1}\right),$$

where $\Sigma = \mathbb{E}\left[\nabla g(X,m)\nabla g(X,m)^T\right]$.

The proof is given in Section 5.

3.3 Comparison of the methods

We perform a numerical experiment in order to compare the performances of the Stochastic Newton algorithm (SN), the Averaged Stochastic Newton algorithm (ASN), the Weighted Averaged stochastic Newton Algorithm (WASN) and the averaged stochastic gradient descent (ASGD) proposed in [6]. For WASN, we choose $\tau_n = \frac{\ln(n+1)^2}{\sum_{k=0}^{n}\ln(k+1)^2}$. In this experiment, we generate samples of Gaussian random vector $X \sim \mathcal{N}(0_p, \overline{\Sigma})$ with $p = 10$, and we consider two structures of covariance matrix $\overline{\Sigma}$ defined by

(i) $\overline{\Sigma}_{ij} = 0.5^{|i-j|}$;

(ii) $\overline{\Sigma}$ is diagonal with $\overline{\Sigma}_{1,1} = 1000$ and $\overline{\Sigma}_{i,i} = 1$ for $i \neq 1$.

To evaluate the performances of algorithms, we compute the following mean squared error:

$$MSE(\hat{m}) = \mathbb{E}\left[\|m - \hat{m}\|^2\right],$$

where $\hat{m}$ is an estimate of the median. We estimate this error through Monte-Carlo experiments with $N = 400$ samples, for each sample we generate $n = 15000$ copies of X. In order to see the impact of the initialization of $\hat{m}$, we consider four different initializations : $m_0 = rU$ with $U \sim \mathcal{N}_p(0, I_p)$ and $r = 1$, 5, 10 or 15.

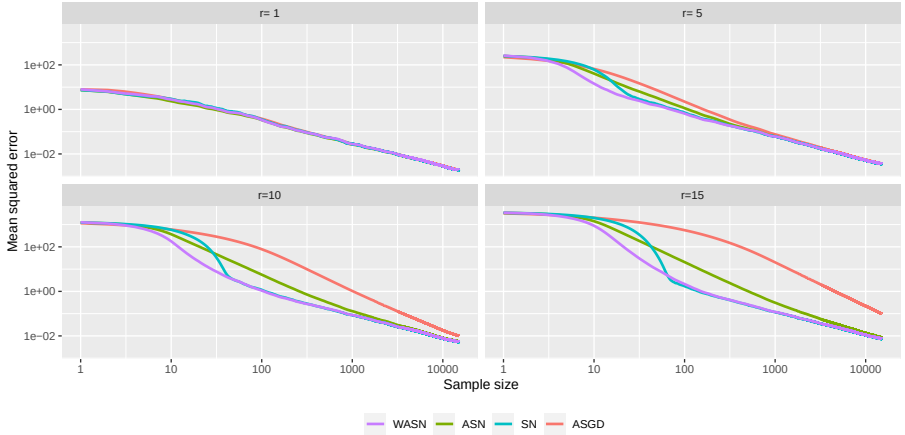

Figure 1: Evolution of the mean squared error with respect to the sample size for structure (i).

Considering the structure (i), the performances of four algorithms are identical for a good initialization. However, when initialization get worse, we can see that second order methods converge faster than ASGD.

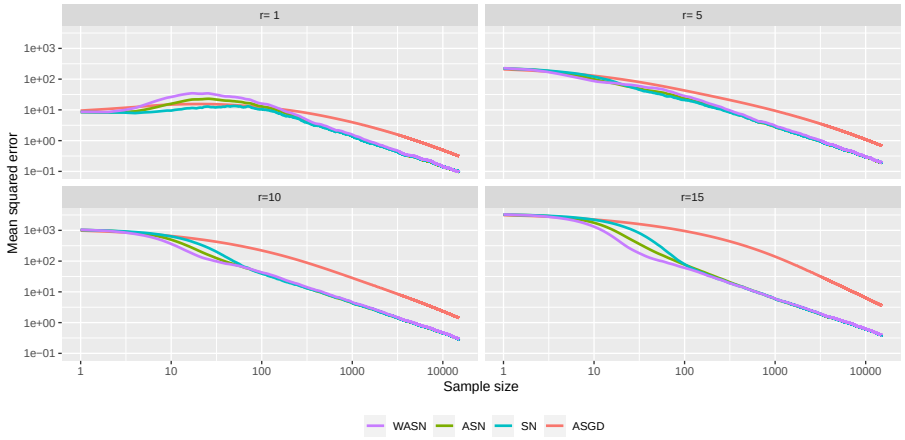

Figure 2: Evolution of the mean squared error with respect to the sample size for structure (ii).

When considering the structure (ii), we observe that the Newton type algorithms perform much better than ASGD. Even with a good initialization, the convergence of ASGD is clearly slower than WASN, ASN and SN. Thus ASGD is more sensitive to the structure of the Hessian. Note that for bad initializations, WASN estimators seem to achieve converge faster, and that this phenomenon can be accentuated in the case of even worse conditioned problems, i.e. for even worse Hessian structures [3].

4 Confidence intervals and tests

In this section, we shall propose confidende intervals and statistical tests for the median. These results are obtained from Theorems 1 and 2, and therefore require recursive estimates of the covariance matrix Σ defined by

$$\Sigma = \mathbb{E}\left[\nabla g(X,m)\nabla g(X,m)^T\right]$$

supposed here positive. In the sequel of the section, $\widetilde{m}_n$ will denote any asymptotically efficient estimate of the geometric median. For example, $\widetilde{m}_n$ can be the ASGD estimate defined by (2), or the SN estimate defined by (5), or the WASN estimate defined by (8).

4.1 Estimating the covariance

A natural recursive estimate of Σ is given by

$$\overline{\Sigma}_n = \frac{1}{n+1}\left(\sum_{k=1}^{n} \frac{(X_k - \widetilde{m}_{k-1})}{\|X_k - \widetilde{m}_{k-1}\|}\frac{(X_k - \widetilde{m}_{k-1})^T}{\|X_k - \widetilde{m}_{k-1}\|} + \Sigma_0\right),$$

where Σ_0 is symmetric positive. As well as for H_n, the Riccati's formula ([8]) allows us to recursively update matrix $W_{n+1}^{-1} = (n+1)^{-1}\overline{\Sigma}_n^{-1}$:

$$W_{n+1}^{-1} = W_n^{-1} - \left(1 + \frac{\left(\widetilde{X}_{n+1}\right)^T}{\left\|\widetilde{X}_{n+1}\right\|} W_n^{-1} \frac{\left(\widetilde{X}_{n+1}\right)}{\left\|\widetilde{X}_{n+1}\right\|}\right)^{-1} W_n^{-1} \frac{\left(\widetilde{X}_{n+1}\right)}{\left\|\widetilde{X}_{n+1}\right\|} \frac{\left(\widetilde{X}_{n+1}\right)^T}{\left\|\widetilde{X}_{n+1}\right\|} W_n^{-1}, \qquad (10)$$

where $\widetilde{X}_{n+1} := X_{n+1} - \widetilde{m}_n$. This property will be of particular interest to build online tests (see section 4.2). The following theorem gives the rate of convergence of $\overline{\Sigma}_n$.

Theorem 3 *Let $\widetilde{m}_n$ be an estimate defined by* (2), (5) *or* (8). *Suppose Assumptions 1 and 2 hold, then for any $\delta > 0$,*

$$\left\|\overline{\Sigma}_n - \Sigma\right\|^2 = o\left(\frac{(\ln n)^{1+\delta}}{n}\right) a.s.$$

The proof is given in section 5.

4.2 Confidence intervals and statistical hypothesis tests

Let us recall that under Assumptions 1 and 2

$$\sqrt{n}\left(\widetilde{m}_n - m\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, H^{-1}\Sigma H^{-1}\right).$$

Thus, we have for any $x_0 \in \mathbb{R}^p \backslash \{0\}$

$$\frac{\sqrt{n}}{\sqrt{x_0^T \overline{S}_n^{-1} \overline{\Sigma}_n \overline{S}_n^{-1} x_0}} \left(x_0^T \widetilde{m}_n - x_0^T m\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1),$$

where

$$\overline{S}_n = \frac{1}{n+1}\left(\sum_{k=1}^{n} \frac{\|X_k - \widetilde{m}_{k-1}\|}{\alpha_k^2} \widetilde{\phi}_k \widetilde{\phi}_k^T + S_0\right),$$

with S_0 symmetric positive, $(Z_k)_k$ standard independent Gaussian vectors, and $\widetilde{\phi}_k$ defined by $\widetilde{\phi}_k = \nabla g(X_k, \widetilde{m}_{k-1} + \alpha_k Z_k) - \nabla g(X_k, \widetilde{m}_{k-1})$. As $\overline{S}_n^{-1}$ and $\overline{\Sigma}_n$ can be recursively calculated (see Section 3.1.1 for the update of $\overline{S}_n^{-1}$), we can then compute an online confidence interval of $x_0^T m$, which means that we can determine the confidence interval of the median in any designated direction. Moreover, since $\widetilde{m}_n$ is asymptotically efficient, one has

$$n\left(\widetilde{m}_n - m\right)^T \overline{H}_n^* \overline{\Sigma}_n^{-1} \overline{H}_n^* \left(\widetilde{m}_n - m\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{X}_p^2,$$

where

$$\overline{H}_n^* = \frac{1}{n+1}\left(\sum_{k=1}^{n} \frac{1}{\|X_k - \widetilde{m}_{k-1}\|}\left(I_p - \frac{(X_k - \widetilde{m}_{k-1})(X_k - \widetilde{m}_{k-1})^T}{\|X_k - \widetilde{m}_{k-1}\|^2}\right) + \overline{H}_0^*\right)$$

with $\overline{H}_0^*$ symmetric positive. Thus $\overline{H}_n^*$ can be computed in a recursive way. Recall that $\overline{\Sigma}_n^{-1}$ can also be recursively updated with (10), so that we can perform an online statistical hypothesis test with significance level $\alpha \in (0,1)$: H_0 : "$m = m_{test}$" versus H_1 : "$m \neq m_{test}$". We calculate the test statistic Z_n by

$$Z_n = n \left(\tilde{m}_n - m_{test}\right)^T \overline{H}_n^* \overline{\Sigma}_n^{-1} \overline{H}_n^* \left(\tilde{m}_n - m_{test}\right),$$

and we reject the null hypothesis if $Z_n > \zeta_{1-\alpha,p}$, where $\zeta_{1-\alpha,p}$ is the quantile of order $1-\alpha$ of the chi-squared distribution with p degrees of freedom.

4.3 Simulations

We now evaluate performances of the different algorithms by studying the empirical levels under H_0. To this aim, we generate samples of size $n = 3000$ of a Gaussian random vector $X \sim \mathcal{N}(0_p, \overline{\Sigma})$ with $p = 10$, where we consider two structures of the covariance matrix $\overline{\Sigma}$ defined in Section 3.3. We compute the empirical levels through experiments with $N = 1000$ samples. We consider two different initializations : $m_0 = rU$ with $U \sim \mathcal{N}_p(0, I_p)$ and $r = 1$ or 5.

Structure of $\overline{\Sigma}$	m_0	Algorithm	Empirical level (%)
(i)	U	WASN	6.1
		ASN	5.2
		SN	5.8
		ASGD	6.3
	$5U$	WASN	5.9
		ASN	10.7
		SN	5.4
		ASGD	23.0
(ii)	U	WASN	18.9
		ASN	16.8
		SN	44.3
		ASGD	56.8
	$5U$	WASN	19.1
		ASN	22.0
		SN	50.4
		ASGD	97.2

Table 1: Empirical levels achieved by different algorithms under H_0

The performance of WASN is noticeable, it gives lower empirical level and closer to the 5% theoretical level even if m_0 is not well initialized. Same as what we observed in previous experiments, the algorithm ASN is sensitive to the initializations. We can also observe that the empirical level achieved by ASGD is the highest in every considered case. In general, proposed second-order methods achieve better results than ASGD, and the improvements are more significant in the case where we consider the structure (ii).

In addition, as the statistic test has a chi-squared limit distribution under the null hypothesis, we are now interested in the closeness between the simulated distribution of the test statistic and

the theoretical distribution. For this purpose, we plot the estimated probability densities obtained from different algorithms and the chi-square probability density. We can see that with second-order estimates, the estimated distributions are closer to the theoretical distribution, especially with WASN estimates.

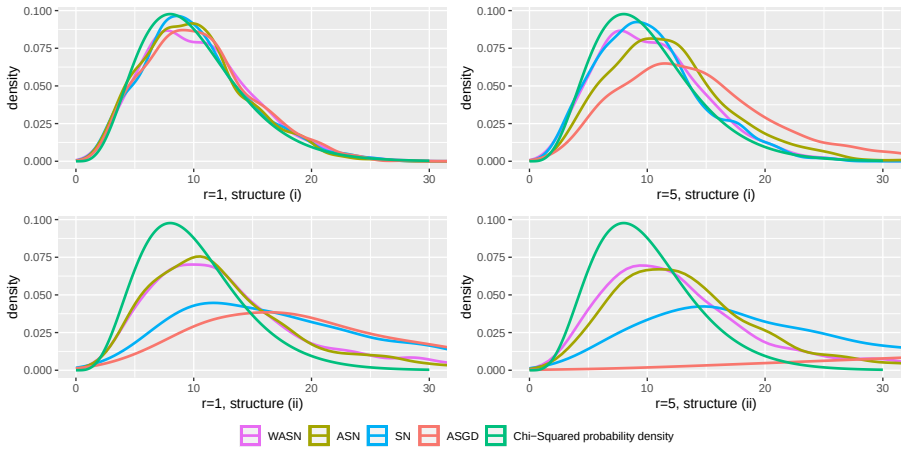

Figure 3: Simulated distributions of the test statistic and the theoretical distribution under H_0

5 Proofs

In the following, $\|\cdot\|$ indicates the Euclidean norm for vectors or the spectral norm for matrices.

5.1 Proof of Theorems 1 and 2

Remark that the proofs of Theorems 1 and 2 are very close. We therefore give the proof of Theorem 1 and just highlight differences with the help of remarks. Our objective is to apply Theorem 3.3 (or Theorem 4.3) in [3]. To do so, we are going to verify that the hypotheses given in [3], termed **(A1b)**, **(A1c)**, **(A2a)**, **(A2b)**, **(A2c)**, **(H1)**, **(H2a)** and **(H2b)** are satisfied.

5.1.1 Verification of conditions on the function to minimize

First we are going to verify the hypotheses that the function should be met.

Verification of (A1a). Assumption 1 ensures that the median m is the unique solution (see [14] and [6]) of the equation

$$\nabla G(h) = 0,$$

so that Hypothesis **(A1a)** is satisfied.

Verification of (A1b). Recall that $\nabla g(X,h) = -\frac{X-h}{\|X-h\|}$, so that for all $h \in \mathbb{R}^p$,

$$\|\nabla g(X,h)\| \leq 1,$$

Hypothesis **(A1b)** is then satisfied.

Verification of (A1c). We have for all $h \in \mathbb{R}^p$

$$\Sigma(h) = \mathbb{E}\left[\nabla g(X,h)\nabla g(X,h)^T\right] = \mathbb{E}\left[\frac{(X-h)(X-h)^T)}{\|X-h\|^2}\right].$$

The function Σ is continuous on $\mathbb{R}^p$, thus Hypothesis **(A1c)** is satisfied.

Verification of (A2a). For all $h \in \mathbb{R}^p$,

$$\left\|\nabla^2 G(h)\right\| \leq \mathbb{E}\left[\frac{1}{\|X-h\|}\left\|I_p - \frac{(X-h)(X-h)^T)}{\|X-h\|^2}\right\|\right] \leq \mathbb{E}\left[\frac{1}{\|X-h\|}\right],$$

and Assumption 1 ensures that

$$\mathbb{E}\left[\frac{1}{\|X-h\|}\right] \leq C_6^{\frac{1}{6}},$$

so that Hypothesis **(A2a)** is satisfied.

Verification of (A2b). Under Assumption 1 and Assumption 2, for all $h \in \mathbb{R}^p$ the Hessian $\nabla^2 G(h)$ is positive definite (see Section 2.2 in [6]), thus Hypothesis **(A2b)** is satisfied.

Verification of (A2c). Under Assumption 1, the Hessian $\nabla^2 G(h)$ is $6C_6^{\frac{1}{3}}$-Lipschitz (see Lemma 1), so that Hypothesis **(A2c)** is satisfied.

5.1.2 Controllability of eigenvalues of the Hessian estimator and consistency

Verification of (H1). We are going to verify if eigenvalues of the Hessian estimator are well controlled. We recall that

$$\widetilde{H}_n = \overline{H}_n + \frac{1}{n+1}\sum_{k=1}^{n} \tilde{\beta}_k Z_k Z_k^T,$$

with $\tilde{\beta}_k = \frac{c_\beta}{k^\beta}$ where $0 < \beta < \frac{1}{2}$ and $c_\beta > 0$.

Remark 5.1 *For WASN the condition on β should be $\beta < \gamma - \frac{1}{2}$ instead of $\beta < \frac{1}{2}$.*

Thus we have

$$\lambda_{min}(\tilde{H}_n) \geq \frac{\lambda_{min}(H_0)}{n+1} + \frac{1}{n+1}\lambda_{min}\left(\sum_{k=1}^{n} \tilde{\beta}_k Z_k Z_k^T\right).$$

In addition,

$$\left(\frac{c_\beta}{1-\beta} n^{1-\beta}\right)^{-1} \sum_{k=1}^{n} \tilde{\beta}_k Z_k Z_k^T \xrightarrow[n\to+\infty]{a.s.} I_p,$$

so that $\lambda_{max}\left(\widetilde{H}_n^{-1}\right) = \mathcal{O}\left(n^{\beta}\right)$. For the largest eigenvalue of $\overline{H}_n$, we have

$$\begin{aligned}
\left\|\overline{H}_n\right\| &\leq \frac{1}{n+1}\sum_{k=1}^{n}\frac{\|X_k - m_{k-1}\|}{\alpha_k^2}\|\phi_k\|^2 + \frac{1}{n+1}\|H_0\| \\
&= \frac{1}{n+1}\sum_{k=1}^{n}\frac{\|X_k - m_{k-1}\|}{\alpha_k^2}\|\nabla g(X_k, m_{k-1}) - \nabla g(X_k, m_{k-1} + \alpha_k Z_k)\|^2 + \frac{1}{n+1}\|H_0\| \\
&= \frac{1}{n+1}\sum_{k=1}^{n}\frac{\|X_k - m_{k-1}\|}{\alpha_k^2}\left\|\frac{X_k - m_{k-1}}{\|X_k - m_{k-1}\|} - \frac{X_k - (m_{k-1} + \alpha_k Z_k)}{\|X_k - (m_{k-1} + \alpha_k Z_k)\|}\right\|^2 + \frac{1}{n+1}\|H_0\|.
\end{aligned}$$

Since $\left\|\frac{C-A}{\|C-A\|} - \frac{B-A}{\|B-A\|}\right\| \leq 2\frac{\|C-B\|}{\|B-A\|}$ (see [5] page 25), we have

$$\begin{aligned}
\left\|\overline{H}_n\right\| &\leq \frac{4}{n+1}\sum_{k=1}^{n}\frac{\|X_k - m_{k-1}\|}{\alpha_k^2}\left\|\frac{\alpha_k Z_k}{\|X_k - m_{k-1}\|}\right\|^2 + \frac{1}{n+1}\|H_0\| \\
&= \frac{1}{n+1}\sum_{k=1}^{n}\frac{4\|Z_k\|^2}{\|X_k - m_{k-1}\|} + \frac{1}{n+1}\|H_0\|.
\end{aligned}$$

Thanks to Assumption 1, by independence between Z_k and X_k and by Hölder's inequality, we have

$$\mathbb{E}\left[\frac{4\|Z_k\|^2}{\|X_k - m_{k-1}\|}\,|\,\mathcal{F}_{k-1}\right] = 4\mathbb{E}\left[\|Z_k\|^2\right]\mathbb{E}\left[\|X_k - m_{k-1}\|^{-1}\right] \leq 4pC_6^{\frac{1}{6}},$$

so that

$$\frac{1}{n+1}\sum_{k=1}^{n}\mathbb{E}\left[\frac{4\|Z_k\|^2}{\|X_k - m_{k-1}\|}\,|\,\mathcal{F}_{k-1}\right] \leq 4pC_6^{\frac{1}{6}}.$$

Moreover, with analogous calculs, one has

$$\mathbb{E}\left[\left(\frac{4\|Z_k\|^2}{\|X_k - m_{k-1}\|}\right)^2\,|\,\mathcal{F}_{k-1}\right] = 16\mathbb{E}\left[\|Z_k\|^4\right]\mathbb{E}\left[\|X_k - m_{k-1}\|^{-2}\right] \leq 16p(p+2)C_6^{\frac{1}{3}},$$

With the help of law of large numbers for martingales, for all $\delta > 0$,

$$\left(\frac{1}{n}\sum_{k=1}^{n}\frac{4\|Z_k\|^2}{\|X_k - m_{k-1}\|} - \mathbb{E}\left[\frac{4\|Z_k\|^2}{\|X_k - m_{k-1}\|}\,|\,\mathcal{F}_{k-1}\right]\right)^2 = o\left(\frac{(\ln n)^{1+\delta}}{n}\right) a.s.$$

Thus,

$$\begin{aligned}
\|\overline{H}_n\| \leq{}& \left\|\frac{1}{n+1}\sum_{k=1}^{n}\mathbb{E}\left[\frac{4\|Z_k\|^2}{\|X_k - m_{k-1}\|}\,|\,\mathcal{F}_{k-1}\right]\right\| + \left\|\frac{1}{n+1}\sum_{k=1}^{n}\frac{4\|Z_k\|^2}{\|X_k - m_{k-1}\|} - \mathbb{E}\left[\frac{4\|Z_k\|^2}{\|X_k - m_{k-1}\|}\,|\,\mathcal{F}_{k-1}\right]\right\| \\
&+ \frac{1}{n+1}\|H_0\| \\
={}& \mathcal{O}(1)\ a.s.
\end{aligned}$$

Then,

$$\left\|\widetilde{H}_n\right\| \leq \left\|\overline{H}_n\right\| + \frac{1}{n+1}\sum_{k=1}^{n}\tilde{\beta}_k = \mathcal{O}(1)\ a.s.$$

The largest eigenvalue of $\widetilde{H}_n^{-1}$ and $\widetilde{H}_n$ can be controlled, according to Theorem 3.1 in [3], the stochastic Newton estimator satisfies

$$m_n \xrightarrow[n\to+\infty]{a.s.} m.$$

Remark 5.2 *For WASN, according to Theorem 4.1 in [3], the estimator $\hat{m}_n$ converges almost surely to m, which implies the almost sure convergence of $m_{n,\tau}$.*

5.1.3 Convergence of the Hessian estimator and rate of convergence

Verification of (H2a). We verify now if the Hessian estimator converges towards $\nabla^2 G(m)$. We define

$$X_{k,t} := X_k - (m_{k-1} + t\alpha_k Z_k),$$

and

$$w_{k,t} := \frac{1}{\|X_{k,t}\|}\left(I_p - \frac{X_{k,t}X_{k,t}^T}{\|X_{k,t}\|^2}\right).$$

We then have

$$\begin{aligned}
\overline{H}_n &= \frac{1}{n+1}\sum_{k=1}^{n}\frac{\|X_k - m_{k-1}\|}{\alpha_k^2}\int_0^1 w_{k,t}dt\alpha_k Z_k\alpha_k Z_k^T\int_0^1 w_{k,t}dt + \frac{1}{n+1}H_0 \\
&= \frac{1}{n+1}\sum_{k=1}^{n}\|X_k - m_{k-1}\|\int_0^1 w_{k,t}dtZ_kZ_k^T\int_0^1 w_{k,t}dt + \frac{1}{n+1}H_0 \\
&= \overbrace{\frac{1}{n+1}\sum_{k=1}^{n}\|X_k - m_{k-1}\|\int_0^1 w_{k,t} - w_{k,0}dtZ_kZ_k^T\int_0^1 w_{k,t}dt}^{\mathcal{M}_{1,n}} \\
&+ \overbrace{\frac{1}{n+1}\sum_{k=1}^{n}\|X_k - m_{k-1}\| w_{k,0}Z_kZ_k^T\int_0^1 w_{k,t} - w_{k,0}dt}^{\mathcal{M}_{2,n}} \\
&+ \overbrace{\frac{1}{n+1}\sum_{k=1}^{n}\|X_k - m_{k-1}\| w_{k,0}Z_kZ_k^T w_{k,0}}^{\mathcal{M}_{3,n}} + \frac{1}{n+1}H_0
\end{aligned}$$

Convergence of $\mathcal{M}_{3,n}$. We define

$$\mathcal{Y}_k := \|X_k - m_{k-1}\| w_{k,0}Z_kZ_k^T w_{k,0},$$

remark that one has

$$\mathcal{M}_{3,n} = \frac{1}{n+1}\sum_{k=1}^{n} \mathcal{Y}_k = \frac{1}{n+1}\sum_{k=1}^{n} \mathbb{E}\left[\mathcal{Y}_k \,|\, \mathcal{F}_{k-1}\right] + \frac{1}{n+1}\sum_{k=1}^{n} \mathcal{Y}_k - \mathbb{E}\left[\mathcal{Y}_k \,|\, \mathcal{F}_{k-1}\right].$$

First we prove that

$$\frac{1}{n+1}\sum_{k=1}^{n} \mathbb{E}\left[\mathcal{Y}_k \,|\, \mathcal{F}_{k-1}\right] \xrightarrow[n\to+\infty]{a.s.} \nabla^2 G(m).$$

We have

$$w_{k,0} := \frac{1}{\|X_k - m_{k-1}\|}\left(I_p - \left(\frac{X_k - m_{k-1}}{\|X_k - m_{k-1}\|}\right)\left(\frac{X_k - m_{k-1}}{\|X_k - m_{k-1}\|}\right)^T\right) = \nabla^2 g(X_k, m_{k-1}),$$

so that by equation (3)

$$\|X_k - m_{k-1}\|\, w_{k,0}^2 = w_{k,0}.$$

In addition, as the estimator of the median satisfies

$$m_n \xrightarrow[n\to+\infty]{a.s.} m,$$

we have by continuity

$$\mathbb{E}\left[w_{n,0} \,|\, \mathcal{F}_{n-1}\right] = \nabla^2 G(m_n) \xrightarrow[n\to+\infty]{a.s.} \nabla^2 G(m).$$

Therefore, as $\{Z_k\}_k$ are standard independent Gaussian vectors, by law of large numbers, we have

$$\begin{aligned}\frac{1}{n+1}\sum_{k=1}^{n} E\left[\mathcal{Y}_k \,|\, \mathcal{F}_{k-1}\right] &= \frac{1}{n+1}\sum_{k=1}^{n} E\left[\|X_k - m_{k-1}\|\, w_{k,0}^2 \,|\, \mathcal{F}_{k-1}\right] \\ &= \frac{1}{n+1}\sum_{k=1}^{n} E\left[w_{k,0} \,|\, \mathcal{F}_{k-1}\right] \xrightarrow[n\to+\infty]{a.s.} \nabla^2 G(m).\end{aligned}$$

Moreover, thanks to Assumption 1 and by independence,

$$\begin{aligned}\mathbb{E}\left[\|\mathcal{Y}_k\|^2 \mid \mathcal{F}_{k-1}\right] &\le \mathbb{E}\left[\|X_k - m_{k-1}\|^2 \|w_{k,0}\|^4 \|Z_k\|^4 \mid \mathcal{F}_{k-1}\right] \\ &\le \mathbb{E}\left[\|\mathcal{Z}_k\|^4 \mid \mathcal{F}_{k-1}\right] \mathbb{E}\left[\left\|\frac{1}{X_k - m_{k-1}}\right\|^2 \mid \mathcal{F}_{k-1}\right] \\ &\le p(p+2)C_6^{\frac{1}{3}},\end{aligned}$$

which results in, with the help of law of large numbers for martingales, that for all $\delta > 0$,

$$\left\|\frac{1}{n}\sum_{k=1}^{n} \mathcal{Y}_k - \mathbb{E}\left[\mathcal{Y}_k \,|\, \mathcal{F}_{k-1}\right]\right\|^2 = o\left(\frac{(\ln n)^{1+\delta}}{n}\right).$$

Thus,

$$\mathcal{M}_{3,n} \xrightarrow[n\to+\infty]{a.s.} \nabla^2 G(m).$$

Convergence of $\mathcal{M}_{2,n}$: In order to get the rate of convergence of $\mathcal{M}_{2,n}$, let us first introduce a generalization of Lemma 5.1 in [4].

Lemma 1 *For all $h, h' \in \mathbb{R}^p$ and $0 < q \leq 3$,*

$$\left(\mathbb{E}\left[\left\|\nabla^2 g(X,h') - \nabla^2 g(X,h)\right\|^q\right]\right)^{\frac{1}{q}} \leq 6C_6^{\frac{1}{3}} \left\|h' - h\right\|.$$

In our case, for all $t \in (0,1)$ and $q \in (0,3]$, and considering the filtration $\mathcal{F}'_k = \sigma\left(X_1, \ldots, X_{k-1}, Z_1, \ldots, Z_k\right)$, we have

$$\mathbb{E}\left[\|w_{k,t} - w_{k,0}\|^q \,|\mathcal{F}_k\right] \leq 6^q C_6^{\frac{q}{3}} \alpha_k^q \left\|Z_k\right\|^q.$$

We define

$$\mathcal{W}_k := \left\|X_k - m_{k-1}\right\| w_{k,0} Z_k Z_k^T \int_0^1 w_{k,t} - w_{k,0} dt.$$

Then,

$$\left\|\mathcal{M}_{2,n}\right\| \leq \frac{1}{n+1}\sum_{k=1}^n \mathbb{E}\left[\|\mathcal{W}_k\| \mid \mathcal{F}_{k-1}\right] + \frac{1}{n+1}\sum_{k=1}^n \|\mathcal{W}_k\| - \mathbb{E}\left[\|\mathcal{W}_k\| \mid \mathcal{F}_{k-1}\right].$$

Remark that

$$\begin{aligned}
\mathbb{E}\left[\|\mathcal{W}_k\| \mid \mathcal{F}_{k-1}\right] &\leq \mathbb{E}\left[\left\|X_k - m_{k-1}\right\| \|w_{k,0}\| \left\|Z_k\right\|^2 \left\|\int_0^1 w_{k,t} - w_{k,0}\right\| dt \,|\, \mathcal{F}_{k-1}\right] \\
&\leq \mathbb{E}\left[\left\|Z_k\right\|^2 \int_0^1 \|w_{k,t} - w_{k,0}\| \, dt \,|\, \mathcal{F}_{k-1}\right] \\
&= \mathbb{E}\left[\left\|Z_k\right\|^2 \int_0^1 \mathbb{E}\left[\|w_{k,t} - w_{k,0}\| \, dt \,|\, \mathcal{F}'_k\right] \mid \mathcal{F}_{k-1}\right],
\end{aligned}$$

where $\mathcal{F}_{k'} = \sigma\{X_1, ..., X_{k-1}, Z_1, ..., Z_k\}$. Therefore, thanks to Lemma 1, we have

$$\mathbb{E}\left[\|\mathcal{W}_k\| \mid \mathcal{F}_{k-1}\right] \leq 6\alpha_k C_6^{\frac{1}{3}} \mathbb{E}\left[\left\|Z_k\right\|^3 \mid \mathcal{F}_{k-1}\right].$$

Since $\alpha_k = \frac{1}{k \ln k + 1}$, it comes $\mathbb{E}\left[\|\mathcal{W}_k\| \mid \mathcal{F}_{k-1}\right] = \mathcal{O}\left(\frac{1}{k \ln k}\right)$, which leads to

$$\frac{1}{n+1}\sum_{k=1}^n \mathbb{E}\left[\|\mathcal{W}_k\| \mid \mathcal{F}_{k-1}\right] = \mathcal{O}\left(\frac{\ln n}{n}\right) a.s.$$

In addition, according to Lemma 1, and with the help of Hölder's inequality,

$$\begin{aligned}
\mathbb{E}\left[\|\mathcal{W}_k\|^2 \mid \mathcal{F}_{k-1}\right] &\leq \mathbb{E}\left[\left\|Z_k\right\|^4 \left(\int_0^1 \|w_{k,t} - w_{k,0}\| \, dt\right)^2 \mid \mathcal{F}_{k-1}\right]^2 \\
&\leq \mathbb{E}\left[\mathbb{E}\left[\left\|Z_k\right\|^4 \left(\int_0^1 \|w_{k,t} - w_{k,0}\| \, dt\right)^2 \mid \mathcal{F}'_k\right] \mid \mathcal{F}_{k-1}\right] \\
&\leq \mathbb{E}\left[\left\|Z_k\right\|^4 \int_0^1 \mathbb{E}\left[\|w_{k,t} - w_{k,0}\|^2 \,|\mathcal{F}'_k\right] dt | \mathcal{F}_{k-1}\right] \\
&\leq 36\alpha_k^2 \mathbb{E}\left[\left\|Z_k\right\|^6 \mid \mathcal{F}_{k-1}\right] C_6^{\frac{2}{3}}.
\end{aligned}$$

Thus, with the help of law of large numbers for martingales,

$$\left\| \frac{1}{n} \sum_{k=1}^{n} \mathcal{W}_k - \mathbb{E}\left[\|\mathcal{W}_k\| \mid \mathcal{F}_{k-1} \right] \right\|^2 = o\left(\frac{(\ln n)^{1+\delta}}{n} \right) a.s.$$

Therefore, we obtain

$$\mathcal{M}_{2,n} = o\left(\frac{(\ln n)^{1+\delta}}{n} \right) a.s.$$

Convergence of $\mathcal{M}_{1,n}$. We define

$$\mathcal{V}_k := \|X_k - m_{k-1}\| \int_0^1 w_{k,t} - w_{k,0} dt Z_k Z_k^T \int_0^1 w_{k,t} dt.$$

Remark that

$$\begin{aligned}
\mathbb{E}\left[\|\mathcal{V}_k\| \mid \mathcal{F}_{k-1} \right] &\leq \mathbb{E}\left[\|X_k - m_{k-1}\| \left\| \int_0^1 w_{k,t} dt \right\| \|Z_k\|^2 \left\| \int_0^1 w_{k,t} - w_{k,0} dt \right\| \mid \mathcal{F}_{k-1} \right] \\
&\leq \mathbb{E}\left[\|Z_k\|^2 \int_0^1 \frac{\|X_{k,t}\| + t\alpha_k \|Z_k\|}{\|X_{k,t}\|} dt \int_0^1 \|w_{k,t} - w_{k,0}\| \, dt \, | \, \mathcal{F}_{k-1} \right] \\
&\leq \mathbb{E}\left[\mathbb{E}\left[\|Z_k\|^2 \left(1 + \int_0^1 \frac{\alpha_k \|Z_k\|}{\|X_{k,t}\|} dt \right) \int_0^1 \|w_{k,t} - w_{k,0}\| \, dt \, | \, \mathcal{F}'_k \right] \mid \mathcal{F}_{k-1} \right]
\end{aligned}$$

where $\mathcal{F}'_k = \sigma\left(X_1, ..., X_{k-1}, Z_1, ..., Z_k\right)$. Thus, according to Lemma 1 and Assumption 1, one has with the help of Hölder's inequality

$$\begin{aligned}
\mathbb{E}\left[\|\mathcal{V}_k\| \mid \mathcal{F}_{k-1} \right] &\leq \mathbb{E}\left[\|Z_k\|^2 \int_0^1 \mathbb{E}\left[\|w_{k,t} - w_{k,0}\| \, | \mathcal{F}'_k \right] dt \right] \\
&\quad + \mathbb{E}\left[\alpha_k \|Z_k\|^3 \left(\int_0^1 \mathbb{E}\left[\frac{1}{\|X_{k,t}\|^2} | \mathcal{F}'_k \right] dt \right)^{\frac{1}{2}} \left(\int_0^1 \mathbb{E}\left[\|w_{k,t} - w_{k,0}\|^2 | \mathcal{F}'_k \right] dt \right)^{\frac{1}{2}} | \mathcal{F}_{k-1} \right] \\
&\leq 6\alpha_k C_6^{\frac{1}{3}} \mathbb{E}\left[\|Z_k\|^3 \mid \mathcal{F}_{k-1} \right] + 6\alpha_k^2 \mathbb{E}\left[\|Z_k\|^4 | \mathcal{F}_{k-1} \right] C_6^{\frac{1}{2}}.
\end{aligned}$$

We have $\alpha_k = \frac{1}{k \ln k + 1}$, so that $\mathbb{E}\left[\|\mathcal{V}_k\| \mid \mathcal{F}_{k-1} \right] = \mathcal{O}\left(\frac{1}{k \ln k} \right)$, which leads to

$$\frac{1}{n+1} \sum_{k=1}^{n} \mathbb{E}\left[\|\mathcal{V}_k\| \mid \mathcal{F}_{k-1} \right] = \mathcal{O}\left(\frac{\ln n}{n} \right).$$

Furthermore, we have by Hölder's inequality

$$\begin{aligned}
\mathbb{E}\left[\|\mathcal{V}_k\|^2 \mid \mathcal{F}_{k-1} \right] &\leq \mathbb{E}\left[\|Z_k\|^4 \left(1 + \int_0^1 \frac{\alpha_k \|Z_k\|}{\|X_{k,t}\|} dt \right)^2 \left(\int_0^1 \|w_{k,t} - w_{k,0}\| \, dt \right)^2 \mid \mathcal{F}_{k'} \mid \mathcal{F}_{k-1} \right] \\
&\leq 2\mathbb{E}\left[\|Z_k\|^4 \int_0^1 \|w_{k,t} - w_{k,0}\|^2 \, dt \, | \, \mathcal{F}_{k-1} \right] + 2\alpha_k^2 \mathbb{E}\left[\|Z_k\|^6 \int_0^1 \int_0^1 \frac{1}{\|X_{k,t}\|^2} \left\| w_{k,t'} - w_{k,0} \right\|^2 dt dt' \, | \, \mathcal{F}_{k-1} \right] =: (*)
\end{aligned}$$

and we therefore have, applying Hölder's inequality,

$$(*) \leq 2\mathbb{E}\left[\|Z_k\|^4 \int_0^1 \mathbb{E}\left[\|w_{k,t} - w_{k,0}\|^2 \,|\mathcal{F}'_k\right] dt \,|\, \mathcal{F}_{k-1}\right]$$
$$+ 2\alpha_k^2 \mathbb{E}\left[\|Z_k\|^6 \int_0^1 \int_0^1 \left(\mathbb{E}\left[\frac{1}{\|X_{k,t}\|^6} |\mathcal{F}'_k\right]\right)^{\frac{1}{3}} \left(\mathbb{E}\left[\left\|w_{k,t'} - w_{k,0}\right\|^3 |\mathcal{F}'_k\right]\right)^{\frac{2}{3}} dtdt' \,|\, \mathcal{F}_{k-1}\right].$$

Then, thanks to Assumption 1 and Lemma 1,

$$\mathbb{E}\left[\|\mathcal{V}_k\|^2 \,|\, \mathcal{F}_{k-1}\right] \leq 72\alpha_k^2 C_6^{2/3} \mathbb{E}\left[\|Z_k\|^6\right] + 72\alpha_k^4 C_6 \mathbb{E}\left[\|Z_k\|^8\right].$$

With the help of law of large numbers for martingales, one then has

$$\left\|\frac{1}{n}\sum_{k=1}^n \mathcal{V}_k - \mathbb{E}\left[\mathcal{V}_k \,|\, \mathcal{F}_{k-1}\right]\right\|^2 = o\left(\frac{(\ln n)^{1+\delta}}{n}\right) a.s$$

and

$$\mathcal{M}_{1,n} = o\left(\frac{(\ln n)^{1+\delta}}{n}\right) a.s.$$

Finally, we have

$$\overline{H}_n \xrightarrow[n\to+\infty]{a.s.} \nabla^2 G(m).$$

Notice that

$$\widetilde{H}_n = \overline{H}_n + \frac{1}{n+1}\sum_{k=1}^n \tilde{\beta}_k Z_k Z_k^T,$$

and

$$\frac{1}{n+1}\sum_{k=1}^n \tilde{\beta}_k Z_k Z_k^T \xrightarrow[n\to+\infty]{a.s.} 0.$$

Therefore,

$$\widetilde{H}_n \xrightarrow[n\to+\infty]{a.s.} \nabla^2 G(m).$$

According to Theorem 3.2 in [3], the stochastic Newton estimator satisfies

$$\|m_n - m\|^2 = \mathcal{O}\left(\frac{\ln n}{n}\right) a.s.$$

Remark 5.3 *For WASN, according to Theorem 4.2 in [3], we have*

$$\|\hat{m}_n - m\|^2 = \mathcal{O}\left(\frac{\ln n}{n^\gamma}\right) a.s., \quad \textit{which implies that} \quad \|m_{n,\tau} - m\|^2 = \mathcal{O}\left(\frac{\ln n}{n^\gamma}\right) a.s.$$

5.1.4 Rate of convergence of the Hessian estimator and asymptotic efficiency

We now give the rate of convergence of $\widetilde{H}_n$. We recall that

$$w_{k,0} := \frac{1}{\|X_k - m_{k-1}\|}\left(I_p - \left(\frac{X_k - m_{k-1}}{\|X_k - m_{k-1}\|}\right)\left(\frac{X_k - m_{k-1}}{\|X_k - m_{k-1}\|}\right)^T\right),$$

which means that

$$\mathbb{E}\left[w_{k,0} \,|\, \mathcal{F}_{k-1}\right] = \nabla^2 G(m_{k-1}).$$

Note that $\nabla^2 G(h)$ is $6C_6^{\frac{1}{3}}$-Lipschitz, so that we have

$$\mathbb{E}\left[\left\|w_{k,0} - \nabla^2 G(m)\right\|^2 \,|\, \mathcal{F}_{k-1}\right] = \left\|\nabla^2 G(m_{k-1}) - \nabla^2 G(m)\right\|^2 \leq 6C_6^{\frac{1}{3}} \left\|m_{k-1} - m\right\|^2.$$

As the estimator satisfies

$$\|m_n - m\|^2 = \mathcal{O}\left(\frac{\ln n}{n}\right) a.s.,$$

we have

$$\left\|\mathbb{E}\left[w_{k,0} \,|\, \mathcal{F}_{k-1}\right] - \nabla^2 G(m)\right\|^2 = \mathcal{O}\left(\frac{\ln k}{k}\right) a.s.$$

Remark 5.4 *For WASN, here we have* $\|m_{n,\tau} - m\|^2 = \mathcal{O}\left(\frac{\ln n}{n^\gamma}\right) a.s.$*, so that* $\left\|\mathbb{E}\left[w_{k,0} \,|\, \mathcal{F}_{k-1}\right] - \nabla^2 G(m)\right\|^2 = \mathcal{O}\left(\frac{\ln k}{k^\gamma}\right) a.s.$

Therefore, as $\{Z_k\}_k$ are standard independent Gaussian vectors, we have

$$\left\|\frac{1}{n+1}\sum_{k=1}^{n} E\left[\mathcal{Y}_k \,|\, \mathcal{F}_{k-1}\right] - \nabla^2 G(m)\right\|^2 = \mathcal{O}\left(\frac{\ln n}{n}\right) a.s.$$

We have proved that for all $\delta > 0$

$$\left\|\frac{1}{n}\sum_{k=1}^{n} \mathcal{Y}_k - \mathbb{E}\left[\mathcal{Y}_k \,|\, \mathcal{F}_{k-1}\right]\right\|^2 = o\left(\frac{(\ln n)^{1+\delta}}{n}\right) a.s.,$$

so that

$$\left\|\mathcal{M}_{3,n} - \nabla^2 G(m)\right\|^2 = o\left(\frac{(\ln n)^{1+\delta}}{n}\right) a.s.$$

Finally, the Hessian estimator satisfies for all $\delta > 0$

$$\left\|\widetilde{H}_n - H\right\|^2 = \mathcal{O}\left(\max\left\{\frac{(\ln n)^{1+\delta}}{n}, \frac{c_\beta}{n^{2\beta}}\right\}\right) a.s.$$

According to Theorem 3.3 in [3], the stochastic Newton estimator satisfies

$$\sqrt{n}\left(m_n - m\right) \sim \mathcal{N}\left(0, H^{-1}\Sigma H^{-1}\right),$$

where $\Sigma = \mathbb{E}\left[\nabla g(X, m) \nabla g(X, m)^T\right]$.

Remark 5.5 *For WASN, we have first for all* $\delta > 0$

$$\left\| \widetilde{H}_{n,\tau} - H \right\|^2 = \mathcal{O}\left(\max\left\{ \frac{(\ln n)^{1+\delta}}{n^{\gamma}}, \frac{c_\beta}{n^{2\beta}} \right\} \right) a.s.$$

Then according to Theorem 4.3 in [3], we have

$$\left\| m_{n,\tau} - m \right\|^2 = \mathcal{O}\left(\frac{\ln n}{n} \right) a.s.,$$

which results in

$$\left\| \widetilde{H}_{n,\tau} - H \right\|^2 = \mathcal{O}\left(\max\left\{ \frac{(\ln n)^{1+\delta}}{n}, \frac{c_\beta}{n^{2\beta}} \right\} \right) a.s. \quad \textit{and} \quad \sqrt{n}\left(m_{n,\tau} - m \right) \sim \mathcal{N}\left(0, H^{-1}\Sigma H^{-1} \right).$$

5.2 Proof of Theorem 3

We define

$$\mathcal{T}_k := \frac{(X_k - \widetilde{m}_{k-1})}{\|X_k - \widetilde{m}_{k-1}\|} \frac{(X_k - \widetilde{m}_{k-1})^T}{\|X_k - \widetilde{m}_{k-1}\|},$$

then one has

$$\Sigma_n = \frac{1}{n+1}\sum_{k=1}^{n} \mathcal{T}_k + \frac{1}{n+1}\Sigma_0 = \frac{1}{n+1}\sum_{k=1}^{n} \mathbb{E}\left[\mathcal{T}_k \,|\, \mathcal{F}_{k-1} \right] + \frac{1}{n+1}\sum_{k=1}^{n} \mathcal{T}_k - \mathbb{E}\left[\mathcal{T}_k \,|\, \mathcal{F}_{k-1} \right] + \frac{1}{n+1}\Sigma_0.$$

Note that $\mathbb{E}\left[\mathcal{T}_k \,|\, \mathcal{F}_{k-1} \right] = \Sigma(\widetilde{m}_{k-1})$. In addition, we have

$$\Sigma(h) = \mathbb{E}\left[\nabla g(X,h) \nabla g(X,h)^T \right]$$

Thus, thanks to Hypothesis **(A2c)**, $\Sigma(h)$ is $6C_6^{\frac{1}{6}}$-Lipschitz (see [11] section 6.2), which means

$$\left\| \mathbb{E}\left[\mathcal{T}_k \,|\, \mathcal{F}_{k-1} \right] - \Sigma \right\|^2 = \left\| \Sigma(\widetilde{m}_{k-1}) - \Sigma(m) \right\|^2 \leq 6C_6^{\frac{1}{6}} \left\| \widetilde{m}_{k-1} - m \right\|^2.$$

As the estimator $\widetilde{m}_k$ satisfies

$$\left\| \widetilde{m}_k - m \right\|^2 = \mathcal{O}\left(\frac{\ln n}{n} \right) a.s.,$$

we obtain

$$\left\| \mathbb{E}\left[\mathcal{T}_k \,|\, \mathcal{F}_{k-1} \right] - \Sigma \right\|^2 = \mathcal{O}\left(\frac{\ln n}{n} \right) a.s.$$

Moreover, it is obvious that

$$\mathbb{E}\left[\left\| \mathcal{T}_k \right\|^2 \,|\, \mathcal{F}_{k-1} \right] \leq 1,$$

which leads to, with the help of law of large numbers for martingales,

$$\left\| \frac{1}{n}\sum_{k=1}^{n} \mathcal{T}_k - \mathbb{E}\left[\mathcal{T}_k \,|\, \mathcal{F}_{k-1} \right] \right\|^2 = o\left(\frac{(\ln n)^{1+\delta}}{n} \right) a.s.$$

Finally we have

$$\|\Sigma_n - \Sigma\|^2 \leq \left\| \frac{1}{n+1} \sum_{k=1}^{n} \mathcal{T}_k - \mathbb{E}\left[\mathcal{T}_k \,|\, \mathcal{F}_{k-1}\right] \right\|^2 + \left\| \frac{1}{n+1} \sum_{k=1}^{n} \Sigma_n - \mathbb{E}\left[\mathcal{T}_k \,|\, \mathcal{F}_{k-1}\right] \right\|^2 + \left\| \frac{1}{n+1} \Sigma_0 \right\|^2$$
$$= o\left(\frac{(\ln n)^{1+\delta}}{n} \right) a.s.$$

References

[1] Beck, A. and Sabach, S. (2015). Weiszfeld's method: Old and new results. *Journal of Optimization Theory and Applications*, 164(1):1–40.

[2] Bercu, B., Godichon, A., and Portier, B. (2020). An efficient stochastic newton algorithm for parameter estimation in logistic regressions. *SIAM Journal on Control and Optimization*, 58(1):348–367.

[3] Boyer, C. and Godichon-Baggioni, A. (2022). On the asymptotic rate of convergence of stochastic newton algorithms and their weighted averaged versions. *Computational Optimization and Applications*, pages 1–52.

[4] Cardot, H., Cénac, P., and Godichon-Baggioni, A. (2017). Online estimation of the geometric median in hilbert spaces: Nonasymptotic confidence balls. *The Annals of Statistics*, 45(2):591–614.

[5] Cardot, H., Cénac, P., and Monnez, J.-M. (2012). A fast and recursive algorithm for clustering large datasets with k-medians. *Computational Statistics & Data Analysis*, 56(6):1434–1449.

[6] Cardot, H., Cénac, P., and Zitt, P.-A. (2013). Efficient and fast estimation of the geometric median in hilbert spaces with an averaged stochastic gradient algorithm. *Bernoulli*, 19(1):18–43.

[7] Church, J. C., Chen, Y., and Rice, S. V. (2008). A spatial median filter for noise removal in digital images. In *IEEE SoutheastCon 2008*, pages 618–623. IEEE.

[8] Duflo, M. (1997). Random iterative models, stochastic modelling and applied probability, vol. 34.

[9] Gini, C. and Galvani, L. (1929). Di talune estensioni dei concetti di media ai caratteri qualitativi. *Metron*, 8(1-2):3–209.

[10] Godichon-Baggioni, A. (2016). Estimating the geometric median in hilbert spaces with stochastic gradient algorithms: Lp and almost sure rates of convergence. *Journal of Multivariate Analysis*, 146:209–222.

[11] Godichon-Baggioni, A. (2019). Online estimation of the asymptotic variance for averaged stochastic gradient algorithms. *Journal of Statistical Planning and Inference*, 203:1–19.

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

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

[14] Kemperman, J. (1987). The median of a finite measure on a banach space. *Statistical data analysis based on the L1-norm and related methods (Neuchâtel, 1987)*, pages 217–230.

[15] Koltchinskii, V. I. (1997). M-estimation, convexity and quantiles. *The annals of Statistics*, pages 435–477.

[16] Lopuhaa, H. P. and Rousseeuw, P. J. (1991). Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *The Annals of Statistics*, pages 229–248.

[17] Lu, X., Chen, W., and Schaefer, S. (2017). Robust mesh denoising via vertex pre-filtering and l1-median normal filtering. *Computer Aided Geometric Design*, 54:49–60.

[18] Vardi, Y. and Zhang, C.-H. (2000). The multivariate l 1-median and associated data depth. *Proceedings of the National Academy of Sciences*, 97(4):1423–1426.

[19] Weiszfeld, E. (1937). Sur le point pour lequel la somme des distances de n points donnés est minimum. *Tohoku Mathematical Journal, First Series*, 43:355–386.

[20] Zhang, Y., Wang, X., Li, Q., and Yan, F. (2021). A new section line extraction method of ring forgings based on normal vector and l1-median. *Measurement*, 184:109894.