SORBONNE UNIVERSITÉ
CRÉATEURS DE FUTURS DEPUIS 1257

LPSM
Laboratoire de probabilités statistique & modélisation

HABILITATION À DIRIGER DES RECHERCHES

Discipline : Mathématiques
Spécialité : Statistique

présentée par

Antoine Godichon-Baggioni

Online stochastic algorithms and applications

Au vu des rapports établis par Jérémie Bigot, Sébastien Gadat
et Irène Gijbels

Soutenue le 14 mars 2023 devant le Jury composé de:

Gérard BIAU	Sorbonne Université	Président du Jury
Jérémie BIGOT	Université de Bordeaux	Rapporteur
Yohann DE CASTRO	Ecole Centrale de Lyon	Examinateur
Sébastien GADAT	Université Toulouse I Capitole	Rapporteur
Irène GIJBELS	Katholieke Universiteit Leuven	Rapportrice
Davy PAINDAVEINE	Université Libre de Bruxelles	Examinateur

Remerciements

Malgré le stress que cela génère, j'ai pris énormément de plaisir à rédiger ce manuscrit, et mes premiers remerciements vont donc à mes rapporteurs qui ont eu la gentillesse et surtout la patience de le lire. Un grand merci pour tous vos commentaires constructifs. Je remercie également Gérard Biau, Yohann De Castro et Davy Paindaveine qui ont accepté de faire partie de mon jury.

L'habilitation est censée signifier que l'on a atteint la maturité (scientifique) nécessaire pour pouvoir s'occuper seul d'un enfant scientifique (doctorant-e). Avant d'atteindre cette maturité, le chemin n'a pas été de tout repos, et je tiens à remercier Peggy pour tout ce qu'elle a fait pour moi. Un grand merci également à Hervé de m'avoir pris en thèse à l'époque et pour toute l'aide apportée depuis. Je profite de ce paragraphe au sujet de ma naissance scientifique pour remercier Bruno Portier, qui a été en quelque sorte un parrain scientifique, pour tous ces nombreux (bons ?) conseils et pour tout le travail effectué ensemble, ainsi que la co-paternité de mon deuxième enfant. Pour finir de dresser le tableau familial, je remercie Olivier Wintenberger pour la co-paternité de mon premier enfant.

Je voudrais également remercier toutes les autres personnes avec qui j'ai eu la chance de collaborer, en commençant par Cathy Maugis-Rabusseau et Andrea Rau grâce à qui j'ai pu découvrir le monde merveilleux du code. Je remercie également Sofiane Saadane, Bernard Bercu, Claire Boyer, Nicklas Werge, Olivier Wintenberger, Sobihan Surendran, Stéphane Robin, Wei Lu et Pierre Tarrago d'avoir supporté mes errements, approximations,... lors de nos collaborations.

Ce n'est un secret pour personne que je suis arrivé à Paris à reculons. Malgré cela, j'ai eu la chance de rencontrer de nombreux collègues sympathiques au sein du LPSM. Je pense tout d'abord au club des cruciverbistes (Anna, Anna, Pierre, Arnaud, Claire, Sylvain, Stéphane, Erwan (désolé si j'en ai oublié)). A cela, j'ajouterai l'équipe Scientifique (qu'il va falloir renommer au vu des trop nombreuses infidélités faites aux Sciences). Je ne prendrai pas la peine de nommer cette équipe car il me semble qu'elle est sensiblement la même que la précédente, avec quelques greffes ponctuelles. Je citerai également les équipes RU et Salle Café qui me permettent de mentionner Maxime, Etienne, les (post-)doctorants et ATER présents et passés, collègues de passage (désolé vous êtes trop nombreux, je préfère ne pas essayer de vous citer tous). Pour finir, un grand merci aux Béotiens pour votre patience, votre aide, et tous les moments passés ensemble. Un grand merci à toutes les personnes sus-mentionnées (ainsi que toutes celles que j'ai oublié) de supporter mon caractère doux et délicat, et de m'avoir aidé à progressivement apprécier ma vie au laboratoire et

à Paris.

Le dernier paragraphe de ces remerciements est dédié à ma famille (génétique, celle que l'on ne choisit pas). Un grand merci à mes parents pour les choix et sacrifices qu'ils ont faits et qui m'ont amené jusque là. Une pensée également pour ma sœur et mon frère, qui ont bénéficié des allocations familiales obtenues grâce à ma naissance, mais ont également hérité de mon humeur radieuse durant toutes ces années. Un grand merci à Simin pour sa patience, de m'avoir supporté et aidé durant toute cette année, qui n'a clairement pas été reposante ("qu'est-ce que tu veux, mener deux carrières sportives et une carrière professionnelle de front, c'est épuisant"). Je finirai par une grosse pensée pour mon village et tous mes proches.

Contents

Publications

Papers accepted in peer-reviewed journals

- Godichon-Baggioni, A. and Saadane, S. (2020): On the rates of convergence of Parallelized Averaged Stochastic Gradient Algorithms, *Statistics*
- Bercu, B., Godichon-Baggioni, A., Portier, B. (2020): An efficient stochastic Newton algorithm for parameter estimation in logistic regressions, *SIAM, Journal on Control and Optimization*
- Godichon-Baggioni, A., C. Maugis-Rabusseau, C., Rau, A. (2020): Multi-view cluster agregation and splitting with an application to multi-omic breast cancer data, *Annals of Applied Statistics*
- Godichon-Baggioni, A. (2019): Lp and almost sure rates of convergence of averaged stochastic gradient algorithms: locally strongly convex objective, *ESAIM PS*
- Godichon-Baggioni, A. (2019): Online estimation of the asymptotic variance for averaged stochastic gradient algorithms, *Journal of Statistical Planning and Inference*
- Godichon-Baggioni, A., Maugis-Rabusseau, C. and Rau, A. (2018): Clustering transformed compositional data using K-means, with applications in gene expression and bicyle sharing system data, *Journal of Applied Statistics*
- Godichon-Baggioni, A. and Portier, B. (2017): An averaged projected Robbins-Monro algorithm for estimating the parameters of a truncated spherical distribution, *Electronic Journal of Statistics*
- Cardot, H., Godichon-Baggioni, A. (2017): Fast Estimation of the Median Covariation Matrix with Application to Online Robust Principal Components Analysis, *TEST*
- Cardot, H., Cénac, P., Godichon-Baggioni, A. (2017): Online estimation of the geometric median in Hilbert spaces : non asymptotic confidence balls, *The Annals of Statistics*
- 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*

Submitted papers

- Godichon-Baggioni, A., Lu, W. and Portier, B.: Recursive ridge regression using second-order stochastic algorithms
- Godichon-Baggioni, A. and Robin, S.: A robust model-based clustering based on the geometric median and the Median Covariation Matrix
- Godichon-Baggioni, A. and Surendran, S.: A penalized criterion for selecting the number of clusters for K-medians
- Godichon-Baggioni, A., Werge, N. and Wintenberger, O.: Learning from time-dependent streaming data with online stochastic algorithms
- Godichon-Baggioni, A., Werge, N. and Wintenberger, O.: Non Asymptotic Analysis of Stochastic Approximation Algorithms for Streaming Data
- Godichon-Baggioni, A.: Convergence in quadratic mean of averaged stochastic gradient algorithms without strong convexity nor bounded gradient
- Cénac, P., Godichon-Baggioni, A. and Portier, B.: An efficient Averaged Stochastic Gauss-Newtwon algorithm for estimating parameters of non linear regressions models
- Boyer, C. and Godichon-Baggioni, A.: On the asymptotic rate of convergence of Stochastic Newton algorithms and their Weighted Averaged versions

R packages

- RGMM: Algorithms for estimating robustly the parameters of a Gaussian, Student, or Laplace Mixture Model, https://cran.r-project.org/web/packages/RGMM/index.html
- Kmedians: K-medians algorithms, https://cran.r-project.org/web/packages/Kmedians/index.html
- maskmeans: Multi-view aggregation/splitting K-means clustering algorithm, https://github.com/andreamrau/maskmeans
- coseq: Co-expression analysis of sequencing data, https://bioconductor.org/packages/coseq

Introduction

This manuscript is based on most of my research on online stochastic optimization and its applications to robust statistics. It is composed of five chapters that are described quickly there.

Chapter 1: Stochastic Gradient algorithms

A usual stochastic optimization problem, encountered for estimating the parameters of logistic regression [Bac14, CNS17], the geometric median and quantiles [CCZ13, GB16a, CCGB15], or superquantiles [CG20, BCG20] for instance, is to estimate the minimizer of a convex function $G : \mathcal{H} \longrightarrow \mathbb{R}$ of the form

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

where X is a random variable and $\mathcal{H}$ is a separable Hilbert space. A regular method is to approximate the minimizer of the empirical function generated by a sample with the help of deterministic optimization methods. Nevertheless, it often necessitates high computational costs if we deal with large samples taking values in high dimensional spaces. Then, stasticians have studied more and more mini-batch alternatives [AHA+20, KLRT15]. In any case, this kind of method necessitates to store all the data into memory and do not enable to easily update the estimates if the data arrive sequentially or in a streaming set. In order to overcome this, we focus in this chapter on online stochastic gradient algorithms that have been introduced by [RM51]. These algorithms have become hegemonic by a low computational cost per iteration, they allow performing machine learning tasks on large datasets, processing each observation only once (see [BCN18, Pel98, BM13, GLQ+19, Bac14, GP17, NJLS09, JN+14, NND+18]). In Chapter 1, we will focus on the obtaining of theoretical guarantees such that almost sure and L^2 rates of convergence under weak assumptions in possibly infinite dimensional spaces. All the theoretical results are illustrated on three applications: the estimation of the parameter of linear and logistic regressions as well as the estimation of p-means.

Chapter 2: Averaged Stochastic Gradient algorithms

Most of the time, it is almost impossible for stochastic gradient estimates to achieve the usual rate of convergence in quadratic mean $\frac{1}{n}$ (where n is the sample size). Worse, the estimates are not

asymptotically efficient. Anyway, a usual way to accelerate the convergence of gradient estimates has been introduce by [Rup88] and [PJ92]. This consists in considering the averaged stochastic gradient algorithm, i.e it consists in taking the averaging of all the estimates obtained with the help of the stochastic gradient algorithm. Remark that here again, these estimates have been deeply studied these last decades (see [Pel00, BM13, Bac14, GP17] for instance). In Chapter 2, we go on the theoretical study of online estimates in possibly infinite dimensional spaces and give a weak framework for each the averaged estimates are asymptotically efficient and for each we are able to uniformly bound the quadratic mean errors of the estimates. Here again, all the theoretical results are illustrated through three applications: linear and logistic regressions as well as the estimation of p-means.

Chapter 3: Online Stochastic Newton algorithms

The averaged stochastic gradient estimates are known to be asymptotically efficient [PJ92, Pel00, GB17] and to achieve, under mild assumptions, the Cramer-Rao bound (up to rest terms) [GP17, BM13]. However, these first-order online algorithms can be shown, in practice, to be very sensitive to the Hessian structure of the risk they are supposed to minimize [BGBP19, LP20, BGB20]. To address this issue, (quasi) online second-order optimization has been also considered in the literature (see [DHS11, Zei12, BHNS16, LP20] for instance). In Chapter 3, we consider a unified and general framework that includes various applications of machine learning tasks, for which we propose a stochastic Newton algorithm as well as a weighted averaged version. In addition, one the main problem for online Newton methods is to propose online estimates of the inverse of the Hessian, and we will see all along Chapter 3, through examples (linear logistic and softmax regressions), how the estimates of the Hessian can be constructed and updated over iterations using genuine second-order information.

Chapter 4: Stochastic Streaming Gradient algorithms

Although averaged stochastic gradient/Newton algorithms are known to be asymptotically efficient, the studied framework cannot be directly applied to the case where the data are not independent and/or identically distributed. In order to overcome this, we focus in Chapter 4 on streaming methods. More precisely, we consider data arriving sequentially by (non independent) blocs and introduce new Stochastic Streaming Gradient algorithms and their averaged version [GBWW21, GBWW22]. We then give a framework where the data are not supposed to be independent nor identically distributed and prove that under conditions, the Averaged Stochastic Streaming Gradient estimates achieve the Cramer-Rao bound.

Chapter 5: Applications to robust statistics

The acquisition of massive data lying in high dimensional spaces is unfortunately often accompanied by a contamination of these last ones. In this context of contaminated data, even few individuals may corrupt simple statistical indicators such as the mean or the variance. Detecting these atypical data automatically is not straightforward and considering robust techniques is an interesting alternative [Sma90, RL05, FM01, CFF07].

In Chapter 5, we first focus on the geometric median (also called L^1-median or spatial median) introduced by [Hal48]. Several iterative methods based on Weiszfeld algorithm [Wei37] have been developped [VZ00]. Nevertheless, for all the reasons mentioned above, we focus on the online estimates of the median obtained with the help of an averaged stochastic gradient algorithm [CCZ13]. We then give an example of application to unsupervised robust clustering. One of the most usual method for hard clustering is probably the K-means algorithm [For65, Mac67], and one can refer to [CAGM97, GEG99] for the robust version obtained with the help of Trimmed K-means. In Chapter 5, we focus on K-medians algorithms [Mac67, KR09, CCM12], and more precisely, we propose a method for selecting the number of clusters based on a penalized criterion [Fis11] whose penalty is calibrated with the help of a slope heuristic [BMM12, AM09].

Finally, we focus on the recursive estimation of the Median Covariation Matrix (MCM), which is a new robust dispersion indicator [KP12, CGB15], and its applications to online robust Principal Components analysis (PCA) and robust mixture models. PCA is one of the most useful statistical tool to extract information by reducing the dimension when one has to analyze large samples of multivariate data [Jol02, RS05, Ver06, HPV14]. Nevertheless, principal components, which are derived from the spectral analysis of the covariance matrix, can be very sensitive to outliers and many robust procedures for principal components analysis have been considered in the literature (see [HRVA08, HR09, Ger08] among others). We focus here on a new approach based on the MCM, which has, under conditions [KP12], the same eigenvectors as the usual covariance matrix. Finally, in the case where the law of the sample is known, one can rebuild robustly the covariance matrix from the estimates of the MCM [GBR22], and this approach is so applied to the development of robust methods for model based clustering. This represents an interesting alternative to usual robust methods which often necessitates to modelize the contamination (see [BR93, CH16, CH17, FP20] for instance).

Chapter 1

Stochastic Gradient algorithms

This chapter is based on [GB16b, GB17, GB21].

Contents

1.1 Introduction

A usual stochastic optimization problem is to estimate the minimizer of a convex function $G : \mathcal{H} \longrightarrow \mathbb{R}$ of the form

$$G(h) = \mathbb{E}\left[g(X,h)\right] \tag{1.1}$$

where X is a random variable. This problem is encountered, for instance, for estimating the parameters of logistic regression [Bac14, CNS17], the geometric median and quantiles [CCZ13, GB16a, CCGB15], or superquantiles [CG20, BCG20]. Nevertheless, since most of the time it is not possible to calculate explicitly the gradient or the Hessian of G, one cannot apply usual optimization

methods such that gradient or Newton algorithms amoung others [BV04, DGN14, N+18, NNG19]. Then, a solution is to consider a sample $X_1, \ldots, X_n$ and consider the empirical function

$$G_n(h) = \frac{1}{n} \sum_{k=1}^{n} g(X_k, h)$$

as well as its minimizer that we will denote $\hat{m}_n$. Even if in some cases, such that linear regression or the estimation of the mean, we can explicitly calculate $\hat{m}_n$, it is not possible in most of cases. Then, a solution is to approximate $\hat{m}_n$ with the help of usual optimization methods. Nevertheless, it often necessitates high computational costs if we deal with large samples taking values in high dimensional spaces. One solution to reduce the calculation time is to consider iterative mini-batch gradient algorithms of the form

$$m_{t+1} = m_t - \gamma_t \sum_{i \in S_t} \nabla_h g(X_i, m_t)$$

where $S_t \subset \{1, \ldots, n\}$ is the mini-batch considered at time t [AHA+20, KLRT15]. Nevertheless, this kind of methods necessitates to store all the data into memory and do not enable to easily update the estimates if the data arrive sequentially or in a streaming set. In order to overcome this, we focus in this chapter on the online stochastic gradient algorithms.

Stochastic gradient algorithms have been introduced by [RM51] and are more and more studied nowadays. It is hardly ever possible to cite all the recent results, but we focus particularly on the almost sure rates of convergence obtained by [Pel98] in the case where $\mathcal{H} = \mathbb{R}^d$. Always in a finite dimensional set, non asymptotic rate of convergence of stochastic gradient estimates were given in the strongly convex case [BM13, GLQ+19]. Nevertheless, the loss of strong convexity leads the results to be harder to obtain. In recent work, [Bac14] and [GP17] succeeded in obtaining the rate of convergence in quadratic mean of the estimates without supposing G to be strongly convex, but supposing that the gradient of g admits exponential moments or is bounded. Remark that we will often refer to the aforementioned papers for non asymptotic rates of convergence, but several other results exist in the literature [NJLS09, JN+14, BCN18, NND+18].

Observe that we decide here to focus on the original stochastic gradient algorithm but it is no less important to mention that several improvements of these estimates have been introduced [BCN18, Rud16]. For instance, momentum methods have been introduced to give more weights for coordinates whose gradients point in the same direction, and so reduce oscillations [Qia99, LR20]. In addition, the Nesterov acceleration method is a modification of the momentum method which allows to take into account an anticipation of the next step of the algorithm [MJ19, EBB+21]. Finally, several methods have been developed to try to adapt the stepsequences to the different corrdinates [DHS11, Zei12, LP20, LVLLJ21, KB14].

In this chapter, we first ensure that all the asymptotic results (almost sure rates of convergence and convergence in law) given by [Pel98] remain true even if $\mathcal{H}$ is not of finite dimension. In a second time, we will focus on the obtaining of explicit upper bounds of the quadratic mean

error of stochastic gradient estimates, and so, under weak assumptions, i.e without supposing that G is strongly convex nor supposing that the gradient of g is uniformly bounded. Finally, quick results on the L^p rates of convergence will be given. All the theoretical results are illustrated on three applications: the estimation of the parameter of linear and logistic regressions as well as the estimation of p-means.

1.2 Definition and framework

In what follows, we consider a separable Hilbert space $\mathcal{H}$ (not necessarily of finite dimension) and we denote by $\|.\|$ the euclidean norm and by $\langle .,. \rangle$ the associated inner product. Let us recall that the aim of this chapter is to estimate m, where m is the minimizer of the function $G : \mathcal{H} \longrightarrow \mathbb{R}$ defined for all $h \in \mathcal{H}$ by

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

with $g : \mathcal{X} \times \mathcal{H} \longrightarrow \mathbb{R}$, where X is a random variable lying in a measurable space $\mathcal{X}$. In the sequel, we suppose that for almost every $x \in \mathcal{X}$, the functional $g(x,.)$ is differentiable. Furthermore, we consider i.i.d random variables $X_1, \ldots, X_n, X_{n+1}, \ldots$ with the same law as X and arriving sequentially. The stochastic gradient algorithm is defined recursively for all $n \geq 0$ by [RM51]

$$m_{n+1} = m_n - \gamma_{n+1} \nabla_h g\left(X_{n+1}, m_n\right) \tag{1.2}$$

where $\nabla_h g\left(X_{n+1},.\right)$ is the gradient of g with respect to the second variable, and γ_n is a positive step sequence satisfying

$$\sum_{n\geq 0} \gamma_{n+1} = +\infty \qquad \text{and} \qquad \sum_{n\geq 0} \gamma_{n+1}^2 < +\infty.$$

Remark that it necessitates few operations to update the estimates. Furthermore, the algorithm can also be written as

$$m_{n+1} = m_n - \gamma_{n+1} \nabla G\left(m_n\right) + \gamma_{n+1} \xi_{n+1} \tag{1.3}$$

where $\xi_{n+1} := \nabla G\left(m_n\right) - \nabla_h g\left(X_{n+1}, \theta_n\right)$. Considering the filtration $\left(\mathcal{F}_n\right)_{n\geq 0}$ generated by the sample, one has, since m_n is $\mathcal{F}_n$-measurable, that (ξ_n) is a sequence of martingale differences adapted to $(\mathcal{F}_n)$, i.e $\mathbb{E}\left[\xi_{n+1} | \mathcal{F}_n\right] = 0$. Then, this online algorithm can be seen as a noisy gradient algorithm.

1.3 Almost sure rate of convergence

In all the following, we assume that G admits a minimizer m.

1.3.1 Convergence results

We first recall a usual theorem [Duf97] giving the strong consistency of stochastic gradient estimates under weak assumptions. In this aim, let us first introduce a new assumption:

(A1a) There are non-negative constants C_1, C_2 such that for all $h \in \mathcal{H}$,

$$\mathbb{E}\left[\left\|\nabla_h g\left(X,h\right)\right\|^2\right] \leq C_1 + C_2\left\|h-m\right\|^2.$$

This assumption is quite usual and just means that we have at worse a linear increasing of the gradient of g (up to the expectation), and so, at worse a quadratic increasing of the functional G. We can now give the strong consistency of stochastic gradient estimates.

Theorem 1.3.1. *Suppose that Assumption **(A1a)** is fulfilled and that for all $h \in \mathcal{H}$ such that $h \neq m$,*

$$\left\langle \nabla G\left(h\right), h-m\right\rangle > 0.$$

Then

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

Remark that the conditions on the step sequence are due to the use of Robbins-Siegmund Theorem for obtaining the strong consistency of the estimates. Furthermore, for the sake of simplicity, we have chosen a deterministic stepsequence, but previous theorem remains true taking a random stepsequence $(\Gamma_n)_{n\geq 1}$ verifying

$$\sum_{n\geq 0}\Gamma_{n+1} = +\infty \quad a.s \qquad \text{and} \qquad \sum_{n\geq 0}\Gamma_{n+1}^2 < +\infty \quad a.s,$$

and such that for all $n \geq 0$, Γ_{n+1} is $\mathcal{F}_n$-measurable. This possible choice is crucial to prove the convergence of Stochastic Newton estimates in Chapter 3.

We now focus on the almost sure rates of convergence of the estimates obtained with the help of stochastic gradient algorithm. In this aim, let us suppose from now that we have a stepsequence (γ_n) satisfying $\gamma_n = c_\gamma n^{-\gamma}$ with $c_\gamma > 0$ and $\gamma \in (1/2, 1)$. Furthermore, we introduce the following assumptions:

(A1η) There are positive constants $\eta > \frac{1}{\gamma} - 1$ and C_η such that for all $h \in \mathcal{H}$,

$$\mathbb{E}\left[\left\|\nabla_h g\left(X,h\right)\right\|^{2+2\eta}\right] \leq C_\eta\left(1 + \left\|h-m\right\|^{2+2\eta}\right)$$

(A2) The functional G is twice continuously differentiable on a neighborhood V_m of m and

$$\liminf_{h\in V_m} \lambda_{\min}\left(\nabla^2 G(h)\right) > 0.$$

Note that Assumption **(A1η)** is verified, for instance, since $\nabla_h g(X,.)$ admits a fourth order moment while Assumption **(A2)** implies that the functional G is locally strongly convex and we will denote $\lambda_{\min} := \lambda_{\min}\left(\nabla^2 G(m)\right)$. We can now give the almost sure rate of convergence of the estimates.

Theorem 1.3.2 ([GB16b]). *Suppose Assumptions **(A1η)** and **(A2)** hold. Then*

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

Remark that this result was already given by [Pel98] in the finite dimensional case.

Sketch of the proof. The first idea is to linearize the gradient in equality (1.3), i.e one has

$$m_{n+1} - m = \left(I_{\mathcal{H}} - \gamma_{n+1} H\right)(m_n - m) + \gamma_{n+1}\xi_{n+1} - \gamma_{n+1}\delta_n \tag{1.4}$$

where $H := \nabla^2 G(m)$ and $\delta_n := \nabla G(\theta_n) - H(m_n - m)$ is the remainder term in the Taylor's expansion of the gradient. Since m_n converges almost surely to m and thanks to Assumption **(A2)**, one has $\|\delta_n\| = o\left(\|m_n - m\|\right)$ a.s. Furthermore, with the help of an induction, one has

$$m_n - m = \beta_{n,0}(m_0 - m) + \sum_{k=0}^{n-1} \beta_{n,k+1}\gamma_{k+1}\xi_{k+1} - \sum_{k=0}^{n-1} \beta_{n,k+1}\gamma_{k+1}\delta_k \tag{1.5}$$

with $\beta_{n,n} = I_{\mathcal{H}}$ and $\beta_{n,k} = \prod_{j=k+1}^{n}\left(I_{\mathcal{H}} - \gamma_j H\right)$. With the help of some usual calculus, one can prove that the first term on the right-hand side of equality (1.5) converges exponentially fast while the third one converges at least at the same rate as the second one. Then one has to focus on this second term, in each one can make appear a martingale term. The proof in [Pel98] consists in writing this term in the basis of $\mathcal{H}$ composed of eigenvectors of H and to apply the law of the iterated logarithm to each coordinate. Nevertheless, this could not be applied in the infinite dimensional case and we so propose a proof based on the obtaining of some exponential inequalities for "nearly" martingales (see Lemma 6.1 in the arxiv version of [GB16b]).

1.3.2 Some applications

Application to linear model

Let us consider (X, Y) a couple of random variables taking values in $\mathbb{R}^d \times \mathbb{R}$ such that

$$Y = X^T\theta + \epsilon$$

with $\theta \in \mathbb{R}^d$ deterministic and ϵ is a random variable taking values in $\mathbb{R}$ independent from X. If the matrix $\mathbb{E}\left[XX^T\right]$ is positive, θ is the unique minimizer of the functional $G_{LM} : \mathbb{R}^d \longrightarrow \mathbb{R}_+$ defined for all $h \in \mathbb{R}^d$ by

$$G_{LM}(h) = \frac{1}{2}\mathbb{E}\left[\left(Y - X^T h\right)^2\right].$$

Then, the stochastic gradient algorithm for estimating θ is defined recursively for all $n \geq 0$ by

$$\theta_{n+1} = \theta_n + \gamma_{n+1} \left(Y_{n+1} - X_{n+1}^T \theta_n \right) X_{n+1}. \tag{1.6}$$

The following result gives the almost sure rates of convergence of the estimates and is a direct corollary of Theorem 1.3.2.

Corollaire 1.3.1. *Suppose there exists $\eta > \frac{1}{\gamma} - 1$ such that X admits a moment of order $4 + 4\eta$ and such that ϵ admits a moment of order $2 + 2\eta$. Let us also suppose that $\mathbb{E}\left[XX^T\right]$ is positive. Then, gradient estimates defined by (1.6) satisfy*

$$\|\theta_n - \theta\|^2 = O\left(\frac{\ln n}{n^\gamma}\right) \quad a.s.$$

In Figure 1.1, we focus on the quadratic error of the estimates with respect to the sample size for different values of γ. More precisely we consider $\gamma = 0.5, 0.66, 0.75, 1$ (although $\gamma = 0.5$ or 1 is out of our framework). One can see that larger γ is, faster the algorithm converges. In addition, one can remark that in the case where $\gamma = 1$, it seems to be more stable, but (one of) the price to pay is an increased sensitivity to a possible bad initialization.

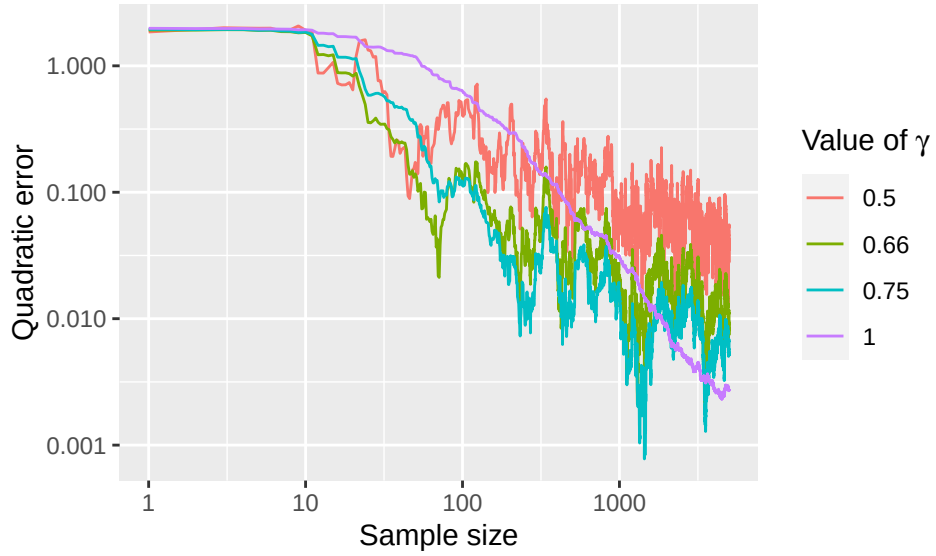


Figure 1.1 – Evolution of the quadratic error of θ_n with respect to the sample size n for different choices of γ in the linear regression case.

Application to logistic regression

Let (X, Y) a couple of random variables taking values in $\mathbb{R}^d \times \{0, 1\}$ such that $Y|X \sim \mathcal{B}\left(\pi\left(\theta^T X\right)\right)$ where $\theta \in \mathbb{R}^d$ and $\pi(x) = \frac{e^x}{1+e^x}$. Under assumptions, θ is the unique minimizer of the function $G_{\log} : \mathbb{R}^d \longrightarrow \mathbb{R}$ defined for all $h \in \mathbb{R}^d$ by

$$G_{\log}(h) = \mathbb{E}\left[\log\left(1 + \exp\left(h^T X\right)\right) - h^T XY\right].$$

Then, the stochastic gradient algorithm for estimating θ is defined recursively for all $n \geq 0$ by

$$\theta_{n+1} = \theta_n + \gamma_{n+1} \left(Y_{n+1} - \pi \left(X_{n+1}^T \theta_n \right) \right) X_{n+1}. \tag{1.7}$$

The following result gives the almost sure rate of convergence of the estimates and is a direct corollary of Theorem 1.3.2.

Corollaire 1.3.2. *Suppose there exists $\eta > \frac{1}{\gamma} - 1$ such that X admits a moment of order $2 + 2\eta$ and assume that $\nabla^2 G_{\log}(\theta) = \mathbb{E}\left[\pi\left(\theta^T X\right)\left(1 - \pi\left(\theta^T X\right)\right) X X^T\right]$ is positive. Then, the gradient estimates defined by (1.7) satisfy*

$$\|\theta_n - \theta\|^2 = O\left(\frac{\ln n}{n^\gamma}\right) \quad a.s.$$

In Figure 1.2, we focus on the quadratic error of the estimates with respect to the sample size for different values of γ. More precisely we consider $\gamma = 0.5, 0.66, 0.75, 1$. For $\gamma = 0.5, 0.66$ and 0.75, one can see again that larger γ is, faster the algorithm converges. Nevertheless, in the case where $\gamma = 1$, it does not converge faster at all. We will see later that it is due to a bad calibration of the parameter c_γ.

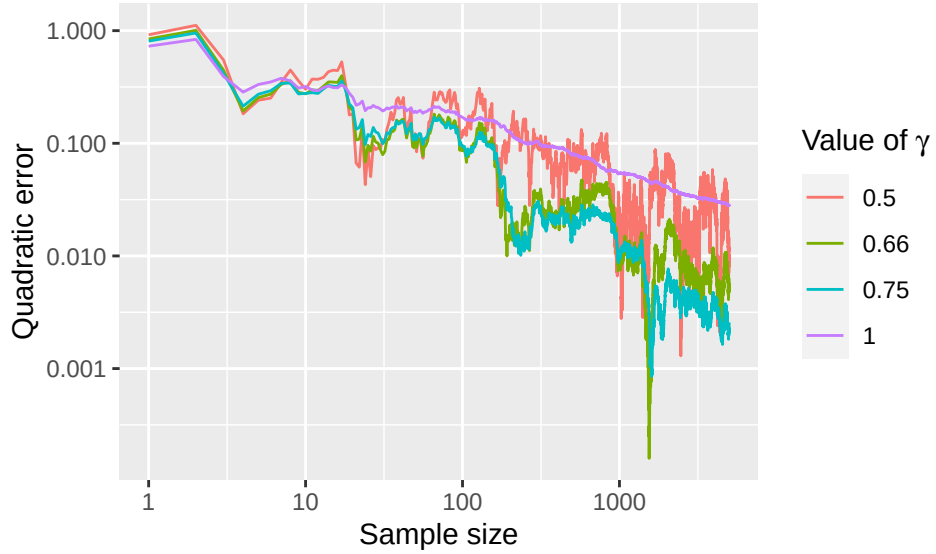


Figure 1.2 – Evolution of the quadratic error of θ_n with respect to the sample size n for different choices of γ in the logistic regression case.

Application to p-means

Let X be a random variable taking values in a separable Hilbert space $\mathcal{H}$ and let $p \in (1, 2)$. Then, the p-mean of X (denoted by m_p) is, under conditions, the unique minimizer of the functional $G_p : \mathcal{H} \longrightarrow \mathbb{R}$ defined for all $h \in \mathcal{H}$ by

$$G_p(h) = \frac{1}{p} \mathbb{E}\left[\|X - h\|^p\right].$$

Remark that taking $p = 2$ would have leaded to the mean while the case $p = 1$ corresponds to the geometric median. Let us suppose from now that the following assumptions are fulfilled:

($\mathbf{H_{p\text{-means}}1}$) X is not concentrated around single points: there is a positive constant C_p such that for all $h \in \mathcal{H}$,

$$\mathbb{E}\left[\|X - h\|^{p-2}\right] \leq C_p.$$

Under this assumption, G is locally strongly convex and m_p is the unique minimizer of G. Furthermore, the stochastic gradient algorithm for estimating m_p is defined recursively for all $n \geq 0$ by

$$m_{p,n+1} = m_{p,n} + \gamma_{n+1} \frac{(X_{n+1} - m_{p,n})}{\|X_{n+1} - m_{p,n}\|^{2-p}}. \tag{1.8}$$

The following result gives the almost sure rates of convergence of the gradient estimates and is a direct corollary of Theorem 1.3.2.

Corollaire 1.3.3. *Suppose that Assumption* ***($H_{p\text{-means}}1$)*** *holds. Suppose also that there is* $\eta > \frac{1}{\gamma} - 1$ *such that* X *admits a moment of order* $(p-1)(2+2\eta)$. *Then the gradient estimates defined by* (1.8) *satisfy*

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

In Figure 1.3, we focus on the evolution of the quadratic error of $m_{p,n}$ (with $p = 1.5$) with respect to the sample size n for different choices of γ. Remark that here again, larger γ is, faster the algorithm converges, and so, even for $\gamma = 1$.

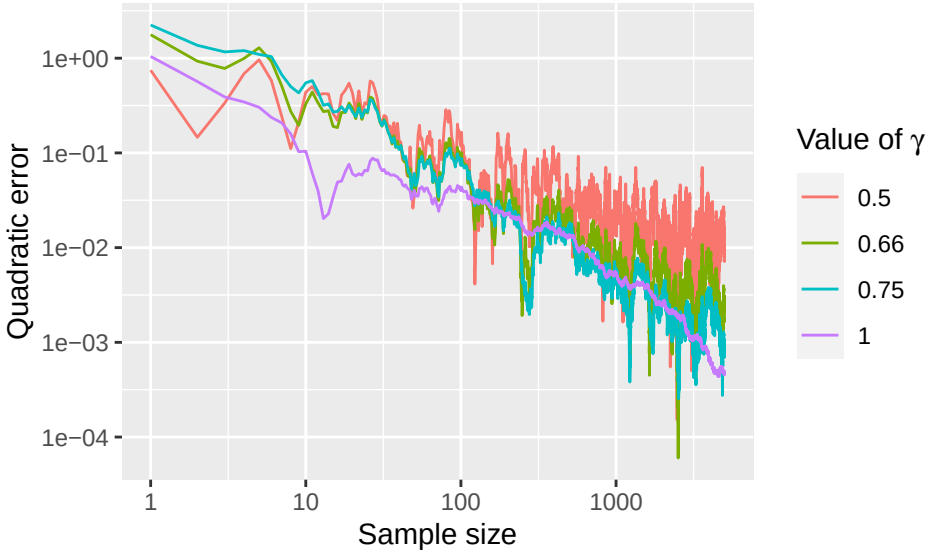


Figure 1.3 – Evolution of the quadratic error of $m_{p,n}$ with respect to the sample size n for different choices of γ.

1.4 Convergence in law

1.4.1 Convergence result

We now focus on the convergence in law of the gradient estimates. In this aim, let us now introduce some usual assumptions:

(A3a) The Hessian of G is bounded on a neighborhood of m.

(A4a) There are a neighborhood V_m of m and a non-negative constant C_{V_m} such that for all $h \in V_m$,

$$\left\|\nabla G(h) - \nabla^2 G(m)(h-m)\right\| \leq C_{V_m} \|h-m\|^2 .$$

(A5a) The function Σ defined for all $h \in \mathcal{H}$ by

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

is continuous at m.

Let us now comment these hypothesis. First, Assumption **(A3a)** implies that the gradient of G is locally Lipschitz. This is verified, in the case of linear and logistic regression since X admits a moment of order 2 while in the case of p-means, this is verified since **($\mathbf{H}_{\text{p-means}}$1)** is fulfilled. Assumption **(A4a)** ensures that the Hessian is locally Lipschitz and is crucial to give the rate of convergence of the rest term in the Taylor's expansion of the gradient. In the case of linear regression $C_{V_m} = 0$ while in the case of logistic regression, this hypothesis is verified since X admits a moment of order 3. In the case of the estimation of p-means, **($\mathbf{H}_{\text{p-means}}$2)** (see Section 1.4.2) ensures that **(A4a)** is fulfilled. Finally, **(A5a)** is crucial to get the convergence in law, and is verified since X admits a second order moment for linear and logistic regressions and a moment of order $2p-2$ for p-means. In all the following, we will denote $H := \nabla^2 G(m)$ and $\Sigma := \Sigma(m)$.

Theorem 1.4.1 ([GB17]). *Suppose assumptions **(A1η)**, **(A2)**, **(A3a)**, **(A4a)** and **(A5a)** hold, then*

$$\frac{1}{\sqrt{\gamma_n}}(m_n - m) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0, \Sigma_{RM})$$

with

$$\Sigma_{RM} = \int_0^{+\infty} e^{-sH} \Sigma e^{-sH} ds.$$

Then the stochastic gradient algorithm converges in law at a rate $\sqrt{\gamma_n}$. Furthermore, remark that Σ_{RM} is the solution of the Lyapunov equation

$$AH + HA = \Sigma.$$

Note that this result was already given in [Pel98] but here again, the proofs were not adapted to the infinite dimensional case. The proof of Theorem 1.4.1 relies on the use of a martingale central

limit theorem in Hilbert spaces due to [Jak88] that we apply on the second term on the right-hand side of equality (1.5).

Remark 1.4.1. *Note that assumptions given in Theorem 1.4.1 differs to the ones in [GB17]. More precisely, in [GB17], the gradient of g was supposed to admit a moment of order four since a uniform bound of $\mathbb{E}\left[\|m_n - m\|^4\right]$ was used to prove inequality (29), i.e to prove that for all $\epsilon > 0$,*

$$\mathbb{P}\left[\sup_{0\leq k\leq n-1} \frac{1}{\sqrt{\gamma_n}} \|\beta_{n,k+1}\gamma_{k+1}\xi_{k+1}\| > \epsilon\right] \xrightarrow[n\to+\infty]{} 0.$$

Nevertheless, since

$$\begin{aligned}\mathbb{P}\left[\sup_{0\leq k\leq n-1} \frac{1}{\sqrt{\gamma_n}} \|\beta_{n,k+1}\gamma_{k+1}\xi_{k+1}\| > \epsilon\right] &\leq \sum_{k=0}^{n-1} \mathbb{P}\left[\frac{\gamma_{k+1}}{\sqrt{\gamma_n}} \|\beta_{n,k+1}\|_{op} \|\xi_{k+1}\| \mathbf{1}_{\|m_k-m\|\leq 1} > \epsilon\right] \\ &+ \sum_{k=0}^{n-1} \mathbb{P}\left[\frac{\gamma_{k+1}}{\sqrt{\gamma_n}} \|\beta_{n,k+1}\|_{op} \|\xi_{k+1}\| \mathbf{1}_{\|m_k-m\|>1} > \epsilon\right]\end{aligned}$$

*and since $\|m_n - m\|$ converges almost surely to 0, one can "easily" prove that the second term on the right-hand side of previous inequality converges exponentially fast to 0 (almost surely). For the first term, applying Markov inequality and thanks to Assumption **(A1η)**, it comes*

$$\begin{aligned}\sum_{k=0}^{n-1} \mathbb{P}\left[\frac{\gamma_{k+1}}{\gamma_n} \|\beta_{n,k+1}\|_{op} \|\xi_{n+1}\| \mathbf{1}_{\|m_n-m\|\leq 1} > \epsilon\right] &\leq \frac{1}{\gamma_n^{1+\eta}\epsilon^{2+2\eta}} \sum_{k=0}^{n-1} \gamma_{k+1}^{2+2\eta} \|\beta_{n,k+1}\|_{op}^{2+2\eta} \mathbb{E}\left[\|\xi_{n+1}\|^{2+2\eta} \mathbf{1}_{\|m_n-m\|\leq 1}\right] \\ &\leq \frac{1}{\gamma_n^{1+\eta}\epsilon^{2+2\eta}} \sum_{k=0}^{n-1} \gamma_{k+1}^{2+2\eta} \|\beta_{n,k+1}\|_{op}^{2+2\eta} 2^{2+2\eta} C_\eta = O\left(\gamma_n^\eta\right),\end{aligned}$$

and inequality (29) in [GB17] is so satisfied.

1.4.2 Some applications

Application to linear regression

The following result gives the convergence in law of the gradient estimates defined by (1.6) and is a direct corollary of Theorem 1.4.1.

Corollaire 1.4.1. *Suppose there exists $\eta > \frac{1}{\gamma} - 1$ such that X admits a moment of order $4 + 4\eta$ and such that ϵ admits a moment of order $2 + 2\eta$. Let us also suppose that $\mathbb{E}\left[XX^T\right]$ is positive. Then, stochastic gradient estimates defined by (1.6) satisfy*

$$\frac{1}{\sqrt{\gamma_n}}(\theta_n - \theta) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \frac{\mathbb{E}\left[\epsilon^2\right]}{2} I_d\right).$$

Remark that we give here a "strict" corollary of Theorem 1.4.1, but one can probably obtain the same results with less restrictive assumptions on the moments of X and ϵ. In Figure 1.4, we con-

sider the case where $\mathbb{E}\left[\epsilon^2\right] = 1$, and rewrite Corollary 1.4.1 as

$$C_n := \frac{2}{\gamma_n} \|\theta_n - \theta\|^2 \xrightarrow[n\to+\infty]{\mathcal{L}} \chi^2_d.$$

We then focus on the distribution of C_n for a sample size $n = 5000$ and for different choices of γ ($\gamma = 0.5, 0.66, 0.75$). Remark that in both cases, the distribution function of C_n is very close to the one of a Chi-square law with d degrees of freedom, which seems to confirm Corollary 1.4.1.

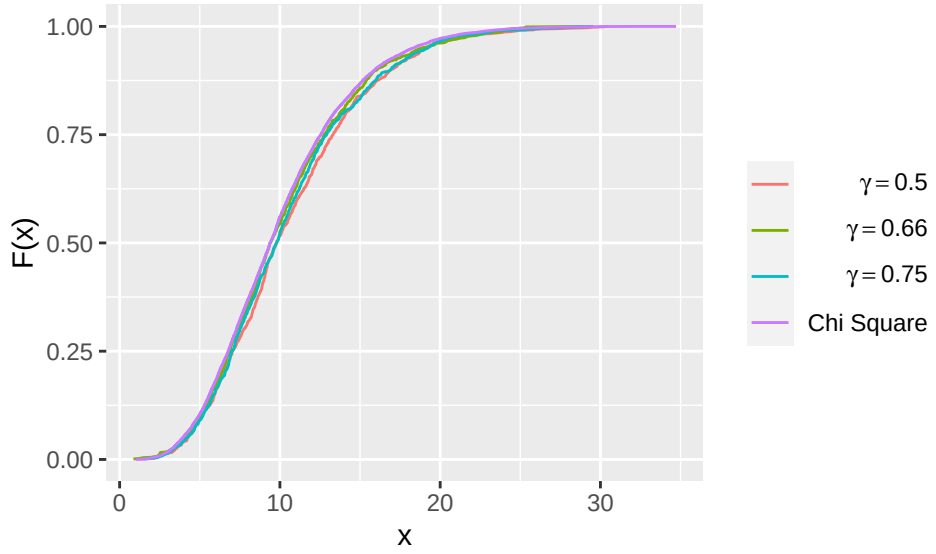


Figure 1.4 – Comparison of the distribution function of C_n (with $n = 5000$ and for $\gamma = 0.5, 0.66$ and 0.75) with the distribution function of a Chi-square law with d degrees of freedom.

Application to logistic regression

The following result gives the convergence in law of the gradient estimates defined by (1.7) and is a direct corollary of Theorem 1.4.1.

Corollaire 1.4.2. *Suppose that there exists* $\eta > \frac{1}{\gamma} - 1$ *such that* X *admits a moment of order* $\max\{3, 2+2\eta\}$. *Let us also suppose that* $\mathbb{E}\left[\pi\left(X^T\theta\right)\left(1-\pi\left(X^T\theta\right)\right)XX^T\right]$ *is positive. Then, stochastic gradient estimates defined by* (1.7) *satisfy*

$$\frac{1}{\sqrt{\gamma}_n}\left(\theta_n - \theta\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \frac{1}{2} I_d\right).$$

One can rewrite Corollary 1.4.2 as

$$C_n := \frac{2}{\gamma_n} \|\theta_n - \theta\|^2 \xrightarrow[n\to+\infty]{\mathcal{L}} \chi^2_d.$$

In Figure 1.5, we focus on the distribution of C_n for a sample size $n = 5000$ and for different choices of γ ($\gamma = 0.5, 0.66, 0.75$). Remark that in both cases, the distribution function of C_n is close to the

one of a Chi-square law with d degrees of freedom but not enough to tell that, at time $n = 5000$, we have achieved convergence.

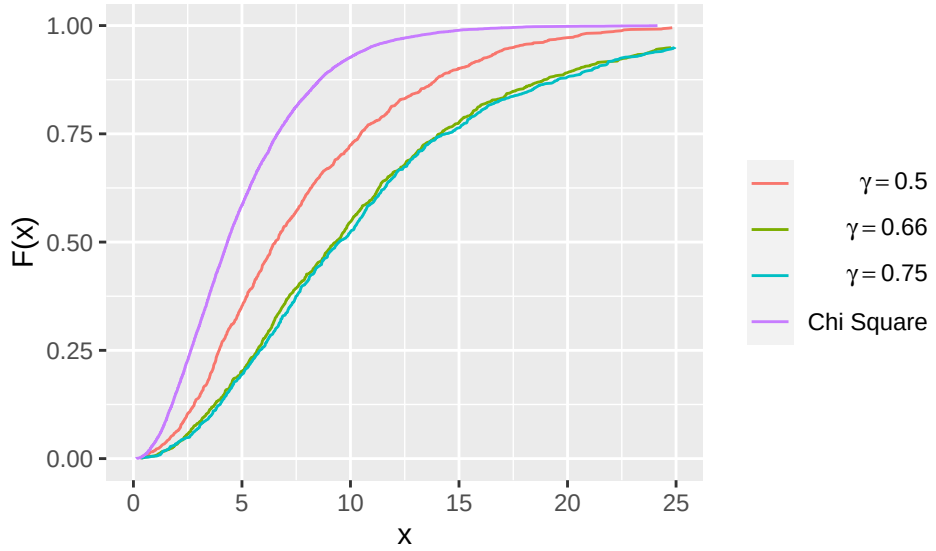


Figure 1.5 – Comparison of the distribution function of C_n (with $n = 5000$ and for $\gamma = 0.5, 0.66$ and 0.75) with the distribution function of a Chi-square law with d degrees of freedom.

Application to the estimation of p-means

In order to get the convergence in law of the gradient estimates of p-means, let us first introduce a new assumption:

($\mathbf{H_{p\text{-means}}2}$) X is not concentrated around single points: there is a positive constant C_p such that for all $h \in \mathcal{H}$,

$$\mathbb{E}\left[\|X - h\|^{p-3}\right] \leq C_p.$$

This assumption implies **($\mathbf{H_{p\text{-means}}1}$)**, and we denote the constant in the same way for the sake of simplicity. This hypothesis is crucial to verify **(A4a)**. Furthermore, note that in the case of p-means, one has

$$H_{(m_p)} := \nabla^2 G_p\left(m_p\right) = \mathbb{E}\left[\frac{1}{\left\|X - m_p\right\|^{2-p}}\left(I_{\mathcal{H}} + (p-2)\frac{\left(X - m_p\right)\left(X - m_p\right)^T}{\left\|X - m_p\right\|^2}\right)\right], \tag{1.9}$$

and one has $\lambda_{\min}\left(H_{(m_p)}\right) \geq (p-1)\mathbb{E}\left[\frac{1}{\left\|X-m_p\right\|^{2-p}}\right] > 0$, and **(A2)** is so verified. Then, one can obtain the convergence in law of the estimates of the p-means defined by (1.8).

Corollaire 1.4.3. *Suppose that Assumption ($\mathbf{H_{p\text{-means}}2}$) holds. Suppose also that there exists $\eta > \frac{1}{\gamma} - 1$ such that X admits a moment of order $(p-1)(2+2\eta)$. Then the stochastic gradient estimates defined by*

(1.8) *satisfy*

$$\frac{1}{\sqrt{\gamma_n}}\left(m_{n,p}-m_p\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0,\Sigma_{RM}^{(m_p)}\right)$$

with

$$\Sigma_{(m_p)} = \mathbb{E}\left[\frac{\left(X-m_p\right)\left(X-m_p\right)^T}{\left\|X-m_p\right\|^{4-2p}}\right] \qquad \textit{and} \qquad \Sigma_{RM}^{(m_p)} = \int_0^{+\infty} e^{-sH_{(m_p)}}\Sigma_{(m_p)}e^{-sH_{(m_p)}}ds.$$

1.4.3 Remarks

Remark that we only get a rate of convergence in law of order $\sqrt{\gamma_n}$ with $\gamma < 1$, i.e one cannot not obtain an "optimal" rate of order $\frac{1}{\sqrt{n}}$. Intuitively, we can try to take $\gamma = 1$ to obtain a good rate of convergence. Nevertheless, this implies to take $c_\gamma > \frac{1}{2\lambda_{\min}}$, with $\lambda_{\min} = \lambda_{\min}\left(\nabla^2 G(m)\right)$. For instance, in Figure 1.1, this assumption was satisfied and one can see that estimates converge at the good rate. Nevertheless, in Figure 1.2, this condition was not fulfilled and one has observed that the estimates did not converge at a good rate. Anyway, this approach generates two main problems: (i) one has to calibrate the stepsequence according to the smallest unknown eigenvalue of the Hessian, (ii) even if the stepsequence is well calibrated, and although one can obtain a convergence in law of the form

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

where

$$\Sigma'_{RM} = \int_0^{+\infty} e^{-s\left(H-\frac{1}{2c_\gamma}I_d\right)}\Sigma e^{-s\left(H-\frac{1}{2c_\gamma}I_d\right)}ds,$$

the asymptotic variance Σ'_{RM} is not optimal. For instance, considering a M-estimate $\hat{m}_n$, under regularity assumptions, one can prove that (see Proposition 2.2.1 for instance)

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

and $H^{-1}\Sigma H^{-1}$ is a better variance in the sense that $\Sigma'_{RM} - H^{-1}\Sigma H^{-1}$ is non-negative. This represents the main disadvantage of stochastic gradient algorithms, but we will see how to modify the estimates in order to achieve the asymptotic efficiency.

1.5 Non asymptotic rates of convergence

As explained before, non asymptotic rates of convergence for stochastic gradient estimates have been deeply studied in the strongly convex case (see [BM13] for instance). We focus here on the case where the functional G is locally strongly convex. Some results where already given by [Bac14] or [GP17] but under slightly restrictive assumptions on g, i.e supposing that the gradient of g admits exponential moments or is uniformly bounded.

1.5.1 Rate of convergence in quadratic mean

In order to give an explicit upper bound of the risk error, let us now give some additional assumptions:

(A1a') There are non negative constants $\tilde{C}_1, \tilde{C}_2$ such that for all $h \in \mathcal{H}$,

$$\mathbb{E}\left[\|\nabla_h g(X,h)\|^2\right] \leq \tilde{C}_1 + \tilde{C}_2 \left(G(h) - G(m)\right).$$

(A1b') There are non negative constants $\tilde{C}_1', \tilde{C}_2'$ such that for all $h \in \mathcal{H}$,

$$\mathbb{E}\left[\|\nabla_h g(X,h)\|^4\right] \leq \tilde{C}_1' + \tilde{C}_2' \left(G(h) - G(m)\right)^2.$$

(A3b) The functional G is twice continuously differentiable on $\mathcal{H}$ and there is a positive constant $L_{\nabla G}$ such that for all $h \in \mathcal{H}$,

$$\|\nabla G(h)\|_{op} \leq L_{\nabla G}.$$

(A4a') There are positive constants λ_0, r_{λ_0} and a non-negative constant C_{λ_0} such that for all $h \in \mathcal{B}(m, r_{\lambda_0})$,

$$\lambda_{\min}\left(\nabla^2 G(h)\right) \geq \lambda_0 \qquad \text{and} \qquad \left\|\nabla G(h) - \nabla^2 G(m)(h-m)\right\| \leq C_{\lambda_0} \|h-m\|^2.$$

Note that Assumption **(A1a')** is very closed to **(A1a)** since if the function G is strongly convex, **(A1a')** implies **(A1a)** and one has the contrary if the gradient of G is Lipschitz, i.e if **(A3b)** is verified for instance. This new assumption is crucial to obtain an uniform upper bound of the risk error. In addition, remark that we introduce Assumption **(A4a')** only for fixing some notations in the sens that if Assumptions **(A2)** and **(A4a)** are fulfilled, then Assumption **(A4a')** is also satisfied. In addition, coupled with **(A3b)**, it enables to give an upper bound of the rest term in the Taylor's expansion of the gradient, which will be crucial to obtain the rate of convergence in quadratic mean of the gradient estimates.
Remark that in [GB21], two cases are differentiated: if ∇G is uniformly bounded or not, i.e if $\tilde{C}_2 = \tilde{C}_2' = 0$ or not. Nevertheless, in what follows, we do not give explicit constants and so decide not to differentiate the cases. Whatever, one can read the Appendix to see the detailed version of the results of this section. Let us now give the rate of convergence in quadratic mean of $G(m_n)$.

Lemma 1.5.1 ([GB21]). *Suppose Assumptions **(A1b')**, **(A2)**, **(A3)** and **(A4a')** hold. Then, there are positive constants A_0', A_1' such that for all $n \geq 1$,*

$$\mathbb{E}\left[\left(G(m_n) - G(m)\right)^2\right] \leq A_0' e^{-\frac{1}{4} c_\gamma a_0 n^{1-\gamma}} + A_1' n^{-2\gamma}$$

with $a_0 := \frac{\lambda_0^2 \min\left\{1, r_{\lambda_0}^2\right\}}{L_{\nabla G}}$.

Note that constants A_0' and A_1' are explicitly given in Lemma A.1.1 and A.1.2. In other words, this lemma ensures that we have the usual rate of convergence $\mathbb{E}\left[\left(G\left(m_n\right)-G(m)\right)\right]=O\left(n^{-\gamma}\right)$. Remark that the first term is "generated" by the initialization error. In addition, if the functional G is μ-strongly convex, one can take $a_0=\frac{\mu^2}{L_{\nabla G}}$, meaning that this term can eventually encounter some troubles in the ill specified case, i.e if the eigenvalues of the Hessian are at different scales. We will see in Chapter 3 the possible negative influence of this case on the estimates, and how to solve it. We can now give an uniform bound of the quadratic mean error of the gradient estimates.

Theorem 1.5.1 ([GB21]). *Suppose Assumptions **(A1b')**, **(A2)**, **(A3)** and **(A4a')** hold. Then, there are positive constant A_0, A_1, A_2 such that for all $n \geq 1$,*

$$\mathbb{E}\left[\left\|m_n-m\right\|^2\right] \leq A_0 e^{-\frac{1}{4}\lambda_{\min} c_\gamma n^{1-\gamma}}+A_1 e^{-\frac{1}{8}a_0 c_\gamma n^{1-\gamma}}+A_2 n^{-2\gamma}+\frac{2^{1+\gamma}\tilde{C}_1}{\lambda_{\min}} c_\gamma n^{-\gamma},$$

with $a_0:=\frac{\lambda_0^2 \min\left\{1, r_{\lambda_0}^2\right\}}{L_{\nabla G}}$.

Remark that constants A_0, A_1 and A_2 are explicitly given in Theorems A.1.1 and A.1.2. In other words, we get the usual L^2 rate of convergence for gradient estimates given by [BM13, Bac14, GP17] and so, with weaker assumptions. Furthermore, note that every constants can be calculated or recursively estimated. Let us now speak about the different terms in the upper bound of the quadratic mean error. First, note that the main term $\frac{2^{1+\gamma}\tilde{C}_1}{\lambda_{\min}} c_\gamma n^{-\gamma}$ is analogous to the one in the strongly convex case given by [BM13]. In addition, the term $A_0 e^{-\frac{1}{4}\lambda_{\min} c_\gamma n^{1-\gamma}}$ is due to the initialization error while the terms $A_1 e^{-\frac{1}{8}a_0 c_\gamma n^{1-\gamma}}$ and $A_2 \frac{1}{\lambda_{\min}^2} n^{-2\gamma}$ are due to the error of approximation of $\nabla^2 G(m)\left(m_n-m\right)$ by the gradient $\nabla G\left(m_n\right)$, and are negligible.

Sketch of the proof. The proof relies on the induction relation:

$$\mathbb{E}\left[\left\|m_{n+1}-m\right\|^2 \left|\mathcal{F}_n\right.\right] \leq \left\|\theta_n-\theta\right\|^2-2\gamma_{n+1}\left\langle \nabla G\left(m_n\right), \theta_n-\theta\right\rangle+\gamma_{n+1}^2 \mathbb{E}\left[\left\|\nabla_h g\left(X_{n+1}, m_n\right)\right\|^2 \left|\mathcal{F}_n\right.\right]$$

which can be written, thanks to Assumptions **(A1a')** and **(A3)** as

$$\mathbb{E}\left[\left\|m_{n+1}-m\right\|^2 \left|\mathcal{F}_n\right.\right] \leq \left(1+\frac{1}{2} L_{\nabla G} \tilde{C}_2 \gamma_{n+1}\right)\left\|m_n-m\right\|^2-2\gamma_{n+1}\left\langle \nabla G\left(m_n\right), m_n-m\right\rangle+\tilde{C}_1 \gamma_{n+1}^2$$

Remark that usually, one can "easily" conclude thanks to this induction relation when the functional G is strongly convex. In our case, one has to linearize the gradient, i.e one can rewrite previous inequality as

$$\begin{aligned}\mathbb{E}\left[\left\|m_{n+1}-m\right\|^2 \left|\mathcal{F}_n\right.\right] &\leq \left(1+\frac{1}{2} L_{\nabla G} \tilde{C}_2 \gamma_{n+1}\right)\left\|m_n-m\right\|^2-2\gamma_{n+1}\left\langle H\left(m_n-m\right)-\delta_n, m_n-m\right\rangle+\tilde{C}_1 \gamma_{n+1}^2 \\ &\leq \left(1-2\lambda_{\min}\gamma_{n+1}+\frac{1}{2} L_{\nabla G} \tilde{C}_2 \gamma_{n+1}\right)\left\|m_n-m\right\|^2+2\gamma_{n+1}\left\langle \delta_n, m_n-m\right\rangle+\tilde{C}_1 \gamma_{n+1}^2.\end{aligned}$$

Observing that under Assumptions **(A3)** and **(A4a')** one has $\|\delta_n\| \leq L_\delta \left(G\left(m_n\right) - G(m)\right)$, it comes

$$\mathbb{E}\left[\|m_{n+1} - m\|^2 \,|\mathcal{F}_n\right] \leq \left(1 - \lambda_{\min}\gamma_{n+1} + \frac{1}{2}L_{\nabla G}\tilde{C}_2\gamma_{n+1}\right)\|m_n - m\|^2 + \tilde{C}_1\gamma_{n+1}^2 + \frac{L_\delta^2}{\lambda_{\min}}\gamma_{n+1}\left(G\left(m_n\right) - G(m)\right)^2.$$

Then, in order to use Proposition A.5 in [GBWW21], one has to upper bound $\mathbb{E}\left[\left(G\left(m_n\right) - G(m)\right)^2\right]$. For the sake of simplicity, we only explain here how to upper bound $\mathbb{E}\left[G\left(m_n\right) - G(m)\right]$ since the reasoning is quite analogous. With the help of a Taylor's expansion, one has thanks to Assumption **(A3)**

$$\mathbb{E}\left[G\left(m_{n+1}\right)|\mathcal{F}_n\right] = G\left(m_n\right) - \gamma_{n+1}\left\|\nabla G\left(m_n\right)\right\|^2 + \frac{1}{2}L_{\nabla G}\gamma_{n+1}\mathbb{E}\left[\left\|\nabla_h g\left(X_{n+1}, m_n\right)\right\|^2 |\mathcal{F}_n\right].$$

Furthermore, thanks to Assumption **(A4a')**, $\left\|\nabla G\left(m_n\right)\right\|^2 \geq 2a_0\left(G\left(m_n\right) - G(m)\right)$. Then, with the help of **(A1a')**, one has

$$\mathbb{E}\left[G\left(m_{n+1}\right) - G(m)|\mathcal{F}_n\right] \leq \left(1 - 2a_0\gamma_{n+1} + \frac{1}{2}L_{\nabla G}\tilde{C}_2\gamma_{n+1}^2\right)\left(G\left(m_n\right) - G(m)\right) + \tilde{C}_1\gamma_{n+1}^2,$$

and the upper bound is derived from Proposition A.5 in [GBWW21].

1.5.2 L^p rates of convergence

In this section, we focus on the L^p rates of convergence of the estimates, for any $p > 0$. In this aim, let us introduce a new assumption:

(A1p) There are positive constants p, C_p such that for all $h \in \mathcal{H}$,

$$\mathbb{E}\left[\left\|\nabla_h g\left(X, h\right)\right\|^{2p}\right] \leq C_p\left(1 + \|h - m\|^{2p}\right).$$

We can now give the L^p rates of convergence of the stochastic gradient estimates.

Theorem 1.5.2 ([GB16b]). *Suppose Assumption **(A1p)** holds for any $p > 0$ and that Assumptions **(A2)**, **(A3)** and **(A4a')** hold too. Then*

$$\mathbb{E}\left[\|m_n - m\|^{2p}\right] = O\left(\gamma_n^p\right).$$

Remark that contrary to the L^2 rate of convergence, we were not able to exhibit an explicit upper bound of the L^p error. Nevertheless, leading up to the L^p rates of convergence can be crucial to obtain the rate of convergence of the recursive estimates of the Median Covariation Matrix for instance (see Chapter 5 or [CGB15]).

1.5.3 Some applications

Application to logistic regression

Let us consider the logistic regression model. The following corollary gives an upper bound of the quadratic mean error of the estimates obtained with the help of the stochastic gradient algorithm defined by (1.7).

Corollaire 1.5.1. *Suppose that X admits a fourth order moment and that there are positive constants $r_{\log}, \lambda_{\log}$ such that for all $h \in \mathcal{B}\left(\theta, r_{\log}\right)$, $\lambda_{\min}\left(\nabla^2 G(h)\right) \geq \lambda_{\log}$. Then, there are positive constants $A_{0,\log}, A_{1,\log}, A_{2,\log}$ such that for all $n \geq 1$,*

$$\mathbb{E}\left[\|\theta_n - \theta\|^2\right] \leq A_{0,\log} e^{-\lambda_{\log} c_\gamma n^{1-\gamma}} + A_{1,\log} e^{-\frac{1}{4} a_{\log} c_\gamma n^{1-\gamma}} + A_{2,\log} n^{-2\gamma} + \frac{2^\gamma \mathbb{E}\left[\|X\|^2\right] c_\gamma}{\lambda_{\log}} n^{-\gamma}$$

where $a_{\log} = \frac{4\lambda_{\log}^2 \min\left\{1, r_{\log}^2\right\}}{\mathbb{E}\left[\|X\|^2\right]}$.

Remark that constants $A_{0,\log}, A_{1,\log}$ and $A_{2,\log}$ are explicitly given in Corollary A.1.1. Note also that the convergence of projected estimates in the particular case where X is bounded can be easily derived from Theorem 3 in [BM13]. One can then check that the bounds are analogous, up to the term $A_{1,\log} e^{-\frac{1}{4} a_{\log} c_\gamma n^{1-\gamma}} + A_{2,\log} n^{-2\gamma}$, which is the "price to pay" to avoid projecting. Remark that this result is also analogous to the one in [GP17], but without supposing that X is bounded.
In Figure 1.6, we focus on the evolution of the quadratic mean error of the estimates θ_n with respect to the sample size n for $\gamma = 0.66$ and $\gamma = 0.75$. We also compare it to the main term of theoretical bound $\frac{2^{1+\gamma}\tilde{C}_1}{\lambda_{\min}} c_\gamma n^{-\gamma} = \frac{2^{1+\gamma}\mathbb{E}\left[\|X\|^2\right]}{\lambda_{\min}} c_\gamma n^{-\gamma}$, where $\lambda_{\min}$ has been estimated with the help of a Monte Carlo method. One can remark that he slope are analogous, meaning that we have the good rate of convergence. Nevertheless, Figure 1.6 shows that the bound is quite rough. Indeed, we could have derived from the convergence in law that the bound should have been, in the case of the logistic regression, of the form $\frac{2d}{n^\gamma}$ in this case.

Application to p-means

In what follows, let us consider positive constants K, c_K such that $\mathbb{P}\left[\|X\| \leq K\right] \geq c_K$. Then, for all $h \in \mathcal{B}\left(m_p, 1\right)$,

$$\lambda_{\min}\left(\nabla^2 G\left(h\right)\right) \geq \frac{1}{\left(K + \left\|m_p\right\| + 1\right)^{2-p}} (p-1) c_K =: \lambda_K. \tag{1.10}$$

The following corollary gives the rate of convergence in quadratic mean of the recursive estimates of the p-mean defined by (1.8).

Corollaire 1.5.2. *Suppose Assumption* ***($H_{p\text{-means}}2$)*** *holds and that X admits a $2p$-th order moment. Then,*

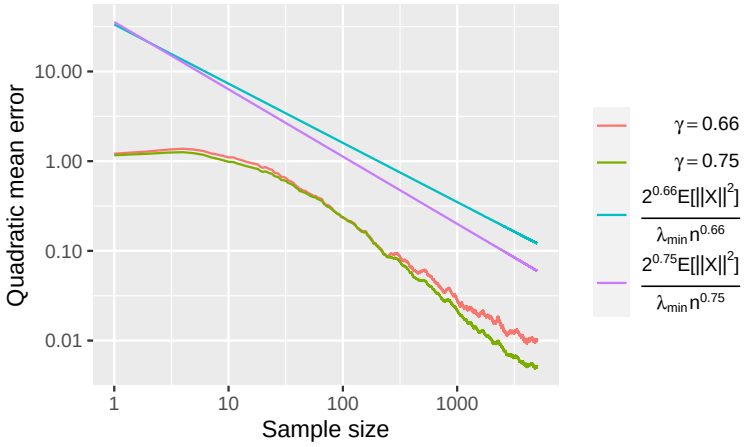


Figure 1.6 – Comparison of the evolution of the quadratic mean error of estimates θ_n (with respect to the sample size n with $\gamma = 0.66, 0.75$) with the main term of the theoretical bound given by Corollary 1.5.1

there are positive constants $A_{0,p}$, $A_{1,p}$ *and* $A_{2,p}$ *such that for all* $n \geq 1$,

$$\mathbb{E}\left[\left\|m_{p,n} - m_p\right\|^2\right] \leq A_{0,p} e^{-\frac{1}{4}\lambda_K c_\gamma n^{1-\gamma}} + A_{1,p} e^{-\frac{1}{8}\frac{\lambda_K^2}{C_p} c_\gamma n^{1-\gamma}} + A_{2,p} n^{-2\gamma} + \frac{2^{1+\gamma}\left(1 + 2G_p\left(m_p\right)\right)}{\lambda_K} c_\gamma n^{-\gamma}.$$

Remark that constants $A_{0,p}$, $A_{1,p}$ and $A_{2,p}$ are explicitly given in Corollary A.1.2.
In Figure 1.7, we focus on the evolution of the quadratic mean error of the estimates $m_{p,n}$ with respect to the sample size n for $\gamma = 0.66$ and $\gamma = 0.75$. We also compare it to the main term of theoretical bound $\frac{2^{1+\gamma}\tilde{C}_1}{\lambda_{\min}} c_\gamma n^{-\gamma} = \frac{2^{1+\gamma}\left(1+2G\left(m_p\right)\right)}{\lambda_{\min}} c_\gamma n^{-\gamma}$, where $\lambda_{\min}$ has been estimated with the help of a Monte Carlo method. One can remark that he slope are analogous, meaning that we have the good rate of convergence but Figure 1.7 suggests that the bound is quite rough.

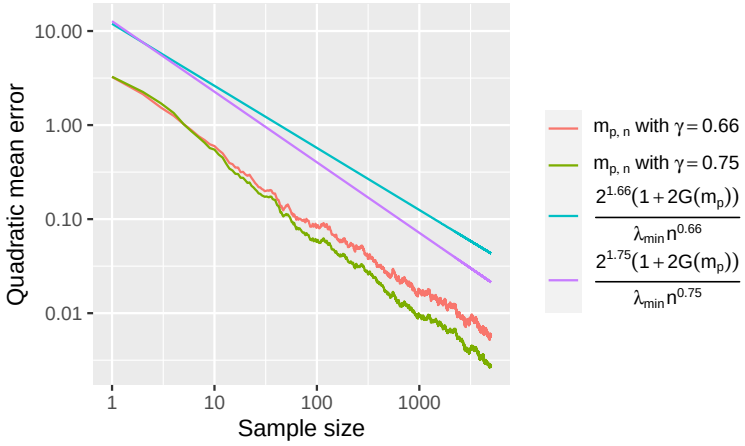


Figure 1.7 – Comparison of the evolution of the quadratic mean error of estimates $m_{p,n}$ (with respect to the sample size n with $\gamma = 0.66, 0.75$) with the main term of the theoretical bound given by Corollary 1.5.2

Chapter 2

Averaged Stochastic Gradient algorithm

This chapter is based on [GB16b, GB17, GB21].

Contents

2.1 Introduction

We have seen in previous chapter that it is nearly impossible to obtain an optimal asymptotic behavior for stochastic gradient estimates. A usual way to accelerate the convergence of gradient estimates has been introduce by [Rup88] and [PJ92], and consists in considering the averaged stochastic gradient algorithm. More precisely, this method consists in taking the averaging of all the estimates obtained with the help of the stochastic gradient algorithm at time n, i.e to consider for all $n \geq 0$,

$$\overline{m}_n = \frac{1}{n+1} \sum_{k=0}^{n} m_k. \tag{2.1}$$

Remark that we are still speaking about online estimates that necessitates only few operations to be updated since they can be written recursively for all $n \geq 0$ as

$$m_{n+1} = m_n - \gamma_{n+1} \nabla_h g\left(X_{n+1}, m_n\right)$$
$$\overline{m}_{n+1} = \overline{m}_n + \frac{1}{n+2}\left(m_{n+1} - \overline{m}_n\right),$$

with $m_0 = \overline{m}_0$. Remark that asymptotic results such that the almost sure rates of convergence as well as the asymptotic efficiency of the averaged estimates are given in the finite dimensional case by [Pel00] for instance. Furthermore, as for gradient estimates, some L^2 rates of convergence were given by [BM13] for the strongly convex case, and [Bac14, GP17] for the strictly convex case with gradient admitting exponential moments or bounded.

We focus here on the original averaged algorithm but it is no less important to mention that several modifications of these estimates exist in the literature. For instance, in order to give more importance to the last iterates of the gradient algorithm, a weighted averaged version can be considered [MP11]. In addition, a parallelized/distributed architecture of these algorithms has been studied to deal with the case where the sample is split into subsamples which are dealt with independently by different agents (cores, processors, computer servers,...) [GBS20, BFHJ11, BFH13, PD19].

In this chapter, we ensure that all the asymptotic results given by [Pel00] remain true even if $\mathcal{H}$ is an infinite dimensional space. In addition, we establish an uniform upper bound of the quadratic mean error of averaged estimates under weak conditions before giving their L^p rates of convergence. All the theoretical results are illustrated on three applications: the estimation of the parameter of linear and logistic regressions as well as the estimation of p-means.

2.2 Asymptotic rates of convergence

2.2.1 Almost sure rates of convergence

Note that by definition of the averaged algorithm and thanks to Toeplitz lemma, the convergence of gradient estimates imply the convergence of their averaged version. We can now focus on the almost sure rates of convergence of the averaged estimates.

Theorem 2.2.1 ([GB16b]). *Suppose Assumptions **(A1η)**, **(A2)**, **(A3a)** and **(A4a)** hold. Then, for all $\delta > 0$,*

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

Remark that up to the log term, we have a $1/n$ rate of convergence. Observe that an analogous result was already given by [Pel00] for finite dimensional spaces, but depending on the ones given by [Pel98] and not available for infinite dimensional spaces. Furthermore, note that one could obtain a $\ln n$ term (instead of $\ln n^{1+\delta}$) supposing that **(A5a)** holds.

Sketch of the proof. First, one has to remark that decomposition (1.4) can be written as

$$\gamma_{n+1} H (m_n - m) = (m_n - m) - (m_{n+1} - m) + \gamma_{n+1}\xi_{n+1} - \gamma_{n+1}\delta_n$$

leading, dividing by γ_{n+1}, to

$$H (m_n - m) = \frac{(m_n - m) - (m_{n+1} - m)}{\gamma_{n+1}} + \xi_{n+1} - \delta_n.$$

Summing these inequalities and dividing by $n+1$, it comes

$$H (\overline{m}_n - m) = \frac{1}{n+1}\sum_{k=0}^{n} \frac{(m_k - m) - (m_{k+1} - m)}{\gamma_{k+1}} + \frac{1}{n+1}\sum_{k=0}^{n} \xi_{k+1} - \frac{1}{n+1}\sum_{k=0}^{n} \delta_k \tag{2.2}$$

and one can conclude by giving the rate of convergence of each term on the right-hand side of previous equality. More precisely, one should first use an Abel's transform on the first term on the right-hand side of equality (2.2). Then, with the help of Theorem 1.3.2, one can prove that the first and third term on the right-hand side are negligible before using a law of large numbers for martingales to obtain the rate of convergence of the second term.

2.2.2 Asymptotic efficiency

Let us now establish the asymptotic efficiency of the averaged estimates.

Theorem 2.2.2 ([GB17]). *Suppose Assumptions **(A1η)**, **(A2)**, **(A3a)**, **(A4a)** and **(A5a)** hold. Then,*

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

The proof of Theorem 2.2.2 consists in applying a Central Limit Theorem in Hilbert spaces given by [Jak88] on the second term on the right-hand side of equality (2.2). Note that we give quite different assumptions compare to [GB17]. More precisely, the gradient of g was supposed to admit a moment of order four since an uniform bound of $\mathbb{E}\left[\|m_n - m\|^4\right]$ was used to prove that for all $\epsilon > 0$,

$$\mathbb{P}\left[\sup_{0\le k\le n} \frac{1}{\sqrt{n}} \|\xi_{k+1}\| > \epsilon\right] \xrightarrow[n\to+\infty]{} 0. \tag{2.3}$$

Nevertheless, one has

$$\begin{aligned}\mathbb{P}\left[\sup_{0\le k\le n} \frac{1}{\sqrt{n}} \|\xi_{k+1}\| > \epsilon\right] &\le \sum_{k=0}^{n} \mathbb{P}\left[\frac{1}{\sqrt{n}} \|\xi_{k+1}\| \mathbf{1}_{\|m_n - m\|\le 1} > \epsilon\right] + \sum_{k=0}^{n} \mathbb{P}\left[\frac{1}{\sqrt{n}} \|\xi_{k+1}\| \mathbf{1}_{\|m_n - m\|>1} > \epsilon\right] \\ &\le \frac{2^{2+2\eta}}{n^{\eta}\epsilon^{2+2\eta}} C_\eta + \frac{C_1}{n\epsilon^2}\sum_{k=0}^{n} \mathbb{P}\left[\|m_n - m\| > 1\right] + \frac{C_2}{n\epsilon^2}\sum_{k=0}^{n} \mathbb{E}\left[\|m_n - m\|^2 \mathbf{1}_{\|m_n - m\|>1}\right]\end{aligned}$$

and since the sequence $\left(\mathbb{E}\left[\|m_n - m\|^2\right]\right)$ is uniformly bounded [GB16b], $\mathbb{E}\left[\|m_n - m\|^2 \mathbf{1}_{\|m_n - m\|>1}\right]$

converges to 0. Then, condition (2.3) is satisfied.

Remark that under regularity assumptions, it is not possible to find estimates with a better convergence, and so, whatever the used methods. For instance, Proposition 2.2.1 ensures that M-estimates converge at the same rate as the averaged estimates.

Proposition 2.2.1. *Let us suppose that* ***(A2)*** *and the following assumptions are fulfilled:*

- *The M-estimate $\hat{m}_n$ converges in probability to m.*
- *For almost every x, the function $g(x,.)$ is twice continuously differentiable.*
- *For almost every x, the Hessian $\nabla^2_{hg}(x,.)$ is $L(x)$-Lipschitz, i.e for $h, h' \in \mathcal{H}$,*

$$\left\|\nabla^2_{hg}\left(x,h\right) - \nabla^2_{hg}\left(x,h'\right)\right\|_{op} \leq L(x)\left\|h-h'\right\|.$$

- *$L(X)$ and $\nabla^2_{hg}(X,m)$ admit a first order moment.*

Then

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

Then, under regularity assumptions, it is not possible to achieve a better rate of convergence than averaged estimates. The main possible gain of iterative methods compare to online methods will be on the rest term. More precisely, the rate of convergence of the main terms for iterative methods will depend on the sample size and on the number of iterations while it will only depend on the sample size for recursive algorithms. Then, with an infinite computational cost, i.e doing as much iterations as necessary, it is (nearly) impossible to beat iterative algorithms (including mini-batch versions).

2.2.3 Some applications

Application to linear model

The averaged version of the stochastic gradient algorithm defined by (1.6) is given recursively for all $n \geq 0$ by

$$\begin{aligned}\theta_{n+1} &= \theta_n + \gamma_{n+1}\left(Y_{n+1} - X_{n+1}^T\theta_n\right)X_{n+1} \\ \overline{\theta}_{n+1} &= \overline{\theta}_n + \frac{1}{n+2}\left(\theta_{n+1} - \overline{\theta}_n\right),\end{aligned} \tag{2.4}$$

with $\overline{\theta}_0 = \theta_0$. The following corollary gives the almost sure rates of convergence of the estimates as well as their asymptotic efficiency.

Corollaire 2.2.1. *Suppose there is* $\eta > \frac{1}{\gamma} - 1$ *such that* X *and* ϵ *admit respectively moments of order* $4+4\eta$ *and* $2+2\eta$. *Furthermore, suppose that* $H_{(LM)} := \mathbb{E}\left[XX^T\right]$ *is positive. Then, for all* $\delta > 0$, *the averaged estimates* $\overline{\theta}_n$ *defined by* (2.4) *satisfy*

$$\left\|\overline{\theta}_n - \theta\right\|^2 = o\left(\frac{(\ln n)^{1+\delta}}{n}\right) \quad a.s \quad \text{and} \quad \sqrt{n}\left(\overline{\theta}_n - \theta\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \mathbb{E}\left[\epsilon^2\right] H_{(LM)}^{-1}\right).$$

In Figure 2.1, we focus on the evolution of the quadratic error of the estimates $\theta_n, \overline{\theta}_n$ with respect to the sample size for $\gamma = 0.66$ and $\gamma = 0.75$. Whatever the choice of γ, one can remark that since the gradient estimates achieve convergence, the averaging enables to accelerate this last one. In addition, note that the slope of the quadratic error of the averaged estimates is close to -1 for n large enough, which seems to confirm Corollary 2.2.1. Finally, remark that for $\gamma = 0.66$, the gradient estimates achieve convergence earlier that for $\gamma = 0.75$, leading the averaged estimates to converge faster for moderate sample size.

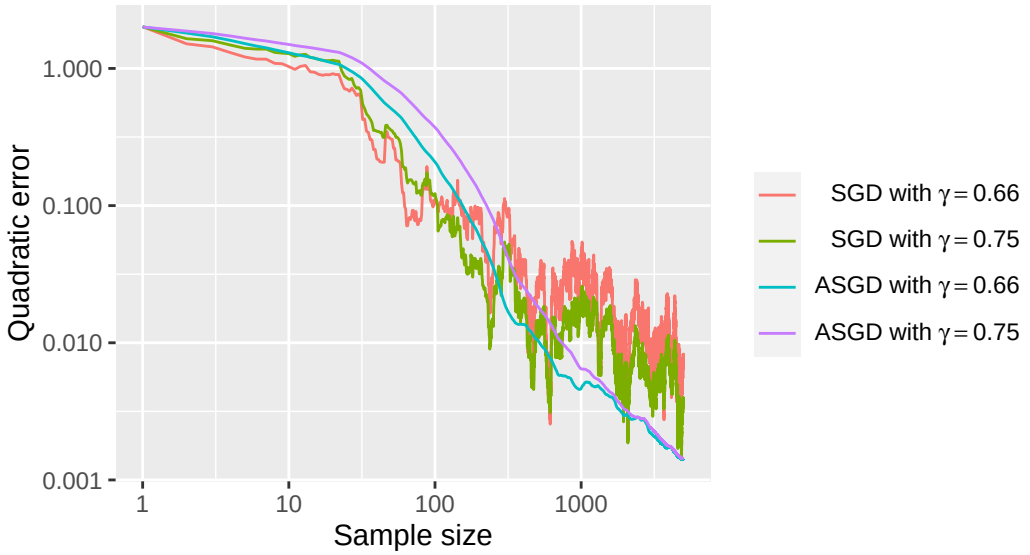


Figure 2.1 – Evolution of the quadratic error of gradient estimates θ_n (SGD) and their averaged version $\overline{\theta}_n$ (ASGD) with respect to the sample size n for different choices of γ in the case of the linear regression.

Remark that one can also recursively estimate $H_{(LM)}$ and $\mathbb{E}\left[\epsilon^2\right] =: \sigma^2_{(LM)}$ as

$$\overline{H}_{(LM),n+1} = \overline{H}_{(LM),n} + \frac{1}{n+2}\left(X_{n+1}X_{n+1}^T - H_{(LM),n}\right)$$
$$\hat{\sigma}^2_{(LM),n+1} = \hat{\sigma}^2_{(LM),n} + \frac{1}{n+2}\left(\left(Y_{n+1} - X_{n+1}^T\overline{\theta}_n\right)^2 - \hat{\sigma}^2_{(LM),n}\right)$$

and check that the estimates are strongly consistent. Then, thanks to Corollary 2.2.1, one has

$$C_n := \frac{n\left(\overline{\theta}_n - \theta\right)^T \overline{H}_n \left(\overline{\theta}_n - \theta\right)}{\hat{\sigma}_n^2} \xrightarrow[n\to+\infty]{\mathcal{L}} \chi_d^2.$$

In Figure 2.2, we focus on the distribution of C_n for a sample size $n = 5000$ and for different choices of γ ($\gamma = 0.66, 0.75$). Remark that in both cases, the distribution function of C_n is close to the one of a Chi-square law with d degrees of freedom but not enough to tell that at time $n = 5000$, we have achieved convergence. Remark that this is also probably due to the cumulative error estimation of θ, H and σ^2.

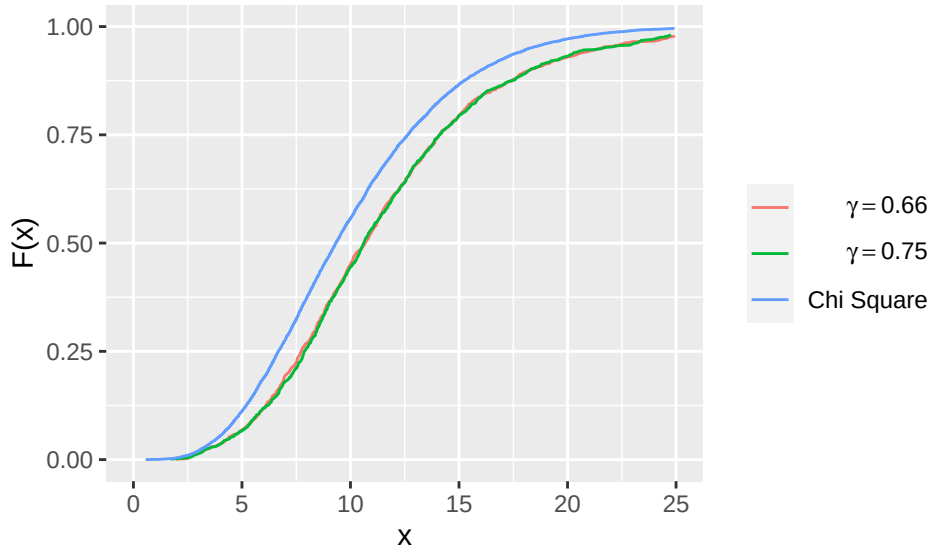


Figure 2.2 – Comparison of the distribution function of C_n (with $n = 5000$ and for $\gamma = 0.66$ and 0.75) with the distribution function of a Chi-square law with d degrees of freedom.

Application to logistic regression

The averaged version of the stochastic gradient defined by (1.7) is given recursively for all $n \geq 0$ by

$$\begin{aligned} \theta_{n+1} &= \theta_n + \gamma_{n+1}\left(Y_{n+1} - \pi\left(X_{n+1}^T\theta_n\right)\right) X_{n+1} \\ \overline{\theta}_{n+1} &= \overline{\theta}_n + \frac{1}{n+2}\left(\theta_{n+1} - \overline{\theta}_n\right) \end{aligned} \tag{2.5}$$

with $\pi(x) = \frac{\exp(x)}{1+\exp(x)}$ and $\theta_0 = \overline{\theta}_0$. The following corollary gives the almost sure rates of convergence of the estimates as well as their asymptotic efficiency.

Corollaire 2.2.2. *Suppose that X admits a moment of order 4 and that* $H_{(\log)} = \mathbb{E}\left[\pi\left(X^T\theta\right)\left(1 - \pi\left(X^T\theta\right)\right) XX^T\right]$

is positive. Then, the averaged estimates defined by (2.5) *satisfy for all* $\delta > 0$,

$$\left\|\overline{\theta}_n - \theta\right\|^2 = o\left(\frac{(\ln n)^{1+\delta}}{n}\right) \quad a.s \quad \text{and} \quad \sqrt{n}\left(\overline{\theta}_n - \theta\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, H_{(\log)}^{-1}\right).$$

In Figure 2.3, we focus on the evolution of the quadratic error of the estimates $\theta_n, \overline{\theta}_n$ with respect to the sample size for $\gamma = 0.66$ and $\gamma = 0.75$. Note that gradient estimates spend much time to achieve convergence, so that their averaged version spend much more time to accelerate the convergence. Nevertheless, for n large enough, the slope of the quadratic error of the averaged estimates is close to -1, which seems to confirm Corollary 2.2.1.

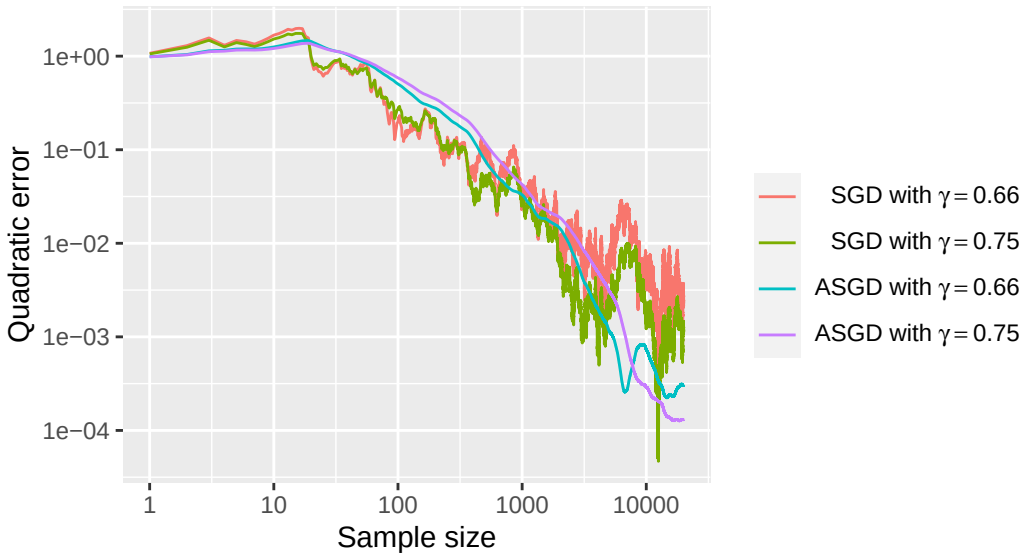


Figure 2.3 – Evolution of the quadratic error of gradient estimates θ_n (SGD) and their averaged version $\overline{\theta}_n$ (ASGD) with respect to the sample size n for different choices of γ in the case of the logistic regression.

Remark that one can estimate $H_{(\log)}$ recursively as

$$\overline{H}_{\log,n+1} = \overline{H}_{\log,n} + \frac{1}{n+2}\left(\pi\left(X_{n+1}^T\overline{\theta}_n\right)\left(1-\pi\left(X_{n+1}^T\overline{\theta}_n\right)\right)X_{n+1}X_{n+1}^T - \overline{H}_{\log,n}\right),$$

i.e for all $n \geq 0$,

$$\overline{H}_{\log,n} = \frac{1}{n+1}\left(\overline{H}_{\log,0} + \sum_{k=1}^{n}\pi\left(X_k^T\overline{\theta}_{k-1}\right)\left(1-\pi\left(X_k^T\overline{\theta}_{k-1}\right)\right)X_kX_k^T\right).$$

Then, one can easily prove that it is strongly consistent and thanks to Corollary 2.2.2, it comes

$$C_n := n\left(\overline{\theta}_n - \theta\right)^T \overline{H}_{\log,n}\left(\overline{\theta}_n - \theta\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \chi_d^2.$$

In Figure 2.4, we focus on the distribution of C_n for a sample size $n = 5000$ and for different choices

of γ ($\gamma = 0.660.75$). Remark that in both cases, the distribution function of C_n approaches the one of a Chi-square law with d degrees of freedom but not enough to tell that at time $n = 5000$, we have achieved convergence. In addition, it seems that taking γ small leads the gradient estimates to converge faster, which leads to have the distribution of C_n closer to the one of the Chi-square distribution in this case.

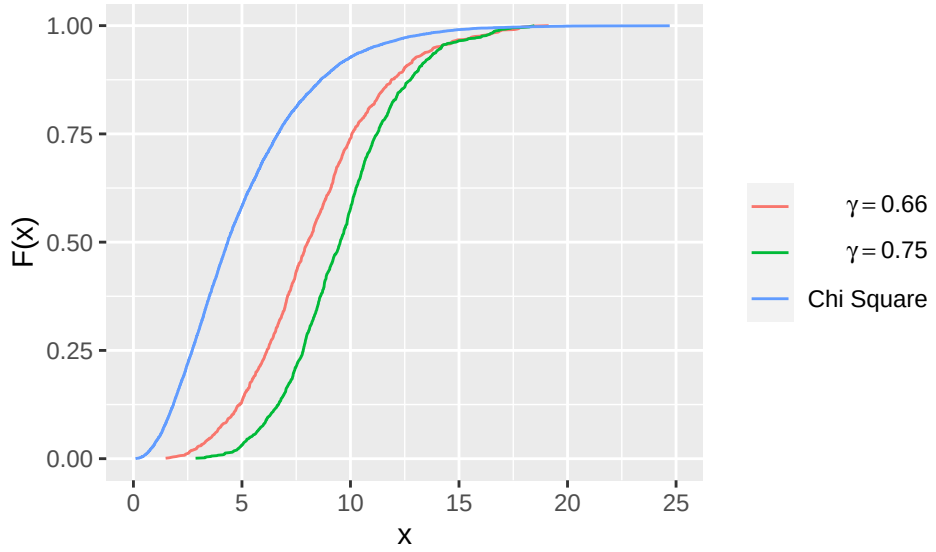


Figure 2.4 – Comparison of the distribution function of C_n (with $n = 5000$ and for $\gamma = 0.66$ and 0.75) with the distribution function of a Chi-square law with d degrees of freedom.

Application to p-means

The averaged version of the stochastic gradient algorithm defined by (1.8) is given recursively for all $n \geq 0$ by

$$\begin{aligned} m_{p,n+1} &= m_{p,n} + \gamma_{n+1} \frac{X_{n+1} - m_{p,n}}{\left\| X_{n+1} - m_{p,n} \right\|^{2-p}} \\ \overline{m}_{p,n+1} &= \overline{m}_{p,n} + \frac{1}{n+2} \left(m_{p,n+1} - \overline{m}_{p,n} \right) \end{aligned} \tag{2.6}$$

with $\overline{m}_{p,0} = m_{p,0}$. Then, the following corollary gives the almost sure rates of convergence of the averaged estimates as well as their asymptotic efficiency.

Corollaire 2.2.3. *Suppose that Assumption* ***($H_{p\text{-means}}2$)*** *holds. Suppose also that there exists* $\eta > \frac{1}{\gamma} - 1$ *such that* X *admits a moment of order* $(p-1)(2+2\eta)$. *Then, the averaged estimates defined by* (2.6) *satisfy for all* $\delta > 0$

$$\left\| \overline{m}_{p,n} - m_p \right\|^2 = o\left(\frac{(\ln n)^{1+\delta}}{n} \right) \quad a.s \quad \text{and} \quad \sqrt{n} \left(\overline{m}_{p,n} - m_p \right) \xrightarrow[n \to +\infty]{\mathcal{L}} \mathcal{N}\left(0, H_{(m_p)}^{-1} \Sigma_{(m_p)} H_{(m_p)}^{-1} \right)$$

with $H_{(m_p)}$ *defined by* (1.9) *and* $\Sigma_{(m_p)} = \mathbb{E}\left[\frac{(X-m_p)(X-m_p)^T}{\|X-m_p\|^{4-2p}}\right]$.

In Figure 2.5, we focus on the evolution of the quadratic error of the estimates $m_{p,n}, \overline{m}_{p,n}$ with respect to the sample size for $\gamma = 0.66$ and $\gamma = 0.75$. The conclusions in this case are the same as the ones for the linear regression case.

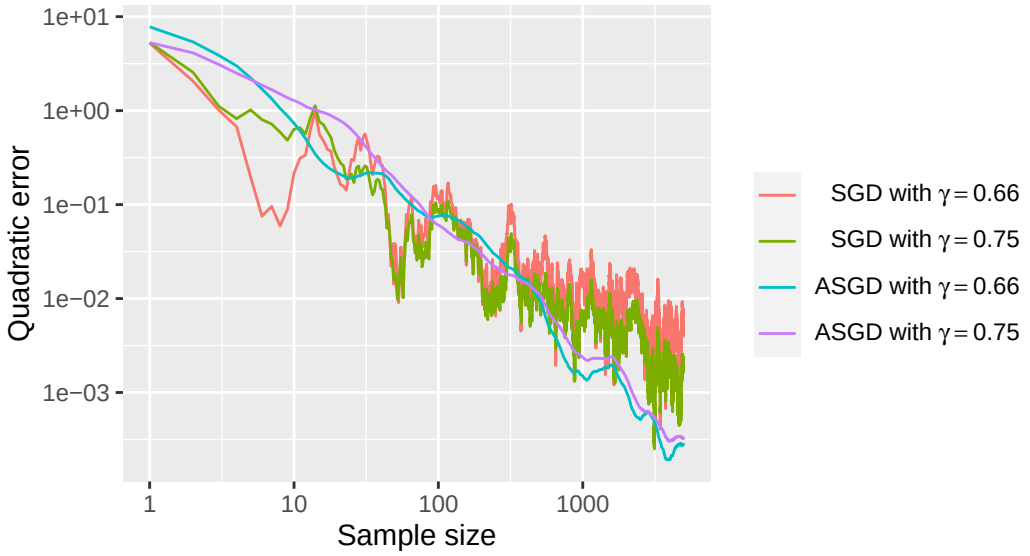


Figure 2.5 – Evolution of the quadratic error of gradient estimates $m_{p,n}$ (SGD) and their averaged version $\overline{m}_{p,n}$ (ASGD) with respect to the sample size n for different choices of γ.

2.3 Non-asymptotic rates of convergence

In this section, we focus on the non asymptotic rates of convergence of averaged stochastic gradient estimates under weak assumptions, i.e under the framework given in Section 1.5.

2.3.1 Rates of convergence in quadratic mean

As for the stochastic gradient estimates, the results for the averaged estimates where split into two cases in [GB21]: ∇G bounded or not. Nevertheless, in what follows, we do not give explicit constants and so decide not to differentiate the cases. Whatever, one can read Appendix to see the detailed version of the results of this section. The following theorem gives a first upper bound of the quadratic mean error of the averaged estimates.

Theorem 2.3.1 ([GB21]). *Suppose Assumptions* ***(A1b')***, ***(A2)***, ***(A3)*** *and* ***(A4a')*** *hold. Then, there are positive constants* A_{av} *and* B_{av} *such that for all* $n \geq 1$,

$$\lambda_{\min}\sqrt{\mathbb{E}\left[\|\overline{m}_n - m\|^2\right]} \leq \frac{\tilde{C}_1}{\sqrt{n+1}} + \frac{A_{av}}{(n+1)^{\gamma}} + \frac{2^{\frac{1+\gamma}{2}}5\sqrt{\tilde{C}_1}}{\sqrt{c_\gamma}\sqrt{\lambda_{\min}}}\frac{1}{(n+1)^{1-\frac{\gamma}{2}}} + \frac{B_{av}}{(n+1)^{\frac{1+\gamma}{2}}}.$$

Constants A_{av} and B_{av} are explicitly given in Theorems A.2.1 or A.2.3. In other words, we achieve the usual rate of convergence $\frac{1}{n}$ and so, under weak assumptions. Remark that the two main rest terms are of order $\frac{1}{n^{\gamma}}$ and $\frac{1}{n^{1-\gamma/2}}$ suggesting that an optimal choice of γ should be $\gamma = 2/3$. Nevertheless, in [GP17], the authors consider the case where ∇g admits exponential moments and give upper bounds of the quadratic mean errors for each the best rate is achieved for $\gamma = 3/4$. Anyway, all these upper bounds can be considered as quite rough, so that it is quite complicated to answer theoretically the question: what a good choice of γ is?
In order to get a (quasi) optimal rate of convergence for the averaged estimates, let us suppose from now that the following assumption is fulfilled:

(A5b) The function Σ defined for all $h \in \mathcal{H}$ by $\Sigma(h) = \mathbb{E}\left[\nabla_h g\left(X,h\right) \nabla_h g\left(X,h\right)^T\right]$ is L_Σ-Lipschitz.

The following theorem ensures that the averaged estimates achieve the Cramer-Rao bound under our framework.

Theorem 2.3.2 ([GB21]). *Suppose Assumptions **(A1b')**, **(A2)**, **(A3)**, **(A4a')** and **(A5b)** hold. Then, there are positive constants A_{av} and B'_{av} such that for all $n \geq 1$,*

$$\sqrt{\mathbb{E}\left[\left\|\overline{m}_n - m\right\|^2\right]} \leq \frac{\sqrt{Tr\left(H^{-1}\Sigma H^{-1}\right)}}{\sqrt{n+1}} + \frac{A_{av}}{\sqrt{\lambda_{\min}}(n+1)^{\gamma}} + \frac{2^{\frac{1+\gamma}{2}} 5\sqrt{\tilde{C}_1}}{\sqrt{c_\gamma}\,\lambda_{\min}} \frac{1}{(n+1)^{1-\frac{\gamma}{2}}} + \frac{B'_{av}}{\sqrt{\lambda_{\min}}(n+1)^{\frac{1+\gamma}{2}}}.$$

Note that constants A_{av} and B'_{av} are explicitly given in Theorems A.2.2 or A.2.4. Remark also that up to rest terms, we achieve the Cramer-Rao bound, i.e $\mathbb{E}\left[\left\|m_n - m\right\|^2\right] \simeq \frac{\text{Tr}\left(H^{-1}\Sigma H^{-1}\right)}{n+1}$. Indeed, under regularity assumptions, any estimates $\tilde{m}_n$ should verify for almost any $m \in \mathcal{H}$:

$$\liminf_n n\mathbb{E}\left[\left\|\tilde{m}_n - m\right\|^2\right] \geq \text{Tr}\left(H^{-1}\Sigma H^{-1}\right).$$

Sketch of the proofs The proofs rely on decomposition (2.2). Indeed, thanks to triangular inequality, one has

$$\sqrt{\mathbb{E}\left[\left\|\overline{m}_n - m\right\|^2\right]} \leq \frac{1}{n+1}\sqrt{\mathbb{E}\left[\left\|\sum_{k=0}^{n} \frac{(m_k - m) - (m_{k+1} - m)}{\gamma_{k+1}}\right\|^2\right]} + \frac{1}{n+1}\sqrt{\mathbb{E}\left[\left\|\sum_{k=0}^{n} \xi_{k+1}\right\|^2\right]} + \frac{1}{n+1}\sum_{k=0}^{n}\sqrt{\mathbb{E}\left[\left\|\delta_k\right\|^2\right]}.$$

Then, one can use Theorem 1.5.1 and Lemma 1.5.1 to get an upper bound of each term on the right-hand side of previous inequality.

2.3.2 L^p rates of convergence

We now focus on L^p rates of convergence of the averaged estimates, for any $p > 0$.

Theorem 2.3.3 ([GB16b]). *Suppose Assumption **(A1p)** holds for any $p > 0$ and that Assumptions **(A2)**, **(A3)** and **(A4a')** hold too. Then*

$$\mathbb{E}\left[\|\overline{m}_n - m\|^{2p}\right] = O\left(\frac{1}{n^p}\right).$$

Remark that contrary to the L^2 rate of convergence, we were not able to exhibit an explicit upper bound of the L^p error. Nevertheless, as explained before, leading up to the L^p rates of convergence can be crucial to obtain the convergence of the recursive estimates of the Median Covariation Matrix for instance (see Chapter 5 or [CGB15]).

2.3.3 Some applications

Logistic regression

Let us consider the logistic regression model. The following corollary gives an upper bound of the quadratic mean error of the averaged estimates defined by (2.5).

Corollaire 2.3.1. *Suppose X admits a moment of order 4 and that there are positive constants $r_{\log}, \lambda_{\log}$ such that for all $h \in \mathcal{B}(\theta, r_{\log})$, $\lambda_{\min}\left(\nabla^2 G_{\log}(h)\right) \geq \lambda_{\log}$. Then, there are positive constants $A_{av,\log}, B_{av,\log}$ such that for all $n \geq 1$,*

$$\sqrt{\mathbb{E}\left[\left\|\overline{\theta}_n - \theta\right\|^2\right]} \leq \frac{\sqrt{Tr\left(H_{\log}^{-1}\right)}}{\sqrt{n+1}} + \frac{A_{av,\log}}{(n+1)^{\gamma}} + \frac{2^{\frac{\gamma}{2}} 5 \sqrt{\mathbb{E}\left[\|X\|^4\right]}}{\sqrt{c_\gamma}\lambda_{\log}(n+1)^{1-\frac{\gamma}{2}}} + \frac{B_{av,\log}}{(n+1)^{\frac{1+\gamma}{2}}}$$

Remark that constants $A_{\text{av,log}}$ and $B_{\text{av,log}}$ are explicitly given in Corollary A.2.1.
In Figure 2.6, we focus on the evolution of the quadratic mean error of the estimates $\overline{\theta}_n$ with respect to the sample size n for $\gamma = 0.66$ and $\gamma = 0.75$. We also compare it to the main term of the theoretical bound $\frac{\text{Tr}\left(H_{\log}^{-1}\right)}{n}$ given by Corollary 2.3.1. One can remark that the curves of the quadratic mean errors seem to tend to the theoretical bound, meaning that the remainder terms are nearly negligible.

Application to the estimation of p-means

We now focus on the estimation of p means. The following corollary gives an upper bound of the quadratic mean error of the averaged estimates obtained with (2.6).

Corollaire 2.3.2. *Suppose Assumption **($H_{p\text{-means}}$2)** holds and that X admits a $2p$-th order moment. Then, there are positive constants $A_{av,p}$ and $B_{av,p}$ such that for all $n \geq 1$,*

$$\sqrt{\mathbb{E}\left[\left\|\overline{m}_{n,p} - m_p\right\|^2\right]} \leq \frac{\sqrt{Tr\left(H_{(m_p)}^{-1}\Sigma_{(m_p)}H_{(m_p)}^{-1}\right)}}{\sqrt{n+1}} + \frac{A_{av,p}}{(n+1)^{\gamma}} + \frac{2^{\frac{1+\gamma}{2}} 5\sqrt{1+2G\left(m_p\right)}}{\sqrt{c_\gamma}\sqrt{\lambda_K}(n+1)^{1-\frac{\gamma}{2}}} + \frac{B_{av,p}}{(n+1)^{\frac{1+\gamma}{2}}}.$$

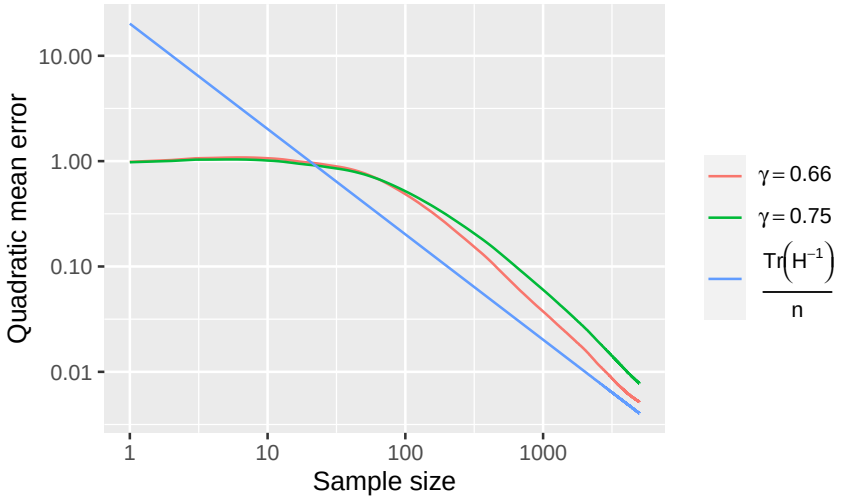


Figure 2.6 – Comparison of the evolution of the quadratic mean error of estimates $\overline{\theta}_n$ (with respect to the sample size n with $\gamma = 0.66, 0.75$) with the main term of the theoretical bound given by Corollary 2.3.1

Remark that constants $A_{\text{av,log}}$ and $B_{\text{av,log}}$ are explicitly given in Corollary A.2.2.
In Figure 2.7, we focus on the evolution of the quadratic mean error of the estimates $m_{p,n}$ with respect to the sample size n for $\gamma = 0.66$ and $\gamma = 0.75$. We also compare it to the main term of theoretical bound $\frac{\text{Tr}\left(H_{(m_p)}^{-1}\Sigma_{(m_p)}H_{(m_p)}^{-1}\right)}{n}$ given by Corollary 2.3.2. One can remark that the curves of the quadratic mean errors are very closed to the theoretical bound, meaning that the remainder terms are negligible, i.e we achieve convergence.

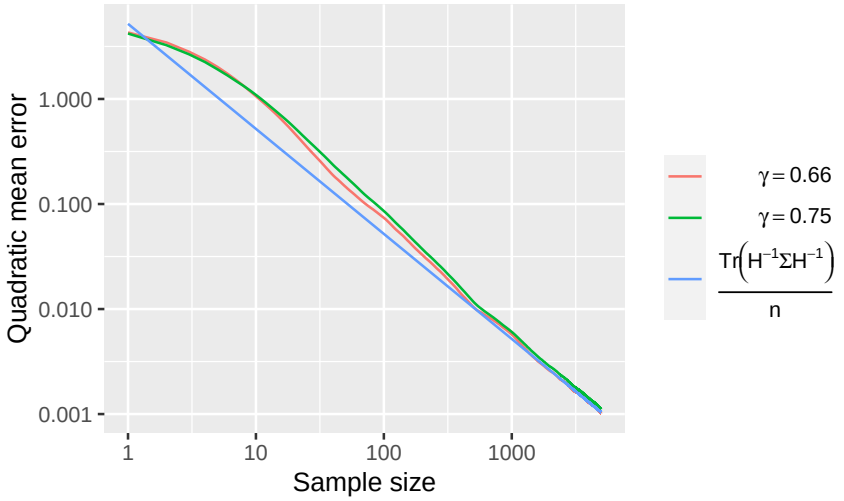


Figure 2.7 – Comparison of the evolution of the quadratic mean error of estimates $\overline{m}_{p,n}$ (with respect to the sample size n with $\gamma = 0.66, 0.75$) with the main term of the theoretical bound given by Corollary 2.3.2

Chapter 3

Online Stochastic Newton algorithms

This chapter is based on [BGBP19, CGBP20, BGB20, GBPL22].

Contents

3.1 Introduction

We have seen that first-order online algorithms have become hegemonic: by a low computational cost per iteration, they allow performing machine learning tasks on large datasets, processing each observation only once. Furthermore, as explained in previous chapters, stochastic gradient methods and their averaged versions are known to be asymptotically efficient [PJ92, Pel00, GB17] and it was proven that, under mild assumptions, averaged estimates can achieve the Cramer-Rao bound (up to rest terms) [GP17, BM13]. However, these first-order online algorithms can be shown in practice to be very sensitive to the Hessian structure of the risk they are supposed to minimize.

For example, when the spectrum of local Hessian matrices shows large variations among their eigenvalues, the stochastic gradient algorithm may be stuck far from the optimum (see for instance the application of [BGBP19, Section 5.2] or Section 3.2).

To address this issue, (quasi) online second-order optimization has been also considered in the literature. In view of avoiding highly costly iterations, most online (quasi) second-order algorithms rely on approximating the Hessian matrix by only using the informations given by the gradient or assuming a diagonal structure of the Hessian. These methods result in choosing a different step size with respect to the components of the current gradient estimate, hence the name of adaptive stochastic gradient algorithms, such as the Adagrad [DHS11] or Adadelta [Zei12] methods.

For more general structures, a Stochastic Quasi-Newton method was introduced in [BHNS16], relying on limited-memory BFGS updates. Specifically, local curvature is captured through (sub-sampled) Hessian-vector products, instead of differences of gradients which enables to provide a stochastic Quasi-Newton algorithm which cost is close to the one of standard SGDs. Nevertheless, two main problems are encountered: the first theroetical one is that the convergence study in [BHNS16] requires the boundedness from above and from below of the spectrum of the estimated Hessian inverses, uniformly over the space of parameters, which can be very restrictive. The second technical one it that in the framework considered in [BHNS16] the stochastic BFGS algorithm can be seen as a refinement of mini-batches gradient algorithms, which is not explicitly derived for online purposes.

In [LP20], the authors introduced a conditioned SGD based on a preconditioning of the gradient direction. The preconditioning matrix is typically an estimate of the inverse Hessian at the optimal point, for which they obtain the asymptotic efficiency. Therefore, the proposed conditioned SGD entails a full inversion of the estimated Hessian, requiring $O(d^3)$ operations per iteration in general, which is less compatible with large-scale data.

In this chapter, we consider a unified and general framework that includes various applications of machine learning tasks, for which we propose a stochastic Newton algorithm. For simplicity, a first version of this algorithm is studied choosing the step size $\frac{1}{n}$. Under suitable and standard assumptions, we define in Section 3.3 the Stochastic Newton algorithm before giving asymptotic results such as almost sure rates of convergence and the asymptotic efficiency.

Nevertheless, considering step sequences of order $1/n$ can lead to poor results in practice [CGBP20]. In order to overcome this problem, we introduce in Section 3.4 a Weighted Averaged Stochastic Newton Algorithm (WASNA) which consists in taking a stepsequence of order $\frac{1}{n^\gamma}$ before weighted averaging over the iterates.

We will see all along this chapter, through examples (linar logistic and softmax regressions for instance) how the estimates of the Hessian can be constructed and *easily* updated over iterations using genuine second-order information. Indeed, given a particular structure of the Hessian estimates that will be encountered in various applications, the Sherman-Morrison formula enables to directly update the inverse of the Hessian matrix at each iteration in $O(d^2)$ operations[1].

[1]Remark that "only" the examples of linear, logistic and softmax regressions are given here. Nevertheless, one can

3.2 Why Stochastic Newton Algorithms?

Contrary to deterministic optimization, one can not use stochastic Newton algorithms with the purpose to improve the rate of convergence. Indeed, we have seen that under regularity assumptions, averaged estimates have already an optimal asymptotic behavior. The idea is to take into account the second order information given by the Hessian to generate stepsequences adapted to each direction of the gradient. This can enable to give better results, in practice, for ill-conditioned problems, i.e in the case where the eigenvalues of the Hessian are at different scale for instance. To illustrate it, let us recall that stochastic gradient estimates $(m_n)_n$ satisfy

$$\mathbb{E}\left[m_{n+1}|\mathcal{F}_n\right] = m_n - \gamma_{n+1}\nabla G\left(m_n\right).$$

Then, if $m_n \simeq m$, one has $\nabla G\left(m_n\right) \simeq \nabla^2 G(m)\left(m_n - m\right)$, leading to

$$\mathbb{E}\left[m_{n+1} - m|\mathcal{F}_n\right] \simeq m_n - m - \gamma_{n+1}\nabla^2 G(m)\left(m_n - m\right) = \left(I_d - \gamma_{n+1}\nabla^2 G\left(m\right)\right)\left(m_n - m\right).$$

Then, if the eigenvalues of $\nabla^2 G(m)$ are at different scales, it is not possible to tune the parameter c_γ to have a step adapted to each direction. To be convinced, let us take the simple example of the linear regression given by

$$Y = X^T\theta + \epsilon$$

with $\epsilon \sim \mathcal{N}(0,1)$, $\theta \in \mathbb{R}^2$ and

$$X \sim \mathcal{N}\left(0, \begin{pmatrix} 10^{-2} & 0 \\ 0 & 10^2 \end{pmatrix}\right).$$

For all $h \in \mathbb{R}^2$, one so has

$$\nabla^2 G(h) = \mathbb{E}\left[XX^T\right] = \begin{pmatrix} 10^{-2} & 0 \\ 0 & 10^2 \end{pmatrix}.$$

Then, since we exactly have $\nabla G\left(m_n\right) = \nabla^2 G(m)\left(\theta_n - \theta\right)$, denoting by $\theta^{(1)}$ and $\theta^{(2)}$ (resp. $\theta_n^{(1)}$ and $\theta_n^{(2)}$) the coordinates of θ (resp. θ_n), it comes

$$\mathbb{E}\left[\theta_{n+1}^{(1)} - \theta^{(1)}|\mathcal{F}_n\right] = \left(1 - \frac{c_\gamma 10^{-2}}{(n+1)^\gamma}\right)\left(\theta_n^{(1)} - \theta^{(1)}\right)$$
$$\mathbb{E}\left[\theta_{n+1}^{(2)} - \theta^{(2)}|\mathcal{F}_n\right] = \left(1 - \frac{c_\gamma 10^{2}}{(n+1)^\gamma}\right)\left(\theta_n^{(2)} - \theta^{(2)}\right).$$

Then, choosing c_γ close to 10^2 would allows to have a step adapted to the first coordinate but would make explode the second coordinate, in the sense that for the first steps, we would have steps of order 10^4. Oppositely, choosing $c_\gamma = 10^{-2}$ would allow to take a step adapted to the second coordinate, but we would have a too small step for the first coordinate. Then, the estimates of the first coordinates should not move. Taking one in between, i.e taking c_γ close to 1 would lead

also deal with non-linear regression [CGBP20] or ridge regression [GBPL22].

to a bad behavior for the two components. This seems to be confirmed by Figure 3.1. Note that in Figure 3.1, a less ill conditioned context has been chosen, i.e the eigenvalues of the Hessian have been chosen equal to 0.1 and 10.

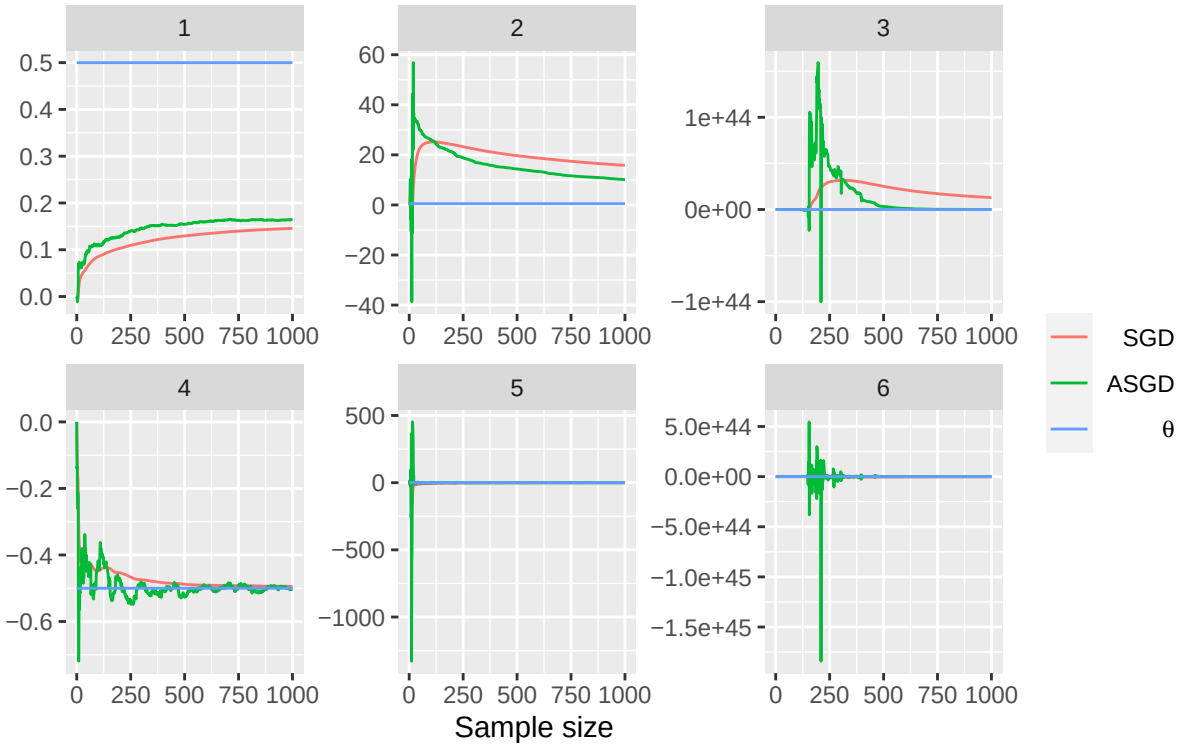


Figure 3.1 – Evolution of the estimates of the first coordinate (first line) and of the second one (second line) with, from the left to the right, $c_\gamma = 0.1, c_\gamma = 1$ and $c_\gamma = 10$.

Then, a solution to overcome this problem is to suppose the Hessian $\nabla^2 G(m)$ to be invertible and to consider a stochastic Newton algorithm, i.e to consider an algorithm of the form

$$m_{n+1} = m_n - \frac{1}{n+1} \nabla^2 G(m)^{-1} \nabla_h g\left(X_{n+1}, m_n\right).$$

In the case of the linear regression, we would have

$$\mathbb{E}\left[\theta_{n+1} - \theta | \mathcal{F}_n\right] = \theta_n - \theta - \frac{1}{n+1} \nabla^2 G(\theta)^{-1} \nabla^2 G(\theta) \left(\theta_n - \theta\right) = \left(1 - \frac{1}{n+1}\right) \left(\theta_n - \theta\right).$$

Nevertheless, since we do generally not know the Hessian of G at m (and less access the inverse), one should replace it by a recursive estimate. We will see all along this chapter how to build such estimates as well as their inverse, and so, with reduced computational costs.

3.3 The stochastic Newton algorithm

3.3.1 Definition

In what follows, let us denote $H := \nabla^2 G(m)$ and suppose that Assumption **(A2)** holds. The Stochastic Newton algorithm (SN for short) is defined recursively for all $n \geq 0$ by [BGB20]

$$\tilde{m}_{n+1} = \tilde{m}_n - \frac{1}{n+1+c'_\gamma} \overline{H}_n^{-1} \nabla_h g\left(X_{n+1}, \tilde{m}_n\right) \tag{3.1}$$

with $\tilde{m}_0$ bounded and $c'_\gamma \geq 0$. Furthermore, $\overline{H}_n^{-1}$ is a recursive estimate of H^{-1}, symmetric and positive, and suppose that there is a filtration $(\mathcal{F}_n)$ satisfying

- $\overline{H}_n^{-1}$ and $\tilde{m}_n$ are $\mathcal{F}_n$-measurable.
- X_{n+1} is independent from $\mathcal{F}_n$.

Note that if we consider the filtration generated by the sample, if $\overline{H}_n^{-1}$ only depends on $X_1, \ldots, X_n$ and $\tilde{m}_0, \ldots, \tilde{m}_n$, the hypothesis on the filtration are so verified. We will see in Section 3.3.5 how to build such recursive estimates as well as their inverse for several examples.

3.3.2 Strong consistency

In order to obtain the almost sure rate of convergence, let us suppose from now that the recursive estimates of the Hessian verify the following assumption:

(H1) One can control the eigenvalues of $\overline{H}_n$: there is $\beta \in (0, 1/2)$ such that

$$\lambda_{\max}\left(\overline{H}_n\right) = O(1) \quad a.s. \qquad \text{and} \qquad \lambda_{\max}\left(\overline{H}_n^{-1}\right) = O\left(n^\beta\right) \quad a.s.$$

This assumption implies that, without knowing if $\tilde{m}_n$ converges, we are able to control the behavior of the smallest and largest eigenvalue of $\overline{H}_n$. Indeed, **(H1)** implies that $\liminf \lambda_{\min}\left(\overline{H}_n\right) > 0$ a.s. We will see in Section 3.3.5 how to modify natural recursive estimates of the Hessian in order to get new estimates satisfying this assumption. We can now give the strong consistency of the stochastic Newton estimates.

Theorem 3.3.1 ([BGB20]). *Suppose Assumptions **(A1a')**, **(A2)**, **(A3b)** and **(H1)** hold. Then*

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

Sketch of the proof. A Taylor's expansion of the functional G coupled with Assumptions **(A2)** and **(A3b)** leads to

$$\mathbb{E}\left[V_{n+1}|\mathcal{F}_n\right] \leq \left(1 + \frac{\tilde{C}_2 L_{\nabla G}}{2}\frac{1}{(n+1)^2}\left\|\overline{H}_n^{-1}\right\|_{op}^2\right) V_n - \frac{1}{n+1}\lambda_{\min}\left(\overline{H}_n^{-1}\right)\left\|\nabla G\left(\tilde{m}_n\right)\right\|^2 + \frac{\tilde{C}_1 L_{\nabla G}}{2}\frac{1}{(n+1)^2}\left\|\overline{H}_n^{-1}\right\|_{op}^2$$

with $V_n = G\left(\tilde{m}_n\right) - G(m)$. Thanks to Assumption **(H1)**, one has $\sum_{n\geq 0}\frac{1}{(n+1)^2}\left\|\overline{H}_n^{-1}\right\|_{op}^2 < +\infty$ a.s, and applying Robbins-Siegmund theorem, it comes that V_n converges almost surely to a finite random variable. In addition $\sum_{n\geq 0}\frac{1}{n+1}\lambda_{\min}\left(\overline{H}_n^{-1}\right)\left\|\nabla G\left(\tilde{m}_n\right)\right\|^2 < +\infty$ a.s and one can conclude with the help of Assumption **(H1)**. Remark that Assumption **(H1)** is purely theoretical and is only necessary to apply Robbins-Siegmund theorem. A possibility to avoid it could be to find a better Lyapunov function, which is, as far as we now, an open question.

3.3.3 Almost sure rate of convergence

In order to get the rate of convergence of the estimates, we unfortunately need the strong consistency of the estimates of the Hessian. In this aim, let suppose that the following assumption is fulfilled:

(H2) The estimate $\overline{H}_n$ converges almost surely to H.

This assumption is satisfied since having the almost sure convergence of the estimates $\tilde{m}_n$ leads to have the strong consistency of the estimates of the Hessian. We will see in Section 3.3.5 how to verify such hypothesis. We can now give the rate of convergence of the stochastic Newton estimates.

Theorem 3.3.2 ([BGB20]). *Suppose Assumptions **(A1a')**, **(A2)**, **(A3b)**, **(H1)** and **(H2)** hold. Then, for all $\delta > 0$,*

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

*Furthermore, if there exists $p > 1$ such that Assumption **(A1p)** is fulfilled and if Assumption **(A5a)** holds,*

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

Sketch of the proof. First, remark that one can rewrite stochastic Newton algorithm as

$$\tilde{m}_{n+1} - m = \tilde{m}_n - m - \frac{1}{n+1}\overline{H}_n^{-1}\nabla G\left(\tilde{m}_n\right) + \frac{1}{n+1}\overline{H}_n^{-1}\tilde{\xi}_{n+1}, \tag{3.2}$$

with $\tilde{\xi}_{n+1} := \nabla_h g(X_{n+1}, \tilde{m}_n) - \nabla G(\tilde{m}_n)$. Then, $(\tilde{\xi}_n)$ is a sequence of martingale differences adapted to the filtration $(\mathcal{F}_n)$. Linearizing the gradient, it comes

$$\tilde{m}_{n+1} - m = \tilde{m}_n - m - \frac{1}{n+1}\overline{H}_n^{-1} H(\tilde{m}_n - m) - \frac{1}{n+1}\overline{H}_n^{-1}\tilde{\delta}_n + \frac{1}{n+1}\overline{H}_n^{-1}\tilde{\xi}_{n+1}$$

where $\tilde{\delta}_n := \nabla G(\tilde{m}_n) - H(\tilde{m}_n - m)$ is the remainder term in the Taylor's decomposition of the gradient. This can also be written as

$$\begin{aligned}\tilde{m}_{n+1} - m = \left(1 - \frac{1}{n+1}\right)(\tilde{m}_n - m) - \frac{1}{n+1}\left(\overline{H}_n^{-1} - H^{-1}\right) H(\tilde{m}_n - m) - \frac{1}{n+1}\overline{H}_n^{-1}\tilde{\delta}_n \\ + \frac{1}{n+1}\overline{H}_n^{-1}\tilde{\xi}_{n+1}. \end{aligned} \tag{3.3}$$

and by induction, for all $n \geq 1$,

$$\tilde{m}_n - m = \underbrace{-\frac{1}{n}\sum_{k=0}^{n-1}\left(\overline{H}_k^{-1} - H^{-1}\right) H(\tilde{m}_k - m) - \frac{1}{n}\sum_{k=0}^{n-1}\overline{H}_k^{-1}\tilde{\delta}_k}_{=:\tilde{\Delta}_n} + \underbrace{\frac{1}{n}\sum_{k=0}^{n-1}\overline{H}_k^{-1}\tilde{\xi}_{k+1}}_{=:\tilde{M}_n}. \tag{3.4}$$

Then, one can apply a law of large numbers to the martingale term $\tilde{M}_n$ and prove that $\tilde{\Delta}_n$ is negligible.

3.3.4 Asymptotic efficiency

In order to get the asymptotic efficiency of the stochastic Newton estimates, it is often necessary to have a first rate of convergence of the estimates of the Hessian. In this aim, we suppose from now that the following assumption is fulfilled:

(H3) There exists $p_H > 0$ such that

$$\left\|\overline{H}_n - H\right\|_{op}^2 = O\left(\frac{1}{n^{p_H}}\right) \quad a.s.$$

Remark that this assumption is satisfied since having the almost sure rate of convergence of the estimates $\tilde{m}_n$ leads to have a rate of convergence of the estimates of the Hessian. We can now establish the asymptotic efficiency of the estimates.

Theorem 3.3.3 ([BGB20]). *Suppose Assumptions **(A1a')**, **(A2)**, **(A3b)**, **(A4a)**, **(A5a)**, and **(H1)** to **(H3)** hold. Suppose also that there exists $p > 1$ such that Assumption **(A1p)** is fulfilled. Then*

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

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

Sketch of the proof. The proof consists in proving that $\tilde{\Delta}_n$ in equation (3.4) is negligible thanks to Theorem 3.3.2 coupled with Assumption **(H3)**, before applying a Central Limit Theorem to the martingale term $\tilde{M}_n$.

3.3.5 Applications

All the applications given here rely on the use of the Riccati's formula (also called Sherman-Morrison formula) for matrix inversion given by

$$\left(A + uv^T\right)^{-1} = A^{-1} - \left(1 + v^T A^{-1} u\right)^{-1} A^{-1} u v^T A^{-1}$$

where $A \in \mathbb{R}^{n \times n}$ is invertible, $u, v \in \mathbb{R}^d$ and $1 + v^T A^{-1} u \neq 0$.

Application to the linear model

We now focus on the linear model case. Let us recall that the Hessian is defined for all $h \in \mathbb{R}^d$ by $\mathbb{E}\left[XX^T\right]$ and we suppose from now that it is positive. Then, a natural estimate is defined by

$$\overline{H}_n = \frac{1}{n+1}\left(\sum_{k=1}^{n} X_k X_k^T + H_0\right)$$

where H_0 is a matrix chosen positive (one can take $H_0 = I_d$ for instance). Note that one can rewrite the sequence $\left(\overline{H}_n\right)$ recursively as

$$\overline{H}_{n+1} = \overline{H}_n + \frac{1}{n+2}\left(X_{n+1} X_{n+1}^T - \overline{H}_n\right).$$

We now focus on the inversion of $\overline{H}_n$. In this aim, let us denote $H_n = (n+1)\overline{H}_n$, i.e one has the recursive relation

$$H_{n+1} = H_n + X_{n+1} X_{n+1}^T.$$

Then, with the help of Ricatti's formula, one can update the inverse of the Hessian matrix with only $O\left(d^2\right)$ operations, i.e for all n,

$$H_{n+1}^{-1} = H_n^{-1} - \left(1 + X_{n+1}^T H_n^{-1} X_{n+1}\right)^{-1} H_n^{-1} X_{n+1} X_{n+1}^T H_n^{-1}$$

This leads to the following Stochastic Newton algorithm

$$\begin{aligned} \tilde{\theta}_{n+1} &= \tilde{\theta}_n + H_n^{-1}\left(Y_{n+1} - \tilde{\theta}_n^T X_{n+1}\right) X_{n+1} \\ H_{n+1}^{-1} &= H_n^{-1} - \left(1 + X_{n+1}^T H_n^{-1} X_{n+1}\right)^{-1} H_n^{-1} X_{n+1} X_{n+1}^T H_n^{-1}. \end{aligned} \tag{3.5}$$

We can now give the rate of convergence of the estimates, which can be seen as a corollary of Theorems 3.3.2 and 3.3.3.

Corollaire 3.3.1 ([BGB20]). *Suppose there is $p > 0$ such that X and ϵ admit moment of order $4 + 4p$ and $2 + 2p$, and suppose that $H := \mathbb{E}\left[XX^T\right]$ is positive. Then, stochastic Newton estimates defined by (3.5) satisfy*

$$\left\|\tilde{\theta}_n - \theta\right\|^2 = O\left(\frac{\ln n}{n}\right) \quad a.s. \qquad and \qquad \sqrt{n}\left(\tilde{\theta}_n - \theta\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \mathbb{E}\left[\epsilon^2\right] H_{(LM)}^{-1}\right).$$

In Figure 3.2,we consider the linear model with

$$\theta = (-4,-3,-2,-1,0,1,2,3,4,5)^T \in \mathbb{R}^{10}, \qquad X \sim \mathcal{N}\left(0, \mathrm{diag}\left(\sigma_i^2\right)\right), \qquad \epsilon \sim \mathcal{N}(0,1) \tag{3.6}$$

where for all $i = 1, \ldots, d$, $\sigma_i^2 = \frac{i^2}{d^2}$. Remark that the largest eigenvalue of the Hessian is so 100 times larger than the smallest one. In Figure 3.2, one can see that gradient estimates do not achieve convergence, so that their averaged version cannot converge too, and even fewer accelerate the convergence. On the other hand, one can observe that stochastic Newton estimates converge very quickly, despite a lack of stability for the first steps. Nevertheless, this can be overcome tuning the parameter c'_γ (chosen equal to 0 here).

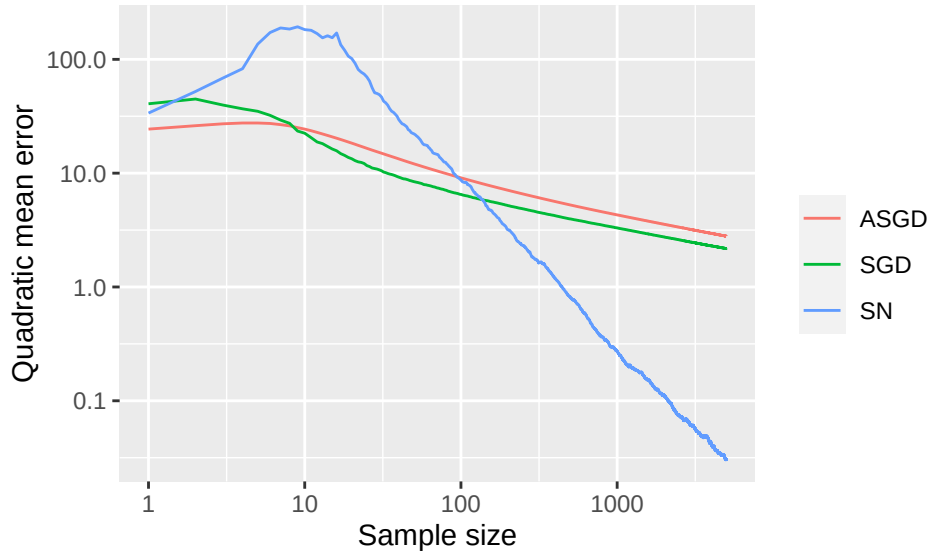


Figure 3.2 – Evolution of the quadratic mean error of the stochastic gradient estimates θ_n (SGD), their averaged version $\overline{\theta}_n$ (ASGD), and the stochastic Newton estimates $\tilde{\theta}_n$ (SN) with respect to the sample size n in the case of the linear model.

We have already seen that for the averaged estimates, one has

$$C_n := \frac{1}{\sigma^2} n \left(\overline{\theta}_n - \theta\right)^T \overline{H}_n \left(\overline{\theta}_n - \theta\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \chi_d^2,$$

and in a same way, Corollary 3.3.1 can be written as

$$K_n := \frac{1}{\sigma^2} n \left(\tilde{\theta}_n - \theta\right)^T \overline{H}_n \left(\tilde{\theta}_n - \theta\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \chi^2_d.$$

In Figure 3.3, we focus on the distribution functions of C_n and K_n. One can see that even in this ill-conditioned case, the distribution of K_n is close to the one of the Chi-square law, contrary C_n.

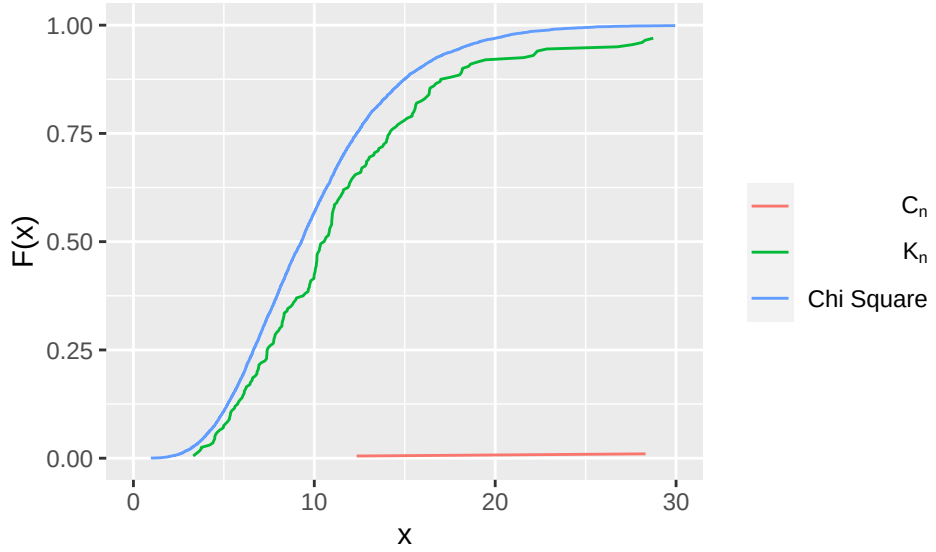


Figure 3.3 – Comparison of the distribution function of C_n and K_n (with $n = 5000$) with the distribution function of a Chi-square law with d degrees of freedom.

Application to logistic regression

We now focus on the estimation of the parameters of the logistic regression. Let us recall that the Hessian of the function to minimize is defined for all $h \in \mathbb{R}^d$ by

$$\nabla^2 G_{\log}(h) = \mathbb{E}\left[\pi\left(h^T X\right)\left(1 - \pi\left(h^T X\right)\right) X X^T\right],$$

with $\pi(x) = \frac{e^x}{1+e^x}$. A natural recursive estimate of the Hessian would have been

$$\overline{S}_n = \frac{1}{n+1}\left(\overline{S}_0 + \sum_{k=1}^{n} \pi\left(X_k^T \hat{\theta}_{k-1}\right)\left(1 - \pi\left(X_k^T \hat{\theta}_{k-1}\right)\right) X_k X_k^T\right)$$

with S_0 positive. Nevertheless, it is not easy (possible?) to prove that this estimate satisfies Assumption **(H1)**. In order to overcome this, we propose [BGBP19] a truncated version, leading to

the following Stochastic Newton algorithm

$$\begin{aligned}
\alpha_{n+1} &= \pi\left(\tilde{\theta}_n^T X_{n+1}\right)\left(1-\pi\left(\tilde{\theta}_n^T X_{n+1}\right)\right)\\
\tilde{\theta}_{n+1} &= \tilde{\theta}_n + H_n^{-1}\left(Y_{n+1}-\pi\left(\tilde{\theta}_n^T X_{n+1}\right)\right)X_{n+1}\\
H_{n+1}^{-1} &= H_n^{-1} - a_{n+1}\left(1+a_{n+1}X_{n+1}^T H_n^{-1} X_{n+1}\right)^{-1} H_n^{-1} X_{n+1} X_{n+1}^T H_n^{-1}
\end{aligned}$$

with H_0 symmetric and positive, $\tilde{\theta}_0$ bounded, $a_{n+1} = \max\left\{\alpha_{n+1}, \frac{c_\beta}{(n+1)^\beta}\right\}$ with $c_\beta > 0$ and $\beta \in (0,1/2)$. Remark that with the help of Riccati's formula, it comes

$$(n+1)\overline{H}_n := H_n = H_0 + \sum_{k=1}^{n} a_k X_k X_k^T$$

The truncation term a_n enables us to control the smallest eigenvalue of the estimates since, supposing that X admits a second order moment, one has

$$\frac{1}{\sum_{k=1}^n \frac{c_\beta}{k^\beta}} \sum_{k=1}^{n} \frac{c_\beta}{k^\beta} X_k X_k^T \xrightarrow[n\to+\infty]{a.s} \mathbb{E}\left[XX^T\right]$$

Then, supposing that $\mathbb{E}\left[XX^T\right]$ is invertible, it comes

$$\lambda_{\max}\left(\left(H_0 + \sum_{k=1}^{n} \frac{c_\beta}{k^\beta} X_k X_k^T\right)^{-1}\right) = O\left(n^{\beta-1}\right) \quad a.s,$$

i.e Assumption **(H1)** is satisfied. The following corollary (of Theorems 3.3.2 and 3.3.3) gives the rates of convergence of the truncated Stochastic Newton algorithm.

Corollaire 3.3.2 ([BGBP19]). *Suppose X admits a second order moment and that $H := \nabla^2 G(\theta)$ is invertible. Then $\tilde{\theta}_n$ converges almost surely to θ. Furthermore, if X admits a moment of order 4,*

$$\left\|\tilde{\theta}_n - \theta\right\|^2 = O\left(\frac{\ln n}{n}\right) \quad a.s. \qquad \textit{and} \qquad \sqrt{n}\left(\tilde{\theta}_n - \theta\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, H_{(\log)}^{-1}\right).$$

We now consider the model

$$\theta = (1,\ldots,1)^T \in \mathbb{R}^5 \qquad \text{and} \qquad X \sim \mathcal{N}\left(0, \operatorname{diag}\left(\sigma_i^2\right)\right)$$

where for all $i = 1,\ldots,d$, $\sigma_i^2 = \frac{i^2}{d^2}$. One can observe in Figure 3.4 that here again, estimates of the gradient do not achieve convergence while Stochastic Newton estimates converge very quickly, even with an Hessian with a complicated structure. Note that under assumptions, $\overline{H}_n$ converges almost surely to H, and it can be derived from Corollary 3.3.2 that

$$K_n := n\left(\tilde{\theta}_n - \theta\right)^T \overline{H}_n \left(\tilde{\theta}_n - \theta\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \chi_d^2.$$

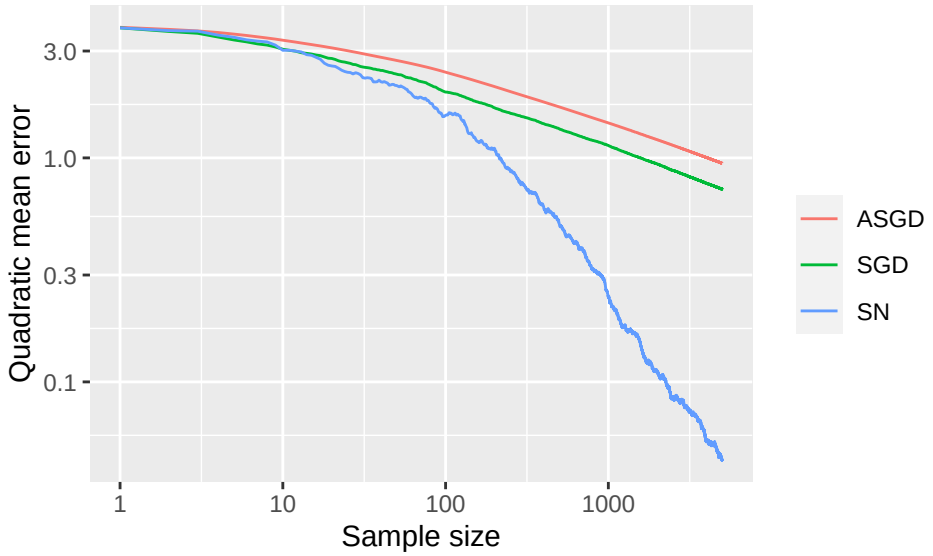


Figure 3.4 – Evolution of the quadratic mean error of the stochastic gradient estimates θ_n (SGD), their averaged version $\overline{\theta}_n$ (ASGD), and the stochastic Newton estimates $\tilde{\theta}_n$ (SN) with respect to the sample size n in the case of the logistic regression.

In Figure 3.5, we focus on the distribution functions of K_n (remark that as for the linear case, ASGD estimates do not converge at all, so that we only focus here on the behavior of K_n). One can see that even in this ill-conditioned case, the distribution of K_n is close to the one of the Chi-square law, and is surprisingly outperforming.

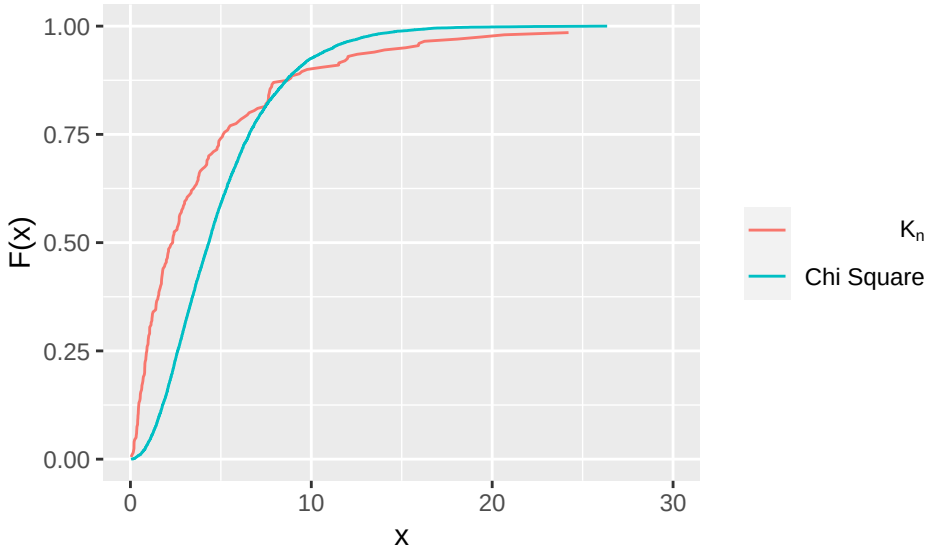


Figure 3.5 – Comparison of the distribution function of K_n (with $n = 5000$) with the distribution function of a Chi-square law with d degrees of freedom.

3.4 The Weighted Averaged Stochastic Newton algorithm

3.4.1 Definition

We have seen in the previous section that considering stochastic Newton algorithms can be helpful in the case where the problem is ill-conditioned. In addition, we have chosen a step sequence of the form $\frac{1}{n+1}$, which enables us to obtain asymptotically efficient estimates. Nevertheless, this can generate troubles in case of bad initialization, since it does not enable the estimates to "move" quickly [CGBP20]. In order to overcome this, the aim was first to propose an averaged stochastic Newton algorithm, which consists in using a step sequence of the form $\frac{1}{(n+1)^\gamma}$ to "move" faster, before using an averaging step to maintain the asymptotic efficiency. Nevertheless, averaging is known to be sensitive to a bad initialization and cannot be entirely considered to solve this problem. An alternative way is to consider a Weighted averaged version (WASNA), which consists in giving more weight to the last estimates in the averaging step, leading to the following recursive algorithm: for all $n \geq 0$,

$$\hat{m}_{n+1} = \hat{m}_n - \gamma_{n+1} \overline{S}_n^{-1} \nabla_h g\left(X_{n+1}, \hat{m}_n\right) \tag{3.7}$$

$$m_{n+1,\tau} = (1 - \tau_{n+1})\, m_{n,\tau} + \tau_{n+1} \hat{m}_{n+1}, \tag{3.8}$$

given $m_{\tau,0} = \hat{m}_0$, $\gamma_n = c_\gamma \left(n + c'_\gamma\right)^{-\gamma}$ with $c_\gamma > 0$, $c'_\gamma \geq 0$ and $\gamma \in (1/2, 1)$. Furthermore $\overline{S}_n^{-1}$ is a recursive estimates of H^{-1} symmetric and positive such that there is a filtration $(\mathcal{F}_n)$ verifying that $\overline{S}_n^{-1}$ and $\hat{m}_n$ are $\mathcal{F}_n$-measurable and X_{n+1} is independent from $\mathcal{F}_n$. Finally, the weighted averaging sequence (τ_n) should satisfy:

- (τ_n) is $\mathcal{GS}(-1)$ (see [MP11]), i.e

$$n\left(1 - \frac{\tau_{n-1}}{\tau_n}\right) \xrightarrow[n\to+\infty]{} -1$$

- There is a constant $\tau > 1/2$ such that

$$n\tau_n \xrightarrow[n\to+\infty]{} \tau.$$

As mentioned before, by choosing different sequences $(\tau_n)_n$, one can play more or less on the strength given to the last iterates of $\hat{m}_n$. For instance, choosing $\tau_n = \frac{1}{n+1}$ leads to the Averaged Stochastic Newton algorithm (ASN for short), i.e

$$m_{n,0} = \frac{1}{n+1} \sum_{k=0}^{n} \hat{m}_k.$$

Considering a sequence $\tau_n = \frac{(n+1)^\omega}{\sum_{k=0}^{n}(k+1)^\omega}$ enables to give much more weights to the last estimates,

and more precisely, this leads to

$$m_{n,\omega} = \frac{1}{\sum_{k=0}^{n}(k+1)^{\omega}} \sum_{k=0}^{n} (k+1)^{\omega} \hat{m}_k$$

Nevertheless, we will see later that this strategy, although it limited the effect of a bad initialization, generates a loss of efficiency. Then, a good trade-off is to consider a weighted averaging sequence of the form $\tau_n = \frac{\log(n+1)^{\omega}}{\sum_{k=0}^{n}\log(k+1)^{\omega}}$, with $\omega > 0$, which leads to

$$m_{n,\log,\omega} = \frac{1}{\sum_{k=0}^{n}\log(k+1)^{\omega}} \sum_{k=0}^{n} \log(k+1)^{\omega} \hat{m}_k.$$

3.4.2 Almost sure rate of convergence

In this section, we focus on the almost sure rate of convergence of the WASNA. In this aim, we first introduce a first assumption which enables to control the behavior of the eigenvalues of $\overline{S}_n^{-1}$, which is a derivative of Assumption **(H1)**.

(H1') One can control the eigenvalues of $\overline{S}_n^{-1}$: there exists $\beta \in (0, \gamma - 1/2)$ such that

$$\lambda_{\max}\left(\overline{S}_n\right) = O(1) \quad a.s \qquad \text{and} \qquad \lambda_{\max}\left(\overline{S}_n^{-1}\right) = O\left(n^{\beta}\right) \quad a.s.$$

Note that here again, this assumption ensures that $\liminf \lambda_{\min}\left(\overline{S}_n^{-1}\right) > 0$ end enables to apply Robbins-Siegmund Theorem since under **(H1')**, $\sum_{n\geq 0} \gamma_{n+1}^2 \left\|\overline{S}_n^{-1}\right\|_{op}^2 < +\infty$ a.s, which enables to prove the consistency of the estimates.

Theorem 3.4.1 ([BGB20]). *Suppose Assumptions **(A1a')**, **(A2)**, **(A3b)** and **(H1)'** hold. Then*

$$\hat{m}_n \xrightarrow[n\to+\infty]{a.s} m \qquad \text{and} \qquad m_{n,\tau} \xrightarrow[n\to+\infty]{a.s} m.$$

Sketch of the proof: The Taylor's decomposition of $V_{n+1} := G\left(\hat{m}_{n+1}\right) - G(m)$ leads to

$$\mathbb{E}\left[V_{n+1}|\mathcal{F}_n\right] \leq \left(1 + \frac{\tilde{C}_2 L_{\nabla G}}{2}\gamma_{n+1}^2 \left\|\overline{S}_n^{-1}\right\|_{op}^2\right) V_n - \gamma_{n+1}\lambda_{\min}\left(\overline{S}_n^{-1}\right) \left\|\nabla G\left(\hat{m}_n\right)\right\|^2 + \frac{\tilde{C}_1 L_{\nabla G}}{2}\gamma_{n+1}^2 \left\|\overline{S}_n^{-1}\right\|_{op}^2$$

and applying Robbins-Siegmund theorem, V_n converges almost surely to a finite random variable while $\sum_{n\geq 0} \gamma_{n+1}\lambda_{\min}\left(\overline{S}_n^{-1}\right) \left\|\nabla G\left(\hat{m}_n\right)\right\|^2 < +\infty$ a.s and Assumption **(H1')** enables to conclude.

As for the Stochastic Newton algorithm, we now have to suppose that $\overline{S}_n$ converges in order to get the rate of convergence of the WASNA. More precisely, we suppose from now the following assumption is fulfilled:

(H2') The estimate $\overline{S}_n$ converges almost surely to H.

This hypothesis just means that obtaining the almost sure convergence of the WASNA estimates leads to to the strong consistency of the estimates of the Hessian, and enables to prove the following theorem:

Theorem 3.4.2 ([BGB20]). *Suppose Assumptions **(A1**η**)**, **(A2)**, **(A3b)**, **(H1')** and **(H2')** hold. Then,*

$$\|\hat{m}_n - m\|^2 = O\left(\frac{\ln n}{n^\gamma}\right) \quad a.s.$$

Then, we obtain the usual rate of convergence $\frac{1}{n^\gamma}$ (up to the log term) for this kind of step sequence, and so, with weak assumptions.

Sketch of the proof. Remark that one can rewrite the algorithm as

$$\hat{m}_{n+1} - m = \hat{m}_n - m - \gamma_{n+1}\overline{S}_n^{-1}\nabla G\left(\hat{m}_n\right) + \gamma_{n+1}\overline{S}_n^{-1}\hat{\xi}_{n+1} \tag{3.9}$$

where $\hat{\xi}_{n+1} := \nabla G\left(\hat{m}_n\right) - \nabla_h g\left(X_{n+1}, \hat{m}_n\right)$ is a martingale difference for the filtration $(\mathcal{F}_n)$. Furthermore, linearizing the gradient, it comes

$$\hat{m}_{n+1} - m = \hat{m}_n - m - \gamma_{n+1}\overline{S}_n^{-1}H\left(\hat{m}_n - m\right) + \gamma_{n+1}\overline{S}_n^{-1}\hat{\xi}_{n+1} - \gamma_{n+1}\overline{S}_n^{-1}\hat{\delta}_n$$

where $\hat{\delta}_n := \nabla G\left(\hat{m}_n\right) - H\left(\hat{m}_n - m\right)$ is the rest term in the Taylor's decomposition of the gradient. Introducing H^{-1}, it comes

$$\hat{m}_{n+1} - m = \left(1 - \gamma_{n+1}\right)\left(\hat{m}_n - m\right) + \left(H^{-1} - \overline{S}_n^{-1}\right)\left(\hat{m}_n - m\right) + \gamma_{n+1}\overline{S}_n^{-1}\hat{\xi}_{n+1} - \gamma_{n+1}\overline{S}_n^{-1}\hat{\delta}_n. \tag{3.10}$$

Then, with the help of an induction, one can prove that

$$\begin{aligned}\hat{m}_n - m = \hat{\beta}_{n,0}\left(\hat{m}_0 - m\right) + \sum_{k=0}^{n-1}\hat{\beta}_{n,k+1}\gamma_{k+1}\left(H^{-1} - \overline{S}_k^{-1}\right)\left(\hat{m}_k - m\right) + \sum_{k=0}^{n-1}\hat{\beta}_{n,k+1}\gamma_{k+1}\overline{S}_k^{-1}\hat{\xi}_{k+1} \\ - \sum_{k=0}^{n-1}\beta_{n,k+1}\gamma_{k+1}\overline{S}_k^{-1}\hat{\delta}_k\end{aligned} \tag{3.11}$$

with $\hat{\beta}_{n,n} = 1$ and $\hat{\beta}_{n,k} = \prod_{j=k+1}^{n}\left(1 - \gamma_j\right)$. Then, one can easily prove that the first term on the right hand side of previous equality converges exponentially fast, before applying Theorem 6.1 in [CGBP20] to the third one and prove that the other ones converge at least at the same rate as the third one.

Remark 3.4.1. *Remark that equality (3.11) represents an important difference with the usual stochastic gradient algorithm. Indeed, in the case of the gradient, one has an analogous equality but with* $\beta_{n,k} = \prod_{j=k+1}^{n}\left(1 - \gamma_j H\right)$. *For a fixed* k*, this term converges exponentially fast to* 0 *and more precisely at a rate* $O\left(e^{-\frac{\lambda_{\min}}{1-\alpha}c_\gamma n^{1-\alpha}}\right)$ *but this convergence only begins when* $\lambda_{\max}(H)\gamma_k \leq 1$*. Then, one could take* c_γ *close to*

$\lambda_{\max}^{-1}$ to begin to converge quickly but this would lead to a convergence of order $O\left(e^{-\frac{1}{1-\alpha}\frac{\lambda_{\min}}{\lambda_{\max}}n^{1-\alpha}}\right)$ which is a bad convergence when the eigenvalues of H are at very different scales or when n is not large enough. On the contrary, one could take c_γ large to accelerate the convergence, but it would mean that it could spend a lot of time before this term starts converging. This seems to confirm that having $1-\gamma_n$ enables to take c_γ close to one and quickly converge whatever happens with the eigenvalues of H.

In order to obtain the rate of convergence of its weighted averaged version, let us now introduce a last assumption, which is a derivative of **(H3)**.

(H3') There is a positive constant $p_S > \frac{1}{2} - \frac{\gamma}{2}$ such that

$$\left\|\overline{S}_n^{-1} - H^{-1}\right\|^2 = O\left(\frac{1}{n^{p_S}}\right) \quad a.s.$$

In other words, this assumptions means that obtaining the rate of convergence of the stochastic Newton estimates with stepsequence γ_n leads to obtain a rate of convergence for the estimates of the Hessian of order at least $\frac{1-\gamma}{2} < \frac{1}{4}$. We can now give the rate of convergence of WASNA estimates.

Theorem 3.4.3 ([BGB20]). *Suppose Assumptions **(A1η)**, **(A2)**, **(A3b)**, **(A4a)**, **(A5a)** and **(H1')** to **(H3')** hold. Then,*

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

We so achieve the usual rate of convergence, and so, choosing any stepsequence τ_n verifying our assumptions.

Sketch of the proof. First, with the help of an induction, one can rewrite $m_{\tau,n}$ as

$$m_{n,\tau} - m = \kappa_{n,0}\left(m_{n,\tau} - m\right) + \sum_{k=0}^{n-1} \kappa_{n,k+1}\tau_{k+1}\left(\hat{m}_k - m\right)$$

with $\kappa_{n,n} = 1$ and $\kappa_{n,k} = \prod_{i=k+1}^{n}(1-\tau_i)$. Furthermore, let us remark that as for averaged gradient algorithms, one can rewrite equality (3.10) as

$$\hat{m}_k - m = \frac{\hat{m}_k - \hat{m}_{k+1}}{\gamma_{k+1}} + \left(H^{-1} - \overline{S}_n^{-1}\right)\left(\hat{m}_k - m\right) + \overline{S}_k^{-1}\hat{\xi}_{k+1} - \overline{S}_k^{-1}\hat{\delta}_k$$

Multiplying these equalities by $\kappa_{k+1,0}^{-1}\tau_{k+1}$, and summing it, one has

$$\begin{aligned}\sum_{k=0}^{n-1}\tau_{k+1}\left(\hat{m}_k - m\right) = &\sum_{k=0}^{n-1}\kappa_{k+1,0}^{-1}\tau_{k+1}\frac{\hat{m}_k - \hat{m}_{k+1}}{\gamma_{k+1}} + \sum_{k=0}^{n-1}\kappa_{k+1,0}^{-1}\tau_{k+1}\left(H^{-1} - \overline{S}_n^{-1}\right)\left(\hat{m}_k - m\right)\\ &+ \sum_{k=0}^{n-1}\kappa_{k+1,0}^{-1}\tau_{k+1}\overline{S}_k^{-1}\hat{\xi}_{k+1} - \sum_{k=0}^{n-1}\kappa_{k+1,0}^{-1}\tau_{k+1}\overline{S}_k^{-1}\hat{\delta}_k.\end{aligned}$$

Adding $\hat{m}_0 - m$ and multiplying by $\kappa_{n,0}$, it comes

$$m_{n,\tau} - m = \kappa_{n,0}\left(\hat{m}_0 - m\right) + \sum_{k=0}^{n-1} \kappa_{n,k+1}\tau_{k+1}\frac{\hat{m}_k - \hat{m}_{k+1}}{\gamma_{k+1}} + \sum_{k=0}^{n-1} \kappa_{n,k+1}\tau_{k+1}\left(H^{-1} - \overline{S}_n^{-1}\right)\left(\hat{m}_k - m\right)$$
$$+ \sum_{k=0}^{n-1} \kappa_{n,k+1}\tau_{k+1}\overline{S}_k^{-1}\hat{\xi}_{k+1} - \sum_{k=0}^{n-1} \kappa_{n,k+1}\tau_{k+1}\overline{S}_k^{-1}\hat{\delta}_k \tag{3.12}$$

Then, one has to apply a law of large numbers to the fourth term on the right-hand side of previous equality before proving that the other ones are negligible (with the help of Theorem 3.4.2 and Assumption **(H3')**).

3.4.3 Asymptotic normality

We now focus on the asymptotic normality of WASNA estimates.

Theorem 3.4.4. *[[BGB20]] Suppose Assumptions **(A1η)**, **(A2)**, **(A3b)**, **(A4a)**, **(A5a)** and **(H1')** to **(H3')** hold. Then,*

$$\sqrt{n}\left(m_{n,\tau} - m\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \frac{\tau^2}{2\tau - 1}H^{-1}\Sigma H^{-1}\right).$$

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

In order to prove this results, one just has to apply a Central Limit Theorem for martingales to the fourth term on the right-hand side of equality (3.12). Note that in the case where $\tau_n = \frac{1}{n+1}$, i.e for the usual averaging, one has $\tau = 1$, i.e the averaged stochastic Newton algorithm is asymptotically efficient:

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

Nevertheless, if one chooses $\tau_n = \frac{(n+1)^\omega}{\sum_{k=0}^n (k+1)^\omega}$ with $\omega > 0$, one has $\tau = \omega + 1$, leading to

$$\sqrt{n}\left(m_{n,\omega} - m\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \frac{(1+\omega)^2}{2\omega + 1}H^{-1}\Sigma H^{-1}\right).$$

and since $\frac{(1+\omega)^2}{2\omega+1} > 1$, the estimates are not asymptotically efficient. Finally, if one chooses $\tau_n = \frac{\log(n+1)^\omega}{\sum_{k=0}^n \log(k+1)^\omega}$, the weighted averaged estimates are asymptotically efficient, i.e

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

Note that in most of cases, Assumption **(H3')** can be easily verified and one can so apply Theorem 3.4.4. Nevertheless, we now propose a way to by-pass this assumption in the case where the estimates of the Hessian have a particular form (when we are able to use Riccati's formula).

Theorem 3.4.5 ([BGB20]). *Suppose that the Hessian estimates* $\left(\overline{S}_n\right)$ *are of the form*

$$\overline{S}_n = \frac{1}{n+1}\left(S_0 + \sum_{k=1}^{n} \hat{u}_k \hat{\Phi}_k \hat{\Phi}_k^T + \sum_{k=1}^{n} \frac{c_\beta}{k^\beta} Z_k Z_k^T\right)$$

with S_0 *symmetric and positive,* $c_\beta \geq 0$ *and* $\beta \in (0, \gamma - 1/2)$, $(Z_k)_k$ *are standard Gaussian vectors in dimension* d,

$$\hat{u}_k = u_k\left(X_k, m_{\tau,k-1}\right) \in \mathbb{R}, \qquad \text{and} \qquad \hat{\Phi}_k = \Phi_k\left(X_k, m_{\tau,k-1}\right) \in \mathbb{R}^d.$$

Furthermore, assume that

- *for all* $\delta > 0$, *there is a positive constant* C_δ *such that for all* k,

$$\mathbb{E}\left[\left\|\hat{u}_k \hat{\Phi}_k \hat{\Phi}_k^T\right\| \mathbf{1}_{\left\{\|m_{\tau,k-1} - m\| \leq (\ln k)^{1/2+\delta}\sqrt{\gamma_k}\right\}} |\mathcal{F}_{k-1}\right] \leq C_\delta$$

- *There is* $\alpha \in (1/2, 1)$ *and* $\delta > 0$ *such that*

$$\sum_{k\geq 0} (k+1)^{2\alpha} \frac{\tau_{k+1}^2}{\gamma_{k+1}} \frac{(\ln(k+1))^{1+\delta}}{(k+1)^2} \mathbb{E}\left[\left\|\hat{u}_k \hat{\Phi}_k \hat{\Phi}_k^T\right\|^2 \mathbf{1}_{\left\{\|m_{\tau,k-1} - m\| \leq (\ln k)^{1/2+\delta}\sqrt{\gamma_k}\right\}} |\mathcal{F}_{k-1}\right] < +\infty.$$

Let us also suppose that Assumptions ***(A1η)***, ***(A2)***, ***(A3b)***, ***(A4a)***, ***(A5a)***, ***(H1')*** *and* ***(H2')*** *hold. Then*

$$\|m_{n,\tau} - m\|^2 = O\left(\frac{\ln n}{n^\gamma}\right) \quad a.s \qquad \text{and} \qquad \sqrt{n}\left(m_{n,\tau} - m\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \frac{\tau^2}{2\tau - 1} H^{-1} \Sigma H^{-1}\right).$$

Remark that usually, to prove that Assumption **(H3')** is fulfilled, one has to prove that the functional

$$h \longmapsto \mathbb{E}\left[u_k\left(X_k, h\right) \Phi_k\left(X_k, h\right) \Phi_k\left(X_k, h\right)^T\right]$$

is Lipschitz on a neighborhood of m. Previous theorem enables to obtain the convergence replacing this by, for instance, if $\alpha \leq \frac{3-\gamma}{2}$, one can "just" prove that the functional

$$h \longmapsto \mathbb{E}\left[\left\|u_k\left(X_k, h\right) \Phi_k\left(X_k, h\right) \Phi_k\left(X_k, h\right)^T\right\|^2 \mathbf{1}_{\left\{\|h - m\| \leq (\ln k)^{1/2+\delta}\sqrt{\gamma_k}\right\}} |\mathcal{F}_{k-1}\right]$$

is uniformly bounded.

3.4.4 Applications and comparison with other methods

Simulation scheme

In this section, we assess the numerical performance of the weighted stochastic Newton algorithm and compare it to that of second-order online algorithms:

- the stochastic Newton algorithm (SN) defined in (3.1) with a step in $1/n$, similar to the one studied in [BGBP19] specifically for the logistic regression;

- the stochastic Newton algorithm (SN) defined in (3.7) with a step in $n^{-3/4}$;
- the averaged stochastic Newton algorithm (SNA) given in (3.8), with standard weighting ($\tau_n = 1/(n+1)$);
- the weighted averaged stochastic Newton algorithm (WASNA) given in (3.8) with logarithmic weighting ($\tau_n = \frac{\log(n+1)^\omega}{\sum_{k=0}^{n}\log(k+1)^\omega}$ and $\omega = 2$);

with first-order online methods:

- the stochastic gradient algorithm (SGD) with step $n^{-3/4}$;
- the averaged Stochastic Gradient Algorithm (ASGD);

and finally with first-order online algorithms mimicking second-order ones:

- the Adagrad algorithm [DHS11], which uses adaptive step sizes using only first-order information,
- the averaged Adagrad algorithm, with standard weighting.

We illustrate their performance in two different learning tasks, the case of linear and logistic regressions, for simple and more complex structured input data.

Application to the linear model

Let us recall that a natural estimate of the Hessian is defined by

$$\overline{H}_n = \frac{1}{n+1}\left(\sum_{k=1}^{n} X_k X_k^T + H_0\right)$$

where H_0 is a matrix chosen positive and can update it recursively with the help of Riccati's formula. Then, the Weighted Stochastic Newton algorithm is defined by

$$\hat{\theta}_{n+1} = \hat{\theta}_n + \gamma_{n+1}(n+1)H_n^{-1}\left(Y_{n+1} - \tilde{\theta}_n^T X_{n+1}\right)X_{n+1} \tag{3.13}$$

$$\theta_{n+1,\tau} = (1-\tau_{n+1})\,\theta_{n,\tau} + \hat{\theta}_{n+1} \tag{3.14}$$

$$H_{n+1}^{-1} = H_n^{-1} - \left(1 + X_{n+1}^T H_n^{-1} X_{n+1}\right)^{-1} H_n^{-1} X_{n+1} X_{n+1}^T H_n^{-1}.$$

One can now obtain the rate of convergence of the estimates, which can be seen as a corollary of Theorems 3.4.2, 3.4.3 and 3.4.4.

Corollaire 3.4.1 ([BGB20]). *Suppose there is $p > 0$ such that X and ϵ moment of order $4+4p$ and $2+2p$, and suppose that $H := \mathbb{E}\left[XX^T\right]$ is positive. Then, WASNA estimates defined by (3.13) and (3.14) satisfy*

$$\left\|\hat{\theta}_n - \theta\right\|^2 = O\left(\frac{\ln n}{n^\gamma}\right) \quad a.s. \qquad \text{and} \qquad \|\theta_{n,\tau} - \theta\|^2 = O\left(\frac{\ln n}{n}\right) \quad a.s.$$

Furthermore,

$$\sqrt{n}\left(\theta_{n,\tau}-\theta\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \frac{\tau^2}{2\tau-1}\mathbb{E}\left[\epsilon^2\right] H_{(LM)}^{-1}\right).$$

We now consider the model given by (3.6). Let us recall that that in such a case the Hessian associated to this model is equal to $\mathrm{diag}\left(\frac{i^2}{d^2}\right)_{i=1,\dots,10}$, meaning that the largest eigenvalue is 100 times larger than the smallest one. Therefore, considering stochastic gradient estimates leads to a step sequence which cannot be adapted to each direction. In Figure 3.6, we monitor the quadratic mean error of the different estimates, for three different type of initializations. One can see that both averaged Newton methods and the stochastic Newton method with step size of the order $1/n$ outperform all the other algorithms, specially for far intializations (right). The faster convergence of Newton methods or of the Adagrad algorithm compared to the one of standard SGD can be explained by their ability to manage the diagonal structure of the Hessian matrix, with eigenvalues at different scales.

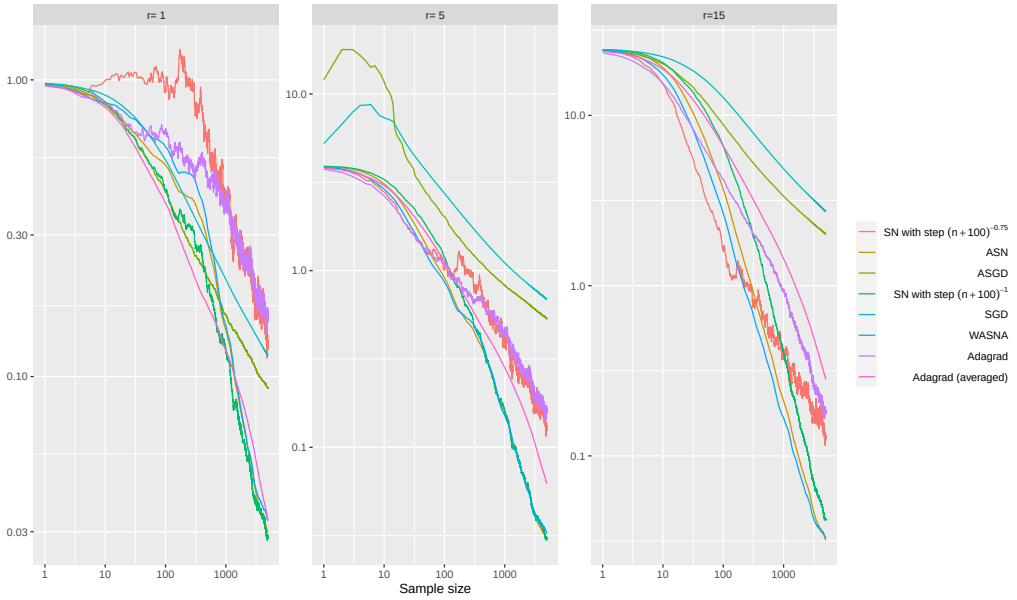


Figure 3.6 – Quadratic mean error of the estimates with respect to the sample size for different initializations: $\theta_0 = \theta + rU$, where U is a uniform random variable on the unit sphere of $\mathbb{R}^d$ with $r = 1$ (left), $r = 2$ (middle) or $r = 5$ (right).

Consider now a more complex covariance structure of the data, such as follows

$$X \sim \mathcal{N}\left(0, A\,\mathrm{diag}\left(\frac{i^2}{d^2}\right)_{i=1,\dots,d} A^T\right)$$

where A is a random orthogonal matrix. This particular choice of the covariates distribution, by the action of A, allows to consider strong correlations between the coordinates of X. In Figure 3.7, one can notice that the choice of adaptive step size used in the Adagrad algorithm is no longer sufficient to give the best convergence result in the presence of highly correlated data. In such

a case, both averaged Newton algorithms remarkably perform, showing their ability to handle complex second-order structure of the data, and all the more so for bad initializations (right).

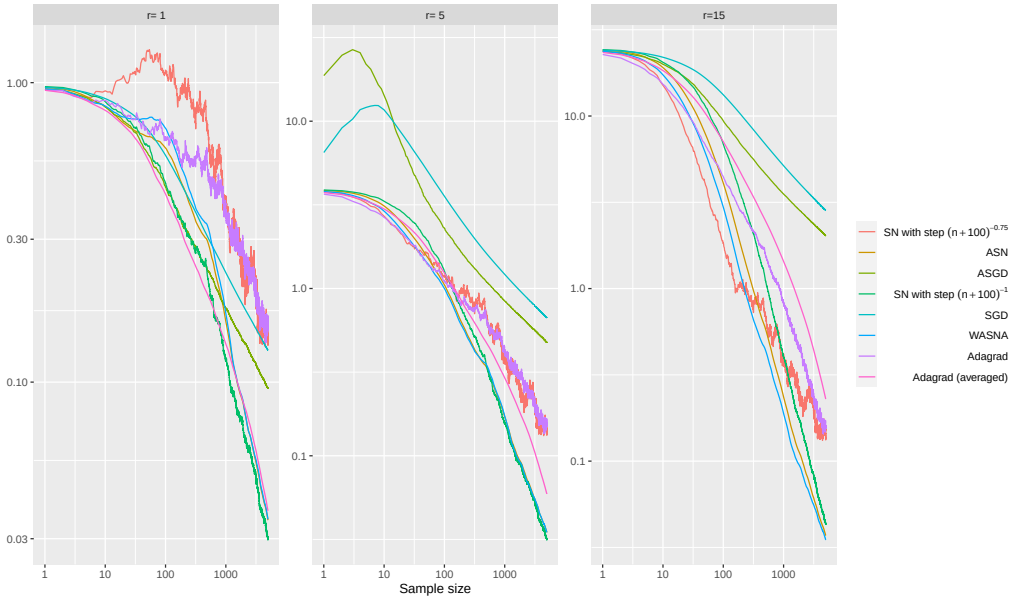


Figure 3.7 – Quadratic mean error of the estimates with respect to the sample size for different initializations: $\theta_0 = \theta + rU$, where U is a uniform random variable on the unit sphere of $\mathbb{R}^d$ with $r = 1$ (left), $r = 2$ (middle) or $r = 5$ (right).

Logistic regression

We now focus on the estimation of the parameter of the logistic regression. The Weighted Stochastic Newton algorithm can be written as

$$\begin{aligned}
\overline{\alpha}_{n+1} &= \pi\left(\theta_{n,\tau}^T X_{n+1}\right)\left(1-\pi\left(\theta_{n,\tau}^T X_{n+1}\right)\right) \\
\hat{\theta}_{n+1} &= \hat{\theta}_n + \gamma_{n+1}\overline{S}_n^{-1}\left(Y_{n+1}-\pi\left(\hat{\theta}_n^T X_{n+1}\right)\right)X_{n+1} \qquad (3.15)\\
\theta_{n+1,\tau} &\left(1-\tau_{n+1}\right)\theta_{n,\tau}+\tau_{n+1}\hat{\theta}_{n+1} \qquad (3.16)\\
S_{n+1}^{-1} &= S_n^{-1} - \overline{a}_{n+1}\left(1+\overline{a}_{n+1}X_{n+1}^T S_n^{-1} X_{n+1}\right)^{-1} S_n^{-1} X_{n+1} X_{n+1}^T S_n^{-1}
\end{aligned}$$

with S_0 symmetric and positive, $\theta_{\tau,0} = \hat{\theta}_0$ bounded, $\overline{a}_{n+1} = \max\left\{\overline{\alpha}_{n+1}, \frac{c_\beta}{(n+1)^\beta}\right\}$ with $c_\beta > 0$ and $\beta \in (0, \gamma - 1/2)$. Remark that with the help of Riccati formula, it comes

$$(n+1)\overline{S}_n := S_n = S_0 + \sum_{k=1}^{n} \overline{a}_k X_k X_k^T$$

and the truncation so ensures that Assumption **(H1')** is verified. Remark that we inject the weighted averaged estimates in the estimates of the Hessian. The following corollary (of Theorems 3.4.2,

3.4.3 and 3.4.4.) gives the rates of convergence of the Weighted truncated Stochastic Newton algorithm.

Corollaire 3.4.2 ([BGBP19]). *Suppose X admits a fourth order moment and that $H_{(\log)} := \nabla^2 G_{\log}(\theta)$ is invertible. Then, WASNA estimates defined by (3.15) and (3.16) satisfy*

$$\left\|\hat{\theta}_n - \theta\right\|^2 = O\left(\frac{\ln n}{n^\gamma}\right) \quad a.s. \qquad and \qquad \left\|\theta_{n,\tau} - \theta\right\|^2 = O\left(\frac{\ln n}{n}\right) \quad a.s.$$

Furthermore,

$$\sqrt{n}\left(\theta_{n,\tau} - \theta\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \frac{\tau^2}{2\tau - 1} H_{(\log)}^{-1}\right).$$

We consider the setting given in [BGBP19] where $\theta = (9,0,3,9,4,9,15,0,7,1,0)^T \in \mathbb{R}^{11}$, with an intercept and standard Gaussian input variables, i.e $X = \left(1, \Phi^T\right)^T$ with $\Phi \sim \mathcal{N}(0, I_{10})$. In Figure 3.8, we display the evolution of the quadratic mean error of the different estimates, for three different initializations. The Newton methods converge faster, in terms of distance to the optimum, than online gradient descents, which can be again explained by the Hessian structure of the model: even if the features are standard Gaussian random variables, the non-linearity introduced by the logistic model leads to a covariance structure difficult to apprehend theoretically and numerically by first-order online algorithms. One can see that in case of bad initialization, the step choice for the non-averaged Newton algorithm is crucial: choosing a step sequence of the form $1/n$ slows down the optimization dynamics, whereas a step choice $n^{-3/4}$) allows to reach the optimum much more quickly.

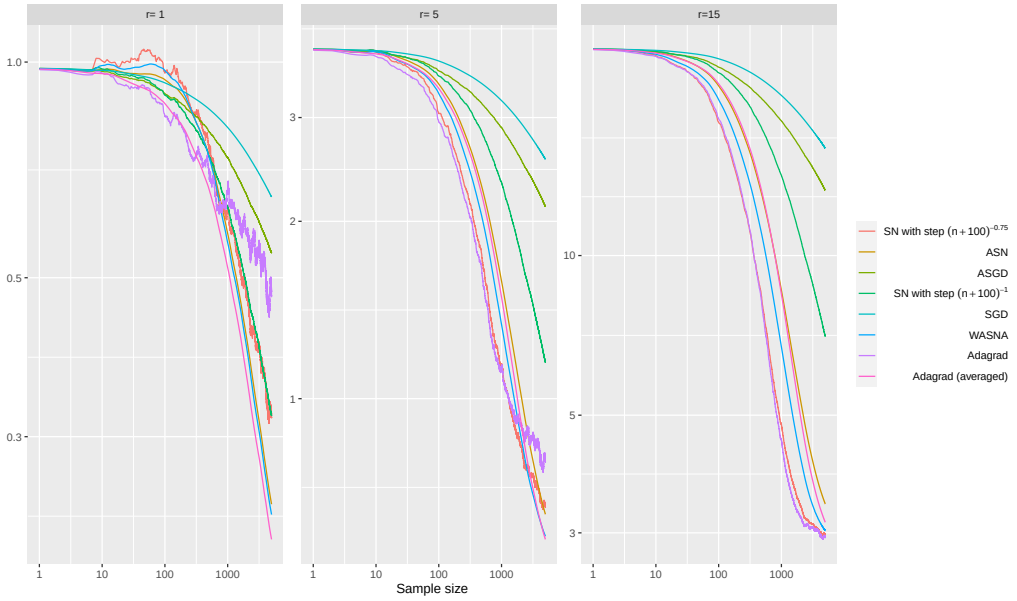


Figure 3.8 – Quadratic mean error of the estimates with respect to the sample size for different initializations: $\theta_0 = \theta + rU$, where U is a uniform random variable on the unit sphere of $\mathbb{R}^d$ with $r = 1$ (left), $r = 2$ (middle) or $r = 5$ (right).

3.4.5 Application to Softmax regression

Let us introduce the Softmax regression model. In this aim, let K be a positive integer larger than 1 and let us consider a couple of random variables (X, Y) lying in $\mathbb{R}^d \times \{1, \ldots, K\}$ verifying for all $k = 1, \ldots, K$,

$$\mathbb{P}\left[Y = k|X\right] = \frac{e^{\theta_k^T X}}{\sum_{k'=1}^{K} e^{\theta_{k'}^T X}}$$

where $\theta_1, \ldots, \theta_K \in \mathbb{R}^d$. In what follows, we denote $\boldsymbol{\theta} := \left(\theta_1^T, \ldots, \theta_K^T\right)^T$. Considering independent couples of random variables $(X_1, Y_1), \ldots, (X_n, Y_n)$ with the same law as (X, Y), the log-likelihood is defined by

$$l_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log\left(\frac{e^{\theta_{Y_i}^T X_i}}{\sum_{k=1}^{K} e^{\theta_k^T X_i}}\right).$$

Then, considering the asymptotic objective function, the aim is to minimize the convex functional $G_S : \mathbb{R}^d \times \ldots \times \mathbb{R}^d \longrightarrow \mathbb{R}$ defined for all h by

$$G_S(h) = -\mathbb{E}\left[\log\left(\frac{e^{h_Y^T X}}{\sum_{k=1}^{K} e^{h_k^T X}}\right)\right] =: \mathbb{E}\left[g_S(X, Y, h)\right].$$

Remark that if X admits a second order moment, the functional G_S is differentiable with

$$\nabla G(h) = \mathbb{E}\left[\begin{pmatrix} X\left(\frac{e^{h_1^T X}}{\sum_{k=1}^{K} e^{h_k^T X}} - \mathbf{1}_{Y=1}\right) \\ \vdots \\ X\left(\frac{e^{h_K^T X}}{\sum_{k=1}^{K} e^{h_k^T X}} - \mathbf{1}_{Y=K}\right) \end{pmatrix}\right] = \mathbb{E}\left[\nabla_h g_S(X, Y, h)\right]$$

and one can easily check that $\boldsymbol{\theta}$ is a zero of the gradient. Note that the functional G is twice continuously differentiable and in order to provide estimates of the Hessian we can easily recursively invert, one has to remark that

$$H_S := \nabla^2 G_S\left(\boldsymbol{\theta}\right) = \mathbb{E}\left[\nabla_h g_S\left(X, Y, \boldsymbol{\theta}\right) \nabla_h g_S\left(X, Y, \boldsymbol{\theta}\right)^T\right].$$

We can now define the WASNA estimates for all $n \geq 0$ as

$$\begin{aligned}
&\overline{\Phi}_{n+1} = \nabla_h g\left(X_{n+1}, Y_{n+1}, \theta_{n,\tau}\right) \\
&\tilde{\theta}_{n+1} = \tilde{\theta}_n - \gamma_{n+1} \overline{S}_n^{-1} \nabla_h g\left(X_{n+1}, Y_{n+1}, \theta_n\right) \qquad (3.17) \\
&\theta_{n+1,\tau} = \left(1 - \tau_{n+1}\right) \theta_{n,\tau} + \tau_{n+1} \tilde{\theta}_{n+1} \qquad (3.18) \\
&S_{n+\frac{1}{2}}^{-1} = S_n^{-1} - \left(1 + \beta_{n+1} Z_{n+1}^T S_n^{-1} Z_{n+1}\right)^{-1} \beta_{n+1} S_n^{-1} Z_{n+1} Z_{n+1}^T S_n^{-1} \\
&S_{n+1}^{-1} = S_{n+\frac{1}{n}}^{-1} - \left(1 + \overline{\Phi}_{n+1}^T S_{n+\frac{1}{n}}^{-1} \overline{\Phi}_{n+1}\right)^{-1} S_{n+\frac{1}{n}}^{-1} \overline{\Phi}_{n+1} \overline{\Phi}_{n+1}^T S_{n+\frac{1}{n}}^{-1},
\end{aligned}$$

with $\overline{S}_n = (n+1)S_n^{-1}$, $\boldsymbol{\theta}_0$ bounded, S_0 symmetric and positive, $\beta_n = c_\beta n^{-\beta}$ with $c_\beta > 0$ and $\beta \in (0, \gamma - 1/2)$. Finally, $Z_1, \ldots, Z_n, Z_{n+1}, \ldots$ are i.i.d with $Z_1 \sim \mathcal{N}(0, I_{d\times K})$. Remark that with the help of Ricatti's formula applied twice, one has

$$\overline{S}_n = \frac{1}{n+1}\sum_{i=1}^{n} \overline{\Phi}_i \overline{\Phi}_i^T + \frac{1}{n+1}\left(S_0 + \sum_{i=1}^{n} \beta_i Z_i Z_i^T\right).$$

Then, the first term on the right-hand side of previous equality is a natural recursive estimates of H_S while the second term enables to ensure that Assumption **(H1')** is fulfilled since by the law of large numbers it comes

$$\frac{1}{\sum_{i=1}^{n}\beta_i}\sum_{i=1}^{n} \beta_i Z_i Z_i^T \xrightarrow[n\to+\infty]{a.s} I_{d\times K}.$$

Remark that one can also consider the trick introduced in [BBGS21], i.e to consider $Z_i = e_{i \bmod d+1}$ where e_j, $j = 1, \ldots, d$ are the element of the canonical basis (see also [GBPL22] for more details). We can now give the rate of convergence of the Newton estimates, which can be seen as a corollary of Theorems 3.4.2, 3.4.3 and 3.4.4.

Corollaire 3.4.3 ([BGB20]). *Suppose X admits a fourth order moment and that H_S is invertible. Then*

$$\left\|\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right\|^2 = O\left(\frac{\ln n}{n^\gamma}\right) \quad a.s \qquad \text{and} \qquad \left\|\boldsymbol{\theta}_{n,\tau} - \boldsymbol{\theta}\right\|^2 = O\left(\frac{\ln n}{n}\right) \quad a.s$$

In addition

$$\sqrt{n}\left(\boldsymbol{\theta}_{n,\tau} - \boldsymbol{\theta}\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \frac{\tau^2}{2\tau - 1} H_S^{-1}\right).$$

We focus here on the MNIST[2] real dataset, in order to illustrate the performance of the WASNA in a context of multi-label classification. It consists in 70000 pictures of 28×28 pixels representing handwritten digits recast into vectors of dimension 784. The goal is to predict the digit $Y \in \{0, \ldots, 9\}$ represented on each vectorized image $X \in \mathbb{R}^{784}$, where each coordinate gives the contrast (between 0 and 255) of each pixel. This is then a multi-label classification setting with 10 different classes. In a preprocessing step, we normalize the features between 0 and 1 before applying the softmax regression. More formally, the model can be defined for any $k \in \{0, \ldots 9\}$ by

$$\mathbb{P}\left[Y = k | X\right] = \frac{e^{\theta_k^T X}}{\sum_{k=0}^{9} e^{\theta_k^T X}}$$

with the parameters $\boldsymbol{\theta} = \left(\theta_0^T, \ldots, \theta_9^T\right)^T$ and the normalized features $X \in [0,1]^{784}$. Despite the simplicity of this model which is not really adapted to imaging data, the obtained performances are noteworthy, even when applied directly on the raw pixels data. The dataset is randomly split into a training set of size 60000 and a test set of size 10000 and the WASNA is run with default parameters, i.e. $\gamma = 0.75$, $c_\gamma = 1$, $c'_\gamma = 0$ and $\omega = 2$ on the training set. The constructed estimates

[2] http://yann.lecun.com/exdb/mnist/

of the parameter $\boldsymbol{\theta}$ are then used to "read" the digit displayed in pictures of the test set, leading to an overall performance of 88% accurate predictions. For completeness, and to understand which digits are mistaken, we provide the resulting confusion matrix in Figure 3.9. Remark that Averaged

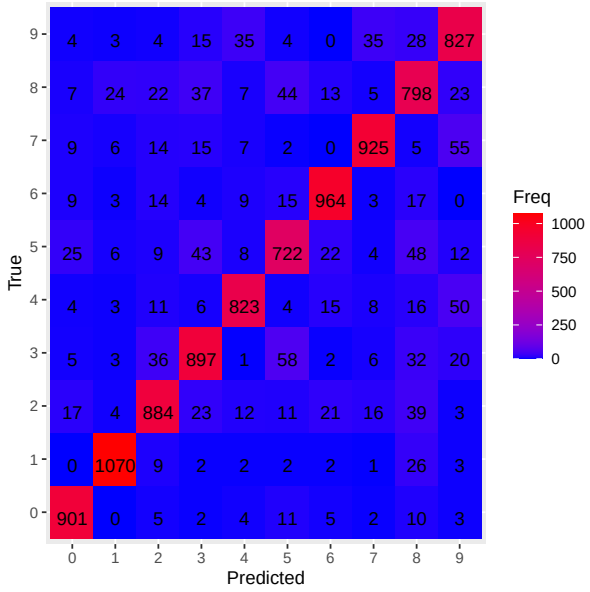


Figure 3.9 – (Softmax regression on the MNIST dataset) Confusion matrix for the predictions given by the default WASNA on a test set of size 10000.

Stochastic gradient algorithms and the Adagrad one leads to analogous (or slightly better) results. The comparison in terms of accuracy may not be totally fair as the hyperparameters of the WASNA have not been optimized at all but chosen as default values. This numerical experiment on the MNIST real dataset proves the proposed WASNA to be a second order online method able to tackle large-scale data. And if the number of hyperparameters can be a legitimate delicate point raised by some readers, it should be noted that a default choice however already leads to very good results on a difficult problem such as the classification of images into 10 classes.

Chapter 4

Stochastic Streaming Gradient algorithms

This chapter is based on [GBWW21, GBWW22].

Contents

4.1 Introduction

In previous chapters, we focus on online stochastic approximation based on the estimation of the gradient obtained with the last data point. Although it was proven that this approach leads to asymptotically efficient estimates, the studied framework cannot be applied to the case where the data are not independent and/or identically distributed. In order to overcome this, we now focus on streaming data.

More precisely, we will focus on the streams to which the data arrives. We will be concerned by two main cases: constant and varying streaming-batch sizes. Since one of the main application of this chapter is the estimation of the parameters of chronological series, we will consider from now

the following notations and problem: the aim is to minimize a function $G : \mathbb{R}^d \longrightarrow \mathbb{R}$ defined for all $h \in \mathbb{R}^d$ by

$$G(h) = \mathbb{E}\left[g_t(h)\right]$$

where $g_t : \mathbb{R}^d \longrightarrow \mathbb{R}$ is a random function [KY03]. Let us consider the sequence of functions $(g_t)_{t\geq 1}$ and let us suppose from now that they are differentiable and that their gradients are estimates of the gradient of G. Typically, we will consider data arriving sequentially and by bloc, i.e at time $t \geq 1$, we have to deal with $n_t \geq 1$ new random functions $g_{t,1}, \ldots, g_{t,n_t}$. A simple example would be to consider the i.i.d case given by (1.1) and considering new i.i.d data $X_{t,1}, \ldots, X_{t,n_t}$. Then

$$g_t(h) = \frac{1}{n_t}\sum_{i=1}^{n_t} g\left(X_{t,i}, h\right).$$

Given $(g_t)_{t\geq 1}$, the Stochastic Streaming Gradient algorithm (SSG for short) is defined recursively for all $t \geq 0$ by

$$m_{t+1} = m_t - \gamma_{t+1}\nabla g_{t+1}\left(m_t\right) = m_t - \frac{\gamma_{t+1}}{n_{t+1}}\sum_{i=1}^{n_{t+1}} \nabla g_{t+1,i}\left(m_t\right), \tag{4.1}$$

where $\nabla g_{t+1,i}(.)$ denotes the gradient of $g_{t+1,i}$, and $(\gamma_t)_{t\geq 1}$ is a decreasing sequence of positive numbers satisfying the usual hypothesis

$$\sum_{t\geq 1}\gamma_t^2 < +\infty \qquad \text{and} \qquad \sum_{t\geq 1}\gamma_t = +\infty.$$

Note that in the usual i.i.d setting defined by (1.1), this algorithm can be assimilated to the Stochastic Gradient algorithm where the last gradient has been calculated with the last n_t data instead of the last one (only). Then, one cannot hope obtaining efficient estimates without an averaging step. In order to define this last one, let us denote from now by N_t the total number of data/functions dealt with at time t, i.e $N_t := \sum_{j=1}^{t} n_j$. Then, the Averaged Stochastic Streaming Gradient algorithm (ASSG for short) is defined for all $t \geq 0$ by

$$\overline{m}_{t+1} = \frac{1}{N_{t+1}}\sum_{j=0}^{t} n_{j+1}\theta_j \tag{4.2}$$

with $\overline{m}_0 = 0$. This can of course be recursively written as $\overline{m}_{t+1} = \frac{N_t}{N_{t+1}}\overline{m}_t + \frac{n_{t+1}}{N_{t+1}}m_t$.

In Section 4.2, we concentrate on the i.i.d settings. The aim is to focus on the behavior of the estimates with respect to the "choice" of streaming-batch sizes $(n_t)_{t\geq 1}$. The aim is to understand each kind of batch settings we can deal with, without affecting too much the rate of convergence in quadratic mean of the SSG and ASSG estimates. With the help of this preliminary work, we will consider a framework where the data are not supposed to be independent nor identically distributed in Section 4.3. More precisely, we will prove that under conditions, the ASSG estimates still achieve the Cramer-Rao bound. The algorithms will be applied to several cases consisting in time-series [BJRL15, BD09, Ham20] or the estimation of the geometric median [Hal48, Kem87].

4.2 Rate of convergence of Averaged Stochastic Streaming Gradient algorithms

4.2.1 Framework

In what follows, let us suppose that m lies in a convex and close subset $\Theta \subset \mathbb{R}^d$. In case of possible constraints on the parameter space Θ, one can consider the Projected Stochastic Streaming Gradient algorithm (PSSG for short) defined for all $t \geq 0$ by

$$m_t = \mathcal{P}_\Theta \left(m_t - \frac{\gamma_{t+1}}{n_{t+1}} \sum_{i=1}^{n_{t+1}} \nabla g_{t,i}\left(m_t\right) \right)$$

where $\mathcal{P}_\Theta$ is the convex projection onto Θ. To shorten notation, let us recall that $\nabla g_t\left(m_t\right) := \frac{1}{n_t}\sum_{i=1}^{n_t} \nabla g_{t,i}\left(m_t\right)$. Furthermore, in order to give the rate of convergence of the estimates, we consider the σ-algebra $\mathcal{F}_{t,i} = \sigma\left(g_{1,1}, \ldots, g_{1,n_1}, \ldots, g_{t,1}, \ldots, g_{t,i}\right)$ (with the convention $\mathcal{F}_{t,0} = \mathcal{F}_{t-1,n_{t-1}}$) and we suppose from now that the following assumptions hold:

($\mathbf{A_{stream}0}$) The functional G is μ quasi-strongly convex on Θ: for all $h, h' \in \Theta$,

$$G(h) \geq G(h') + \left\langle \nabla G(h'), h - h' \right\rangle + \frac{\mu}{2}\left\|h - h'\right\|^2$$

($\mathbf{A_{stream}1}$) The random functions $\nabla g_{t,i}$ are square-integrable and for all $h \in \Theta$,

$$\mathbb{E}\left[\nabla g_{t,i}(h)\right] = \nabla G(h).$$

($\mathbf{A_{stream}2}$) There exists $L_{\nabla_g} \geq 0$ such that for all $h, h' \in \Theta$,

$$\mathbb{E}\left[\left\|\nabla g_{t,i}(h) - \nabla g_{t,i}(h')\right\|^4 | \mathcal{F}_{t,i-1}\right] \leq L_{\nabla g}\left\|h - h'\right\|^4.$$

($\mathbf{A_{stream}3}$) There exists $\tau > 0$ such that

$$\mathbb{E}\left[\left\|\nabla g_{t,i}(m)\right\|^4 | \mathcal{F}_{t,0}\right] \leq \tau^4.$$

Let us now make some comments on these assumptions. First of all, in order to alleviate notations, Assumptions **($\mathbf{A_{stream}2}$)** and **($\mathbf{A_{stream}3}$)** are supposed to be verified for fourth order moments although moments of order two are sufficient to get the rate of convergence of SSG estimates. Furthermore, note that they imply Assumption **($\mathbf{A1}_\eta$)** and **(A1b')** (if the gradient of G is Lipschitz) and are so a bit more restrictive. Nevertheless, these assumptions are often encountered in the literature (see [BM13] for instance) and are often satisfied in practice.

4.2.2 Converge of SSG

We first give an uniform bound of the quadratic mean error of the SSG estimates, and so, for any streaming batch size n_t.

Theorem 4.2.1 ([GBWW21]). *Suppose Assumptions* ***($A_{stream}0$)*** *to* ***($A_{stream}3$)*** *and* ***(A3b)*** *hold. Then, for all* $t \geq 1$

$$\mathbb{E}\left[\|m_t - m\|^2\right] \leq e^{-\mu \sum_{i=\frac{t}{2}}^{t} \gamma_i} e^{L_{\nabla g}^2 \sum_{i=1}^{t} \frac{\gamma_i^2}{n_i}} e^{2L_{\nabla G}^2 \sum_{i=1}^{t} \mathbf{1}_{n_i>1}\gamma_i^2} \left(\mathbb{E}\left[\|m_0 - m\|^2\right] + \frac{2\tau^2}{L_{\nabla g}^2}\right) + \frac{2\tau^2}{\mu} \max_{\frac{t}{2}\leq i\leq t} \frac{\gamma_i}{n_i}.$$

Remark that this theorem holds for any positive and decreasing stepsequence (γ_t) and streaming batch size n_t. Then, it enables the reader to obtain quickly the rate of convergence of the estimates by lower and upper bound the stepsequence. Here, we now consider a stepsequence of the form $\gamma_t = c_\gamma n_t^\beta t^{-\gamma}$ with $c_\gamma > 0$, $\beta \in [0,1]$ and γ has to be calibrated to enable the stepsequence to satisfy the usual conditions. The term n_t^β, allows, when $\beta > 0$, to give more weights to the best estimates of the gradient, i.e to estimates that use more data. The following corollary gives the rate of convergence of the SSG estimates for constant streaming-batch size, i.e when $n_t = C_\rho \in \mathbb{N}^*$.

Corollaire 4.2.1 ([GBWW21]). *Suppose Assumptions* ***($A_{stream}0$)*** *to* ***($A_{stream}3$)*** *and* ***(A3b)*** *hold. Suppose also that* $n_t = C_\rho$ *and* $\gamma_t = c_\gamma C_\rho^\beta t^{-\gamma}$ *with* $\gamma \in (1/2, 1)$. *Then, for all* $t \geq 1$,

$$\mathbb{E}\left[\|m_t - m\|^2\right] \leq \exp\left(-\frac{\mu c_\gamma N_t^{1-\gamma}}{2^{1-\gamma} C_\rho^{1-\gamma-\beta}}\right)\left(\mathbb{E}\left[\|m_0 - m\|^2\right] + \frac{2\tau^2}{L_{\nabla g}^2}\right)\pi_c + \frac{2^{1+\gamma}\tau^2 c_\gamma}{\mu C_\rho^{1-\gamma-\beta} N_t^\gamma} \tag{4.3}$$

with $N_t = \sum_{j=1}^{t} n_j$ *and* $\pi_c = \exp\left(\frac{4\gamma c_\gamma^2 \left(2L_{\nabla g}^2 + C_\rho \mathbf{1}_{C_\rho>1} L_{\nabla G}^2\right)}{(2\gamma-1) C_\rho^{1-2\beta}}\right)$.

Remark that for $n_t = 1$, we have analogous bound to the one in [BM13]. In addition, taking $C_\rho > 0$ and $\gamma + \beta > 1$ leads to a reduction of the variance (the last term on the right-hand side of inequality (4.3)) compare to usual results but increases the first term on the right-hand side of inequality (4.3). Observe that the inverse analysis can be done for γ and β and all the difficulty is to find the good compromise.

Since in practice, constant streaming-batch size are not realistic, we now consider varying streaming-batch size, i.e $n_t = \max\left\{C_\rho t^\rho, 1\right\}$ with $C_\rho \geq 1$ and $\rho \in (-1, 1)$. The case $\rho > 0$ (resp. $\rho < 0$) corresponds to the increasing (resp. decreasing) streaming-batch size. In order to give the rate of convergence for both cases, let us denote $\tilde{\rho} := \rho \mathbf{1}_{\rho \geq 0}$.

Corollaire 4.2.2. *Suppose Assumptions* ***($A_{stream}0$)*** *to* ***($A_{stream}3$)*** *and* ***(A3b)*** *hold. Suppose also that* $\gamma_t = c_\gamma n_t^\beta t^{-\gamma}$ *where* $n_t = \max\left\{c_\rho n_t^\rho, 1\right\}$, $\rho \in (-1, 1)$ *and* $\gamma - \beta\tilde{\rho} \in (1/2, 1)$. *Then, for all* $t \geq 1$,

$$\mathbb{E}\left[\|m_t - m\|^2\right] \leq \exp\left(-\frac{\mu c_\gamma N_t^{1-\phi}}{2^{(2+\rho)(1-\phi)} C_\rho^{1-\beta-\phi}}\right)\left(\mathbb{E}\left[\|m_0 - m\|^2\right] + \frac{2\tau^2}{L_{\nabla g}^2}\right)\pi_\nu + \frac{2^{1+(2+\rho)\phi}\tau^2 c_\gamma}{\mu C_\rho^{(1-\beta)\mathbf{1}_{\rho\geq 0}-\phi} N_t^\phi}$$

where $\phi = \frac{(1-\beta)\tilde{\rho}+\gamma}{1+\tilde{\rho}}$ *and* $\pi_\nu = \exp\left(\frac{4(\gamma-\beta\tilde{\rho})c_\gamma^2 C_\rho^{2\beta}\left(2L_{\nabla g}^2+L_{\nabla G}^2\right)}{2(\gamma-\beta\tilde{\rho})-1}\right)$.

Remark that in the case where ρ is negative, γ just has to verify the usual condition, i.e $\gamma \in (1/2, 1)$. Note also that some choices of γ, β and ρ can eventually improve the usual rate of convergence, but we will see in next section that this will be to the detriment of the performance of the ASSG estimates. Finally, observe that these results can be easily adapted to the case where n_t is random with $C_L t^{\rho_L} \leq n_t \leq C_H t^{\rho_H}$ where $\rho_L, \rho_H \in (-1, 1)$ and $C_L, C_H \geq 1$.

4.2.3 Convergence of ASSG

Let us now focus on the rate of convergence of the ASSG estimates. In this aim, let us first introduce a last assumption:

(A4a") There is a positive constant $L_{\nabla^2 G}$ such that for all $h, h' \in \mathbb{R}^d$,

$$\left\|\nabla G(h) - \nabla^2 G(h')\left(h-h'\right)\right\| \leq L_{\nabla^2 G}\left\|h-h'\right\|^2.$$

Note that this assumption is verified since the function $h \longmapsto \nabla^2 G(h)$ is $L_{\nabla^2 G}$-Lipschitz, and it can be seen as an extension of Assumption **(A4a')**. Let us now give the L^4 rate of convergence of the SSG estimates:

Lemma 4.2.1. *Suppose Assumptions* ***(A_{stream}0)*** *to* ***(A_{stream}3)*** *and* ***(A3b)***, ***(A4a")*** *hold. Then, for all* $t \geq 1$,

$$\mathbb{E}\left[\|m_t - m\|^4\right] \leq e^{-\mu\sum_{i=t/2}^t \gamma_i}\left(\mathbb{E}\left[\|m_0-m\|^4\right] + \frac{2\tau^4}{L_{\nabla g}^4} + \frac{4\tau^4\gamma_1}{\mu L_{\nabla g}^2 n_1}\right)\Pi + \frac{32\tau^4}{\mu^2}\max_{\frac{t}{2}\leq i\leq t}\frac{\gamma_i^2}{n_i^2} + \frac{48\tau^4}{\mu}\max_{\frac{t}{2}\leq i\leq t}\frac{\gamma_i^3}{n_i^3} + \frac{114\tau^4}{n_i^3}\max_{\frac{t}{2}\leq i\leq t}\frac{\gamma_i^3 \mathbf{1}_{n_i>1}}{n_i^2}$$

with Π *given in [GBWW21].*

Let us introduce a last assumption:

(A_{stream}4) There exists a a non-negative self-adjoint matrix Σ such that for all $t \geq 1$,

$$\mathbb{E}\left[\nabla g_{t,i}(h)\nabla g_{t,i}(h)^T | \mathcal{F}_{t,i-1}\right] \preccurlyeq \Sigma.$$

We can now give a first convergence result for the ASSG estimates, available for any choice of streaming-batch or positive decreasing stepsequence verifying the usual assumptions [RM51].

Theorem 4.2.2 ([GBWW21])**.** *Suppose Assumptions* ***(A_{stream}0)*** *to* ***(A_{stream}4)***, ***(A3b)*** *and* ***(A4a")*** *hold.*

Then, for all $t \geq 1$,

$$\sqrt{\mathbb{E}\left[\|\overline{m}_t - m\|^2\right]} \leq \frac{\Lambda^{1/2}}{N_t^{1/2}} + \frac{1}{N_t\mu}\sum_{i=1}^{t-1}\left|\frac{n_{i+1}}{\gamma_{i+1}} - \frac{n_i}{\gamma_i}\right|\sqrt{\mathbb{E}\left[\|m_i - m\|^2\right]} + \frac{n_t}{N_t\gamma_t\mu}\sqrt{\mathbb{E}\left[\|m_t - m\|^2\right]}$$
$$+ \frac{n_1}{N_t\mu}\left(\frac{1}{\gamma_1} + L_{\nabla g}\right)\sqrt{\mathbb{E}\left[\|m_0 - m\|^2\right]} + \frac{L_{\nabla g}}{N_t\mu}\left(\sum_{i=1}^{t-1} n_{i+1}\mathbb{E}\left[\|m_i - m\|^2\right]\right)^{1/2}$$
$$+ \frac{L_{\nabla^2 G}}{N_t\mu}\sum_{i=0}^{t-1} n_{i+1}\sqrt{\mathbb{E}\left[\|m_i - m\|^4\right]}$$

where $\Lambda = Tr\left(\nabla^2 G(m)^{-1}\Sigma\nabla^2 G(m)^{-1}\right)$.

As for the SSG estimates, remark that this theorem enables the reader to obtain quickly the rate of convergence of the estimates by lower and upper bound the stepsequence. We now consider a stepsequence of the form $\gamma_t = c_\gamma n_t^\beta t^{-\gamma}$ with $c_\gamma > 0$, $\beta \in [0,1]$ and γ has to be calibrated to enables the stepsequence to verify usual conditions. We now give the rates of convergence for the two considered case: constant streaming-batch size and varying streaming-batch size. Note that since the bounds given in [GBWW21] are obviously hard to read, we provide here less precise but more readable bounds.

Corollaire 4.2.3 ([GBWW21]). *Suppose Assumptions* ***($A_{stream}0$)*** *to* ***($A_{stream}4$)***, ***(A3b)*** *and* ***(A4a")*** *hold. Suppose also that* $\gamma_t = c_\gamma n_t^\beta t^{-\gamma}$ *where* $n_t = \max\left\{c_\rho n_t^\rho, 1\right\}$, $\rho \in (-1,1)$ *and* $\gamma - \beta\tilde{\rho} \in (1/2,1)$. *Then, for all* $t \geq 1$,

$$\sqrt{\mathbb{E}\left[\|\overline{m}_t - m\|^2\right]} \leq \frac{\Lambda^{1/2}}{N_t^{1/2}} + C_{ASSG}\max\left\{N_t^{-1+\phi/2}, N_t^{-\phi}\right\}, \tag{4.4}$$

where $\Lambda = Tr\left(\nabla^2 G(m)^{-1}\Sigma\nabla^2 G(m)^{-1}\right)$ *and* $\phi = ((1-\beta)\tilde{\rho} + \gamma)/(1+\tilde{\rho})$. *The second term in inequality* (4.4) *is explicitely given in [GBWW21].*

Remark that the case where $\tilde{\rho} = 0$ corresponds to the constant or decreasing streaming-batch size. Note that in these cases, the Cramer-Rao bound is achieved and the two main rest terms converge at the same rate as in [BM13], meaning that considering ASSG does not seems to have a negative impact on the convergence here. Observe that for increasing streaming-batch sizes, the rest terms remain negligible as long as $\phi \in (1/2,1)$.

4.2.4 Simulations

In this section, we consider independent random variables $X_i \sim \mathcal{N}(\theta, I_d)$ with $\theta = (\theta_1, \ldots, \theta_d)^T$ and θ_i taken randomly in the range $[-d,d]$ (and $d = 10$). Moreover, we set $c_\gamma = \sqrt{d}$ and $\gamma = 2/3$. Furthermore, we focus on the estimation of the geometric median of X_1 [Hal48, Kem87, CCZ13, VZ00]. Of course, the function G is not strongly convex, but one can project the estimates on a compact and convex subset containing θ.

Let us now make some comments on Figure 4.1. First, Figure 4.1a shows the variance reduction effect on SSG estimates for different constant streaming batches $C_\rho \in \{1, 8, 64, 128\}$ with $\beta = 0$.

Nevertheless, too large (constant) streaming batch sizes C_ρ hinders the convergence as we make too few iterations, leading to potential bad practical results for ASSG estimates. These findings can be extended to Figures 4.1b, to 4.1e. These figures show an increase in decay of the SSG when the streaming rate ρ increase as mentioned after but Figures 4.1d and 4.1e highlight the fact hat taking $\beta = 0$ for increasing streaming-batch sizes can lead to bad results in practice. In this case, one could chose the following setting: $\gamma = 2/3$ and $\beta = 1/3$ for any positive ρ, which seems to be confirmed by Figure 4.1f.

4.3 Learning from time-dependent streming data

4.3.1 Framework

We no overcome the usual framework where the blocks $(g_t)_{t\geq 1}$ are independent and where the gradients ∇g_t are unbiased. The aim of this section is to analyze the behavior of the SSG and ASSG estimates in this non i.i.d case. In this aim, let us consider a modified version of previous assumptions.

($\mathbf{A_{stream}1'}$) For all $t \geq 1$ and for any $h \in \Theta$ such that h is $\mathcal{F}_{t,0}$-measurable, the random variable $\nabla g_t(h)$ is square-integrable. Furthermore,

$$\mathbb{E}\left[\left\|\mathbb{E}\left[\nabla g_t(h)|\mathcal{F}_{t,0}\right] - \nabla G(h)\right\|^4\right] \leq \nu_t^4\left(D_\nu^4\mathbb{E}\left[\|h-m\|^4\right] + B_\nu^4\right)$$

for some positive sequence $(\nu_t)_{t\geq 1}$ and $D_\nu, B_\nu \geq 0$.

Note that in the case of i.i.d settings, one has of course $B_\nu = D_\nu = 0$. The constant B_ν gives the potential bias of the estimates of the gradient at m, and if it is equal to 0, we will speak about unbiased or well-specified case. Let us now give an alternative formulation of Assumption **($\mathbf{A_{stream}2}$)**.

($\mathbf{A_{stream}2'}$) There exists a positive sequence $(\kappa_t)_{t\geq 1}$ such that for all $h, h' \in \Theta$ and for all $t \geq 1$

$$\mathbb{E}\left[\left\|\nabla g_t(h) - \nabla g_t(h')\right\|^4\right] \leq \kappa_t^4\mathbb{E}\left[\left\|h-h'\right\|^4\right].$$

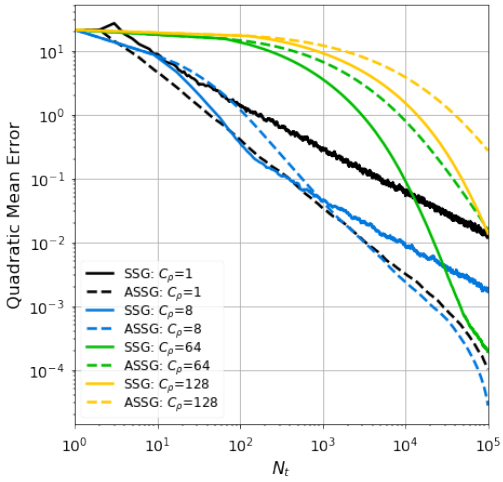
This assumption can be seen as a property of expected smoothness of the gradient of the random functions g_t. Remark that in the i.i.d setting, this assumption is verified with $\kappa_t^4 = L_{\nabla g}^4 n_t^{-2}$. Let us now give an alternative formulation of Assumption **($\mathbf{A_{stream}3}$)**.

($\mathbf{A_{stream}3'}$) There is a positive sequence $(\tau_t)_{t\geq 1}$ such that for all $t \geq 1$,

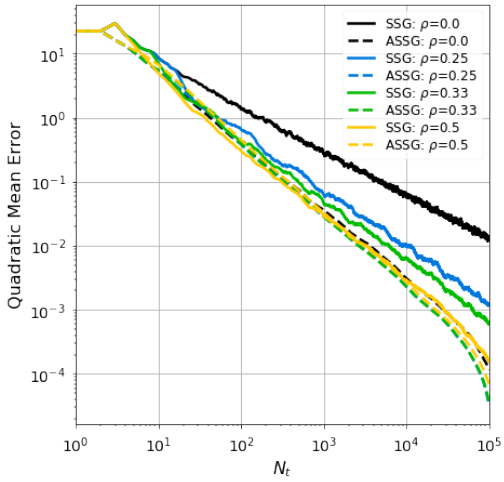
$$\mathbb{E}\left[\left\|\nabla g_t(m)\right\|^4\right] \leq \tau_t^4.$$

Remark that in the i.i.d settings, one has $\tau_t^4 = \tau^4 n_t^{-2}$. Finally, note that for the convergence of SSG, only moments of order 2 are needed for these assumptions (see [GBWW22]).

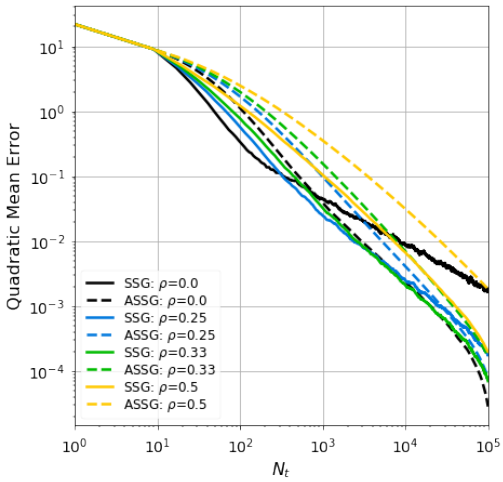
(a) Constant streaming batches, $\rho = 0$, $\beta = 0$



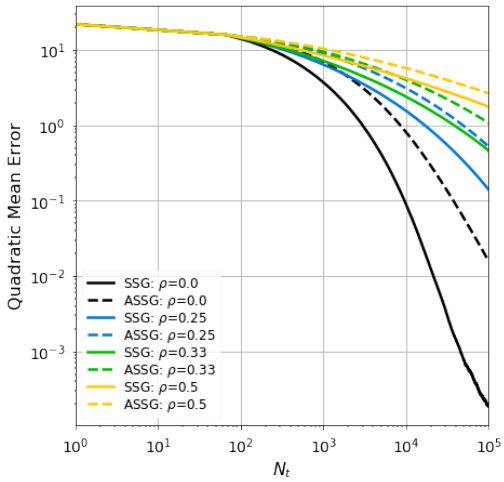
(b) Varying streaming batches, $C_\rho = 1$, $\beta = 0$



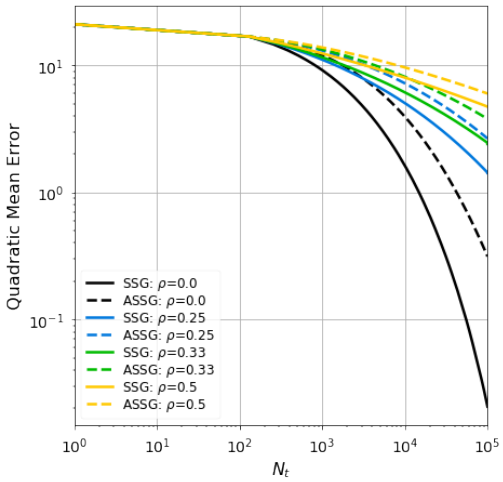
(c) Varying streaming batches, $C_\rho = 8$, $\beta = 0$



(d) Varying streaming batches, $C_\rho = 64$, $\beta = 0$



(e) Varying streaming batches, $C_\rho = 128$, $\beta = 0$



(f) Varying streaming batches, $C_\rho = 8$, $\beta = 1/3$

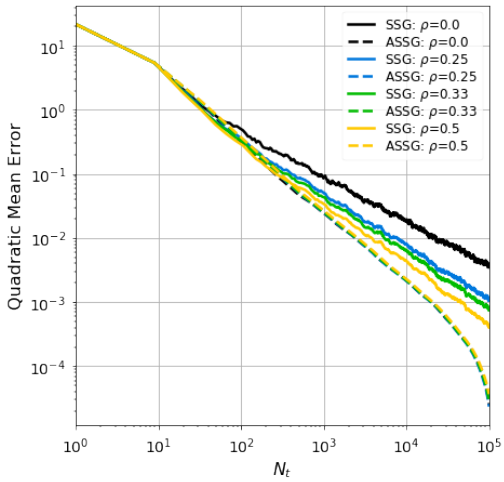


Figure 4.1 – Geometric median for various data streams $n_t = C_\rho t^\rho$.

4.3.2 Convergence of SSG estimates

Let us now consider a streaming-batch size of the form $n_t = C_\rho t^\rho$ with $C_\rho \geq 1$ and $\rho \in [0,1)$ as well as a stepsequence γ_t of the form $\gamma_t = c_\gamma n_t^\beta t^{-\gamma}$ with $c_\gamma > 0$, $\beta \geq 0$ and γ discussed later. In addition, we suppose from now that the sequences $(\nu_t)_{t\geq 1}$, $(\kappa_t)_{t\geq 1}$ and $(\tau_t)_{t\geq 1}$ are under the form $\nu_t = n_t^{-\nu}$, $\kappa_t = C_\kappa t^{-\kappa}$ and $\tau_t = C_\tau t^{-\tau}$ where $C_\kappa, C_\tau > 0$ and $\kappa, \tau \in [0, 1/2]$. Furthermore, $\nu = (0, +\infty)$, and considering the i.i.d case leads to take $\nu \to +\infty$ or $B_\nu = D_\nu = 0$. Let us now give the rate of convergence of the SSG estimates.

Theorem 4.3.1 ([GBWW22]). *Suppose Assumptions* ***($A_{stream}0$)*** *and* ***($A_{stream}1'$)*** *to* ***($A_{stream}3'$)*** *hold. Suppose also that* $\mu_\nu := \mu - \mathbf{1}_{\rho=0} 2 D_\nu C_\rho^{-\nu} > 0$ *and* $\gamma - \rho\beta \in (1/2, 1)$. *Then, for all* $t \geq 0$,

$$\mathbb{E}\left[\|m_t - m\|^2\right] \leq \pi_t + \frac{2^{\frac{2+6\rho\nu}{1+\rho}} B_\nu^2}{\mu\mu_\nu C_\rho^{\frac{2\nu}{1+\rho}} N_t^{\frac{2\rho\nu}{1+\rho}}} + \frac{2^{\frac{7+6\rho\tau}{1+\rho}} C_\tau^2 c_\gamma}{\mu_\nu C_\rho^{\frac{2\tau-\beta-\gamma}{1+\rho}} N_t^{\frac{\rho(2\tau-\beta)+\gamma}{1+\rho}}},$$

where π_t *converges exponentially fast and is defined in Theorem 1 in [GBWW22].*

Remark that for the i.i.d case, i.e taking $B_\nu = D_\nu = 0$ and $\kappa = \tau = 1/2$, this result coincides with the one given by Corollary 4.2.2. Furthermore, the condition $\mu - \mathbf{1}_{\rho=0} 2 D_\nu C_\rho^{-\nu} > 0$ implies that in case of dependency (i.e if $D_\nu > 0$), if the streaming-batch size is constant, i.e if $\rho = 0$, one has to take C_ρ sufficiently large to ensure the convergence of the SSG estimates. In addition, in the unbiased case ($B_\nu = 0$), supposing that $\tau = 1/2$, one can take $\gamma = 2/3$ and $\beta = 1/2$ and get a rate of convergence of order $N_t^{-\gamma}$ for instance. Finally, in the biased case, one can remark that the term induced by the bias converges at a rate of order $N_t^{-\frac{2\rho\nu}{1+\rho}}$ meaning that we still have convergence for increasing streaming-batch sizes, i.e if $\rho > 0$.

4.3.3 Convergence of ASSG estimates

As for the i.i.d case, let us first make an assumption on the variance of the score.

($A_{stream}4'$) There is a non-negative self-adjoint operator Σ such that for all $t \geq 1$,

$$n_t^{2\tau} \mathbb{E}\left[\nabla g_t(m) \nabla g_t(m)^T\right] \preccurlyeq \Sigma + \Sigma_t$$

where Σ_t is a non-negative symmetric matrix with $\mathrm{Tr}\,(\Sigma_t) = C'_\tau n_t^{-2\tau'}$, with $C'_\tau \geq 0$ and $\tau' \in (0, 1/2]$.

Remark that in the i.i.d case, this assumption is verified with $\tau = 1/2$ and $C'_\tau = 0$. Furthermore, in case of short-range dependence, i.e in the case where $\tau = 1/2$, it is possible to achieve the Cramer-Rao bound. We can now give the rate of convergence of the ASSG estimates.

Theorem 4.3.2 ([GBWW21]). *Suppose Assumptions* ***($A_{stream}0$),($A_{stream}1'$)*** *to* ***($A_{stream}4'$)*** *as well as* ***(A3b)*** *and* ***(A4a'')*** *hold. Suppose also that* $\mu_\nu := \mu - \mathbf{1}_{\rho=0} 2 D_\nu C_\rho^{-\nu} > 0$ *and* $\gamma - \beta\rho \in (1/2, 1)$. *Then, for all*

$t \geq 1$,

$$\sqrt{\mathbb{E}\left[\|\overline{m}_t - m\|^2\right]} \leq \frac{\Lambda^{1/2}}{N_t^{1/2}}\mathbf{1}_{\{\tau=1/2\}} + \frac{2^{1/2}\Lambda^{1/2}C_\rho^{\frac{1-2\tau}{2(1+\rho)}}}{N_t^{\frac{1+2\rho\sigma}{2(1+\rho)}}}\mathbf{1}_{\{\tau<1/2\}} + R_t + \mathbf{1}_{\{B_\nu\neq 0\}}\Psi_t, \tag{4.5}$$

with $\delta = \mathbf{1}_{\{B_\nu=0\}}(\rho(2\tau-\beta)+\gamma) + \mathbf{1}_{\{B_\nu\neq 0\}}\min\{\rho(2\tau-\beta)+\gamma, 2\rho\nu\}$ *and* Ψ_t *satisfies*

$$\Psi_t = O\left(\max\left\{N_t^{-\frac{\rho(\tau+\nu)}{2(1+\rho)}}, N_t^{-\frac{1+\rho(\beta+\nu)-\gamma}{1+\rho}}, N_t^{-\frac{1+2\rho\nu}{2(1+\rho)}}, N_t^{-\frac{\delta/2+\rho\nu}{2(1+\rho)}}, N_t^{-\frac{2\rho\nu}{1+\rho}}\right\}\right).$$

Furthermore, R_t *and* Ψ_t *are explicitely given in [GBWW22].*

A first main conclusion is that one can ensure the convexity taking an increasing streaming-batch size, which is sufficient to ensure the convergence of the estimates. In addition, previous theorem claims that it is possible to achieve the Cramer-Rao bound, especially for the unbiased cased (i.e $B_\nu = 0$) and if $\tau = 1/2$. Remark that when $\tau = 1/2$, a judicious choice of parameters seems to be $\gamma = 2/3$ and $\beta = 1/3$, which leads to a result of the form

$$\sqrt{\mathbb{E}\left[\|\overline{m}_t - m\|^2\right]} \leq \frac{\Lambda^{1/2}}{N_t^{1/2}} + O\left(N_t^{-2/3}\right) + \mathbf{1}_{B_\nu\neq 0}O\left(N_t^{-\frac{\rho(1/2+\nu)}{2(1+\rho)}}\right).$$

Then, in case of short-range dependence, i.e for ν large enough, the Cramer-Rao bound is achieved (up to rest terms) and so, even in the biased case (if $\rho > 0$).

4.3.4 Applications

Application to time-series

In what follows, we consider real valued time-series (X_s). More precisely, we will focus on four examples: the AutoRegressive (AR), Moving-Average (MA), AutoRegressive Conditional Heteroskedasticity (ARCH) and AR(1)-ARCH(1) models ([BJRL15, BD09, Ham20]). Let us describe quickly these processes.

AR model. A process (X_s) is called a zero-mean AR(1) process if there exists θ such that $X_s = \theta X_{s-1} + \epsilon_s$ where ϵ_s is a white noise. Remark that in this example, the problem is well specified in the sense that $B_\nu = 0$ (see [GBWW22]).

MA model. An MA(1) process is defined by $X_s = \epsilon_s + \phi\epsilon_{s-1}$, with $\phi \in \mathbb{R}$. We focus here on the misspecification error of fitting an AR(1) model to a MA(1) process, i.e we focus on the minimization of the error

$$L(\theta) = \mathbb{E}\left[(X_s - \theta X_{s-1})^2\right] = \mathbb{E}\left[(\epsilon_s + \phi\epsilon_{s-1} - \theta(\epsilon_{s-1} + \phi\epsilon_{s-2}))^2\right].$$

Remark that it is a misspecified case, i.e $B_v \neq 0$.

ARCH model. A process (ϵ_s) is called an ARCH(1) process with parameters α_0, α_1 if

$$\begin{cases} \epsilon_s = \sigma_s z_s, \\ \sigma_s^2 = \alpha_0 + \alpha_1 \epsilon_{s-1}^2, \end{cases} \tag{4.6}$$

where z_s is a white noise.

AR(1)-ARCH(1) model. A process (X_s) is called an AR(1)-ARCH(1) process of parameters $\theta, \alpha_0, \alpha_1$ if

$$\begin{cases} X_s = \theta X_{s-1} + \epsilon_s, \\ \epsilon_s = \sigma_s z_s, \\ \sigma_s^2 = \alpha_0 + \alpha_1 \epsilon_{s-1}^2. \end{cases} \tag{4.7}$$

where (z_s) is a weak white noise.

Simulations To compare different data streams through the selection of C_ρ and ρ, we fix the parameters $C_\gamma = 1$, $\gamma = 2/3$, and $\beta = 0$. First consider the AR (well- and misspecified) cases in 4.2a,4.2b; these figures show the results for long-range dependent white noise processes. Note that in this case the traditional stochastic gradient method experiences a large amount of noise initially, particularly affecting the average estimate period but not its decay rate. Both methods show a noticeable reduction in variance when C_ρ increases although, without surprise, too large streaming batch sizes C_ρ may hinder the convergence as this leads to too few iterations. Furthermore, 4.2a,4.2b indicates an improved decay of the SSG methods when the streaming rate ρ is increased. Conversely, improvements to the ASSG method do not occur as we do not exploit the potential of using more observations through parameter β, which could accelerate convergence, e.g., see 4.3.4. In Figures 4.2c,4.2d, the lack of convexity when using small streaming batch sizes C_ρ, e.g., the averaged stochastic gradient estimates $(C_\rho = 1, \rho = 0)$ diverges. Remark that the lack of convexity is expressed through the lack positively of μ_v, which only larger streaming batch sizes C_ρ can counteract. Figure 4.2d shows that large ($C_\rho = 64$) and non-decreasing ($\rho \geq 0$) streaming batches can converge under difficult settings.

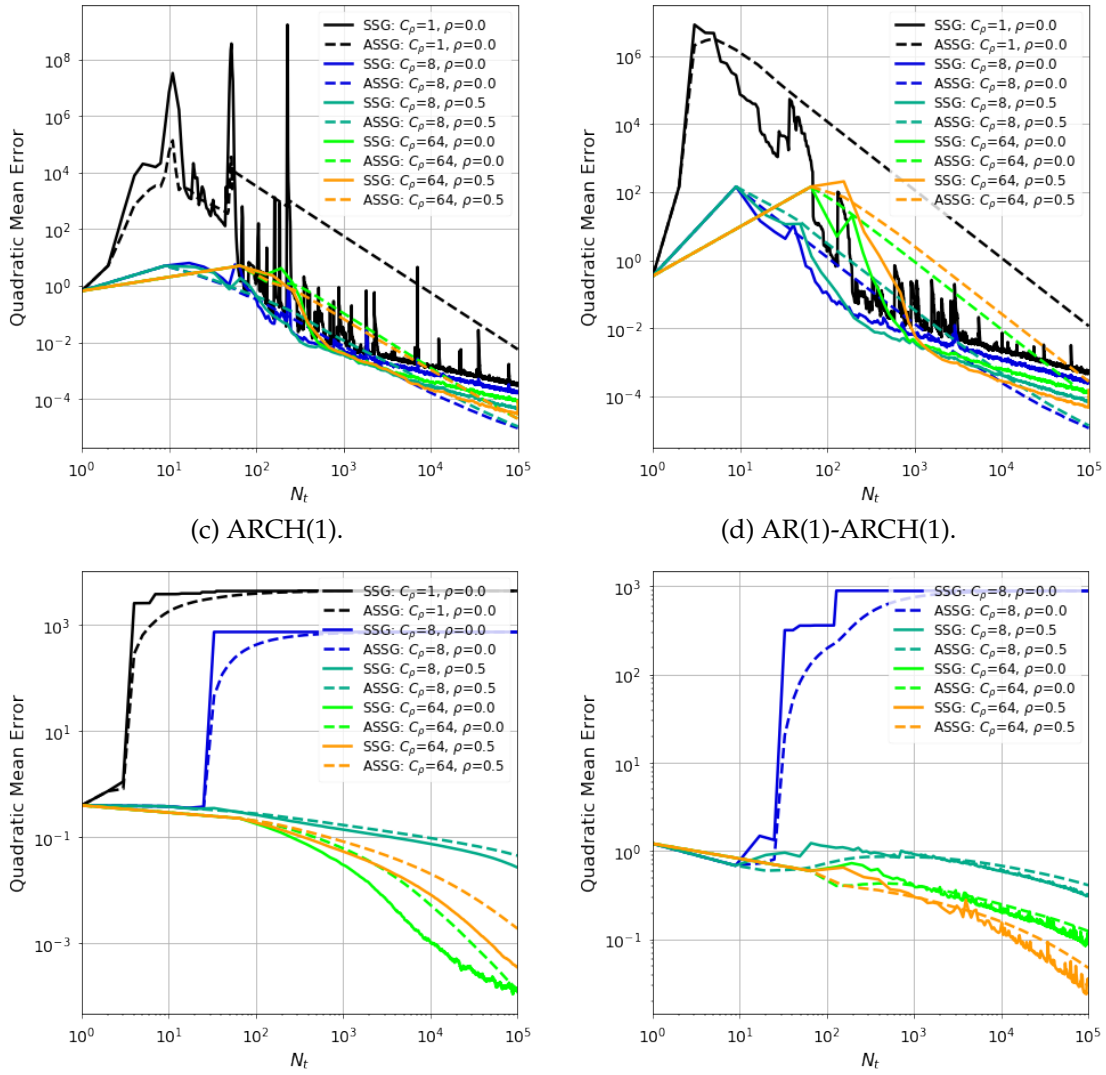
Application to the geometric median

In order to illustrate our method on real-life time-dependent streaming data, we consider some historical hourly weather data[1]. The dataset contains around five years (roughly 45000 data points) of

[1] https://www.kaggle.com/datasets/selfishgene/historical-hourly-weather-data

(a) AR(1): well-specified case.

(b) AR(1): misspecified case.



(c) ARCH(1).

(d) AR(1)-ARCH(1).

Figure 4.2 – Simulation of various data streams $n_t = C_\rho t^\rho$.

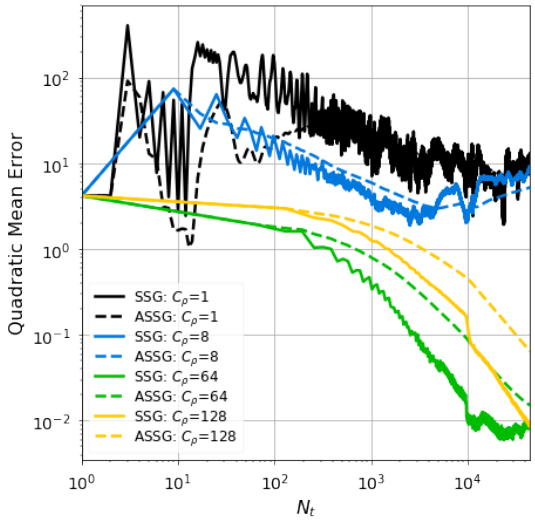
high temporal resolution hourly measurements over various weather attributes, such as temperature, humidity, and air pressure. These measurements are available for thirty-six cities, i.e., $d = 36$. In our study, we consider the hourly temperature measurements, which we filter for monthly and annual seasonality by subtracting the monthly and annual averages.

We first estimate the geometric median with the help of the Weiszfeld's algorithm (see [Wei37] and Section 5.2.4) with a very large number of iterations. Moreover, following the reasoning of [CCZ13], we set $C_\gamma = \sqrt{d}$, and let $\gamma = 2/3$.

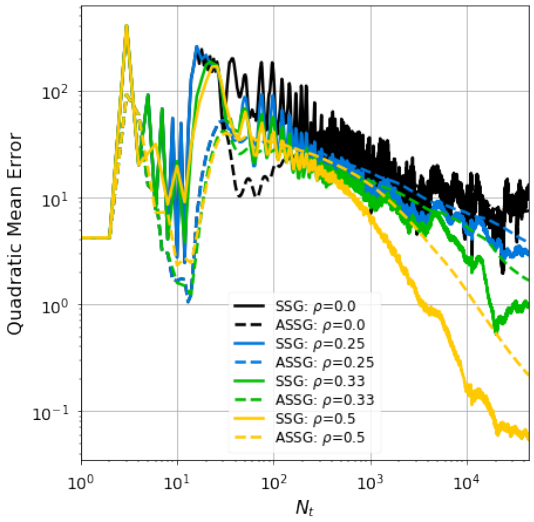
Figure 4.3a shows that it is essential to use a mini-batch C_ρ of a certain size to stabilize the optimization, i.e., ensure convexity through larger streaming batches C_ρ. In addition, to achieve reasonable convergence, we need to have increasing streaming batches, i.e., positive streaming rates $\rho > 0$; this is illustrated in 4.3b and 4.3c. Indeed, these figures leads to think that an increasing ρ leads

Figure 4.3 – Geometric median for various data streams $n_t = C_\rho t^\rho$.

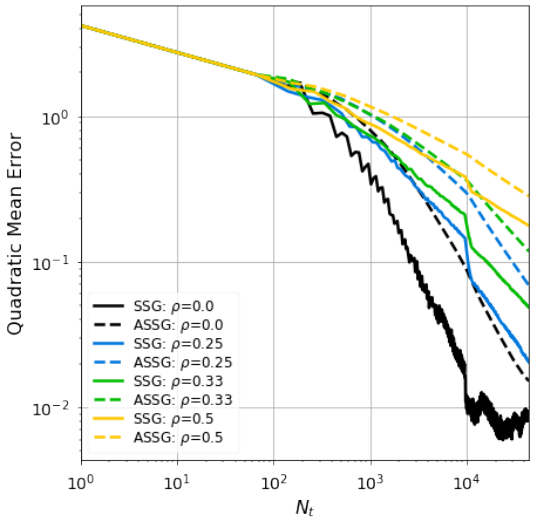
(a) Varying C_ρ, $\rho = 0$, $\beta = 0$



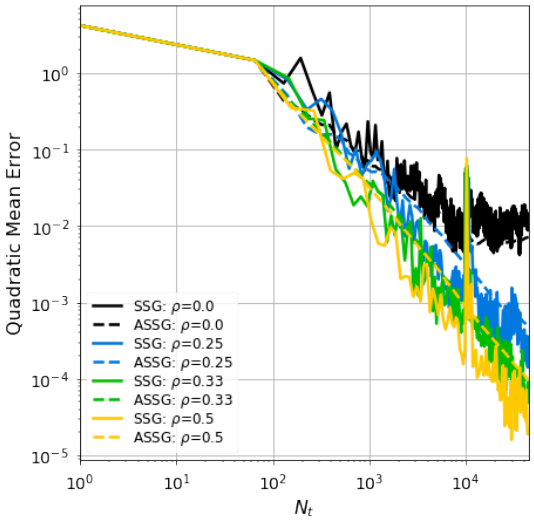
(b) Varying ρ, $C_\rho = 1$, $\beta = 0$



(c) Varying ρ, $C_\rho = 64$, $\beta = 0$



(d) Varying ρ, $C_\rho = 64$, $\beta = 1/3$



to a decay of the SSG methods. However, the lack of convergence improvements in 4.3c comes from $\beta = 0$, which means we do not exploit the potential of using more observations to accelerate convergence. As shown in Figure 4.3d, we can achieve this acceleration by simply taking $\beta = 1/3$. In addition, $\beta = 1/3$ provides optimal convergence robust to any streaming rate ρ. As expected, choosing a proper $\beta > 0$ is particularly important when C_ρ is large. Most surprising is that we can achieve excellent convergence with a final error of only 10^{-5} by combining increasing streaming batches with averaging, e.g., see 4.3d with $C_\rho = 64$, $\rho > 0$ and $\beta = 1/3$.

Chapter 5

Application to robust statistics

This chapter is based on [CCGB15, GB16a, CGB15, GBS22, GBR22].

Contents

5.1 Introduction

The acquisition of massive data lying in high dimensional spaces is unfortunately often accompanied by a contamination of these last ones. In this context of contaminated data, even few individuals may corrupt simple statistical indicators such as the mean or the variance. Detecting these atypical data automatically is not straightforward and considering robust techniques is an interesting alternative [NT15]. There are many robust location indicators [Sma90, GH10, MNO+10, CGP12]. For instance, Trimmed-means [RL05, FM01] consist in taking the averaged of the $(1-\alpha)n$ most central information. Nevertheless, this approach necessitates to have an idea of the proportion of contaminated data and assume that these last ones are necessary far from 0. In addition, these approaches often necessitates high computational efforts, although some procedures have been developed to deal with dimensionality issues [CFF07].
In this chapter, we first focus on the geometric median (also called L^1-median or spatial median) introduced by [Hal48]. Indeed, this location indicator is known to have a 0.5 breakdown point, meaning that even if nearly half of the sample is contaminated, one can control the divergence of the estimates, contrary to the mean which has a 0 breakdown point [Ger08]. Several iterative methods based on Weiszfeld algorithm [Wei37] have been developped [VZ00]. We focus in Section 5.2 on the online estimates of the median obtained with the help of an averaged stochastic gradient algorithm [CCZ13]. More precisely, we establish non asymptotic rates of convergence such that the rates of convergence in quadratic mean as well as confidence balls.
In a second time, we will focus on robust non-supervised clustering. One of the most usual method for hard clustering is probably the K-means algorithm [For65, Mac67], and one can refer to [CAGM97, GEG99] for the robust versions obtained with the help of Trimmed K-means. Since these modified robust version share the same problems as the Trimmed means, we focus in Section 5.3 on K-medians algorithms [Mac67, KR09, CCM12]. More precisely, we propose a method for selecting the number of clusters based on a penalized criterion [Fis11] whose penalty is calibrated with the help of a slope heuristic [BMM12, AM09]. All the proposed methods are available in the R package `Kmedians` accessible on CRAN[1].
In section 5.4, we focus on online robust Principal Components Analysis (PCA). PCA is one of the most useful statistical tools to extract information by reducing the dimension when one has to analyze large samples of multivariate data [Jol02, RS05]. Nevertheless, principal components, which are derived from the spectral analysis of the covariance matrix, can be very sensitive to outliers and many robust procedures for principal components analysis have been considered in the literature (see [HRVA08, HR09, Ger08] among others). We focus here on a new approach based on the Median Covariation Matrix, which is a robust dispersion indicator which has, under conditions [KP12], the same eigenvectors as the usual covariance matrix. All the proposed methods are available in the R package `Gmedian` accessible on CRAN[2].
Finally, Section 5.5 deals with the case where the law of the sample is known. Indeed, one can so

[1] https://cran.r-project.org/package=Kmedians
[2] https://cran.r-project.org/package=Gmedian

rebuild robustly the covariance matrix from the estimates of the MCM [GBR22], and this approach is so applied to the development of robust methods for model based clustering, such as Gaussian Mixtures. This represent an interesting alternative to usual robust methods which often necessitates to modelize the contamination (see [BR93, CH16, CH17, FP20] for instance). All the proposed methods are available in the R package `RGMM` accessible on CRAN[3].

5.2 Online estimation of the geometric median via averaged stochastic gradient algorithms

5.2.1 Definition and algorithms

In what follows, we consider a random variable X taking values in a separable Hilbert space $\mathcal{H}$ (not necessarily of finite dimension). Then, the geometric median of X is defined as the minimizer of the functional $G_{1/2} : \mathcal{H} \longrightarrow \mathbb{R}$ defined for all $h \in \mathcal{H}$ by

$$G_{1/2}(h) = \mathbb{E}\left[\|X - h\| - \|X\|\right].$$

Remark that the term $\|X\|$ just enables not to make any assumption on the existence of the first order moment of X. We suppose from now that the following usual assumptions are fulfilled [Kem87, Cha92, Cha96, CCZ13]:

($\mathbf{A_{median}1a}$) The random variable X is not concentrated around single points: there is a constant C_{med} such that for all $h \in \mathcal{H}$,

$$\mathbb{E}\left[\frac{1}{\|X - h\|}\right] \leq C_{\text{med}}.$$

($\mathbf{A_{median}2}$) The random variable X is not concentrated on a straight line: for all $h \in \mathcal{H}$, there is $h' \in \mathcal{H}$ such that

$$\langle h, h' \rangle \neq 0 \qquad \text{and} \qquad \mathbb{V}\left[\langle X, h \rangle\right] > 0.$$

Remark that Assumption **($\mathbf{A_{median}1a}$)** is closely related to small ball probabilities and is not restrictive since the dimension of $\mathcal{H}$ is larger that 3. This assumption is crucial to ensure that for all $h \in \mathcal{H}$, the functional $G_{1/2}$ is twice continuously differentiable with

$$\nabla G_{1/2}(h) = -\mathbb{E}\left[\frac{X - h}{\|X - h\|}\right] \qquad \text{and} \qquad \nabla^2 G_{1/2}(h) = \mathbb{E}\left[\frac{1}{\|X - h\|}\left(I_{\mathcal{H}} - \frac{(X - h)(X - h)^T}{\|X - h\|^2}\right)\right].$$

Finally, Assumption **($\mathbf{A_{median}2}$)** ensures that the functional $G_{1/2}$ is locally strongly convex on a neighborhood of the median $m_{1/2}$, and so ensures its uniqueness [Kem87]. In what follows, let us consider $X_1, \ldots, X_n, X_{n+1}, \ldots$ i.i.d copies of X arriving sequentially. Then, the stochastic gradient algorithm for estimating the geometric median and its averaged version are defined recursively

[3] https://cran.r-project.org/package=RGMM

for all $n \geq 0$ by [CCZ13]

$$\begin{aligned} m_{1/2,n+1} &= m_{1/2,n} + \gamma_{n+1} \frac{X_{n+1} - m_{1/2,n}}{\|X_{n+1} - m_{1/2,n}\|} \\ \overline{m}_{1/2,n+1} &= \overline{m}_{1/2,n} + \frac{1}{n+2} \left(m_{1/2,n+1} - \overline{m}_{1/2,n} \right) \end{aligned} \tag{5.1}$$

with $\overline{m}_{1/2,0} = m_{1/2,0}$ and $\gamma_n = c_\gamma n^{-\gamma}$ where $c_\gamma > 0$ and $\gamma \in (1/2, 1)$.

5.2.2 Rates of convergence

First, note that it was proven in [CCZ13] that under Assumptions **($\mathbf{A_{median}1a}$)** and **($\mathbf{A_{median}2}$)**, $m_{1/2,n}$ converges almost surely to m. We now give the almost sure rate of convergence of the stochastic gradient estimates:

Theorem 5.2.1. *Suppose Assumptions **($A_{median}1a$)** and **($A_{median}2$)** hold. Then,*

$$\|m_{1/2,n} - m_{1/2}\|^2 = O\left(\frac{\ln n}{n^\gamma} \right) \quad a.s.$$

Remark that this result represents a slight improvement compare to the one in [GB16a], and it is a direct corollary of Theorem 1.3.2. In order to obtain the rate of convergence of the averaged estimates, let us introduce a last assumption:

($\mathbf{A_{median}1b}$) The random variable X is not concentrated around single points: there is a positive constant C_{med} such that for all $h \in \mathcal{H}$,

$$\mathbb{E}\left[\frac{1}{\|X - h\|^2} \right] \leq C_{\text{med}}^2.$$

Note that thanks to Hölder's inequality, this implies Assumption **($\mathbf{A_{median}1a}$)**. Furthermore, this hypothesis is crucial to bound the rest term in the Taylor's decomposition of the gradient, i.e to ensure that Assumption **(A4a)** is fulfilled, and so to prove the following theorem.

Theorem 5.2.2 ([CCZ13, GB16a])**.** *Suppose Assumptions **($A_{median}1b$)** and **($A_{median}2$)** hold. Then, for all $\delta > 0$,*

$$\|\overline{m}_{1/2,n} - m_{1/2}\| = o\left(\frac{(\ln n)^{1+\delta}}{n} \right) \quad a.s \quad \text{and} \quad \sqrt{n} \left(\overline{m}_{1/2,n} - m_{1/2} \right) \xrightarrow[n \to +\infty]{\mathcal{L}} \mathcal{N}\left(0, H_{1/2}^{-1} \Sigma_{1/2} H_{1/2}^{-1} \right)$$

where $H_{1/2} = \nabla^2 G_{1/2}(m_{1/2})$ *and* $\Sigma_{1/2} = \mathbb{E}\left[\left(\frac{X - m_{1/2}}{\|X - m_{1/2}\|} \right) \left(\frac{X - m_{1/2}}{\|X - m_{1/2}\|} \right)^T \right]$.

Then, the averaged estimates are unsurprisingly asymptotically efficient.

5.2.3 Non asymptotic rates of convergence

Let us now focus on the rate of convergence in quadratic mean of the estimates. More precisely, the aim is to apply Theorem 1.5.1. To do so, let us recall two important results. First, under

assumptions **($\mathbf{A_{median}}$1a)** and **($\mathbf{A_{median}}$2)**, it was proven in [CCZ13] that there is K large enough such that

$$c_{\min} := \inf_{\|v\|=1} \mathbb{V}\left[\langle v, X\rangle \mathbf{1}_{\|X\|\leq K}\right] > 0.$$

Then, one has for all $h \in \mathcal{B}(m_{1/2}, 1)$ [CCZ13],

$$\lambda_{\min}\left(\nabla^2 G_{1/2}(h)\right) \geq \frac{1}{(K+1)^3} c_{\min}.$$

In addition, it was proven in [GB16a] that under Assumption **($\mathbf{A_{median}}$1b)**,

$$\left\|\nabla G_{1/2}(h) - \nabla^2 G_{1/2}(m_{1/2})(h - m_{1/2})\right\| \leq C_{\text{med}}^2 \|h - m_{1/2}\|^2.$$

Then, Assumption **(A4a')** is fulfilled. Let us now denote $\lambda_{1/2} := \lambda_{\min}\left(\nabla^2 G_{1/2}(m_{1/2})\right)$ and apply Theorem 1.5.1 to obtain the following rate of convergence for the stochastic gradient estimates of the median.

Theorem 5.2.3. *Suppose Assumptions **(A_{median}1)** and **(A_{median}2)** hold. Then, there are positive constants $A_{0,med}$, $A_{1,med}$ and $A_{2,med}$ such that for all $n \geq 1$,*

$$\mathbb{E}\left[\|m_{1/2,n} - m_{1/2}\|^2\right] \leq A_{0,med} e^{-\lambda_{1/2} c_\gamma n^{1-\gamma}} + A_{1,med} e^{-\frac{(K+1)^3}{4 c_{\min} C_{med}} c_\gamma n^{1-\gamma}} + A_{2,med} n^{-2\gamma} + \frac{2^\gamma c_\gamma}{\lambda_{1/2}} n^{-\gamma}.$$

Note that constants $A_{0,\text{med}}$, $A_{1,\text{med}}$ and $A_{2,\text{med}}$ are explicitly given in the detailed Theorem A.3.1. Without any surprise, we achieve the usual rate of convergence $n^{-\gamma}$. Observe that Figure 5.1 leads to think that this bound can still be improved. Furthermore, remark that under the same assumptions, one can apply Theorem 1.5.2 to prove that for any integer $p > 0$,

$$\mathbb{E}\left[\|m_{1/2,n} - m_{1/2}\|^{2p}\right] = O\left(n^{-\gamma p}\right).$$

This result is of particular interest to obtain the L^p rates of convergence of the averaged estimates, and by extension, to obtain the rate of convergence of the estimates of the Median Covariation Matrix. We now give the rate of convergence of the averaged estimates.

Theorem 5.2.4. *Suppose Assumptions **(A_{median}1)** and **(A_{median}2)** hold. Then, there are positive constants $A_{av,med}$ and $B_{av,med}$ such that for all $n \geq 1$,*

$$\sqrt{\mathbb{E}\left[\|\overline{m}_{1/2,n} - m_{1/2}\|^2\right]} \leq \frac{\sqrt{Tr\left(H^{-1}\Sigma_{1/2}H^{-1}\right)}}{\sqrt{n+1}} + \frac{A_{av,med}}{(n+1)^\gamma} + \frac{2^{\frac{\gamma}{2}} 5}{\sqrt{c_\gamma}\lambda_{1/2}(n+1)^{1-\frac{\gamma}{2}}} + \frac{B_{av,med}}{(n+1)^{\frac{1+\gamma}{2}}}.$$

Note that constants $A_{\text{av,med}}$ and $B_{\text{av,med}}$ are explicitly given in Theorem A.3.2. The averaged estimates so achieve the Cramer-Rao bound (up to the rest terms), which seems to be confirmed by Figure 5.2. In addition, observe that under the same assumptions, one can apply Theorem 2.3.3 to

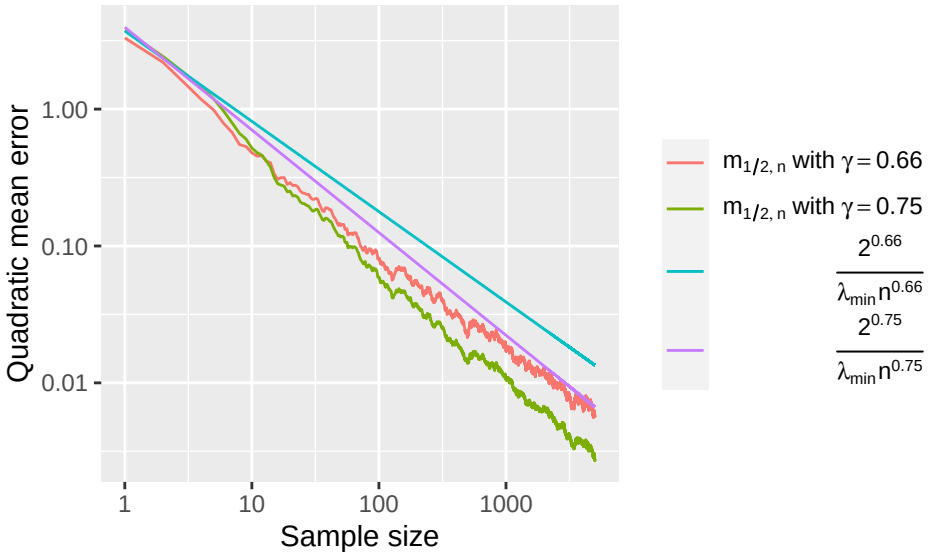


Figure 5.1 – Comparison of the evolution of the quadratic mean error of estimates $m_{1/2,n}$ (with respect to the sample size n with $\gamma = 0.66, 0.75$) with the main term of the theoretical bound given by Theorem 5.2.3

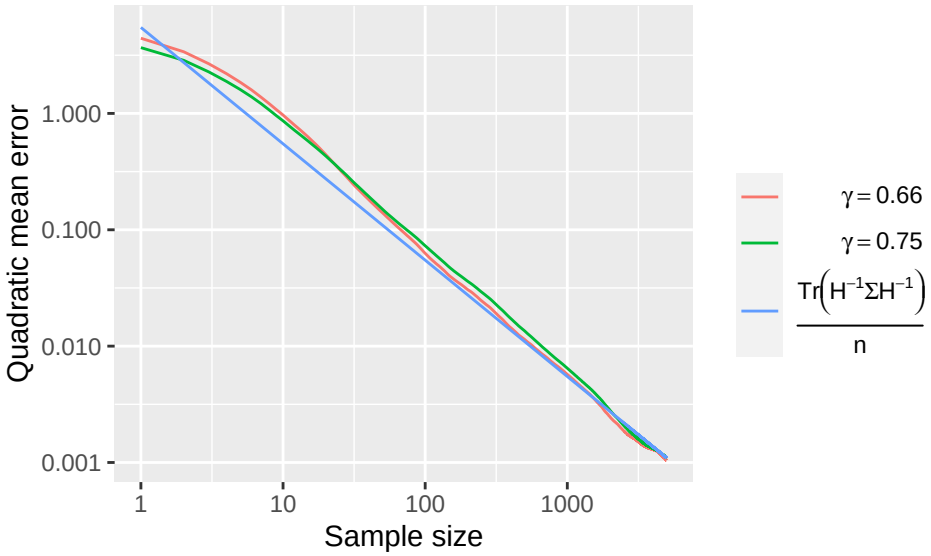


Figure 5.2 – Comparison of the evolution of the quadratic mean error of estimates $m_{1/2,n}$ (with respect to the sample size n with $\gamma = 0.66, 0.75$) with the main term of the theoretical bound given by Theorem 5.2.3

verify that for any positive integer p,

$$\mathbb{E}\left[\|\overline{m}_{1/2,n} - m_{1/2}\|^{2p}\right] = O\left(\frac{1}{n^p}\right),$$

which is of particular interest to obtain the rate of convergence in quadratic mean of the estimates

of the Median Covariation Matrix. We now give non-asymptotic confidence balls of the median.

Theorem 5.2.5 ([CCGB15]). *Suppose Assumptions* ***($A_{median}1$)*** *and* ***($A_{median}2$)*** *hold. Then, for all* $\delta \in (0,1)$ *there is a rank* n_δ *such that for all* $n \geq n_\delta$

$$\mathbb{P}\left[\|\overline{m}_{1/2,n} - m_{1/2}\| \leq \frac{4}{\lambda_{1/2}}\left(\frac{2}{3n} + \frac{1}{\sqrt{n}}\right)\ln\left(\frac{4}{\delta}\right)\right] \geq 1-\delta.$$

Remark that the proof of this theorem rely on the application of an exponential inequality [Pin94] for the martingale term in decomposition (2.2) before finding the rank n_δ such that the other terms of the decomposition are negligible. Then, one can derive constants C_1, C_2, C_3 such that (see the proof of Theorem 4.2 in [CCGB15])

$$n_\delta = \max\left\{\left(\frac{C_1}{\delta\log\left(\frac{4}{\delta}\right)}\right)^{\frac{2}{1-\gamma}}, \left(\frac{C_2}{\delta\log\left(\frac{4}{\delta}\right)}\right)^{\frac{2}{2\gamma-1}}, \left(\frac{C_3}{\delta\log\left(\frac{4}{\delta}\right)}\right)^{\frac{1}{2}}\right\}.$$

5.2.4 Weiszfeld's algorithm

In this section, we make same recalls on the Weiszfeld's algorithm which can be of interest for robust clustering methods developed in Sections 5.3 and 5.5. First, observe that one can see the median as a fix point. Indeed, one has

$$\nabla G(m_{1/2}) = \mathbb{E}\left[\frac{X - m_{1/2}}{\|X - m_{1/2}\|}\right] = 0 \Leftrightarrow m_{1/2} = \frac{\mathbb{E}\left[\frac{X}{\|X-m_{1/2}\|}\right]}{\mathbb{E}\left[\frac{1}{\|X-m_{1/2}\|}\right]}$$

Then, considering $X_1, \ldots, X_n$ with the same law as X, one can use a fix point algorithm with the empirical function generated by the sample, leading to the following Weiszfeld's algorithm [Wei37]

$$m_{1/2,n,t+1} = \frac{\sum_{k=1}^n \frac{X_k}{\|X_k - m_{1/2,n,t}\|}}{\sum_{k=1}^n \frac{1}{\|X_k - m_{1/2,n,t}\|}}. \tag{5.2}$$

Note that this can be written as

$$m_{1/2,n,t+1} = m_{1/2,n,t} + \frac{1}{\sum_{k=1}^n \frac{1}{\|X_k - m_{1/2,n,t}\|}} \sum_{k=1}^n \frac{X_k - m_{1/2,n,t}}{\|X_k - m_{1/2,n,t}\|},$$

i.e Weiszfeld's algorithm can be seen as an iterative gradient algorithm with a stepsequence $\eta_t = \frac{1}{\sum_{k=1}^n \frac{1}{\|X_k - m_{1/2,n,t}\|}}$. Furthermore, under Assumptions **($A_{median}1b$)** and **($A_{median}2$)**, one can check that we have the convergence in law [VZ00]

$$\lim_{n,t\to+\infty} \sqrt{n}\,(m_{1/2,n,t} - m_{1/2}) = \mathcal{N}\left(0, H_{1/2}^{-1}\Sigma_{1/2}H_{1/2}^{-1}\right),$$

i.e one obtains the same asymptotic normality as for the averaged estimates. Nevertheless, although these estimates can be very performing in case of small samples in small dimensional spaces, they necessitate much more computational costs for dealing with big data.

5.3 Application to K-medians

This section is based on [GBS22]

5.3.1 Introduction

Clustering is unsupervised machine learning technique which is defined as the algorithm for grouping the data points into a collection of groups based upon similar features. There is a vast literature on clustering techniques and general references regarding clustering may be found in [Spa80, JD88, Mir96, JMF99, Ber06, KR09]. We focus here on hard clustering methods whose most popular one is the K-means algorithm [For65, Mac67]. More precisely, considering $X_1, ..., X_n$ be random vectors taking values in $\mathbb{R}^d$, the aim of K-means algorithm is to find k centroids $\{c_1, ..., c_k\}$ minimizing the empirical distortion

$$\frac{1}{n}\sum_{i=1}^{n} \min_{j=1,..,k} \left\|X_i - c_j\right\|^2. \tag{5.3}$$

Nevertheless, K-means methods are very sensitive to the presence of outliers. It is then preferable to focus on K-medians clustering [Mac67, KR09]. This can be seen as a variant of K-means clustering where instead of calculating the mean of each cluster to determine its centroid, we calculate instead the geometric median. It consists in considering criteria based on least norms instead of least squared norms. More precisely, considering the same sequence of i.i.d copies $X_1, ..., X_n$, the objective of K-medians clustering is to minimize the empirical L^1-distortion :

$$\frac{1}{n}\sum_{i=1}^{n} \min_{j=1,..,k} \left\|X_i - c_j\right\|,$$

i.e the centroids are now the medians of the clusters. Nevertheless, in practical applications, the number of clusters k is unknown. The aim of this section is to give a method to chose the "optimal" number of clusters for robust clustering. Note that several methods for determining the optimal number of clusters have been studied for K-means algorithms and can be easily adapted for K-medians. In practice, one of the most used method for determining the optimal number of clusters is elbow method. Other methods often used are the Silhouette [KR09] and the Gap Statistic [TWH01].

We propose here a new approach based on a penalized criterion to chose the number of cluster (see [Fis11] for the case of K-means). More precisely, we introduce a penalty function to avoid choosing

too large k and a data-driven calibration algorithm [BM07, AM09] is used to find the constant of this penalty function. All the proposed algorithms are available in the R package `Kmedians` on CRAN[4].

5.3.2 K-medians algorithms

For a positive integer k, a vector quantizer Q of dimension d and codebook size k is a (measurable) mapping of the d-dimensional Euclidean $\mathbb{R}^d$ into a finite set of points $\{c_1, ..., c_k\}$ [Lin00]. More precisely, the points $c_i \in \mathbb{R}^d, i = 1, ..., k$ are called the codepoints and the vector composed of the code points $\{c_1, ..., c_k\}$ is called codebook, denoted by c. Given a d-dimensional random vector X admitting a finite first order moment, the L^1-distortion of a vector quantizer Q with codebook $c = \{c_1, ..., c_k\}$ is defined by

$$W(c) := \mathbb{E}\left[\min_{j=1,..,k} \left\|X - c_j\right\|\right].$$

Let us now consider $X_1, ..., X_n$ random vectors of $\mathbb{R}^d$ i.i.d with the same law as X. Then, one can define the empirical L^1-distortion as :

$$W_n(c) := \frac{1}{n}\sum_{i=1}^{n} \min_{j=1,..,k} \left\|X_i - c_j\right\|.$$

We consider here two kinds of K-medians algorithms : sequential and non sequential algorithm. The non sequential algorithm uses Lloyd-style iteration which alternates between an expectation (E) and maximization (M) step and is precisely described in Algorithm 1:

Inputs : $D = \{x_1, ..., x_n\}$ datapoints, k number of clusters
Output : A set of k clusters : $C_1, ..., C_k$
Randomly choose k centroids : $m_1, ..., m_k$.
while *the clusters change* **do**
  for $1 \leq i \leq n$ **do**
    $r = \arg\min_{1 \leq j \leq k} \left\|x_i - m_j\right\|$
    $C_r \leftarrow x_i$
  end
  for $1 \leq j \leq k$ **do**
    $m_j = \arg\min_m \sum_{i, x_i \in C_j} \|x_i - m\|$
  end
end

Algorithme 1 : Non Sequential K-medians Algorithm .

For $1 \leq j \leq k$, m_j is nothing but the geometric median of the points in the cluster C_j. As m_j is not explicit, we will use Weiszfeld algorithm defined by (5.2) (indicated by "Offline") or the averaged stochastic gradient algorithm defined by 5.1 (indicated by "Semi-online") to estimate it. The Online

[4] https://cran.r-project.org/package=Kmedians

K-median algorithm proposed by [CCM12] based on an averaged Robbins-Monro procedure is described in Algorithm 2:

Inputs : $D = \{x_1, ..., x_n\}$ datapoints, k number of clusters, $c_\gamma > 0$ and $\gamma \in (1/2, 1)$
Output : A set of k clusters : $C_1, ..., C_k$
Randomly choose k centroids : $m_1, ..., m_k$.
$\overline{m}_j = m_j \; \forall \; 1 \leq j \leq k$
$n_j = 1 \; \forall \; 1 \leq j \leq k$
for $1 \leq i \leq n$ **do**
  $r = \arg\min_{1 \leq j \leq k} \left\| x_i - \overline{m}_j \right\|$
  $C_r \leftarrow x_i$
  $m_r \leftarrow m_r + \frac{c_\gamma}{(n_r+1)^\gamma} \frac{x_i - m_r}{\|x_i - m_r\|}$
  $\overline{m}_r \leftarrow \frac{n_r \overline{m}_r + m_r}{n_r + 1}$
  $n_r \leftarrow n_r + 1$
end

Algorithme 2 : Online K-medians Algorithm .

5.3.3 Selecting the number of clusters

In this section, we adapt the results that have been shown for K-means in [Fis11] to K-medians clustering. In this aim, let $X_1, ..., X_n$ be i.i.d random vectors with the same law as X, and we assume that $\|X\| \leq R$ almost surely for some $R > 0$. Let S_k denote the countable set of all $\{c_1, ..., c_k\} \in \mathbb{Q}^k$, where $\mathbb{Q}$ is some grid over $\mathbb{R}^d$. A codebook $\hat{c}_k$ is said empirically optimal codebook if we have $W_n(\hat{c}_k) = \min_{c \in S_k} W_n(c)$. In the sequel, let $\hat{c}_k$ be a minimizer of the criterion $W_n(c)$ over S_k. Our aim is to determine $\hat{k}$ minimizing a criterion of the type

$$\text{crit}(k) = W_n(\hat{c}_k) + \text{pen}(k)$$

where pen : $\{1, ..., n\} \to \mathbb{R}_+$ is a penalty function described later. The following theorem provides an uniform upper bound of the difference between the empirical and the L^1 distortion.

Theorem 5.3.1 ([GBS22]). *Let X a random vector taking values in $\mathbb{R}^d$ such that $\|X\| \leq R$ almost surely for some $R > 0$. Then for all $1 \leq k \leq n$,*

$$\mathbb{E}\left[\sup_{c \in S_k} \{W(c) - W_n(c)\}\right] \leq 48R\sqrt{\frac{kd}{n}}.$$

This theorem shows that the maximum difference between the L^1-distortion and the empirical distortion of any vector quantizer is of order $n^{-1/2}$ and enables to give the following upper bound of the L^1-distortion.

Theorem 5.3.2 ([GBS22]). *Consider non-negative weights $\{x_k\}_{1 \leq k \leq n}$ such that $\sum_{k=1}^n e^{-x_k} = \Sigma$. Suppose*

that $\|X\| \leq R$ *almost surely and that for every* $1 \leq k \leq n$

$$pen(k) \geq R\left(48\sqrt{\frac{kd}{n}} + 2\sqrt{\frac{x_k}{2n}}\right).$$

Then

$$\mathbb{E}\left[W(\tilde{c})\right] \leq \inf_{1\leq k\leq n}\left\{\inf_{c\in S_k} W(c) + pen(k)\right\} + \Sigma R\sqrt{\frac{\pi}{2n}}$$

where $\tilde{c} = \hat{c}_{\hat{k}}$ *minimizer of the penalized criterion.*

Considering the simple situation where one can take $\{x_k\}_{1\leq k\leq n}$ such that $x_k = Lk$ for some positive constant L and $\Sigma = \sum_{k=1}^{n} e^{-x_k} \leq 1$ and taking

$$\text{pen}(k) = R\left(48\sqrt{\frac{kd}{n}} + 2\sqrt{\frac{Lk}{2n}}\right) = R\sqrt{\frac{k}{n}}\left(48\sqrt{d} + 2\sqrt{\frac{L}{2}}\right)$$

we deduce that the penalty shape is $a\sqrt{\frac{k}{n}}$ where a is a constant. From Proposition 3.1 in [GBS22], considering a penalty $\text{pen}(k) = aR\sqrt{\frac{k}{n}}$ where $a \geq \left(48\sqrt{d} + 2\sqrt{\frac{L}{2}}\right)$, we obtain

$$\mathbb{E}\left[W(\tilde{c})\right] \leq R\left(\inf_{1\leq k\leq n}\left\{4k^{-1/d} + a\sqrt{\frac{k}{n}}\right\} + \Sigma\sqrt{\frac{\pi}{2n}}\right).$$

Minimizing the term on the right hand side of previous inequality leads to k of order $n^{\frac{d}{d+2}}$ and

$$\mathbb{E}\left[W(\tilde{c})\right] = \mathcal{O}(n^{-\frac{1}{d+2}}).$$

We now focus on the callibration of the constant a. In this aim, we focus on the data-driven method introduced by [BM07]: the "slope heuristics". This method consists in estimating the constant of penalty function by the slope of the expected linear relation of $-W_n(\hat{c}_k)$ with respect to the penalty shape values $\text{pen}_{\text{shape}}(k) = \sqrt{\frac{k}{n}}$. More precisely, denoting $c^* = \arg\min_{c\in S} W(c)$ and $c_k = \arg\min_{c\in S_k} W(c)$ where S any linear subspace of $\mathbb{R}^d$ and S_k set of predictors. It was shown in [BM07, AM09, BMM12] that under conditions, the optimal penalty satisfies for large n

$$\text{pen}_{\text{opt}}(k) = a_{\text{opt}}\text{pen}_{\text{shape}}(k) \approx 2(W_n(c^*) - W_n(\hat{c}_k)).$$

This gives

$$\frac{a_{opt}}{2}\text{pen}_{\text{shape}}(k) - W_n(c^*) \approx -W_n(\hat{c}_k).$$

The term $-W_n(\hat{c}_k)$ with respect to the penalty shape behaves like a linear function for a large k. The slope $\hat{S}$ of the linear regression of $-W_n(\hat{c}_k)$ on $pen_{shape}(k)$ is estimated by $\frac{a_{opt}}{2}$. Finally, we obtain

$$\text{pen}(k) = 2\hat{S}\text{pen}_{\text{shape}}(k).$$

5.3.4 Simulations

The studied algorithms are available in the R package `Kmedians`[5]. In what follows, the centers initialization is generated from robust hierarchical clustering algorithm with `genieclust`[6] package [GBC16].

Visualization of results with the package `Kmedians`

In Section 5.3.3, we proved that the penalty shape is $a\sqrt{\frac{k}{n}}$ where a is a constant to calibrate. To find the constant a, we will use the data-based calibration algorithm for penalization procedures that is explained at the end of section 5.3.3. This data-driven slope estimation method is implemented in CAPUSHE (CAlibrating Penalty Using Slope HEuristics) [BBM+11] which is available in the R package `capushe`[7]. Remark that this proposed slope estimation method has been built to be robust in order to preserve the eventual undesirable variations of criteria.
In what follows, we consider a random variable X following a Gaussian Mixture Model in $\mathbb{R}^5$ with $k = 6$ classes and we consider $n = 3000$ i.i.d realizations of X. We first focus on some visualization of the slope method.

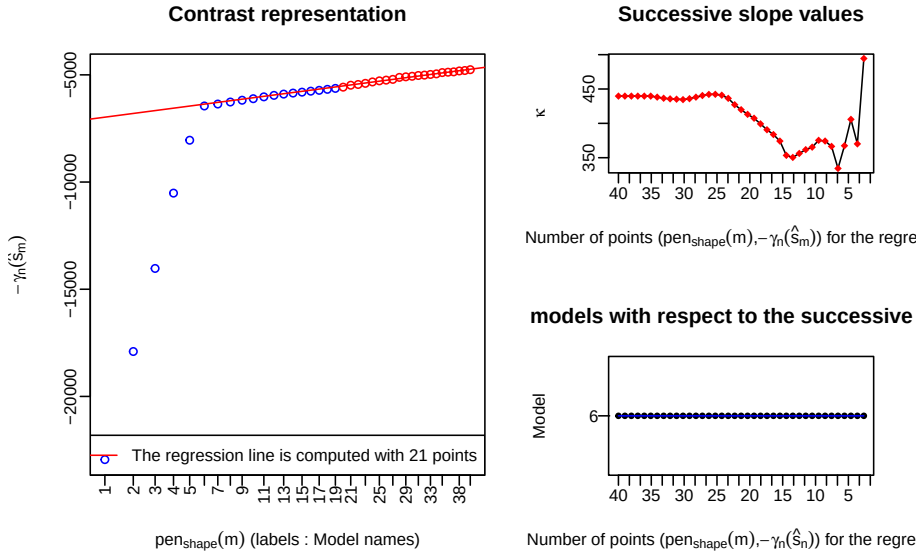


Figure 5.3 – Evolution of $-W_n(\hat{c}_k)$ with respect to k (on the left), Slope values as function of the number of points used to estimate the slope (upper right) and selected number of clusters for each number of points used to estimate the slope (bottom right).

Figure 5.4 (left) shows that there are two possible elbow of this curve so, the elbow method suggests taking 5 or 6 as the number of clusters. In this case, the elbow method is not ideal. In Figures 5.5 and 5.6, in order to visualize data points in dimensions higher than 3, we represent data as

[5] https://cran.r-project.org/package=Kmedians
[6] https://cran.r-project.org/package=genieclust
[7] https://cran.r-project.org/package=capushe

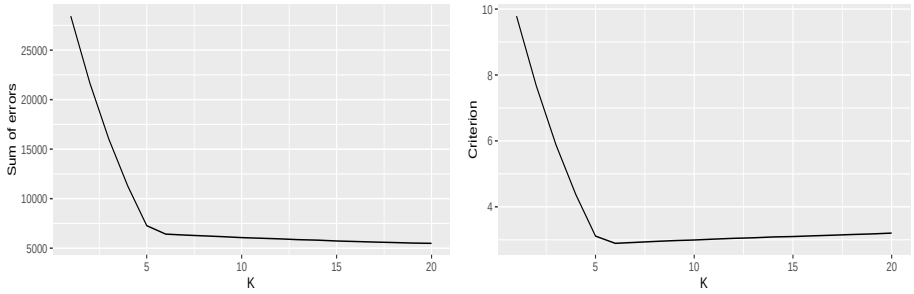


Figure 5.4 – Evolution of $W_n(\hat{c}_k)$ (on the left) and crit(k) (on the right) with respect to k.

curves that we call "profiles", gathered it by cluster, and represented the centers of the groups in red. We also represent the 2 first principal components of the data using robust PCA (see Section 5.4). In order to visualize the robustness of the proposed method, we consider contaminated data with the law $Z = (Z_1, ..., Z_5)$ where Z_i are i.i.d, with $Z_i \sim \mathcal{T}_1$ where $\mathcal{T}_1$ is a Student law with 1 degree of freedom. Applying our method for selecting the number of clusters for K-medians algorithms, we selected the correct number of clusters and the obtained groups are coherent. Nevertheless, in the case of K-means clustering, the method assimilates some far outliers as single clusters (see Figure 5.6). Note that in the case of contaminated data (Figures 5.5 and 5.6), we only represented 95% of the data in order to better visualize them. Then, in Figure, 5.6, Clusters 5, 7, 8, 11 and 12 are not visible since they are "far" outliers.

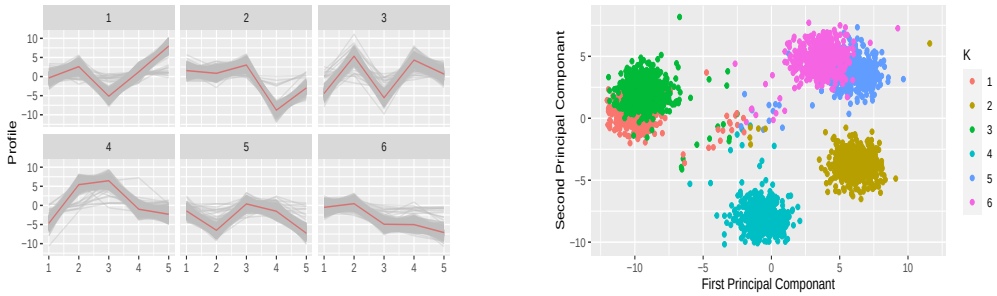


Figure 5.5 – Profiles (on the left) and clustering via K-medians algorithm represented on the first two principal components (on the right) with 5% of contaminated data.

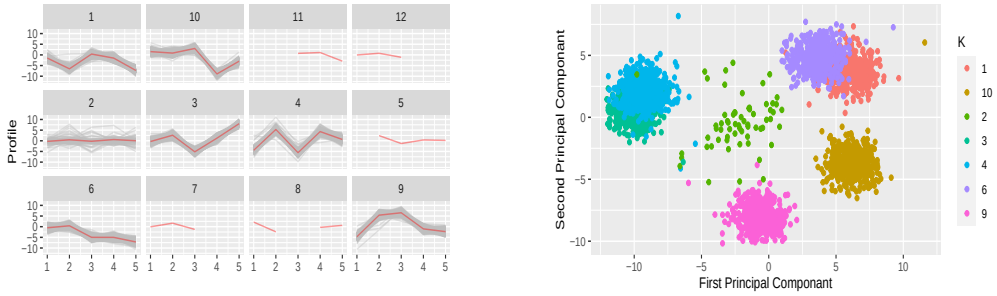


Figure 5.6 – Profiles (on the left) and clustering via K-means algorithm represented on the first two principal components (on the right) with 5% of contaminated data.

Comparison with Gap Statistic and Silhouette

In what follows, we focus on the choice of the number of clusters and compare our results with different methods. For this, we generated some basic data sets in three different scenarios (see [Fis11]) :

(S1) A single cluster in dimension 10 **:** We consider 2000 points uniformly distributed over the unit hypercube in dimension 10.

(S2) 4 clusters in dimension 3 **:** The data are generated by Gaussian mixture centered at $(0,0,0)$, $(0,2,3)$, $(3,0,-1)$, and $(-3,-1,0)$ with variance equal to the identity matrix. Each cluster contains 500 data points.

(S3) 5 clusters in dimension 4 **:** The data are generated by Gaussian mixture centered at $(0,0,0,0)$, $(3,5,-1,0)$, $(-5,0,0,0)$, $(1,1,6,-2)$ and $(1,-3,-2,5)$ with variance equal to the identity matrix. Each cluster contains 500 data points.

For each scenario, we contaminated our data with the law $Z = (Z_1, ..., Z_d)$ where Z_i are i.i.d, with $Z_i \sim \mathcal{T}_1$. We then evaluate our method for the different methods and scenarios by considering:

- N : number of times we get the right value of cluster in 50 repeated trials without contaminated data.
- $\bar{k}$: average of number of clusters obtained over 50 trials without contaminated data.
- $N_{0.1}$: number of times we get the right value of cluster in 50 repeated trials with 10% of contaminated data.
- $\bar{k}_{0.1}$: average of number of clusters obtained over 50 trials with 10% of contaminated data.

In every scenario, Offline, Semi-Online, Online K-medians with the slope method give very competitive (best) results and in the case where the data are contaminated, they clearly over perform other methods (especially the Offline method). As expected, in terms of efficiency, we find the order Offline, Semi-Online, Online since the sample size is moderate, but the Online algorithm is very competitive and is very cheap in term of computational calculus.

Simulations		S1				S2				S3			
	Algorithms	N	$\bar{k}$	$N_{0.1}$	$\bar{k}_{0.1}$	N	$\bar{k}$	$N_{0.1}$	$\bar{k}_{0.1}$	N	$\bar{k}$	$N_{0.1}$	$\bar{k}_{0.1}$
Slope	Offline	**50**	**1**	49	1.04	**50**	**4**	**50**	**4**	**50**	**5**	**50**	**5**
	Semi-Online	46	1.1	44	1.7	50	4	49	4.02	**50**	**5**	46	5.1
	Online	43	1.6	49	1.1	48	**4**	42	**4**	**50**	**5**	40	5.2
	K-means	18	1.6	0	7	**50**	**4**	1	7.9	**50**	**5**	2	6.7
Gap	Offline	**50**	**1**	**50**	**1**	6	1.7	0	1	47	4.8	2	1.2
	Semi-Online	**50**	**1**	**50**	**1**	7	1.7	0	1	47	4.8	2	1.2
	Online	**50**	**1**	**50**	**1**	8	2.4	0	1	47	4.8	2	1.2
	K-means	**50**	**1**	**50**	**1**	0	1.2	0	1.2	12	2	0	1.3
Silhouette	Offline	0	6.4	0	2	0	3	0	2.9	24	4.4	1	3.5
	Semi-Online	0	5.8	0	2	0	3	0	2.9	24	4.4	1	3.5
	Online	0	2.1	0	2.1	0	3	2	3.2	27	4.5	2	4.5
	K-means	0	7.9	0	2.1	0	3	7	3.2	27	4.5	0	6.7

Table 5.1 – Comparison of the number of times we get the right value of clusters and the averaged selected number of clusters obtained with the different methods without contaminated data and with 10% of contaminated data.

Contaminated Data

We now focus on the impact of contaminated data on K-means and K-medians clustering and on the choice of the number of clusters. In this aim, we generate data with a Gaussian mixture model with 10 classes in dimension 5 (whose centers are generated randomly on the sphere of radius 10) and each class contains 500 data points. The data are contaminated with the law $Z = (Z_1, ..., Z_5)$ where Z_i are i.i.d, with 3 possible scenarios:

1. $Z_i \sim \mathcal{T}_1$
2. $Z_i \sim \mathcal{T}_2$
3. $Z_i \sim \mathcal{U}[-10, 10]$

where $\mathcal{T}_m$ is the Student law with m degrees of freedom and $\mathcal{U}[a, b]$ is the continuous uniform distribution on $[a, b]$. In what follows, let us denote by ρ the proportion of contaminated data. In order to compare the different clustering results, we focus on the Adjusted Rand Index (ARI) [Ran71, HA85].

Without contaminated data, the three K-medians algorithms as well as the K-means algorithm have globally found the right number of clusters with an averaged ARI close to 0.99. In addition, in the case of contaminated data (and especially for a contamination following a Student's law with 1 degree of freedom), the proposed slope method for K-medians successfully found more or less the optimal number of clusters up to 28% contamination, and so with competitive ARI, and globally over-perform K-means method. Note that in case of high contamination rate, we usually get 11 clusters, which is logical since most of the contaminated data forms a kind of new cluster around the center of the sphere.

	ρ		0	0.01	0.02	0.03	0.05	0.09	0.16	0.28	0.5
$Z_i \sim \mathcal{T}_1$	Offline	$\hat{k}$	**10**	**10**	**10.2**	**10.2**	**10.7**	**10.8**	11.4	**9.9**	3.1
	Semi-Online		**10**	10.1	**10.2**	10.7	11	11.2	12	10.6	3.2
	Online		**10**	10.1	**10.2**	10.8	11.1	11.7	12.1	11.2	2.8
	K-means		10.6	13.5	14	13.6	12.9	12.3	**8.9**	8.5	**11.5**
	Offline	ARI	**0.99**	**0.99**	**0.98**	**0.99**	**0.98**	**0.98**	**0.97**	0.81	0.15
	Semi-Online		**0.99**	**0.99**	**0.98**	0.98	**0.98**	0.97	**0.97**	**0.91**	**0.19**
	Online		**0.99**	**0.99**	**0.98**	0.98	**0.98**	**0.98**	**0.97**	0.87	0.16
	K-means		0.98	0.94	0.92	0.88	0.79	0.69	0.5	0.33	0.12
$Z_i \sim \mathcal{T}_2$	Offline	$\hat{k}$	**10**	**10**	**10.7**	**11**	**11**	**10.9**	**10.9**	**11.2**	**11.1**
	Semi-Online		**10**	**10**	10.9	**11**	**11**	**10.9**	**10.9**	**11.2**	**11.1**
	Online		**10**	10.1	11.3	**11**	**11**	**10.9**	**10.9**	**11.2**	11.2
	K-means		10.6	11.1	11.5	11.3	11.7	12.1	13	12.7	8
	Offline	ARI	**0.99**	**0.99**	**0.97**	**0.98**	**0.97**	**0.98**	**0.98**	**0.97**	**0.96**
	Semi-Online		**0.99**	**0.99**	**0.97**	**0.98**	**0.97**	**0.98**	**0.98**	**0.97**	**0.96**
	Online		**0.99**	**0.99**	**0.97**	**0.98**	**0.97**	**0.98**	**0.98**	**0.97**	**0.96**
	K-means		0.98	0.98	**0.97**	**0.98**	**0.97**	0.96	0.96	0.95	0.68
$Z_i \sim \mathcal{U}[-10, 10]$	Offline	$\hat{k}$	**10**	**10**	**10.1**	**10.1**	**10**	**10**	10.5	11.9	**10.8**
	Semi-Online		**10**	**10**	**10.1**	**10.1**	**10**	**10**	**10.3**	11.9	**10.8**
	Online		**10**	**10**	**10.1**	**10.1**	**10**	**10**	10.5	**11.1**	11.2
	K-means		10.6	10.7	11.1	11.2	12	11.6	11.8	11.3	**9.2**
	Offline	ARI	**0.99**	**0.99**	**0.97**	**0.98**	**0.97**	**0.98**	**0.98**	**0.97**	**0.96**
	Semi-Online		**0.99**	**0.99**	**0.97**	**0.98**	**0.97**	**0.98**	**0.98**	**0.97**	**0.96**
	Online		**0.99**	**0.99**	**0.97**	**0.98**	**0.97**	**0.98**	**0.98**	**0.97**	**0.96**
	K-means		0.98	0.97	**0.97**	**0.98**	**0.97**	0.96	0.96	0.92	0.79

Table 5.2 – Comparison of the selected number of clusters and the averaged ARI obtained with the different methods with respect to the proportion of contaminated data for $Z_i \sim \mathcal{T}_1$, $Z_i \sim \mathcal{T}_2$ and $Z_i \sim \mathcal{U}[-10, 10]$.

We now define the empirical L^1-error of the centroids estimation by:

$$\sum_{j=1}^{\hat{k}} \min_{j=1,..,k} \left\| \hat{c}_i - c_j \right\| \tag{5.4}$$

with $c = \{c_1, ..., c_k\}$ and $\hat{c} = \left\{\hat{c}_1, ..., \hat{c}_{\hat{k}}\right\}$ where $\hat{k}$ selected number of clusters. The empirical L^1-error of the centroids estimation and the selected number of clusters, for each algorithms, are given in Figure 5.7 and 5.8. In Figure 5.8 (left), only the K-medians algorithms is visible since the empirical L^1-error of the centroid estimation of K-means algorithm totally blows up and varies between the values 10000 and 30000 with a median close to 15000. The K-means algorithm is clearly affected by the presence of outliers and both its L^1-error and its predicted number of clusters are now much larger than for the other algorithms. Other three K-medians algorithms have globally good performances, even if Offline is slightly better.

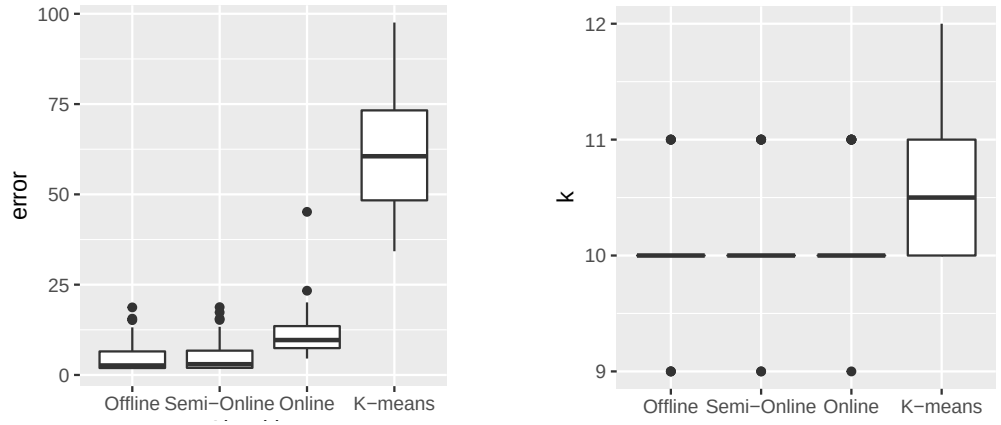


Figure 5.7 – Box plots reflect empirical L^1-error (see (5.4)) of centroid estimation (on the left) and the selected number of clusters k (on the right) for the "Offline", "Semi-Online", "Online" and K-means without contaminated data.

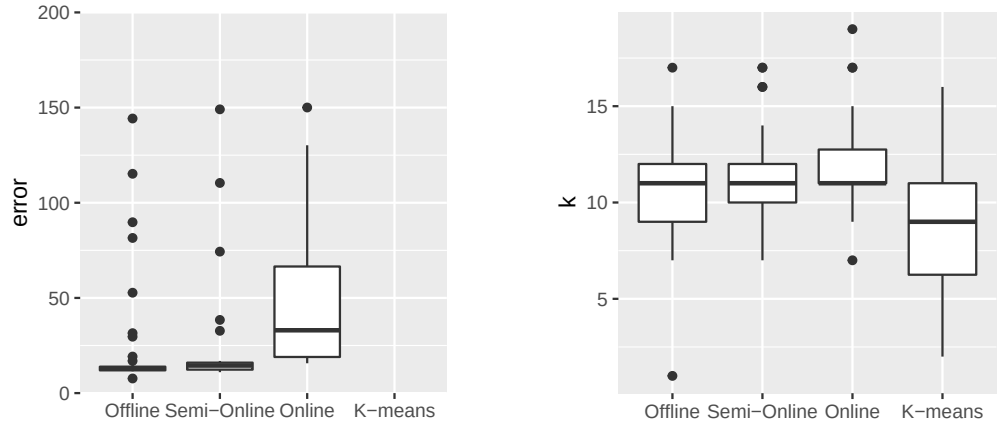


Figure 5.8 – Box plots reflect empirical L^1-error (see (5.4)) of centroid estimation (on the left) and the selected number of clusters k (on the right) for the "Offline", "Semi-Online", "Online" and K-means with 28% of contaminated data.

5.4 Estimating the Median Covariation Matrix with application to online Robust PCA

5.4.1 Introduction

Principal Components Analysis is one of the most useful statistical tool to extract information by reducing the dimension when one has to analyze large samples of multivariate or functional data (see *e.g.* [Jol02, RS05, Ver06, HPV14]). Nevertheless, principal components, which are derived from the spectral analysis of the covariance matrix, can be very sensitive to outliers (see [DGK81]) and many robust procedures for principal components analysis have been considered in the literature (see [HRVA08, HR09, MMY06, RvD99, CRG05, CFO07, HU07, BBT+11, LMS+99, Ger08, TKO12] among others).

We consider in this section another approach for robust PCA based on a new robust dispersion

indicator that we called Median Covariation Matrix (MCM). As shown in [KP12] the MCM has the same eigenspaces as the usual covariance matrix when the distribution of the data is symmetric and the second order moment is finite so that considering the MCM to compute principal components can be of interest. Since the MCM can be seen as the median of the random variable $(X - m_{1/2})(X - m_{1/2})^T$ (where $m_{1/2}$ is the median of X), different algorithms can be considered to get effective estimators of the MCM. When the dimension of the data is not too high and the sample size is not too large, Weiszfeld's algorithm (see [Wei37, VZ00] and Section 5.4.5) can be directly used to estimate effectively both the geometric median and the MCM. When both the dimension and the sample size are large, we will show in this section how the stochastic algorithms for estimating the geometric median can be adapted to estimate recursively and simultaneously the geometric median as well as the MCM without necessity to store all the data. We then highlight the interest of considering the MCM to perform principal components analysis of large samples of high dimensional contaminated data through a simulation study.

5.4.2 Definition and framework

Let us denote by $\mathcal{L}(\mathcal{H})$ the space of linear operators on $\mathcal{H}$. Denoting $\mu = \mathbb{E}[X]$, remark that one can see the covariance of X as

$$\mathrm{Cov}[X] = \mathrm{argmin}_{V \in \mathcal{L}(\mathcal{H})} \mathbb{E}\left[\left\| (X-\mu)(X-\mu)^T - V \right\|_F^2 - \left\| (X-\mu)(X-\mu)^T \right\|_F^2 \right]$$

where $\|.\|_F$ denotes the Frobenius norm. Nevertheless, the covariance as well as the mean are not robust at all. Then, we now focus on the Median Covaration Matrix (MCM for short) which is the minimizer of the functional $G_{m_{1/2}} : \mathcal{L}(\mathcal{H}) \longrightarrow \mathbb{R}$ defined for all $V \in \mathcal{L}(\mathcal{H})$ by

$$G_{m_{1/2}}(V) = \mathbb{E}\left[\left\| (X - m_{1/2})(X - m_{1/2})^T - V \right\|_F - \left\| (X - m_{1/2})(X - m_{1/2})^T \right\|_F \right],$$

where $m_{1/2}$ is the median of X. In other words, the MCM, denoted by V^*, can be seen as the geometric median of the random variable $Y = (X - m_{1/2})(X - m_{1/2})^T$. In order to ensure the existence and uniqueness of the MCM, we suppose from now that the following assumptions are fulfilled:

($\mathbf{A_{MCM}1}$) There is a positive constant C_{MCM} such that for all $h \in \mathcal{H}$ and $V \in \mathcal{L}(\mathcal{H})$,

$$\mathbb{E}\left[\frac{1}{\|(X-h)(X-h)^T - V\|_F^2} \right] \leq C_{\mathrm{MCM}}.$$

($\mathbf{A_{MCM}2}$) For all $V \in \mathcal{L}(H)$, there is $V' \in \mathcal{L}(H)$ such that

$$\langle V, V' \rangle = 0 \qquad \text{and} \qquad \mathbb{V}\left[\left\langle (X - m_{1/2})(X - m_{1/2})^T, V' \right\rangle_F \right] > 0,$$

where $\langle .,. \rangle_F$ is the inner product associated to the Frobenius norm. Since the MCM can be seen as a median, these hypothesis are the mirror of Assumptions **($\mathbf{A_{median}1}$)** and **($\mathbf{A_{median}2}$)**. As for the median, these assumptions ensure that the MCM is uniquely defined and that for any $h \in \mathcal{H}$, the functional G_h defined for all $V \in \mathcal{L}(\mathcal{H})$ by

$$G_h(V) = \mathbb{E}\left[\|Y(h) - V\|_F - \|Y(h)\|_F\right]$$

with $Y(h) := (X-h)(X-h)^T$ is twice continuously differentiable. Indeed, one has

$$\nabla G_h(V) = -\mathbb{E}\left[\frac{Y(h)-V}{\|Y(h)-V\|_F}\right] \quad \text{and} \quad \nabla^2 G_h(V) = \mathbb{E}\left[\frac{1}{\|Y(h)-V\|_F}\left(I_{\mathcal{L}(\mathcal{H})} - \frac{(Y(h)-V)\otimes_F (Y(h)-V)}{\|Y_h - V\|_F^2}\right)\right]$$

where for all $A, B, V \in \mathcal{L}(\mathcal{H})$, $(A \otimes_F B)(V) = \langle A, V\rangle_F B$.

5.4.3 Online estimation of the Median Covariation Matrix

We suppose from now that we have i.i.d copies $X_1, \ldots, X_n, X_{n+1}, \ldots$ of X. Remark that since the MCM can be seen as a median, knowing $m_{1/2}$, one could use the averaged stochastic gradient algorithm for estimating a median, i.e one could consider the recursive estimates defined by

$$\begin{aligned} W_{n+1} &= W_n + \gamma_{n+1}\frac{(X_{n+1}-m_{1/2})(X_{n+1}-m_{1/2})^T - W_n}{\left\|(X_{n+1}-m_{1/2})(X_{n+1}-m_{1/2})^T - W_n\right\|_F} \\ \overline{W}_{n+1} &= \overline{W}_n + \frac{1}{n+2}\left(W_{n+1} - \overline{W}_n\right) \end{aligned}$$

with $\overline{W}_0 = W_0$, and $\gamma_n = c_\gamma n^{-\gamma}$ with $c_\gamma > 0$ and $\gamma \in (1/2, 1)$. Nevertheless, since most of the time the median is unknown, one has to simultaneously estimate $m_{1/2}$ and V^*, leading to the following recursive algorithm:

$$\begin{aligned} m_{1/2,n+1} &= m_{1/2,n} + \gamma_{n+1}^{(m)}\frac{X_{n+1}-m_{1/2,n}}{\|X_{n+1}-m_{1/2,n}\|} \\ \overline{m}_{1/2,n+1} &= \overline{m}_{1/2,n} + \frac{1}{n+2}\left(m_{1/2,n+1} - \overline{m}_{1/2,n}\right) \\ V_{n+1} &= V_n + \gamma_{n+1}\frac{(X_{n+1}-\overline{m}_{1/2,n})(X_{n+1}-\overline{m}_{1/2,n})^T - V_n}{\left\|(X_{n+1}-\overline{m}_{1/2,n})(X_{n+1}-\overline{m}_{1/2,n})^T - V_n\right\|_F} \\ \overline{V}_{n+1} &= \overline{V}_n + \frac{1}{n+2}\left(V_{n+1} - \overline{V}_n\right) \end{aligned}$$

with $\overline{m}_{1/2,0} = m_{1/2,0}$, $\overline{V}_0 = V_0$, $\gamma_n^{(m)} = c_\gamma^{(m)} n^{-\gamma^{(m)}}$ and $\gamma_n = c_\gamma n^{-\gamma}$, where $c_\gamma^{(m)}, c_\gamma > 0$ and $\gamma^{(m)}, \gamma \in (1/2, 1)$. Remark that $m_{1/2,n}$ and $\overline{m}_{1/2,n}$ corresponds to the averaged stochastic gradient algorithm for estimating the median and does not depend on V_n nor $\overline{V}_n$. Furthermore, the difference between V_n and W_n is that we naturally replace the unknown median $m_{1/2}$ by its averaged estimates $\overline{m}_{1/2,n}$.

Remark that choosing V_0 symmetric and positive leads V_n to be symmetric but we cannot ensure that it is positive. In order to overcome this problem, a first solution is to project V_n on the convex cone of non negative operators, which would require to compute each eigenvalues of V_n, which is time consuming in high dimension. An other solution could be to consider a new stepsequence of the form

$$\gamma_{n+1,\text{pos}} = \min\left\{\gamma_{n+1}, \left\|\left(X_{n+1} - \overline{m}_{1/2,n}\right)\left(X_{n+1} - \overline{m}_{1/2,n}\right)^T - V_n\right\|_F\right\}$$

or

$$\gamma_{n+1,\text{pos}} = \gamma_{n+1}\mathbf{1}_{\gamma_{n+1} \leq \left\|\left(X_{n+1}-\overline{m}_{1/2,n}\right)\left(X_{n+1}-\overline{m}_{1/2,n}\right)^T - V_n\right\|_F}.$$

This choice of stepsequence, if V_0 is chosen non negative, would ensure that V_n is non negative for all $n \geq 0$.

5.4.4 Convergence results

In this section, we focus on the rate of convergence of the estimates (V_n), $(\overline{V}_n)$. We first establish the strong consistency of the estimates.

Theorem 5.4.1 ([CGB15]). *Suppose Assumptions* ***($A_{median}2$)***, ***($A_{MCM}1$)*** *and* ***($A_{MCM}2$)*** *hold. Then*

$$V_n \xrightarrow[n\to+\infty]{a.s} V^* \qquad \text{and} \qquad \overline{V}_n \xrightarrow[n\to+\infty]{a.s} V^*.$$

The obtaining of this result relies on the almost sure rate of convergence of the averaged estimates $\overline{m}_{1/2,n}$ coupled with the use of Robbins-Siegmund Theorem. We now give the rate of convergence in quadratic mean of the estimates:

Theorem 5.4.2 ([CGB15]). *Suppose Assumptions* ***($A_{median}2$)***, ***($A_{MCM}1$)*** *and* ***($A_{MCM}2$)*** *hold. Then*

$$\mathbb{E}\left[\|V_n - V^*\|_F^2\right] = O\left(\frac{1}{n^\gamma}\right) \qquad \text{and} \qquad \mathbb{E}\left[\left\|\overline{V}_n - V^*\right\|_F\right] = O\left(\frac{1}{n}\right).$$

Note that we so achieve the usual rate of convergence $\frac{1}{n^\gamma}$ for gradient estimates and achieve the usual rate $\frac{1}{n}$ for their averaged version. Nevertheless, we do not give explicitly the upper bound of the quadratic mean error. Furthermore, injecting the estimates of the median in algorithms avoid the obtaining of the asymptotic efficiency of the estimates.

5.4.5 Remark on the Weiszfeld's algorithm

Note that as in the case of the median, for moderate sample size lying in small dimensional spaces, one could estimate the MCM with the help of Weiszfeld algorithm. More precisely, considering the Weiszfeld estimate of the median $m_{1/2,n,T}$ (see Section 5.2.4), one could consider the iterative algorithm

$$V_{n,t+1} = \frac{\sum_{k=1}^n \frac{\left(X_k - m_{1/2,n,T}\right)\left(X_k - m_{1/2,n,T}\right)^T}{\left\|\left(X_k - m_{1/2,n,T}\right)\left(X_k - m_{1/2,n,T}\right)^T - V_{n,t}\right\|_F}}{\sum_{k=1}^n \frac{1}{\left\|\left(X_k - m_{1/2,n,T}\right)\left(X_k - m_{1/2,n,T}\right)^T - V_{n,t}\right\|_F}}.$$

5.4.6 Application to robust PCA

Application to robust online PCA

As mentioned before, we are interested in the estimation of the MCM since, if the distribution of X is symmetric, the MCM and the usual covariance matrix have the same eigenvectors, but this last one is not robust, i.e it is very sensitive to the presence of outliers. In this aim, we now focus one the recursive estimation of the q eigenvectors of V^* associated to the q largest eigenvalues, and so, without performing a spectral decomposition of $\overline{V}_n$ at each new observation. More precisely, we consider the following recursive strategy

$$u_{j,n+1} = u_{j,n} + \frac{1}{n+1}\left(\overline{V}_{n+1}\frac{u_{j,n}}{\|u_{j,n}\|} - u_{j,n}\right), \quad j = 1, \ldots, q$$

combined with an orthonormalization of $u_{1,n+1}, \ldots, u_{q,n+1}$. Remark that this approach enables to update the main eigenvectors with only $O\left(d^2\right)$ operations at each update.

Protocol

In what follows, we consider independent realizations of a random variable $Y \in \mathbb{R}^d$ where

$$Y = (1 - B(\delta))\, X + B(\delta)\epsilon$$

is a mixture of two distributions, and X, B, ϵ are independent random variables. The random vector X has a centered Gaussian distribution in $\mathbb{R}^d$ with covariance matrix $\Sigma[l, j] = \min(l, j)/d$. The multivariate contamination comes from ϵ, while $B(\delta) \sim \mathcal{B}(\delta)$ controls the rates of contamination. In what follows, we consider three different scenarios:

- The elements of vector ϵ are d independent realizations of a Student t distribution with one degree of freedom. This means that the first moment of Y is not defined when $\delta > 0$.
- The elements of vector ϵ are d independent realizations of a Student t distribution with two degrees of freedom. This means that the second moment of Y is not defined when $\delta > 0$.
- The vector ϵ is distributed has a "reverse time" Brownian motion. It has a Gaussian centered distribution, with covariance matrix $[\Sigma_\epsilon]_{\ell,j} = 2\min(d - \ell, d - j)/d$. The covariance matrix of Y is $(1 - \delta)\Sigma + \delta\Sigma_\epsilon$.

For the averaged recursive algorithms, we have considered $c_\gamma^{(m)} = c_\gamma = 2$ and a speed rate of $\gamma = \gamma^{(m)} = 3/4$. Note that the values of these tuning parameters have not been particularly optimized. The estimation error of the eigenspaces associated to the largest eigenvalues is evaluated by considering the squared Frobenius norm between the associated orthogonal projectors.

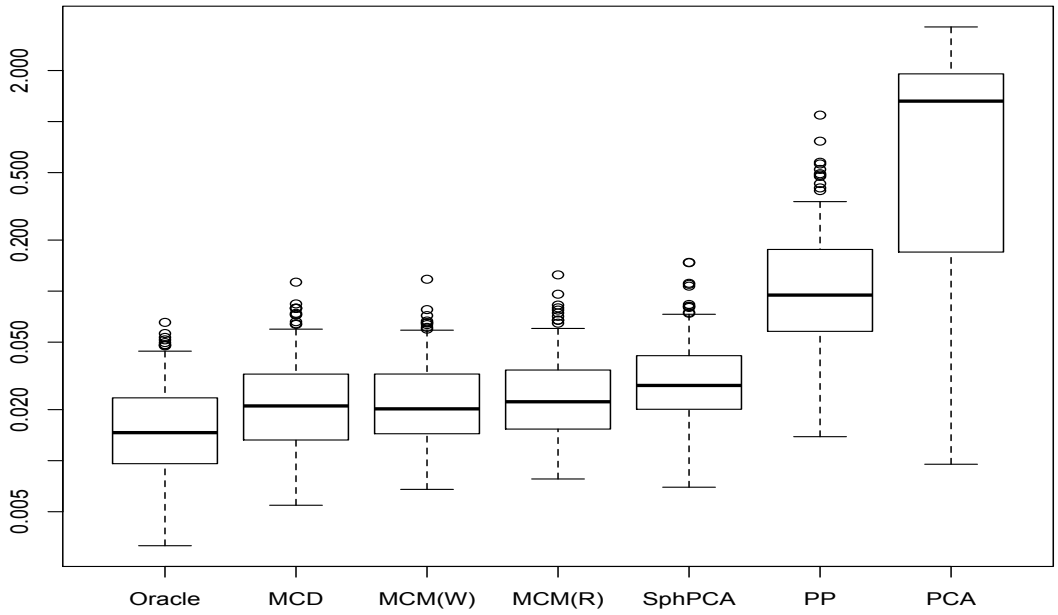


Figure 5.9 – Estimation errors (at a logarithmic scale) over 200 Monte Carlo replications, for $n = 200$, $d = 50$ and a contamination by a t distribution with 2 degrees of freedom with $\delta = 0.02$. MCM(W) stands for the estimation performed by the Weiszfeld's algorithm whereas MCM(R) denotes the averaged recursive approach.

Comparison with usual robust PCA techniques

We first compare the performances of the two estimates of the MCM based on the Weiszfeld's algorithm and the recursive algorithms with more classical robust PCA techniques. We generated samples of Y with size $n = 200$ (the conclusions do not differ much for different sample sizes) and dimension $d \in \{50, 200\}$, over 500 replications. Different levels of contamination are considered : $\delta \in \{0, 0.02, 0.05, 0.10, 0.20\}$. For both dimensions $d = 50$ and $d = 200$, the first eigenvalue of the covariance matrix of X represents about 81 % of the total variance, and the second one about 9 %. The median errors of estimation of the eigenspace generated by the first two eigenvectors ($q = 2$) are given in Table 5.3. In Figure 5.9, the distribution of the estimation error is drawn for the different approaches.

Note that even when the level of contamination is small (2% and 5%), the performances of classical PCA are strongly affected by the presence of outlying values in such (large) dimensions. When $d = 50$, the MCD algorithm and the MCM estimation provide the best estimations of the original two dimensional eigenspace, whereas when d gets larger ($d = n = 200$), the MCD estimator can not be used anymore (by construction) and the MCM estimator remains the most accurate. The performances of the spherical PCA are slightly less accurate whereas the median error of the robust PP is about four times larger. We can also note that the recursive MCM algorithm, which is

δ	Method	t 1 df	t 2 df d = 50	inv. B.	t 1 df	t 2 df d = 200	inv. B.
0%	PCA		0.015			0.015	
2%	PCA	3.13	1.18	0.677	3.95	1.85	0.691
	PP	0.097	0.087	0.090	0.099	0.088	0.093
	MCD	0.022	0.021	0.021	–	–	–
	Sph. PCA	0.029	0.028	0.029	0.031	0.027	0.028
	MCM (Weiszfeld)	0.021	0.021	0.022	0.023	0.021	0.021
	MCM (recursive)	0.023	0.024	0.025	0.026	0.023	0.026
5%	PCA	3.82	1.91	0.884	3.96	1.98	0.925
	PP	0.100	0.099	0.096	0.097	0.091	0.098
	MCD	0.022	0.020	0.024	–	–	–
	Sph. PCA	0.029	0.029	0.033	0.030	0.029	0.038
	MCM (Weiszfeld)	0.022	0.021	0.029	0.023	0.023	0.033
	MCM (recursive)	0.026	0.024	0.033	0.027	0.026	0.038
10%	PCA	3.83	1.95	1.05	3.96	1.99	1.12
	PP	0.107	0.109	0.099	0.100	0.105	0.093
	MCD	0.023	0.022	0.023	–	–	–
	Sph. PCA	0.031	0.031	0.059	0.030	0.028	0.056
	MCM (Weiszfeld)	0.024	0.023	0.059	0.022	0.023	0.056
	MCM (recursive)	0.030	0.027	0.072	0.028	0.026	0.069
20%	PCA	3.84	2.02	1.19	3.96	2.01	1.25
	PP	0.114	0.132	0.134	0.084	0.115	0.132
	MCD	0.025	0.026	0.026	–	–	–
	Sph. PCA	0.038	0.036	0.140	0.033	0.035	0.155
	MCM (Weiszfeld)	0.030	0.029	0.167	0.025	0.026	0.184
	MCM (recursive)	0.040	0.035	0.211	0.035	0.031	0.224

Table 5.3 – Median estimation errors, according to criterion $R(\widehat{\mathbf{P}}_q, \mathbf{P}_q)$ with a dimension $q = 2$, for datasets with a sample size $n = 200$, over 500 Monte Carlo experiments.

designed to deal with very large samples, performs well even for such moderate sample sizes (see also Figure 5.9).

5.5 Application to Robust Mixture Models

This section is based on [GBR22].

5.5.1 Introduction

In Section 5.3, we focused on hard partitioning methods, and in particular on K-medians algorithms. This section is dedicated to model-based clustering, which is one of the most popular soft clustering method [MP00]. It relies on the assumption that the observed data come from a mixture model, so that each cluster is characterized by a specific distribution. One reason for the popularity of these methods is that the maximum likelihood estimates of the parameters can be obtained via the well-known EM algorithm [DLR77], accompanied by statistical guarantees. Nevertheless, these methods are often very sensitive to the presence of outliers.

Several robust approaches have been proposed to overcome this problem. A first track sticks to the parametric framework, but uses emission distributions with heavier tails (see, e.g., [PM00, Wan15, SPIM15, RS19]). Alternatively, a component associated with (possibly improper) parametric distribution can be added, in order to capture outliers ([BR93, CH16, CH17, FP20]). A second approach is to prune the observations, so that the outliers do not weigh too heavily on the estimates [GEGMMI08]. A final approach is to use a dedicated weighted contrast (instead of negative log-likelihood [GYZ19, GMYZ21]).
This section focuses on the robustness of model-based clustering methods to the presence of outliers, meaning that we make no assumptions about how outliers deviate from prescribed emission distributions. To this end, we adopt a fully parametric model-based clustering framework, but modify the EM algorithm (more specifically, the M-step) to ensure robustness. Our proposed method resorts to the estimation of the median vector and the Median Covariation Matrix instead of the mean vector and the covariance matrix. In this section, we first propose methods to get robust estimates of the covariance when the law of the studied variable is known before applying it to robust model-based clustering. All the proposed methods are available in the R package `RGMM` accessible on CRAN[8].

5.5.2 Robust estimation of the variance

The algorithms

Let us suppose from now that X admits a second order moment and let us denote by μ and Σ its mean and variance (supposed to be positive). Let us recall that if the distribution of X is symmetric, the MCM of X denoted by V^* and Σ have the same eigenvectors ([KP12]). Furthermore, denoting $U = (U_1, \dots, U_d)^T := \Sigma^{-1/2}(X - \mu)$ and δ (resp. λ) the vector of eigenvalues (by decreasing order) of V^* (resp. Σ), one has ([KP12]),

$$\delta_k = \lambda_k \mathbb{E}\left[U_k^2 h(\delta, \lambda, U)\right] \left(\mathbb{E}\left[h(\delta, \lambda, U)\right]\right)^{-1} \tag{5.5}$$

where $h(\delta, \lambda, U) := \left(\sum_{i=1}^d \left(\delta_i - \lambda_i U_i^2\right)^2 + \sum_{i \neq j} \lambda_i \lambda_j U_i^2 U_j^2\right)^{-1/2}$. In what follows, we will denote by Ψ_U the function such that

$$\Psi_U(V^*) = \Sigma. \tag{5.6}$$

Let us suppose from now that the law of U is known and that we know how to simulate i.i.d random variables following this law (which is the case for multivariate Gaussian, Student or Laplace laws among others). Let us consider estimates of the eigenvalues of the MCM denoted by $\delta_n = (\delta_{1,n}, \dots, \delta_{d,n})$ and the associated estimates $\left(v_{1,n}, \dots, v_{p,n}\right)$ of the eigenvectors (see Section 5.4 to see how to build such estimates). In order to use a Monte Carlo method to estimate robustly the eigenvalues of the variance, we now consider that we generate $U_1, \dots, U_N$ i.i.d copies of U. A first solution to estimate λ is so to consider the following fix point algorithm: for all $t \in \mathbb{N}$, and

[8] https://cran.r-project.org/package=RGMM

$k = 1, \dots, d$,

$$\lambda_{n,N,t+1}[k] = \delta_n[k] \frac{\sum_{i=1}^N h\left(\delta_n, \lambda_{n,N,t}, U_i\right)}{\sum_{i=1}^N \left(U_i[k]\right)^2 h\left(\delta_n, \lambda_{n,N,t}, U_i\right)}$$

where for all $x = (x_1, \dots, x_d)^T \in \mathbb{R}^d$, $x[k] = x_k$. In order to try to improve the convergence, we now introduce the following gradient algorithm: for all $t \in \mathbb{N}$,

$$\lambda_{n,N,t+1} = \lambda_{n,N,t} - \eta_t \sum_{k=1}^n \lambda_{n,N,t} \left(U_i^2 h\left(\delta_n, \lambda_{n,N,t}, U_i\right) - \delta_n h\left(\delta_n, \lambda_{n,N,t}, U_i\right)\right)$$

where η_t is non-decreasing positive step sequence. Finally, we now give a sequential estimate of the eigenvalues of the variance, which consists in a Robbins-Monro algorithm [RM51] and its weighted averaged version [MP11]: for all $k \leq N-1$, one has

$$\begin{aligned}
\lambda_{n,N,k+1} &= \lambda_{n,N,k} - \gamma_{k+1} \left(\lambda_{n,N,k} U_{k+1}^2 h\left(\delta_n, \lambda_{n,N,k}, U_{k+1}\right) - \delta_n h\left(\delta_n, \lambda_{n,N,k}, U_{k+1}\right)\right) \\
\overline{\lambda}_{n,N,k+1} &= \overline{\lambda}_{n,N,k} + \frac{\log(k+1)^w}{\sum_{l=0}^k \log(l+1)^w} \left(\lambda_{n,N,k+1} - \overline{\lambda}_{n,N,k}\right),
\end{aligned}$$

with $\overline{\lambda}_{n,N,0} = \lambda_{n,N,0}$, $\gamma_k = c_\gamma k^{-\gamma}$ with $c_\gamma > 0$ and $\gamma \in (1/2, 1)$, $\omega \geq 0$.

Simulations

No outlier. We first consider the estimation of the variance and median in absence of outliers. To this aim, we consider $X \sim \mathcal{N}(0, \Sigma)$, with

$$\Sigma = \begin{bmatrix}
4 & 0.86 & 0.83 & 0.29 & 1.35 \\
0.86 & 4 & 1.4 & 0.97 & 1.79 \\
0.83 & 1.4 & 4 & 0.35 & 0.84 \\
0.29 & 0.97 & 0.35 & 4 & 0.86 \\
1.35 & 1.79 & 0.84 & 0.86 & 4
\end{bmatrix}.$$

We first focus on the accuracy of each method to estimate the variance. To do so, we consider $n = 10^5$ i.i.d copies of X and estimate the MCM with the help of the Weiszfeld's algorithm. In Figure 5.10, we show the evolution of the quadratic mean error of the estimates with respect to the sample size. More precisely, we compared the estimates obtained with fix point algorithm, with 10, 20 and 50 iterations, with the iterative gradient algorithm with 10, 20 and 50 iterations and the weighted averaged Robbins-Monro estimates (Robbins-Monro). We also compared the behavior of the methods but with fixed computation budget. We observe that all methods achieve convergence and have similar behaviors when they use samples with same sizes. Nevertheless, for fixed computation budget, the method based on the Robbins-Monro algorithm seems (without surprise) to lead to better results.

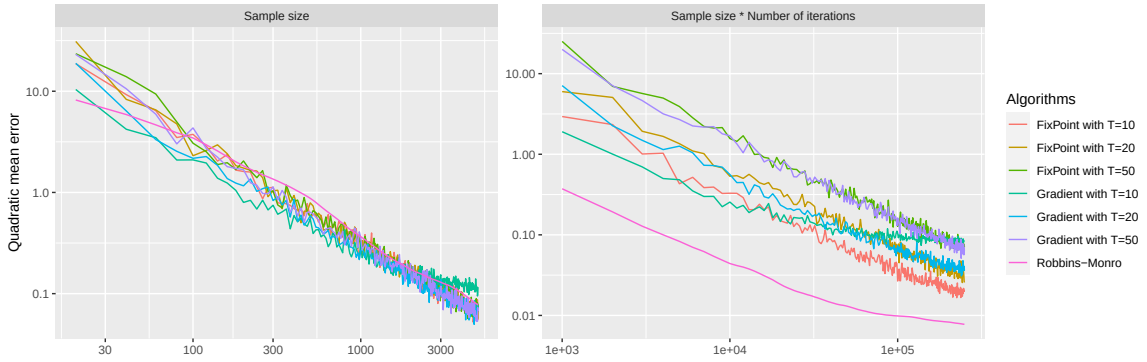


Figure 5.10 – Evolution of the quadratic mean error of the different methods with respect to the sample size (on the left) and to computation time (on the right).

With outliers. We then introduced an increasing fraction δ of outliers following three possible scenarios (a), (b) or (e) (see Section 5.5.4 for more details). We considered samples with size $n = 5000$, and estimated the MCM with the help of the Weiszfeld algorithm (indicated by (W)) or with the ASGD algorithm (indicated by (R)). We then estimated the eigenvalues of the variance with the three proposed methods and with a sample size of $N = 2000$ for the Monte Carlo method before building the variance. For iterative methods, we used $T = 50$ iterations. Remark that the different methods for estimating robustly the variance perform very well and so, even for high contamination. In addition, one can see that even if Robbins-Monro method slightly under perform the other robust alternatives, it performs well any way. Then, since Robbins-Monro procedure is less expansive in term of calculus time, and since with a fixed computational budget it can over performs other methods, it will be the chosen method in the sequel.

5.5.3 Robust Mixture Model

Mixture model

In what follows, we consider a random variable X following a mixture with K classes, i.e

$$X \sim \sum_{k=1}^{K} \pi_k^* Y_k, \tag{5.7}$$

that is $Z \sim \mathcal{M}(1, \pi^*)$ and $(X \mid Z = k) \sim Y_k$, where $\pi^* = (\pi_1^*, \dots, \pi_K^*)$ belongs to $\mathcal{S}^K := \left\{\pi, \pi_k > 0, \sum_{k=1}^K \pi_k = 1\right\}$. Furthermore, we suppose from now that Y_k satisfies the following conditions:

- Y_k admits a second order moment, and we denote by μ_k^* and Σ_k^* its mean and variance;
- the distribution of Y_k is symmetric;
- the variance of Y_k is positive;

	δ (%)	FixPoint (R)	FixPoint (W)	Gradient (R)	Gradient (W)	Robbins (R)	Robbins (W)	Variance
(a): U	0	0.32	0.24	0.34	0.31	0.45	0.36	**0.11**
	2	0.39	**0.34**	0.36	**0.34**	0.40	0.36	39.75
	3	**0.36**	0.39	0.39	**0.36**	0.43	0.38	78.20
	5	0.63	**0.51**	0.59	0.57	0.57	0.59	212.60
	9	1.35	1.36	1.29	1.21	1.28	**1.06**	682.80
	16	4.01	3.88	3.91	3.89	3.41	**3.36**	2.10^3
	28	16.65	17.56	16.21	16.13	13.78	**13.51**	7.10^3
	50	154.52	165.05	133.19	142.32	**109.12**	116.59	2.10^4
(b): T_1	0	0.31	0.29	0.32	0.34	0.38	0.40	**0.10**
	2	0.33	0.31	**0.30**	0.31	0.44	0.37	2.10^8
	3	0.36	**0.28**	0.29	0.35	0.40	0.36	2.10^7
	5	**0.35**	0.36	0.41	0.40	0.43	0.54	10^9
	9	0.49	**0.46**	0.48	0.47	0.67	0.65	7.10^9
	16	0.86	0.77	0.80	**0.76**	0.98	0.93	8.10^{13}
	28	1.74	1.76	**1.64**	1.78	2.01	1.92	5.10^{11}
	50	5.49	**5.28**	5.38	5.52	5.59	5.84	2.10^{13}
(e): T_2	0	0.29	0.28	0.37	0.29	0.46	0.33	**0.12**
	2	0.33	0.33	**0.31**	0.34	0.41	0.48	1.06
	3	**0.35**	0.40	0.42	0.38	0.63	0.41	0.59
	5	0.52	0.60	**0.48**	0.49	0.66	0.76	7.03
	9	0.86	1.02	**0.79**	0.98	1.10	1.20	6.10
	16	**1.99**	2.07	2.08	2.21	2.50	2.54	330.59
	28	5.80	5.59	**5.50**	5.88	5.92	6.20	9.10^6
	50	**14.84**	15.12	14.99	15.16	15.38	15.31	2.10^4

Table 5.4 – Multivariate Gaussian case: Mean quadratic error of the estimates of the variance for the different methods and for different contamination scenarios and fractions δ.

- the random variable Y_k is absolutely continuous with density $\phi_{\mu_k^*,\Sigma_k^*}(.)$ determined by μ_k^*, Σ_k^* and known parameters.

Remark that these conditions are satisfied for multivariate Gaussian, Student and Laplace mixtures (to name a few). The three first conditions enable to build the mean and the variance robustly with the method proposed in previous section, while the last one just ensures that the density only depends on known parameters or on parameters that can be estimated robustly. Of course, one can adapt this work for more specific cases such as Student mixtures with unknown degrees of freedom. In what follows, we will denote $\mu^* = (\mu_1^*, \ldots, \mu_K^*)$, $\Sigma^* = (\Sigma_1^*, \ldots, \Sigma_K^*)$ and $\theta^* = (\pi^*, \mu^*, \Sigma^*)$. The popular EM algorithm ([DLR77]) aims at providing the maximum likelihood

estimates by minimizing the empirical risk

$$R_n(\pi, \mu, \Sigma) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\tau_k(X_i)\left(\log(\pi_k) + \log\left(\phi_{\mu_k,\Sigma_k}(X_i)\right)\right),$$

the theoretical counterpart of which is

$$R(\pi, \mu, \Sigma) = -\mathbb{E}_{\theta^*}\left[\sum_{k=1}^{K}\tau_k(X)\left(\log(\pi_k) + \log\left(\phi_{\mu_k,\Sigma_k}(X)\right)\right)\right],$$

where $\tau_k(X) = \mathbb{P}_{\theta^*}[Z = k \mid X] = \frac{\pi_k^*\phi_{\mu_k^*,\Sigma_k^*}(X)}{\sum_{\ell=1}^{K}\pi_\ell^*\phi_{\mu_\ell^*,\Sigma_\ell^*}(X)}$ Furthermore, we know that

$$\pi^* \in \arg\min_{\pi\in\mathcal{S}^K} -\mathbb{E}_{\theta^*}\left[\sum_{k=1}^{K}\tau_k(X)\log\pi_k\right]$$

while

$$\mu^* = \arg\min_{\mu}\mathbb{E}_{\theta^*}\left[\sum_{k=1}^{K}\tau_k(X)\|X-\mu_k\|^2\right] \quad \Sigma^* = \arg\min_{\Sigma}\mathbb{E}_{\theta^*}\left[\sum_{k=1}^{K}\tau_k(X)\left\|(X-\mu^*)(X-\mu^*)^T - \Sigma_k\right\|_F^2\right].$$

Loss

Consider a mixture model as defined in (5.7) with parameter $\theta^* = (\pi^*, \mu^*, \Sigma^*)$ and let us denote by $m^* = (m_1^*, \ldots, m_K^*)$ and $V^* = (V_1^*, \ldots, V_K^*)$ the medians and MCM of the classes. Intuitively, the idea is to replace, in the usual EM algorithm, the estimates of the mean μ_k and the variance Σ_k of each class by the median m_k^* and the transformation of the MCM $\Psi_U(V_k^*)$ of each class. In this aim, let us introduce the two following functions:

$$G_2(m) = \mathbb{E}_{\theta^*}\left[\sum_{k=1}^{K}\tau_k(X)\|X-m_k\|\right] \quad G_3(m,V) = \mathbb{E}_{\theta^*}\left[\sum_{k=1}^{K}\tau_k(X)\|(X-m_k)(X-m_k)^\intercal - V_k\|_F\right].$$

The following proposition ensures that the minimizers of these functions correspond to m^* and V^*, which will be crucial to construct robust estimates of θ^*.

Proposition 5.5.1 ([GBR22]). *Consider a mixture model as defined in* (5.7) *and parametrized with* $\theta^* = (\pi^*, \mu^*, \Sigma^*)$. *Then*

$$m^* = \arg\min_{m}\mathbb{E}_{\theta^*}[G_2(m)], \qquad \textit{and} \qquad V^* = \arg\min_{V}\mathbb{E}_{\theta^*}[G_3(m^*,V)].$$

Furthermore, $m^* = \mu^*$, $\Psi_U(V^*) := (\Psi_U(V_1^*), \ldots, \Psi_U(V_K^*)) = \Sigma^*$, $\tau_k(X) = \frac{\pi_k^*\phi_{m_k^*,\Psi_U\left(V_k^*\right)}(X)}{\sum_{\ell=1}^{K}\pi_\ell^*\phi_{m_\ell^*,\Psi_U\left(V_\ell^*\right)}(X)}$, *and*

$$R_{\pi^*}(m^*, \Psi(V^*)) = \min_{\mu,\Sigma} R_{\pi^*}(\mu,\Sigma) = R_{\pi^*}(\mu^*,\Sigma^*).$$

In other words, we propose here a new parametrization of the problem where the new parameters correspond to robust indicators.

Fix-point property

The following proposition enables to see (π^*, m^*, V^*) as a fixpoint of a function g^*.

Proposition 5.5.2 ([GBR22]). *Consider a mixture model as defined in (5.7) and parametrized with $\theta^* = (\pi^*, \mu^*, \Sigma^*)$. Then, (π^*, m^*, V^*) (with π^*, m^*, V^* defined in Proposition 5.5.1) satisfy*

$$(\pi^*, m^*, V^*) = g^*(\pi^*, m^*, V^*)$$

where $g^*(\pi, m, V) = \left(g_1^*(\pi), g_{2,1}^*(m_1)), \ldots, g_{2,K}^*(m_K), g_{3,1}^*(V_1), g_{3,K}^*(V_K)\right)$ *with* $g_1(\pi) = (g_{1,1}(\pi), \ldots, g_{1,K}(\pi))$ *and*

$$g_{1,k}(\pi) := \mathbb{E}\left[\frac{\pi_k \phi\left(X, m_k^*, \Psi_U\left(V_k^*\right)\right)}{\sum_{i=1}^K \pi_i \phi\left(X, m_i^*, \Psi_U\left(V_i^*\right)\right)}\right] \quad g_{2,k}(m_k) := \frac{\mathbb{E}\left[\frac{\tau_k(X) X}{\|X - m_k\|}\right]}{\mathbb{E}\left[\frac{\tau_k(X)}{\|X - m_k\|}\right]} \quad g_{3,k}(V_k) := \frac{\mathbb{E}\left[\frac{\tau_k(X)\left(X - m_k^*\right)\left(X - m_k^*\right)^T}{\left\|\left(X - m_k^*\right)\left(X - m_k^*\right)^T - V_k\right\|_F}\right]}{\mathbb{E}\left[\frac{\tau_k(X)}{\left\|\left(X - m_k^*\right)\left(X - m_k^*\right)^T - V_k\right\|_F}\right]}$$

and $\tau_k(X) = \frac{\pi_k^* \phi_{m_k^*, \Psi_U\left(V_k^*\right)}(X)}{\sum_{\ell=1}^K \pi_\ell^* \phi_{m_\ell^*, \Psi_U\left(V_\ell^*\right)}(X)}$.

5.5.4 Simulations

Simulation design

Simulation parameters. We considered random vectors with dimension $p = 5$ and mixture models with $K = 3$ clusters with equal proportions. We defined the three mean vectors μ_1, μ_2 and μ_3, each with their all p coordinates equal to 0, 3 and -3, respectively and consider three covariance matrices Σ_1, Σ_2 and Σ_3 (see [GBR22] for more details). We then considered the Gaussian mixture distribution

$$K^{-1} \sum_{k=1}^K \mathcal{N}_p\left(\cdot; \mu_k^*, \Sigma_k^*\right).$$

Contamination scenarios. A contamination rate δ ranging from 0 (no contamination) to 50% was applied to each cluster. Namely, a same fraction δ of the observations of each cluster $k = 1, \ldots K$ was drawn with one of the five following contaminating distributions:

(*a*) uniform distribution over the hypercube: $\mathcal{U}\{[-20, 20]^p\}$;

(*b*) Student distribution with null location vector, identity scale matrix and degree of freedom 1: $\mathcal{T}(0_p, I_p, 1)$;

(*c*) Student distribution with location vector μ_k^*, identity scale matrix and degree of freedom 1: $\mathcal{T}(\mu_k^*, I_p, 1)$;

(*d*) Student distribution with null location vector, identity scale matrix and degree of freedom 2: $\mathcal{T}(0_p, I_p, 2)$;

(*e*) Student distribution with location vector μ_k^*, identity scale matrix and degree of freedom 2: $\mathcal{T}(\mu_k^*, I_p, 2)$.

The contaminating distribution has no first moments under scenarios (*b*) and (*c*), and no variance under scenarios (*d*) and (*e*). Under scenarios (*c*) and (*e*), the contaminating distribution has the same center as the corresponding cluster so the outliers can be considered as belonging to the cluster, whereas outliers arising from different clusters can not be distinguished under scenarios (*a*), (*b*) and (*d*).

Evaluation criteria. For each simulated dataset, we run the four algorithms (with fixed or selected K) and obtained estimates of the parameters μ_k and Σ_k, as well as a classification of each observation.

Classification: we used the Adjusted Rand Index (ARI) to compare the estimated classification with the simulated one.

Parameter estimates: when considering the true number of cluster K, we computed.

- the mean squared error for the center: $MSE(\mu) = K^{-1} \sum_k \|\mu_k^* - \widehat{\mu}_k\|^2 / p$,
- the mean squared error for the covariance: $MSE(\Sigma) = K^{-1} \sum_k \|\Sigma_k^* - \widehat{\Sigma}_k\|^2 / p^2$.

Model selection: when considering the case of unknown number of cluster, we considered both the BIC [Sch78] and the ICL [BCG00, MP00] criteria.

Initialization of the algorithm: Two kind of initialization are considered:

- One can initialize the algorithm considering the clustering given by the robust hierarchical clustering proposed by [GBC16], which enables to have τ^1, and one can run the end of the algorithm.
- One can choose randomly K centers from the data and take $\Sigma_k = I_d$ and $\pi_k = \frac{1}{K}$ for all k. Remark that this can be done for several random choice, and one can take the initialization leading to the best final log-likelihood.

Remark that one can choose these two kinds of initialization and take the best choice (i.e with the best log-likelihood).

Simulations

We consider here a total sample size of $n = 1500$, i.e there are $n_k = 500$ observations in each group. Since no substantial differences between the results obtained when selecting the number of clusters K with BIC and ICL has been observed, only the results obtained with BIC are presented.

The first two columns of Figure 5.11 compare the results of maximum-likelihood (GMM) inference with the proposed approach (RGMM) in terms of classification. When fixing the number of clusters to its true value $K^* = 3$, we observe a dramatic drop of the classification accuracy of GMM estimation, even for a very moderate fraction of outliers ($\delta = 2\%$), as compared to RGMM, in all scenarios. We observe that estimating the number of clusters with BIC improves the classification performances of GMM, at the price of an increase of the number of clusters. On the contrary, the RGMM approach keeps selecting the right number of clusters, even with a medium fraction of outliers ($\delta \sim 10 - 20\%$). As a consequence, model selection does not improve the classification accuracy of RGMM. Lastly, we observe that the difference between GMM and RGMM is even more obvious when outliers can each be associated with one clusters, that is under scenarios (c) and (e), as opposed to scenarios (b) and (d), respectively.

The last two columns of Figure 5.11 compare the respective accuracy of GMM and RGMM in terms of parameter estimation. The precision achieved by RGMM is several order of magnitude better than this of GMM, and, except under scenario (a), this accuracy remains the same for large contamination fractions (up to $\delta = 50\%$). Again, model selection does not improve the estimation precision of the robust approach.

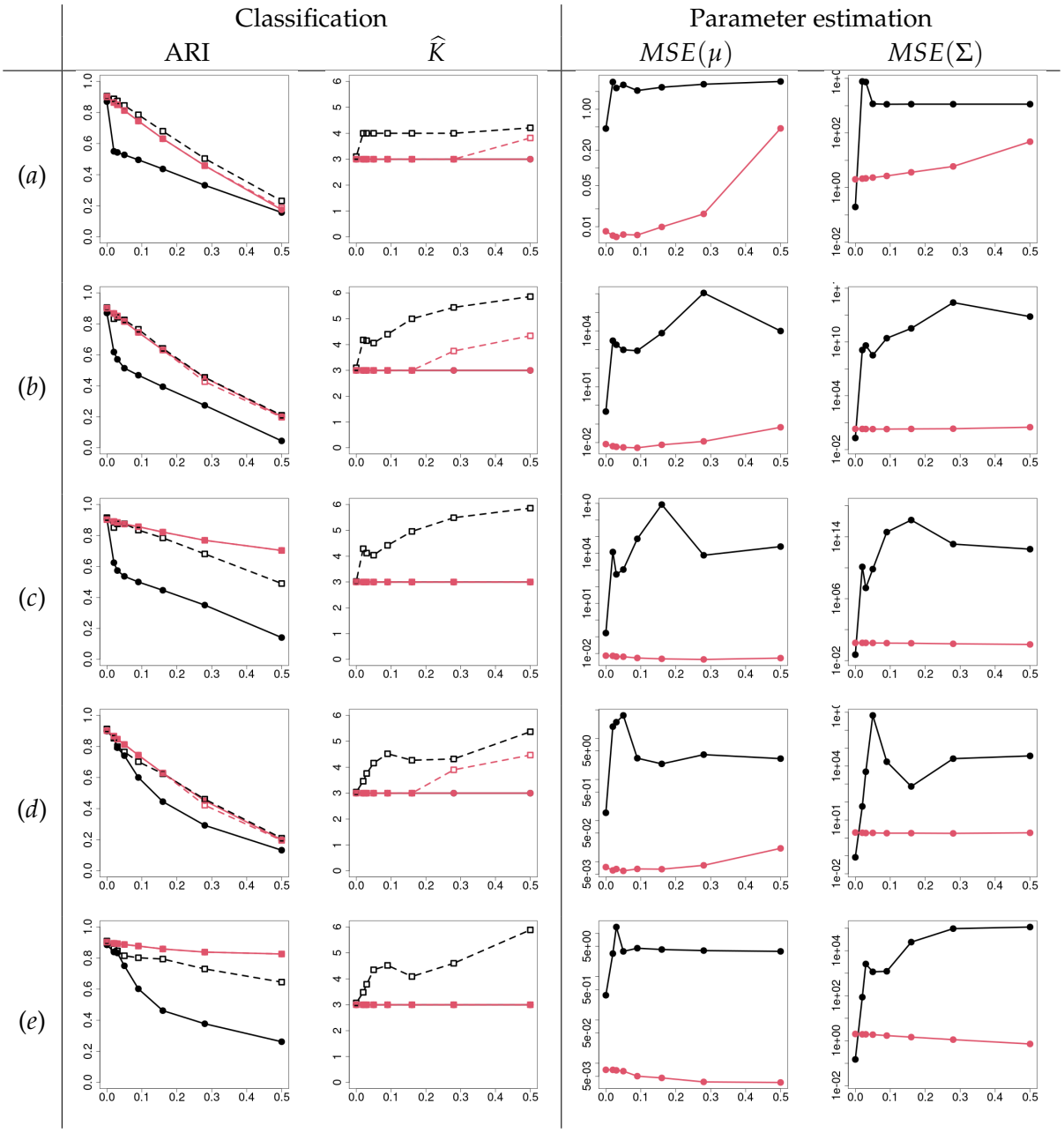


Figure 5.11 – Gaussian mixture model: classification accuracy (ARI), estimated number of clusters $\widehat{K}$, estimation error fu the mean ($MSE(\mu)$) and for the variance ($MSE(\Sigma)$) for scenarios (a) to (e), with $n_k = 500$ observation in each of the K^* clusters ($n = 1500$). Black: maximum likelihood (GMM); red: robust estimation (RGMM). Solid line (•): with true number of clusters K^*; dotted line (□): with number of clusters estimated with BIC.

Perspectives

We give here some perspectives in the continuity of the works presented in this manuscript.

Stochastic Newton algorithms

Part of my research project consists in further developing second order algorithms. In the short term this could consist in obtaining non asymptotic convergence results for adaptive methods such as Adagrad algorithms and stochastic Newton algorithms. This could enable to better understand the theoretical gain of these methods compare to the usual stochastic gradient algorithm. In the longer term, there would be many other challenges regarding online Newton methods. For example, one could propose universal methods to recursively estimate the inverse of the Hessian with a reduced computational cost. Indeed, the methods proposed so far are based on the Sherman-Morrison formula and can only be adapted to some particular cases. A simple example to understand the importance of a general method is to consider the estimation of p-means. In this case, it is not possible to use the Sherman-Morrison formula, so that, as far as we know, no online stochastic Newton algorithm has been proposed yet.
We could also take advantage of the streaming methods developed in Chapter 4 to propose stochastic Newton algorithms with only $O(nd)$ operations (n being the sample size and d the dimension) against $O(nd^2)$ currently. This would allow to have methods comparable to stochastic gradient algorithms in terms of computation time, but more adapted to ill-conditioned problems.
In addition, observe that several modification of usual stochastic gradient algorithms have been introduced. More precisely, the momentum methods have been introduced to give more weights for coordinates whose gradients point in the same direction, and so reduce oscillations [Qia99, LR20]. This has then be improved by the Nesterov acceleration method [MJ19, EBB+21]. Then it could be of particular interest to see how to adapt these procedures to stochastic Newton algorithms.

Parallelization and federated learning

Parallelization consists in distributing the data on several agents (cores, processors, servers,...) which then process these data before centralizing the information. This allows in practice to reduce the computation time but this situation is also encountered in a concrete way when data are collected by different servers which can then process them and send only the main information rather than sending all the data. For instance, [ZWLS10] deals with gradient descent for least

square type functions while [RR13] also deals with stochastic gradient by proposing a rewriting procedure where each processor can rewrite the data of an other. Finally, [BFH13, GBS20] focus on parallelized averaged stochastic gradient algorithms. In the short term, the idea would be to adapt Newton methods for parallelization, with possible communication of informations between agents. Then, an objective would be to build on this work to consider the adaptation of these methods for federated learning, i.e. for the case where the agents do not necessarily minimize the same cost functions [DFMR21, VPD+22].

Stochastic algorithms for Optimal Transport

The Kantorovich formulation of Optimal Transport problem provides a metric (Wasserstein distance) for the spaces of measures. The computation of this distance can then be seen as the minimization of a convex function. One objective would be to understand, based on many recent works (see [GCPB16, BB21] for example), how online methods such as stochastic gradient algorithms can be adapted to this problem. Furthermore, classical optimal transport approaches rely on an entropic regularization of the problem and the regularization is often "fixed" beforehand. It would be possible to think about the implementation of a regularization that would adapt itself over time, which is especially suited in an online context. In addition and in the continuity of [BBGS21] where an online stochastic Gauss-Newton based on the Sherman-Morrison formula was introduced, one could go on developing different second order methods in this context.

Robust statistics

In Chapter 5, we have seen how to build robust estimates of the variance. This approach is based on the spectral decomposition of the estimates of the Median Covariation Matrix, coupled with Monte Carlo and Robbins-Monro methods. Unfortunately, although the simulations are very hopeful, no theoretical guarantees have been given. Then, a first step should be to establish the consistency of the estimates, in the continuity of Chapters 1 and 2. A second step would be to propose a fully online alternative to the proposed method before applying it to the online detection of outliers based on the Mahalanobis distance [RD99]. Finally, one could apply the developped methodoly to the robust estimation of Gaussian means in the case where the variance is unknown (see [DM22] for more details).

Appendix A

Details results for the bounds of the quadratic mean errors

A.1 Detailed results of Chapter 1

A.1.1 Case where ∇G is not uniformly bounded

The following lemma is the detailed version of Lemma A.1.1.

Lemma A.1.1 ([GB21]). *Suppose Assumptions **(A1b')**, **(A2)**, **(A3)** and **(A4a')** are fulfilled. Then, for all $n \geq 1$,*

$$\mathbb{E}\left[\left(G\left(m_n\right) - G(m)\right)^2\right] \leq e^{-\frac{1}{4}c_\gamma a_0 n^{1-\gamma}} e^{2a_1 c_\gamma^2 \frac{2\gamma}{2\gamma-1} + 2a_2 c_\gamma^3 \frac{3\gamma}{3\gamma-1}} \left(u_0 + \sigma^2 c_\gamma^3 \frac{3\gamma}{3\gamma-1}\right) + \frac{2^{1+4\gamma}\sigma^2 c_\gamma^2}{a_0} n^{-2\gamma}$$

with $u_0 = \mathbb{E}\left[\left(G\left(m_0\right) - G(m)\right)^2\right]$, $a_0 = \frac{\lambda_0^2 \min\left\{1, r_{\lambda_0}^2\right\}}{L_{\nabla G}}$, $a_1 = \max\left\{\frac{\lambda_0^4}{4L_{\nabla G}^2}, \tilde{C}_2\left(4L_{\nabla G}+1\right)\right\}$, $a_2 = \frac{1}{2}L_{\nabla G}^2 \tilde{C}_2'$ *and* $\sigma^2 = \frac{\tilde{C}_1^2 (4L_{\nabla G}+1)^2 L_{\nabla G}}{12\lambda_0^2 \min\left\{1, r_{\lambda_0}^2\right\}}$.

The following theorem is the detailed version of Theorem 1.5.1.

Theorem A.1.1 ([GB21]). *Suppose Assumptions **(A1b')**, **(A2)**, **(A3)** and **(A4a')** are fulfilled. Then, for all $n \geq 1$,*

$$\mathbb{E}\left[\left\|m_n - m\right\|^2\right] \leq A e^{-\frac{1}{4}\lambda_{\min} c_\gamma n^{1-\gamma}} + c_1 \frac{2L_\delta^2}{\lambda_{\min}^2} e^{-\frac{1}{8}a_0 c_\gamma n^{1-\gamma}} + \frac{2^{2+8\gamma}\sigma^2 c_\gamma^2}{a_0} \frac{L_\delta^2}{\lambda_{\min}^2} n^{-2\gamma} + \frac{2^{1+\gamma}\tilde{C}_1}{\lambda_{\min}} c_\gamma n^{-\gamma},$$

with a_0, a_1, a_2, σ^2 *defined in Lemma A.1.1,* $v_0 = \mathbb{E}\left[\left\|m_0 - m\right\|^2\right]$, $L_\delta = \max\left\{\frac{2C_{\lambda_0}}{\lambda_0}, \frac{2L_{\nabla G}}{\lambda_0 r_{\lambda_0}}\right\}$, $b_1 = \frac{L_{\nabla G}}{2}\max\left\{\tilde{C}_2, \frac{\lambda_{\min}^2}{2L_{\nabla G}}\right\}$, $c_1 = e^{2a_1 c_\gamma^2 \frac{2\gamma}{2\gamma-1} + 2a_2 c_\gamma^3 \frac{3\gamma}{3\gamma-1}}\left(v_0 + \sigma^2 c_\gamma^3 \frac{3\gamma}{3\gamma-1}\right)$ *and*

$$A = e^{2b_1 c_\gamma^2 \frac{2\gamma}{2\gamma-1}}\left(v_0 + \frac{2c_\gamma^2 \tilde{C}_1}{2\gamma-1} + 2\frac{L_\delta^2}{\lambda_{\min}}\left(u_0 c_\gamma + c_1 + \frac{4c_1}{a_0\left(1-\gamma\right)} e^{-\frac{1}{4}a_0 c_\gamma} + \frac{2^{1+4\gamma}\sigma^2 c_\gamma^3}{a_0}\frac{3\gamma}{3\gamma-1}\right)\right).$$

A.1.2 Case where ∇G is bounded

The following lemma is the detailed version of Lemma 1.5.1 in the case where ∇G is bounded.

Lemma A.1.2 ([GB21]). *Suppose Assumptions **(A1b')**, **(A2)**, **(A3)** and **(A4a')** are fulfilled and that $\tilde{C}_2 = \tilde{C}_2' = 0$. Then, for all $n \geq 1$,*

$$\mathbb{E}\left[\left(G\left(m_n\right) - G(m)\right)^2\right] \leq c_{n_0'} e^{-\frac{1}{2}a_0 c_\gamma n^{1-\gamma}} + \sigma^2 M_0 c_\gamma^2 n^{-2\gamma}$$

where a_0, σ^2 are defined in Lemma A.1.1, $n_0' := \inf\{n, a_0\gamma_{n+1} \leq 1\}$, $c_{n_0'} := \sigma^2\left(e^{\frac{1}{2}a_0 c_\gamma (n_0'+1)^{1-\gamma}}\gamma_{n_0'}^3 + c_\gamma^3 \frac{3\gamma}{3\gamma-1}\right)$ and $M_0 := \max\left\{\frac{2^{4\gamma}}{a_0}, c_\gamma\right\}$.

The following theorem is the detailed version of Theorem 1.5.1 in the case where ∇G is bounded.

Theorem A.1.2 ([GB21]). *Suppose Assumptions **(A1b')**, **(A2)**, **(A3)** and **(A4a')** are fulfilled and that $\tilde{C}_2 = \tilde{C}_2' = 0$. Then, for all $n \geq 1$,*

$$\mathbb{E}\left[\|m_n - m\|^2\right] \leq A' e^{-\lambda_{\min} c_\gamma n^{1-\gamma}} + \frac{c_{n_0'} L_\delta^2}{\lambda_{\min}^2} e^{-\frac{1}{4}a_0 c_\gamma n^{1-\gamma}} + \frac{L_\delta^2 c_\gamma^2 \sigma^2}{\lambda_{\min}^2} M_0 n^{-2\gamma} + \frac{2^\gamma \tilde{C}_1 c_\gamma}{\lambda_{\min}} n^{-\gamma}$$

where a_0, σ^2 are defined in Lemma A.1.1, $c_{n_0'}, M_0$ are defined in Lemma A.1.2, $n_1' = \min\{n, \lambda_{\min}\gamma_{n+1} \leq 1\}$ and

$$A' = e^{\lambda_{\min} c_\gamma (n_1'+1)^{1-\gamma}}\left(C_1 c_\gamma^2 \frac{2\gamma}{2\gamma-1} + c_{n_0'} + c_\gamma u_0 + \frac{2c_{n_0'}}{a_0(1-\gamma)} e^{-\frac{1}{2}a_0 c_\gamma} + \sigma^2 c_\gamma^3 M_0 \frac{3\gamma}{3\gamma-1}\right).$$

A.1.3 Applications

The following corollary is the detailed version of Corollary 1.5.1.

Corollaire A.1.1. *Suppose that X admits a fourth order moment and that there are positive constants $r_{\log}, \lambda_{\log}$ such that for all $h \in \mathcal{B}\left(\theta, r_{\log}\right)$, $\lambda_{\min}\left(\nabla^2 G(h)\right) \geq \lambda_{\log}$. Then, for all $n \geq 1$,*

$$\mathbb{E}\left[\left(G\left(\theta_n\right) - G(\theta)\right)^2\right] \leq c_{\log}' e^{-\frac{1}{2}a_{\log} c_\gamma n^{1-\gamma}} + \sigma_{\log}^2 \max\left\{\frac{2^{4\gamma}}{a_{\log}}, c_\gamma\right\} c_\gamma^2 n^{-2\gamma}$$

where $a_{\log} = \frac{4\lambda_{\log}^2 \min\left\{1, r_{\log}^2\right\}}{\mathbb{E}\left[\|X\|^2\right]}$, $n_{\log}' := \inf\left\{n, a_{\log}\gamma_{n+1} \leq 1\right\}$, $\sigma_{\log}^2 = \frac{\left(\mathbb{E}\left(\|X\|^2\right]\right)^4\left(\mathbb{E}\left[\|X\|^2\right]+1\right)}{48\lambda_{\log}^2 \min\left\{1, r_{\log}^2\right\}}$ and $c_{\log}' := \sigma_{\log}^2\left(e^{\frac{1}{2}a_{\log} c_\gamma\left(n_{\log}'+1\right)^{1-\gamma}}\gamma_{n_{\log}'}^3 + c_\gamma^3 \frac{3\gamma}{3\gamma-1}\right)$. Furthermore, and

$$\mathbb{E}\left[\|\theta_n - \theta\|^2\right] \leq A_{\log} e^{-\lambda_{\log} c_\gamma n^{1-\gamma}} + \frac{c_{\log}' L_{\log}^2}{\lambda_{\log}^2} e^{-\frac{1}{4}a_{\log} c_\gamma n^{1-\gamma}} + \frac{L_{\log}^2 c_\gamma^2 \sigma_{\log}^2}{\lambda_{\log}^2} \max\left\{\frac{2^{4\gamma}}{a_{\log}}, c_\gamma\right\} n^{-2\gamma} + \frac{2^\gamma \mathbb{E}\left[\|X\|^2\right] c_\gamma}{\lambda_{\log}} n^{-\gamma}$$

where $L_{\log} = \max\left\{\frac{\mathbb{E}\left[\|X\|^3\right]}{6\sqrt{3}\lambda_{\log}}, \frac{\mathbb{E}\left[\|X\|^2\right]}{2\lambda_{\log} r_{\log}}\right\}$ *and*

$$A_{\log} = e^{\lambda_{\log} c_\gamma \left(n'_{\log}+1\right)^{1-\gamma}} \left(C_1 c_\gamma^2 \frac{2\gamma}{2\gamma-1} + c'_{\log} + c_\gamma u_0 + \frac{2c'_{\log}}{a_{\log}(1-\gamma)} e^{-\frac{1}{2}a_{\log} c_\gamma} + \sigma_{\log}^2 c_\gamma^3 \max\left\{\frac{2^{4\gamma}}{a_{\log}}, c_\gamma\right\} \frac{3\gamma}{3\gamma-1} \right).$$

Let us consider positive constants K, c_K such that $\mathbb{P}\left[\|X\| \leq K\right] \leq c_K$. Then, for all $h \in \mathcal{B}\left(m_p, 1\right)$,

$$\lambda_{\min}\left(\nabla^2 G\left(h\right)\right) \geq \frac{1}{\left(K + \left\|m_p\right\| + 1\right)^{2-p}} (p-1) c_K =: \lambda_K. \tag{A.1}$$

The following corollary is the detailed version of Corollary 1.5.2.

Corollaire A.1.2. *Suppose Assumption* ***($H_{p\text{-means}}2$)*** *holds and that X admits a $2p$-th order moment. Then, for all $n \geq 1$,*

$$\mathbb{E}\left[\left(G\left(m_{p,n}\right) - G\left(m_p\right)\right)^2\right] \leq e^{-\frac{1}{4}c_\gamma \frac{\lambda_K^2}{C_p} n^{1-\gamma}} c_{1,p} + \frac{2^{1+4\gamma}\sigma_p^2 c_\gamma^2 C_p}{\lambda_K^2} n^{-2\gamma}$$

with $u_0 = \mathbb{E}\left[\left(G\left(m_0\right) - G(m)\right)^2\right]$, $c_{1,p} := e^{\left(16C_p+4\right)c_\gamma^2 \frac{2\gamma}{2\gamma-1} + 2C_p^2 c_\gamma^3 \frac{3\gamma}{3\gamma-1}} \left(u_0 + \sigma_p^2 c_\gamma^3 \frac{3\gamma}{3\gamma-1}\right)$
and $\sigma_p^2 = \frac{\left(1+2G\left(m_p\right)\right)^2 \left(4C_p+1\right)^2 C_p}{12\lambda_K^2}$. *Furthermore, for all $n \geq 1$,*

$$\mathbb{E}\left[\left\|m_{p,n} - m_p\right\|^2\right] \leq A_p e^{-\frac{1}{4}\lambda_K c_\gamma n^{1-\gamma}} + c_{1,p} \frac{288C_p^2}{\lambda_K^4} e^{-\frac{1}{8}\frac{\lambda_K^2}{C_p} c_\gamma n^{1-\gamma}} + \frac{2^{2+8\gamma} 144 \sigma_p^2 c_\gamma^2 C_p^3}{\lambda_K^4} n^{-2\gamma} + \frac{2^{1+\gamma}\left(1 + 2G_p\left(m_p\right)\right)}{\lambda_K} c_\gamma n^{-\gamma},$$

with $v_0 = \mathbb{E}\left[\left\|m_{p,0} - m\right\|^2\right]$ *and*

$$A_p = e^{2C_p c_\gamma^2 \frac{2\gamma}{2\gamma-1}} \left(v_0 + \frac{2\left(1 + 2G\left(m_p\right)\right) c_\gamma^2}{2\gamma-1} + \frac{288C_p^2}{\lambda_K^3} \left(u_0 c_\gamma + c_1 + \frac{4c_{1,p}C_p}{\lambda_K^2(1-\gamma)} e^{-\frac{1}{4}\frac{\lambda_K^2}{C_p} c_\gamma} + \frac{2^{1+4\gamma}\sigma_p^2 c_\gamma^3 C_p}{\lambda_K^2} \frac{3\gamma}{3\gamma-1} \right) \right).$$

A.2 Detailed results of Chapter 2

A.2.1 Case where ∇G is not uniformly bounded

The following theorem is the detailed version of Theorem 2.3.1

Theorem A.2.1 ([GB21]). *Suppose Assumptions **(A1b')**, **(A2)**, **(A3)** and **(A4a')** hold. Then, for all $n \geq 1$,*

$$\lambda_{\min}\sqrt{\mathbb{E}\left[\|\overline{m}_n - m\|^2\right]} \leq \frac{\tilde{C}_1}{\sqrt{n+1}} + \frac{2^{1/2+\gamma}L_\delta\sigma c_\gamma}{\sqrt{a_0}(1-\gamma)}\frac{1}{(n+1)^\gamma} + \frac{2^{\frac{1+\gamma}{2}}5\sqrt{\tilde{C}_1}}{\sqrt{c_\gamma}\sqrt{\lambda_{\min}}}\frac{1}{(n+1)^{1-\frac{\gamma}{2}}}$$
$$+ \frac{\sqrt{\tilde{C}_2}2^{\frac{1}{4}+\gamma}\sqrt{\sigma}\sqrt{c_\gamma}}{a_0^{\frac{1}{4}}\sqrt{1-\gamma}}\frac{1}{(n+1)^{\frac{1+\gamma}{2}}} + \frac{2^{1+4\gamma}\sigma L_\delta}{\sqrt{a_0}\lambda_{\min}}\frac{\ln(n+1)}{n+1} + \frac{\sqrt{A}}{c_\gamma}\frac{e^{-\frac{1}{8}\lambda_{\min}c_\gamma n^{1-\gamma}}}{(n+1)^{1-\gamma}}$$
$$+ \frac{A_\infty + D_\infty + L_\delta B_\infty + \sqrt{\tilde{C}_2}\sqrt{B_\infty} + c_\gamma^{-\frac{1}{2}}\sqrt{v_0}}{n+1} + \frac{\sqrt{2c_1}L_\delta}{c_\gamma\lambda_{\min}}\frac{e^{-\frac{1}{16}a_0c_\gamma n^{1-\gamma}}}{(n+1)^{1-\gamma}},$$

with a_0, a_1, a_2, σ^2 defined in Lemma 1.5.1, $v_0, L_\delta, b_1, c_1, A$ defined in Theorem 1.5.1, $A_\infty := \frac{\sqrt{A}}{c_\gamma}\sum_{n\geq 0} e^{-\frac{1}{8}\lambda_{\min}c_\gamma n^{1-\gamma}}$, $B_\infty := \sum_{n\geq 0} e^{-\frac{1}{8}c_\gamma a_0 n^{1-\gamma}} e^{a_1 c_\gamma^2 \frac{2\gamma}{2\gamma-1} + a_2 c_\gamma^3 \frac{3\gamma}{3\gamma-1}}\left(\sqrt{u_0} + \sigma c_\gamma^{\frac{3}{2}}\sqrt{\frac{3\gamma}{3\gamma-1}}\right)$, and $D_\infty := \frac{\sqrt{2c_1}L_\delta}{\lambda_{\min}c_\gamma}\sum_{n\geq 0} e^{-\frac{1}{16}a_0c_\gamma n^{1-\gamma}}$.

The following theorem is the detailed version of Theorem 2.3.2

Theorem A.2.2 ([GB21]). *Suppose Assumptions **(A1b')**, **(A2)**, **(A3)**, **(A4a')** and **(A5b)** hold. Then, for all $n \geq 1$,*

$$\sqrt{\mathbb{E}\left[\|\overline{m}_n - m\|^2\right]} \leq \frac{\sqrt{Tr\left(H^{-1}\Sigma H^{-1}\right)}}{\sqrt{n+1}} + \frac{2^{1/2+\gamma}L_\delta\sigma c_\gamma}{\sqrt{a_0}(1-\gamma)}\frac{1}{(n+1)^\gamma} + \frac{2^{\frac{1+\gamma}{2}}5\sqrt{\tilde{C}_1}}{\sqrt{c_\gamma}\sqrt{\lambda_{\min}}}\frac{1}{(n+1)^{1-\frac{\gamma}{2}}}$$
$$+ \frac{\sqrt{\tilde{C}_2}2^{\frac{1}{4}+\gamma}\sqrt{\sigma}\sqrt{c_\gamma}}{a_0^{\frac{1}{4}}\sqrt{1-\gamma}}\frac{1}{(n+1)^{\frac{1+\gamma}{2}}} + \frac{2^{1+4\gamma}\sigma L_\delta}{\sqrt{a_0}\lambda_{\min}}\frac{\ln(n+1)}{n+1} + \frac{\sqrt{A}}{c_\gamma}\frac{e^{-\frac{1}{8}\lambda_{\min}c_\gamma n^{1-\gamma}}}{(n+1)^{1-\gamma}}$$
$$+ \frac{A_\infty + D_\infty + L_\delta B_\infty + \sqrt{\tilde{C}_2}\sqrt{B_\infty} + c_\gamma^{-\frac{1}{2}}\sqrt{v_0} + \sqrt{L_\Sigma}\left(\sqrt{v_0} + c_\gamma A_\infty + c_\gamma D_\infty\right) + \frac{2^{1+4\gamma}\sqrt{L_\Sigma}\sigma c_\gamma L_\delta\sqrt{2\gamma}}{\lambda_{\min}\sqrt{a_0}\sqrt{2\gamma-1}}}{n+1}$$
$$+ \frac{\sqrt{2c_1}L_\delta}{c_\gamma\lambda_{\min}}\frac{e^{-\frac{1}{16}a_0c_\gamma n^{1-\gamma}}}{(n+1)^{1-\gamma}} + \frac{2^{\frac{1+\gamma}{2}}\sqrt{\tilde{C}_1 L_\Sigma c_\gamma}}{\lambda_{\min}^{\frac{3}{2}}\sqrt{1-\gamma}}\frac{1}{(n+1)^{\frac{1+\gamma}{2}}}.$$

A.2.2 Case where ∇G is bounded

The following theorem is the detailed version of Theorem 2.3.1 in the case where ∇G is bounded.

Theorem A.2.3. *Suppose Assumptions **(A1b')**, **(A2)**, **(A3)** and **(A4a')** hold and that $\tilde{C}_2 = \tilde{C}_{2'} = 0$. Then, for all $n \geq 1$,*

$$\lambda_{\min}\sqrt{\mathbb{E}\left[\|\overline{m}_n - m\|^2\right]} \leq \frac{\sqrt{\tilde{C}_1}}{\sqrt{n+1}} + \frac{L_\delta c_\gamma\sigma\sqrt{M_0}}{1-\gamma}\frac{1}{(n+1)^\gamma} + \frac{2^{\frac{\gamma}{2}}5\sqrt{\tilde{C}_1}}{\sqrt{c_\gamma}\lambda_{\min}}\frac{1}{(n+1)^{1-\frac{\gamma}{2}}} + \frac{\sigma L_\delta\sqrt{M_0}}{\lambda_{\min}}\frac{\ln(n+1)}{n+1}$$
$$+ \frac{\sigma L_\delta\sqrt{M_0}\lambda_{\min}^{-1} + A'_\infty + L_\delta B'_\infty + D'_\infty}{n+1} + \frac{\sqrt{A'}}{c_\gamma}\frac{e^{-\frac{1}{2}\lambda_{\min}c_\gamma n^{1-\gamma}}}{(n+1)^{1-\gamma}} + \frac{\sqrt{c_{n'_0}}}{c_\gamma\lambda_{\min}}\frac{e^{-\frac{1}{8}a_0c_\gamma n^{1-\gamma}}}{(n+1)^{1-\gamma}},$$

where a_0, u_0, σ^2 are defined in Lemma A.1.1, $c_{n'_0}, M_0$ are defined in Lemma A.1.2, L_δ is defined in Theorem A.1.1, $A', c_{n'_0}$ are defined in Theorem A.1.2, $A'_\infty := \frac{\sqrt{A'}}{c_\gamma}\sum_{n\geq 0} e^{-\frac{1}{2}\lambda_{\min}c_\gamma n^{1-\gamma}}$, $B'_\infty = \left(\sqrt{c_{n'_0}} + \sqrt{u_0}\right)\sum_{n\geq 0} e^{-\frac{1}{4}a_0c_\gamma n^{1-\gamma}}$ and $D'_\infty := \frac{\sqrt{c_{n'_0}}L_\delta}{\lambda_{\min}c_\gamma}\sum_{n\geq 0} e^{-\frac{1}{8}a_0c_\gamma n^{1-\gamma}}$.

The following theorem is the detailed version of Theorem 2.3.2 in the case where ∇G is bounded/

Theorem A.2.4 ([GB21]). *Suppose Assumptions **(A1b')**, **(A2)**, **(A3)**, **(A4a')** and **(A5b)** hold and that $\tilde{C}_2 = \tilde{C}_{2'} = 0$. Then, for all $n \geq 1$,*

$$\sqrt{\mathbb{E}\left[\|\overline{m}_n - m\|^2\right]} \leq \frac{\sqrt{Tr\left(H^{-1}\Sigma H^{-1}\right)}}{\sqrt{n+1}} + \frac{L_\delta c_\gamma \sigma \sqrt{M_0}}{1-\gamma}\frac{1}{(n+1)^\gamma} + \frac{2^{\frac{\gamma}{2}}5\sqrt{\tilde{C}_1}}{\sqrt{c_\gamma}\lambda_{\min}}\frac{1}{(n+1)^{1-\frac{\gamma}{2}}} + \frac{\sigma L_\delta \sqrt{M_0}}{\lambda_{\min}}\frac{\ln(n+1)}{n+1}$$
$$+ \frac{\sqrt{A'}}{c_\gamma}\frac{e^{-\frac{1}{2}\lambda_{\min}c_\gamma n^{1-\gamma}}}{(n+1)^{1-\gamma}} + \frac{\sqrt{c_{n'_0}}}{c_\gamma\lambda_{\min}}\frac{e^{-\frac{1}{8}a_0 c_\gamma n^{1-\gamma}}}{(n+1)^{1-\gamma}} + \frac{\sqrt{L_\Sigma}2^{\frac{\gamma}{2}}\sqrt{\tilde{C}_1}}{\lambda_{\min}^{\frac{3}{2}}\sqrt{1-\gamma}}\frac{1}{(n+1)^{\frac{1+\gamma}{2}}}$$
$$+ \frac{\left(\sigma + \sqrt{L_\Sigma}c_\gamma\sqrt{\frac{2\gamma}{2\gamma-1}}\right)L_\delta\sqrt{M_0}\lambda_{\min}^{-1} + A'_\infty + L_\delta B'_\infty + D'_\infty + \sqrt{L_\Sigma}\left(\sqrt{v_0} + c_\gamma A'_\infty + c_\gamma D'_\infty\right)}{n+1}.$$

A.2.3 Applications

The following corollary is the detailed version of Corollary 2.3.1.

Corollaire A.2.1. *Suppose X admits a moment of order 4 and that there are positive constants $r_{\log}, \lambda_{\log}$ such that for all $h \in \mathcal{B}\left(\theta, r_{\log}\right), \lambda_{\min}\left(\nabla^2 G_{\log}(h)\right) \geq \lambda_{\log}$. Then, for all $n \geq 1$,*

$$\sqrt{\mathbb{E}\left[\left\|\overline{\theta}_n - \theta\right\|^2\right]} \leq \frac{\sqrt{Tr\left(H_{\log}^{-1}\right)}}{\sqrt{n+1}} + \frac{L_{\log}c_\gamma\sigma_{\log}\max\left\{\frac{2^{2\gamma}}{\sqrt{a_{\log}}}, \sqrt{c_\gamma}\right\}}{1-\gamma}\frac{1}{(n+1)^\gamma} + \frac{2^{\frac{\gamma}{2}}5\sqrt{\mathbb{E}\left[\|X\|^4\right]}}{\sqrt{c_\gamma}\lambda_{\log}}\frac{1}{(n+1)^{1-\frac{\gamma}{2}}}$$
$$+ \frac{\sigma_{\log}L_{\log}\max\left\{\frac{2^{2\gamma}}{\sqrt{a_{\log}}}, \sqrt{c_\gamma}\right\}}{\lambda_{\log}}\frac{\ln(n+1)}{n+1} + \frac{\sqrt{A_{\log}}}{c_\gamma}\frac{e^{-\frac{1}{2}\lambda_{\log}c_\gamma n^{1-\gamma}}}{(n+1)^{1-\gamma}} + \frac{\sqrt{c_{n'_{\log}}}}{c_\gamma\lambda_{\log}}\frac{e^{-\frac{1}{8}a_{\log}c_\gamma n^{1-\gamma}}}{(n+1)^{1-\gamma}}$$
$$+ \frac{\sqrt{\mathbb{E}\left[\|X\|^3\right]}2^{\frac{\gamma-1}{2}}\sqrt{\mathbb{E}\left[\|X\|^4\right]}}{\lambda_{\log}^{\frac{3}{2}}\sqrt{1-\gamma}}\frac{1}{(n+1)^{\frac{1+\gamma}{2}}} + \frac{\left(\sigma_{\log} + 2^{\frac{1}{2}}\sqrt{\mathbb{E}\left[\|X\|^3\right]}c_\gamma\sqrt{\frac{2\gamma}{2\gamma-1}}\right)L_{\log}\max\left\{\frac{2^{2\gamma}}{\sqrt{a_{\log}}}, \sqrt{c_\gamma}\right\}\lambda_{\log}^{-1}}{n+1}$$
$$+ \frac{A_\infty^{\log} + L_{\log}B_\infty^{\log} + D_\infty^{\log} + 2^{\frac{1}{2}}\sqrt{\mathbb{E}\left[\|X\|^3\right]}\left(\sqrt{v_0} + c_\gamma A_\infty^{\log} + c_\gamma D_\infty^{\log}\right)}{n+1},$$

with $L_{\log}, \sigma_{\log}, a_{\log}, \lambda_{\log}, A_{\log}, c_{n'_{\log}}$ defined in Corollary A.1.1 and $A_\infty^{\log} := \frac{\sqrt{A_{\log}}}{c_\gamma}\sum_{n\geq 0}e^{-\frac{1}{2}\lambda_{\log}c_\gamma n^{1-\gamma}}$, $B_\infty^{\log} = \left(\sqrt{c_{n'_{\log}}} + \sqrt{u_0}\right)\sum_{n\geq 0}e^{-\frac{1}{4}a_{\log}c_\gamma n^{1-\gamma}}$ and $D_\infty^{\log} := \frac{\sqrt{c_{n'_{\log}}}\max\left\{\frac{2^{4\gamma}}{a_{\log}}, c_\gamma\right\}}{\lambda_{\log}c_\gamma}\sum_{n\geq 0}e^{-\frac{1}{8}a_{\log}c_\gamma n^{1-\gamma}}$.

The following corollary is the detailed version of Corollary 2.3.2.

Corollaire A.2.2. *Suppose Assumption **($H_{p\text{-}means}2$)** holds and that X admits a $2p$-th order moment. Then,*

for all $n \geq 1$,

$$\sqrt{\mathbb{E}\left[\left\|\overline{m}_{n,p} - m_p\right\|^2\right]} \leq \frac{\sqrt{Tr\left(H^{-1}\sigma_p H^{-1}\right)}}{\sqrt{n+1}} + \frac{2^{1/2+\gamma} 6\sqrt{C_p} C'_p \sigma_p c_\gamma}{\lambda_K^2 (1-\gamma)} \frac{1}{(n+1)^\gamma} + \frac{2^{\frac{1+\gamma}{2}} 5\sqrt{1+2G\left(m_p\right)}}{\sqrt{c_\gamma}\sqrt{\lambda_K}} \frac{1}{(n+1)^{1-\frac{\gamma}{2}}}$$
$$+ \frac{2^{\frac{3}{4}+\gamma}\sqrt{\sigma_p} C_p^{\frac{1}{4}} \sqrt{c_\gamma}}{\sqrt{\lambda_K}\sqrt{1-\gamma}} \frac{1}{(n+1)^{\frac{1+\gamma}{2}}} + \frac{2^{1+4\gamma} 6\sigma_p \sqrt{C_p} C'_p}{\lambda_K^3} \frac{\ln(n+1)}{n+1} + \frac{\sqrt{A_p}}{c_\gamma} \frac{e^{-\frac{1}{8}\lambda_K c_\gamma n^{1-\gamma}}}{(n+1)^{1-\gamma}}$$
$$+ \frac{A_\infty^{(p)} + D_\infty^{(p)} + \frac{6C_p B_\infty^{(p)}}{\lambda_K} + \sqrt{2}\sqrt{B_\infty^{(p)}} + c_\gamma^{-\frac{1}{2}}\sqrt{v_0} + \sqrt{L_\Sigma}\left(\sqrt{v_0} + c_\gamma A_\infty^{(p)} + c_\gamma D_\infty^{(p)}\right)}{n+1}$$
$$+ \frac{\frac{2^{1+4\gamma} 6\sqrt{L_\Sigma}\sigma_p c_\gamma \sqrt{C_p} C'_p \sqrt{2\gamma}}{\lambda_K^3 \sqrt{2\gamma-1}}}{n+1} + \frac{\sqrt{12c_1} C_p}{c_\gamma \lambda_K^2} \frac{e^{-\frac{1}{16}\frac{\lambda_K^2}{C_p} c_\gamma n^{1-\gamma}}}{(n+1)^{1-\gamma}} + \frac{2^{\frac{1+\gamma}{2}}\sqrt{C_1 L_\Sigma c_\gamma}}{\lambda_K^{\frac{3}{2}}\sqrt{1-\gamma}} \frac{1}{(n+1)^{\frac{1+\gamma}{2}}}.$$

where $u_0, c_{1,p}, \sigma_p^2$ *and* A_p *are are defined in Corollary 1.5.2 and* λ_K *is given by (1.10). Furthermore,* $A_\infty^{(p)} := \frac{\sqrt{A_p}}{c_\gamma}\sum_{n\geq 0} e^{-\frac{1}{8}\lambda_K c_\gamma n^{1-\gamma}}$, $B_\infty^{(p)} := \sum_{n\geq 0} e^{-\frac{1}{8}c_\gamma \frac{\lambda_K^2}{C_p} n^{1-\gamma}} e^{\left(8C_p+2\right)c_\gamma^2 \frac{2\gamma}{2\gamma-1} + 2C_p^2 c_\gamma^3 \frac{3\gamma}{3\gamma-1}} \left(\sqrt{u_0} + \sigma_p c_\gamma^{\frac{3}{2}}\sqrt{\frac{3\gamma}{3\gamma-1}}\right)$, *and* $D_\infty^{(p)} := \frac{6\sqrt{2c_{1,p}} C'_p}{\lambda_K^2 c_\gamma}\sum_{n\geq 0} e^{-\frac{1}{16}\frac{\lambda_K^2}{C_p} c_\gamma n^{1-\gamma}}$

A.3 Detailed results of Section 5.2.3

Let us now focus on the rate of convergence in quadratic mean of the estimates. More precisely, the aim is to apply Theorem A.1.2. In this aim, let us recall two important results. First, under assumptions **($\mathbf{A}_{\text{median}}$1a)** and **($\mathbf{A}_{\text{median}}$2)**, it was proven in [CCZ13] that there is K large enough such that

$$c_{\min} := \inf_{\|v\|=1} \mathbb{V}\left[\langle v, X\rangle \mathbf{1}_{\|K\|}\right] > 0.$$

Then, one has for all $h \in \mathcal{B}\left(m_{1/2}, 1\right)$ [CCZ13]

$$\lambda_{\min}\left(\nabla^2 G_{1/2}(h)\right) = \frac{1}{(K+1)^3} c_{\min}.$$

In addition, it was proven in [GB16a] that under Assumption **($\mathbf{A}_{\text{median}}$1b)**,

$$\left\|\nabla G_{1/2}(h) - \nabla^2 G_{1/2}\left(m_{1/2}\right)\left(h - m_{1/2}\right)\right\| \leq C_{\text{med}}^2 \left\|h - m_{1/2}\right\|^2.$$

Then, Assumption **(A4a')** is fulfilled. In addition, up to take the max, we will suppose that $C_{\text{med}} \geq 1$. Let us now denote $\lambda_{1/2} := \lambda_{\min}\left(\nabla^2 G_{1/2}\left(m_{1/2}\right)\right)$ and apply Theorem A.1.2 to obtain the following rate of convergence for the stochastic gradient estimates of the median, which is the detailed version of Theorem 5.2.3.

Theorem A.3.1. *Suppose Assumption* ***($A_{median}1$)*** *and Assumption* ***($A_{median}2$)*** *hold. Then, for all* $n \geq 1$,

$$\mathbb{E}\left[\|m_{1/2,n} - m_{1/2}\|^2\right] \leq A' e^{-\lambda_{1/2} c_\gamma n^{1-\gamma}} + \frac{36 c_{n_0'} C_{med}^4 (K+1)^6}{\lambda_{1/2}^2 c_{\min}^2} e^{-\frac{1}{4} a_0 c_\gamma n^{1-\gamma}} + \frac{2^\gamma c_\gamma}{\lambda_{1/2}} n^{-\gamma} + \frac{C_{med}^5 (K+1)^{12} c_\gamma^2 (4C_{med}+1)^2}{3\lambda_{1/2}^2 c_{\min}^4} M_{1/2} n^{-2\gamma}$$

where $n_0' = \inf\left\{n, \frac{(K+1^3}{c_{\min} C_{med}} \gamma_{n+1} \leq 1\right\}$, $n_1' = \min\{n, \lambda_{\min}\gamma_{n+1} \leq 1\}$, $M_{1/2} = \max\left\{c_\gamma, \frac{2^{4\gamma} C_{med}(K+1)^3}{c_{\min}}\right\}$ *and*

$$c_{n_0'} = \frac{(4C_{med}+1)^2 C_{med}(K+1)^6}{c_{\min}^2}\left(e^{\frac{1}{2}\frac{c_{\min}}{(K+1)^3 C_{med}} c_\gamma (n_0'+1)^{1-\gamma}} \gamma_{n_0'}^3 + c_\gamma^3 \frac{3\gamma}{3\gamma-1}\right)$$

$$A' = e^{\lambda_{1/2} c_\gamma (n_1'+1)^{1-\gamma}}\left(\frac{2c_\gamma^2 \gamma}{2\gamma-1} + c_{n_0'} + c_\gamma u_0 + \frac{2c_{n_0'} c_{\min} C_{med}}{(K+1)^3(1-\gamma)} e^{-\frac{(K+1)^3}{2c_{\min} C_{med}} c_\gamma} + \frac{3(4C_{med}+1)^2 C_{med}(K+1)^6 M_{1/2} c_\gamma^3}{c_{\min}^2(3\gamma-1)}\right).$$

We now give the detailed version of Theorem 5.2.4

Theorem A.3.2. *Suppose Assumption* ***($A_{median}1$)*** *and Assumption* ***($A_{median}2$)*** *hold. Then, for all* $n \geq 1$,

$$\begin{aligned}\sqrt{\mathbb{E}\left[\|\overline{m}_{1/2,n} - m_{1/2}\|^2\right]} \leq{}& \frac{\sqrt{Tr\left(H^{-1}\Sigma_{1/2}H^{-1}\right)}}{\sqrt{n+1}} + \frac{6C_{med}^3(K+1)^6(4C_{med}+1)c_\gamma\sqrt{M_{1/2}}}{(1-\gamma)c_{\min}^2(n+1)^\gamma} + \frac{2^{\frac{\gamma}{2}}5}{\sqrt{c_\gamma}\lambda_{1/2}(n+1)^{1-\frac{\gamma}{2}}} \\ &+ \frac{12C_{med}^3(K+1)^6(4C_{med}+1)\sqrt{M_{1/2}}}{\lambda_{1/2}c_{\min}^2}\frac{\ln(n+1)}{n+1} + \frac{\sqrt{A'}}{c_\gamma}\frac{e^{-\frac{1}{2}\lambda_{1/2}c_\gamma n^{1-\gamma}}}{(n+1)^{1-\gamma}} \\ &+ \frac{\sqrt{c_{n_0'}}}{c_\gamma\lambda_{1/2}}\frac{e^{-\frac{c_{\min}c_\gamma}{8(K+1)^3C_{med}}n^{1-\gamma}}}{(n+1)^{1-\gamma}} + \frac{\sqrt{6C_{med}}2^{\frac{\gamma}{2}}}{\lambda_{1/2}^{\frac{3}{2}}\sqrt{1-\gamma}}\frac{1}{(n+1)^{\frac{1+\gamma}{2}}} + \frac{6^{3/2}C_{med}^{3/2}(K+1)^3 c_\gamma\sqrt{2\gamma}}{c_{\min}\sqrt{2\gamma-1}(n+1)} \\ &+ \frac{A_\infty' + +D_\infty' + \sqrt{6C_{med}}\left(\sqrt{v_0} + c_\gamma A_\infty' + c_\gamma D_\infty'\right)}{n+1} + \frac{6C_{med}^2(K+1)^3 B_\infty'}{c_{\min}(n+1)},\end{aligned}$$

with $\Sigma_{1/2} = \mathbb{E}\left[\frac{(X-m_{1/2})(X-m_{1/2})^T}{\|X-m_{1/2}\|^2}\right]$ *and*

$$A_\infty' = \frac{\sqrt{A'}}{c_\gamma}\sum_{n\geq 0} e^{-\frac{\lambda_{1/2}c_\gamma}{2}n^{1-\gamma}}, \qquad B_\infty' = \left(\sqrt{c_{n_0'}'} + \sqrt{u_0}\right)\sum_{n\geq 0} e^{-\frac{c_{\min}}{4(K+1)^3 C_{med}}n^{1-\gamma}},$$

$$D_\infty' = \frac{6\sqrt{c_{n_0'}}C_{med}^2(K+1)^3}{\lambda_{1/2}c_\gamma c_{\min}} e^{-\frac{c_{\min}}{8(K+1)^3C_{med}}n^{1-\gamma}}.$$

List of Figures

Bibliography

[AHA+20] Motasem Alfarra, Slavomir Hanzely, Alyazeed Albasyoni, Bernard Ghanem, and Peter Richtarik. Adaptive learning of the optimal mini-batch size of sgd. *arXiv preprint arXiv:2005.01097*, 2020.

[AM09] Sylvain Arlot and Pascal Massart. Data-driven calibration of penalties for least-squares regression. *Journal of Machine learning research*, 10(2), 2009.

[Bac14] Francis Bach. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *The Journal of Machine Learning Research*, 15(1):595–627, 2014.

[BB21] Bernard Bercu and Jérémie Bigot. Asymptotic distribution and convergence rates of stochastic algorithms for entropic optimal transportation between probability measures. *The Annals of Statistics*, 49(2):968–987, 2021.

[BBGS21] Bernard Bercu, Jérémie Bigot, Sébastien Gadat, and Emilia Siviero. A stochastic gauss-newton algorithm for regularized semi-discrete optimal transport. *arXiv preprint arXiv:2107.05291*, 2021.

[BBM+11] Vincent Brault, Jean-Patrick Baudry, Cathy Maugis, Bertrand Michel, and Maintainer Vincent Brault. Package 'capushe', 2011.

[BBT+11] Juan Lucas Bali, Graciela Boente, David E Tyler, Jane-Ling Wang, et al. Robust functional principal components: A projection-pursuit approach. *The Annals of Statistics*, 39(6):2852–2882, 2011.

[BCG00] C. Biernacki, G. Celeux, and G. Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans. Pattern Anal. Machine Intel.*, 22(7):719–25, 2000.

[BCG20] Bernard Bercu, Manon Costa, and Sébastien Gadat. Stochastic approximation algorithms for superquantiles estimation. *arXiv preprint arXiv:2007.14659*, 2020.

[BCN18] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.

[BD09] Peter J Brockwell and Richard A Davis. *Time series: theory and methods*. Springer Science & Business Media, 2009.

[Ber06] Pavel Berkhin. A survey of clustering data mining techniques. In *Grouping multidimensional data*, pages 25–71. Springer, 2006.

[BFH13] Pascal Bianchi, Gersende Fort, and Walid Hachem. Performance of a distributed stochastic approximation algorithm. *IEEE Transactions on Information Theory*, 59(11):7405–7418, 2013.

[BFHJ11] Pascal Bianchi, Gersende Fort, Walid Hachem, and Jérémie Jakubowicz. Convergence of a distributed parameter estimator for sensor networks with local averaging of the estimates. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3764–3767. IEEE, 2011.

[BGB20] Claire Boyer and Antoine Godichon-Baggioni. On the asymptotic rate of convergence of stochastic newton algorithms and their weighted averaged versions. *arXiv preprint arXiv:2011.09706*, 2020.

[BGBP19] Bernard Bercu, Antoine Godichon-Baggioni, and Bruno Portier. An efficient stochastic newton algorithm for parameter estimation in logistic regressions. *arXiv preprint arXiv:1904.07908*, 2019.

[BHNS16] Richard H Byrd, Samantha L Hansen, Jorge Nocedal, and Yoram Singer. A stochastic quasi-newton method for large-scale optimization. *SIAM Journal on Optimization*, 26(2):1008–1031, 2016.

[BJRL15] George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[BM07] Lucien Birgé and Pascal Massart. Minimal penalties for gaussian model selection. *Probability theory and related fields*, 138(1):33–73, 2007.

[BM13] Francis Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate O (1/n). In *Advances in Neural Information Processing Systems*, pages 773–781, 2013.

[BMM12] Jean-Patrick Baudry, Cathy Maugis, and Bertrand Michel. Slope heuristics: overview and implementation. *Statistics and Computing*, 22(2):455–470, 2012.

[BR93] J.D Banfield and A.E Raftery. Model-based gaussian and non-gaussian clustering. *Biometrics*, pages 803–821, 1993.

[BV04] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[CAGM97] Juan Antonio Cuesta-Albertos, Alfonso Gordaliza, and Carlos Matrán. Trimmed k-means: An attempt to robustify quantizers. *The Annals of Statistics*, 25(2):553–576, 1997.

[CCGB15] Hervé Cardot, Peggy Cénac, and Antoine Godichon-Baggioni. Online estimation of the geometric median in Hilbert spaces: non asymptotic confidence balls. Technical report, arXiv:1501.06930, 2015.

[CCM12] Hervé Cardot, Peggy Cénac, and Jean-Marie Monnez. A fast and recursive algorithm for clustering large datasets with k-medians. *Computational Statistics and Data Analysis*, 56:1434–1449, 2012.

[CCZ13] Hervé Cardot, Peggy Cénac, and Pierre-André Zitt. Efficient and fast estimation of the geometric median in Hilbert spaces with an averaged stochastic gradient algorithm. *Bernoulli*, 19(1):18–43, 2013.

[CFF07] Antonio Cuevas, Manuel Febrero, and Ricardo Fraiman. Robust estimation and classification for functional data via projection-based depth notions. *Computational Statistics*, 22(3):481–496, 2007.

[CFO07] Christophe Croux, Peter Filzmoser, and Maria Rosario Oliveira. Algorithms for projection-pursuit robust principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 87:218–225, 2007.

[CG20] Manon Costa and Sébastien Gadat. Non asymptotic controls on a recursive superquantile approximation. 2020.

[CGB15] Hervé Cardot and Antoine Godichon-Baggioni. Fast estimation of the median covariation matrix with application to online robust principal components analysis. *TEST*, pages 1–20, 2015.

[CGBP20] Peggy Cénac, Antoine Godichon-Baggioni, and Bruno Portier. An efficient averaged stochastic gauss-newtwon algorithm for estimating parameters of non linear regressions models. *arXiv preprint arXiv:2006.12920*, 2020.

[CGP12] Christophe Croux, Irène Gijbels, and Ilaria Prosdocimi. Robust estimation of mean and dispersion functions in extended generalized additive models. *Biometrics*, 68(1):31–44, 2012.

[CH16] P. Coretto and C. Hennig. Robust improper maximum likelihood: tuning, computation, and a comparison with other methods for robust gaussian clustering. *Journal of the American Statistical Association*, 111(516):1648–1659, 2016.

[CH17] P. Coretto and C. Hennig. Consistency, breakdown robustness, and algorithms for robust improper maximum likelihood clustering. *Journal of Machine Learning Research*, 18(142):1–39, 2017.

[Cha92] Probal Chaudhuri. Multivariate location estimation using extension of *R*-estimates through *U*-statistics type approach. *Ann. Statist.*, 20:897–916, 1992.

[Cha96] Probal Chaudhuri. On a geometric notion of quantiles for multivariate data. *J. Amer. Statist. Assoc.*, 91(434):862–872, 1996.

[CNS17] Kobi Cohen, Angelia Nedić, and R Srikant. On projected stochastic gradient descent algorithm with weighted averaging for least squares regression. *IEEE Transactions on Automatic Control*, 62(11):5974–5981, 2017.

[CRG05] Christophe Croux and Anne Ruiz-Gazen. High breakdown estimators for principal components: the projection-pursuit approach revisited. *J. Multivariate Anal.*, 95:206–226, 2005.

[DFMR21] Aymeric Dieuleveut, Gersende Fort, Eric Moulines, and Geneviève Robin. Federated expectation maximization with heterogeneity mitigation and variance reduction. *arXiv preprint arXiv:2111.02083*, 2021.

[DGK81] Susan J. Devlin, Ramanathan Gnanadesikan, and Jon R. Kettenring. Robust estimation of dispersion matrices and principal components. *J. Amer. Statist. Assoc.*, 76:354–362, 1981.

[DGN14] Olivier Devolder, François Glineur, and Yurii Nesterov. First-order methods of smooth convex optimization with inexact oracle. *Mathematical Programming*, 146(1):37–75, 2014.

[DHS11] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

[DLR77] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39:1–38, 1977.

[DM22] Arnak S Dalalyan and Arshak Minasyan. All-in-one robust estimator of the gaussian mean. *The Annals of Statistics*, 50(2):1193–1219, 2022.

[Duf97] Marie Duflo. *Random iterative models*, volume 34 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin, 1997. Translated from the 1990 French original by Stephen S. Wilson and revised by the author.

[EBB+21] Mathieu Even, Raphaël Berthier, Francis Bach, Nicolas Flammarion, Pierre Gaillard, Hadrien Hendrikx, Laurent Massoulié, and Adrien Taylor. A continuized view on nesterov acceleration for stochastic gradient descent and randomized gossip. *arXiv preprint arXiv:2106.07644*, 2021.

[Fis11] Aurélie Fischer. On the number of groups in clustering. *Statistics & Probability Letters*, 81(12):1771–1781, 2011.

[FM01] R. Fraiman and G. Muniz. Trimmed means for functional data. *TEST*, 10:419–440, 2001.

[For65] Edward W Forgy. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *biometrics*, 21:768–769, 1965.

[FP20] A. Farcomeni and A. Punzo. Robust model-based clustering with mild and gross outliers. *Test*, 29(4):989–1007, 2020.

[GB16a] Antoine Godichon-Baggioni. 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, 2016.

[GB16b] Antoine Godichon-Baggioni. Lp and almost sure rates of convergence of averaged stochastic gradient algorithms with applications to online robust estimation. *arXiv preprint arXiv:1609.05479*, 2016.

[GB17] Antoine Godichon-Baggioni. Online estimation of the asymptotic variance for averaged stochastic gradient algorithms. *arXiv preprint arXiv:1702.00931*, 2017.

[GB21] Antoine Godichon-Baggioni. Convergence in quadratic mean of averaged stochastic gradient algorithms without strong convexity nor bounded gradient. *arXiv preprint arXiv:2107.12058*, 2021.

[GBC16] Marek Gagolewski, Maciej Bartoszuk, and Anna Cena. Genie: A new, fast, and outlier-resistant hierarchical clustering algorithm. *Information Sciences*, 363:8–23, 2016.

[GBPL22] Antoine Godichon-Baggioni, Bruno Portier, and Wei Lu. Recursive ridge regression using second-order stochastic algorithms. 2022.

[GBR22] Antoine Godichon-Baggioni and Stéphane Robin. A robust model-based clustering based on the geometric median and the median covariation matrix. *arXiv preprint arXiv:2211.08131*, 2022.

[GBS20] Antoine Godichon-Baggioni and Sofiane Saadane. On the rates of convergence of parallelized averaged stochastic gradient algorithms. *Statistics*, 54(3):618–635, 2020.

[GBS22] Antoine Godichon-Baggioni and Sobihan Surendran. A penalized criterion for selecting the number of clusters for k-medians. *arXiv preprint arXiv:2209.03597*, 2022.

[GBWW21] Antoine Godichon-Baggioni, Nicklas Werge, and Olivier Wintenberger. Non-asymptotic analysis of stochastic approximation algorithms for streaming data. *arXiv preprint arXiv:2109.07117*, 2021.

[GBWW22] Antoine Godichon-Baggioni, Nicklas Werge, and Olivier Wintenberger. Learning from time-dependent streaming data with online stochastic algorithms. *arXiv preprint arXiv:2205.12549*, 2022.

[GCPB16] Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport. *Advances in neural information processing systems*, 29, 2016.

[GEG99] Luis Angel Garcia-Escudero and Alfonso Gordaliza. Robustness properties of k means and trimmed k means. *Journal of the American Statistical Association*, 94(447):956–969, 1999.

[GEGMMI08] L.A García-Escudero, A. Gordaliza, C. Matrán, and A. Mayo-Iscar. A general trimming approach to robust cluster analysis. *The Annals of Statistics*, 36(3):1324–1345, 2008.

[Ger08] Daniel Gervini. Robust functional estimation using the median and spherical principal components. *Biometrika*, 95(3):587–600, 2008.

[GH10] I. Gijbels and M. Hubert. Robust and nonparametric statistical methods. *Comprehensive Chemometrics*, 1:189–211, 01 2010.

[GLQ+19] Robert Mansel Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtárik. Sgd: General analysis and improved rates. In *International Conference on Machine Learning*, pages 5200–5209. PMLR, 2019.

[GMYZ21] J.D Gonzalez, R. Maronna, V.J Yohai, and R.H Zamar. Robust model-based clustering. Technical Report 2102.06851, arXiv, 2021.

[GP17] Sébastien Gadat and Fabien Panloup. Optimal non-asymptotic bound of the ruppert-polyak averaging without strong convexity. *arXiv preprint arXiv:1709.03342*, 2017.

[GYZ19] J.D Gonzalez, V.J Yohai, and R.H Zamar. Robust clustering using tau-scales. (1906.08198), 2019.

[HA85] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

[Hal48] J. B. S. Haldane. Note on the median of a multivariate distribution. *Biometrika*, 35(3-4):414–417, 1948.

[Ham20] James Douglas Hamilton. *Time series analysis*. Princeton university press, 2020.

[HPV14] Marc Hallin, Davy Paindaveine, and Thomas Verdebout. Efficient r-estimation of principal and common principal components. *Journal of the American Statistical Association*, 109(507):1071–1083, 2014.

[HR09] Peter Huber and Elvezio Ronchetti. *Robust Statistics*. John Wiley and Sons, second edition, 2009.

[HRVA08] Mia Hubert, Peter Rousseeuw, and Stefan Van Aelst. High-breakdown robust multivariate methods. *Statistical Science*, 13:92–119, 2008.

[HU07] Robert Hyndman and Shahid Ullah. Robust forecasting of mortality and fertility rates: A functional data approach. *Computational Statistics and Data Analysis*, 51:4942–4956, 2007.

[Jak88] Adam Jakubowski. Tightness criteria for random measures with application to the principle of conditioning in Hilbert spaces. *Probab. Math. Statist.*, 9(1):95–114, 1988.

[JD88] Anil K Jain and Richard C Dubes. *Algorithms for clustering data*. Prentice-Hall, Inc., 1988.

[JMF99] Anil K Jain, M Narasimha Murty, and Patrick J Flynn. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323, 1999.

[JN+14] Anatoli Juditsky, Yuri Nesterov, et al. Deterministic and stochastic primal-dual subgradient algorithms for uniformly convex minimization. *Stochastic Systems*, 4(1):44–80, 2014.

[Jol02] Ian Jolliffe. *Principal Components Analysis*. Springer Verlag, New York, second edition, 2002.

[KB14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[Kem87] Johannes Kemperman. The median of a finite measure on a Banach space. In *Statistical data analysis based on the L_1-norm and related methods (Neuchâtel, 1987)*, pages 217–230. North-Holland, Amsterdam, 1987.

[KLRT15] Jakub Konečnỳ, Jie Liu, Peter Richtárik, and Martin Takáč. Mini-batch semi-stochastic gradient descent in the proximal setting. *IEEE Journal of Selected Topics in Signal Processing*, 10(2):242–255, 2015.

[KP12] David Kraus and Victor M. Panaretos. Dispersion operators and resistant second-order functional data analysis. *Biometrika*, 99:813–832, 2012.

[KR09] Leonard Kaufman and Peter J Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, 2009.

[KY03] Harold J Kushner and George Yin. *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer Science & Business Media, 2003.

[Lin00] Tamás Linder. On the training distortion of vector quantizers. *IEEE Transactions on Information Theory*, 46(4):1617–1623, 2000.

[LMS+99] N. Locantore, J.S. Marron, D.G Simpson, N. Tripoli, J.T. Zhang, and K.L Cohen. Robust principal components for functional data. *Test*, 8:1–73, 1999.

[LP20] Rémi Leluc and François Portier. Towards asymptotic optimality with conditioned stochastic gradient descent. *arXiv preprint arXiv:2006.02745*, 2020.

[LR20] Nicolas Loizou and Peter Richtárik. Momentum and stochastic momentum for stochastic gradient, newton, proximal point and subspace descent methods. *Computational Optimization and Applications*, 77(3):653–710, 2020.

[LVLLJ21] Nicolas Loizou, Sharan Vaswani, Issam Hadj Laradji, and Simon Lacoste-Julien. Stochastic polyak step-size for sgd: An adaptive learning rate for fast convergence. In *International Conference on Artificial Intelligence and Statistics*, pages 1306–1314. PMLR, 2021.

[Mac67] J MacQueen. Classification and analysis of multivariate observations. In *5th Berkeley Symp. Math. Statist. Probability*, pages 281–297, 1967.

[Mir96] Boris Mirkin. *Mathematical classification and clustering*, volume 11. Springer Science & Business Media, 1996.

[MJ19] Michael Muehlebach and Michael Jordan. A dynamical systems perspective on nesterov acceleration. In *International Conference on Machine Learning*, pages 4656–4662. PMLR, 2019.

[MMY06] Ricardo A. Maronna, R. Douglas Martin, and Victor J. Yohai. *Robust statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester, 2006. Theory and methods.

[MNO+10] Jyrki Möttönen, Klaus Nordhausen, Hannu Oja, et al. Asymptotic theory of the spatial median. *Nonparametrics and robustness in modern statistical inference and time series analysis: a Festschrift in honor of Professor Jana Jurecková*, 7:182–193, 2010.

[MP00] G. McLahan and D. Peel. *Finite Mixture Models*. Wiley, 2000.

[MP11] Abdelkader Mokkadem and Mariane Pelletier. A generalization of the averaging procedure: The use of two-time-scale algorithms. *SIAM Journal on Control and Optimization*, 49(4):1523–1543, 2011.

[N+18] Yurii Nesterov et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

[NJLS09] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

[NND+18] Lam Nguyen, Phuong Ha Nguyen, Marten Dijk, Peter Richtárik, Katya Scheinberg, and Martin Takác. Sgd and hogwild! convergence without the bounded gradients assumption. In *International Conference on Machine Learning*, pages 3750–3758. PMLR, 2018.

[NNG19] Ion Necoara, Yu Nesterov, and Francois Glineur. Linear convergence of first order methods for non-strongly convex optimization. *Mathematical Programming*, 175(1):69–107, 2019.

[NT15] Klaus Nordhausen and Sara Taskinen. *Modern nonparametric, robust and multivariate methods*. Springer, 2015.

[PD19] Kumar Kshitij Patel and Aymeric Dieuleveut. Communication trade-offs for synchronized distributed sgd with large step size. *arXiv preprint arXiv:1904.11325*, 2019.

[Pel98] Mariane Pelletier. On the almost sure asymptotic behaviour of stochastic algorithms. *Stochastic processes and their applications*, 78(2):217–244, 1998.

[Pel00] Mariane Pelletier. Asymptotic almost sure efficiency of averaged stochastic algorithms. *SIAM J. Control Optim.*, 39(1):49–72, 2000.

[Pin94] Iosif Pinelis. Optimum bounds for the distributions of martingales in Banach spaces. *The Annals of Probability*, 22:1679–1706, 1994.

[PJ92] Boris Polyak and Anatoli Juditsky. Acceleration of stochastic approximation. *SIAM J. Control and Optimization*, 30:838–855, 1992.

[PM00] D. Peel and G.J McLachlan. Robust mixture modelling using the t distribution. *Statistics and computing*, 10(4):339–348, 2000.

[Qia99] Ning Qian. On the momentum term in gradient descent learning algorithms. *Neural networks*, 12(1):145–151, 1999.

[Ran71] William M Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.

[RD99] Peter J Rousseeuw and Katrien Van Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223, 1999.

[RL05] Peter J Rousseeuw and Annick M Leroy. *Robust regression and outlier detection*. John wiley & sons, 2005.

[RM51] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[RR13] Benjamin Recht and Christopher Ré. Parallel stochastic gradient algorithms for large-scale matrix completion. *Mathematical Programming Computation*, 5(2):201–226, 2013.

[RS05] James O. Ramsay and Bernard W. Silverman. *Functional Data Analysis*. Springer, New York, second edition, 2005.

[RS19] D. Rossell and M.FJ Steel. Continuous mixtures with skewness and heavy tails. In *Handbook of Mixture Analysis*, pages 219–237. Chapman and Hall/CRC, 2019.

[Rud16] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

[Rup88] David Ruppert. Efficient estimations from a slowly convergent robbins-monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988.

[RvD99] Peter Rousseeuw and Katrien van Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41:212–223, 1999.

[Sch78] G Schwarz. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

[Sma90] Christopher G. Small. A survey of multidimensional medians. *International Statistical Review / Revue Internationale de Statistique*, 58(3):263–277, 1990.

[Spa80] Helmuth Spath. *Cluster analysis algorithms for data reduction and classification of objects*. Ellis Horwood Chichester, 1980.

[SPIM15] S. Subedi, A. Punzo, S. Ingrassia, and P.D McNicholas. Cluster-weighted *t* t-factor analyzers for robust model-based clustering and dimension reduction. *Statistical Methods & Applications*, 24(4):623–649, 2015.

[TKO12] Sara Taskinen, Inge Koch, and Hannu Oja. Robustifying principal components analysis with spatial sign vectors. *Statist. and Probability Letters*, 82:765–774, 2012.

[TWH01] Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.

[Ver06] Thomas Verdebout. Common principal components. *Encyclopedia of Environmetrics*, 2006.

[VPD+22] Maxime Vono, Vincent Plassier, Alain Durmus, Aymeric Dieuleveut, and Eric Moulines. Qlsd: Quantised langevin stochastic dynamics for bayesian federated learning. In *International Conference on Artificial Intelligence and Statistics*, pages 6459–6500. PMLR, 2022.

[VZ00] Yehuda Vardi and Cun-Hui Zhang. The multivariate L_1-median and associated data depth. *Proc. Natl. Acad. Sci. USA*, 97(4):1423–1426, 2000.

[Wan15] T-I Wang, W-Land Lin. Robust model-based clustering via mixtures of skew-t distributions with missing information. *Advances in Data Analysis and Classification*, 9(4):423–445, 2015.

[Wei37] Endre Weiszfeld. On the point for which the sum of the distances to n given points is minimum. *Tohoku Math. J.*, 43:355–386, 1937.

[Zei12] Matthew D Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

[ZWLS10] Martin Zinkevich, Markus Weimer, Lihong Li, and Alex Smola. Parallelized stochastic gradient descent. *Advances in neural information processing systems*, 23, 2010.