Algorithmes stochastiques pour la statistique robuste en grande dimension

Antoine Godichon Godichon-Baggioni

Université de Bourgogne , U.F.R Sciences et techniques
Institut de Mathématiques de Bourgogne
Ecole doctorale Carnot-Pasteur

THÈSE

pour l'obtention du grade de

Docteur de l'Université de Bourgogne

en Mathématiques

présentée et soutenue publiquement par

Antoine Godichon-Baggioni

le 17 Juin 2016

Algorithmes stochastiques pour la statistique robuste en grande dimension

Directeurs de thèse : **Hervé Cardot, Peggy Cénac**

Jury composé de

Hervé Cardot	Université de Bourgogne	Directeur
Peggy Cénac	Université de Bourgogne	Co-encadrante
Antonio Cuevas	Universidad Autónoma de Madrid	Rapporteur
Clément Dombry	Université de Franche-Comté	Examinateur
Anatoli Juditsky	Université Joseph Fourier	Rapporteur
Mariane Pelletier	Université de Versailles	Examinatrice
Bruno Portier	INSA de Rouen	Examinateur
Anne Ruiz-Gazen	Université de Toulouse 1	Examinatrice

A Nénette et Toussaint

Remerciements

Je tiens tout d'abord à remercier mes directeurs de thèse, Hervé Cardot et Peggy Cénac. Merci à Peggy pour toute l'aide apportée durant toutes ces années, et de m'avoir appuyé pour que je puisse faire cette thèse. Un grand merci à Hervé d'avoir accepté de diriger cette thèse. Enfin, merci à tous les deux d'avoir été aussi patients, d'avoir parfaitement su orienter mon travail, et de toujours avoir vos portes ouvertes pour mes (très) nombreuses questions.

Je tiens aussi à remercier Bruno Portier, non seulement pour avoir accepté de travailler avec moi, ce qui m'a ouvert de nombreuses perspectives de recherche, mais aussi pour ses nombreux conseils avisés. J'apprécie énormément nos fréquentes discussions grâce auxquelles je peux prendre davantage de recul sur mon travail.

Antonio Cuevas et Anatoli Juditsky m'ont fait l'honneur d'accepter de rapporter ma thèse et de faire parti du Jury, je les remercie pour cela ainsi que pour leurs précieuses remarques, notamment bibliographiques, qui m'ont permis d'améliorer mon manuscrit et qui m'ouvrent de nombreuses perspectives. Je remercie aussi Clément Dombry, Mariane Pelletier, Bruno Portier et Anne Ruiz-Gazen, qui me font l'honneur de faire partie de mon Jury.

Faire une thèse est une chance, mais la faire dans de si bonnes conditions est véritablement un luxe, et cela je le dois à tous les personnels du laboratoire, Anissa, Aziz, Caro, Francis, Magalie, Nadia, Sébastien, Véronique, mais aussi aux enseignants-chercheurs de l'IMB. Merci aussi à tous les doctorants et à l'association des Doctorants en Mathématiques de Dijon pour tous ces bons moments, les exposés, les pots... Enfin merci à tous mes co-bureaux, Adriana, Armand, Bachar, Bruno, Charlie, Jessie, Michael, Rebecca et Simone d'avoir toléré la musique, les ronchonnements.... Merci aussi à mes co-brasseurs, Charlie, Jérémy et Rémi.

Si je suis en train d'écrire ces remerciements, c'est aussi parce que des personnes m'ont aidé à des moments cruciaux, ou m'ont prodigué de très bon conseils. Donc merci à Charlie pour ces précieux avis, et merci à tous ceux qui m'ont gracieusement aidé à préparer l'Agrégation (Peggy Cénac, Shizan Fang, Lucy Moser, Emmanuel Wagner, ...) et m'ont aidé, par la même occasion, à retrouver des bases solides dont j'ai eu besoin tout au long de ma thèse.

Je tiens à remercier toute ma famille, mes parents pour leurs bons (et mauvais) choix grâce auxquels j'en suis là aujourd'hui, mon frère et ma soeur, pour m'avoir supporté toutes ses années, mes grand-mères, mes oncles et tantes, mes cousins... Merci à mon père et à Mamo Paula pour les corrections orthographiques, merci à ma mère et ma tante Louise pour leurs conseils pendant ces trois dernières années.

Grazie a tutti i Zalaninchi pour votre présence et votre soutien. Merci à tous les membres de l'Association Sportive Sociale et Culturelle de Zalana (Antho G, Antho T, Audrey, Clément, Juju, Ludo, Mika, Roland, Romane, Mareva,...), merci à tous mes cousins. Bref, merci à tous et a prestu !

Il y a des moments dans la vie qui sont déterminants, et je remercie Rebecca de m'avoir aidé à saisir les occasions qui se sont offertes à moi ainsi que de m'avoir encouragé à faire cette thèse.

Enfin, mes derniers remerciements vont à mon oncle Toussaint et à ma tante Nénette, à qui je dédie cette thèse. Malgré la distance, vous avez toujours été présents dans ma vie et soucieux de mon avenir. Malheureusement, je ne pourrais pas partager cette dernière ligne droite avec vous. J'ose espérer que vous auriez été fiers de moi.

Résumé

Cette thèse porte sur l'étude d'algorithmes stochastiques en grande dimension ainsi qu'à leur application en statistique robuste. Dans la suite, l'expression grande dimension pourra aussi bien signifier que la taille des échantillons étudiés est grande ou encore que les variables considérées sont à valeurs dans des espaces de grande dimension (pas nécessairement finie). Afin d'analyser ce type de données, il peut être avantageux de considérer des algorithmes qui soient rapides, qui ne nécessitent pas de stocker toutes les données, et qui permettent de mettre à jour facilement les estimations. Dans de grandes masses de données en grande dimension, la détection automatique de points atypiques est souvent délicate. Cependant, ces points, même s'ils sont peu nombreux, peuvent fortement perturber des indicateurs simples tels que la moyenne ou la covariance. On va se concentrer sur des estimateurs robustes, qui ne sont pas trop sensibles aux données atypiques.

Dans une première partie, on s'intéresse à l'estimation récursive de la médiane géométrique, un indicateur de position robuste, et qui peut donc être préférée à la moyenne lorsqu'une partie des données étudiées est contaminée. Pour cela, on introduit un algorithme de Robbins-Monro ainsi que sa version moyennée, avant de construire des boules de confiance non asymptotiques et d'exhiber leurs vitesses de convergence L^p et presque sûre.

La deuxième partie traite de l'estimation de la "Median Covariation Matrix" (MCM), qui est un indicateur de dispersion robuste lié à la médiane, et qui, si la variable étudiée suit une loi symétrique, a les mêmes sous-espaces propres que la matrice de variance-covariance. Ces dernières propriétés rendent l'étude de la MCM particulièrement intéressante pour l'Analyse en Composantes Principales Robuste. On va donc introduire un algorithme itératif qui permet d'estimer simultanément la médiane géométrique et la MCM ainsi que les q principaux vecteurs propres de cette dernière. On donne, dans un premier temps, la forte consistance des estimateurs de la MCM avant d'exhiber les vitesses de convergence en moyenne quadratique.

Dans une troisième partie, en s'inspirant du travail effectué sur les estimateurs de la médiane et de la "Median Covariation Matrix", on exhibe les vitesses de convergence presque

sûre et L^p des algorithmes de gradient stochastiques et de leur version moyennée dans des espaces de Hilbert, avec des hypothèses moins restrictives que celles présentes dans la littérature. On présente alors deux applications en statistique robuste : estimation de quantiles géométriques et régression logistique robuste.

Dans la dernière partie, on cherche à ajuster une sphère sur un nuage de points répartis autour d'une sphère complète où tronquée. Plus précisément, on considère une variable aléatoire ayant une distribution sphérique tronquée, et on cherche à estimer son centre ainsi que son rayon. Pour ce faire, on introduit un algorithme de gradient stochastique projeté et son moyenné. Sous des hypothèses raisonnables, on établit leurs vitesses de convergence en moyenne quadratique ainsi que la normalité asymptotique de l'algorithme moyenné.

Mots-clés : Grande Dimension, Données Fonctionnelles, Algorithmes Stochastiques, Algorithmes Récursifs, Algorithmes de Gradient Stochastiques, Moyennisation, Statistique Robuste, Médiane Géométrique.

Abstract

This thesis focus on stochastic algorithms in high dimension as well as their application in robust statistics. In what follows, the expression high dimension may be used when the the size of the studied sample is large or when the variables we consider take values in high dimensional spaces (not necessarily finite). In order to analyze these kind of data, it can be interesting to consider algorithms which are fast, which do not need to store all the data, and which allow to update easily the estimates. In large sample of high dimensional data, outliers detection is often complicated. Nevertheless, these outliers, even if they are not many, can strongly disturb simple indicators like the mean and the covariance. We will focus on robust estimates, which are not too much sensitive to outliers.

In a first part, we are interested in the recursive estimation of the geometric median, which is a robust indicator of location which can so be preferred to the mean when a part of the studied data is contaminated. For this purpose, we introduce a Robbins-Monro algorithm as well as its averaged version, before building non asymptotic confidence balls for these estimates, and exhibiting their L^p and almost sure rates of convergence.

In a second part, we focus on the estimation of the Median Covariation Matrix (MCM), which is a robust dispersion indicator linked to the geometric median. Furthermore, if the studied variable has a symmetric law, this indicator has the same eigenvectors as the covariance matrix. This last property represent a real interest to study the MCM, especially for Robust Principal Component Analysis. We so introduce a recursive algorithm which enables us to estimate simultaneously the geometric median, the MCM, and its q main eigenvectors. We give, in a first time, the strong consistency of the estimators of the MCM, before exhibiting their rates of convergence in quadratic mean.

In a third part, in the light of the work on the estimates of the median and of the Median Covariation Matrix, we exhibit the almost sure and L^p rates of convergence of averaged stochastic gradient algorithms in Hilbert spaces, with less restrictive assumptions than in the literature. Then, two applications in robust statistics are given : estimation of the geometric quantiles and application in robust logistic regression.

In the last part, we aim to fit a sphere on a noisy points cloud spread around a complete or truncated sphere. More precisely, we consider a random variable with a truncated spherical distribution, and we want to estimate its center as well as its radius. In this aim, we introduce a projected stochastic gradient algorithm and its averaged version. We establish the strong consistency of these estimators as well as their rates of convergence in quadratic mean. Finally, the asymptotic normality of the averaged algorithm is given.

Keywords High Dimension, Functional Data, Stochastic Algorithms, Recursive Algorithms, Stochastic Gradient Algorithms, Averaging, Robust Statistics, Geometric Median.

Table des matières

Introduction

Présentation

Mon travail de thèse se situe à la croisée de deux thématiques assez distinctes, l'optimisation stochastique et la statistique robuste. Il est de plus en plus fréquent en statistique d'avoir à traiter de gros échantillons de variables à valeurs dans des espaces de grande dimension. Dans ce contexte, il est important de repenser les problèmes d'estimation. La construction d'estimateurs repose bien souvent sur la résolution d'un problème d'optimisation et il existe, dans la littérature, de nombreux algorithmes qui sont très efficaces en "petite dimension" mais qui peuvent rencontrer des difficultés pour traiter de gros échantillons à valeurs dans des espaces de grande dimension. Ils nécessitent souvent de stocker en mémoire toutes les données, ce qui peut devenir très compliqué, voir impossible, dans ce contexte de données massives. De plus, les procédures d'estimation classiques, basées par exemple sur des algorithmes de point fixe ou de Newton ([BV04]), ne peuvent pas toujours être mises à jour facilement, et ne permettent donc pas de traiter les données qui arrivent de manière séquentielle. Enfin, l'acquisition de données massives peut s'accompagner d'une contamination de celles-ci, ce qui peut dégrader de manière significative la qualité des estimations. Je développe dans cette thèse des algorithmes qui permettent d'estimer rapidement et en ligne des indicateurs robustes tels que la médiane géométrique ([Hal48], [Kem87]) ou la "Median Covariation Matrix", et j'étudie leurs propriétés mathématiques.

Au Chapitre 3, on fera une synthèse des principaux résultats de cette thèse. L'objectif de cette partie est de permettre au lecteur d'avoir accès aux principaux éléments (hypothèses, résultats,...) de cette thèse, et ce, sans avoir à connaître tous les détails.

On commence, dans le Chapitre 1, par rappeler quelques résultats usuels d'optimisation déterministe avant de s'intéresser à leurs versions stochastiques. Plus précisément, afin de minimiser une fonction on introduit les algorithmes de Robbins-Monro ([RM51]) avant de donner quelques résultats classiques sur leur forte consistance ([Duf97]). On énonce ensuite des résultats de la littérature sur leur vitesse de convergence presque sûre ainsi que sur leur normalité asymptotique ([Pel98], [Pel00]). Enfin, on donne des résultats non asymptotiques tels que des majorations de l'erreur quadratique moyenne ([BM13]). De plus, comme il est

souvent compliqué en pratique d'obtenir la vitesse de convergence paramétrique ($O\left(\frac{1}{n}\right)$) ou bien une variance optimale pour les algorithmes de gradient stochastiques, on introduit leur version moyennée ([PJ92]). De la même façon que pour l'algorithme de Robbins-Monro, on donne des résultats de la littérature sur la vitesse de convergence de ces algorithmes.

L'objectif du Chapitre 2 est de fournir une introduction simple à la statistique robuste ([HR09], [MMY06]). Dans un premier temps, on donne un exemple qui illustre l'intérêt de considérer des indicateurs ou des estimateurs robustes. On introduit une classe importante d'estimateurs appelés M-estimateurs (ces estimateurs consistent à minimiser une fonction, et peuvent être une alternative aux algorithmes de gradient introduits au Chapitre 1 pour traiter des données de taille raisonnable) avant de donner leur comportement asymptotique et des méthodes de construction. De plus, on définit des indicateurs de robustesse comme la fonction d'influence, le biais asymptotique maximum et le point de rupture. Chaque définition et critère abordé est illustré à travers les cas de la moyenne et de la médiane géométrique.

Le Chapitre 4 se focalise sur la médiane géométrique, qui est très utilisée en statistique du fait de sa robustesse. On rappelle une méthode de construction d'algorithmes récursifs qui permettent de l'estimer ([CCZ13]) : un algorithme de type Robbins-Monro et sa version moyennée. Des boules de confiance non-asymptotiques sont déterminées, ainsi que leurs vitesses de convergence en moyenne quadratique. Pour ce dernier résultat, la preuve s'appuie sur une nouvelle technique de démonstration qui peut être considérée comme la pierre angulaire de ce travail. Cette technique consiste à majorer simultanément, à l'aide d'une récurrence, la vitesse de convergence en moyenne quadratique et la vitesse L^4. Les preuves des lemmes techniques sont données dans l'Annexe A.

Le Chapitre 5 établit des majorations des erreurs L^p de l'algorithme d'estimation de la médiane géométrique de type Robbins-Monro ainsi que de sa version moyennée. Ces majorations permettent ensuite d'établir les vitesses de convergence presque sûre des algorithmes. Ce chapitre est particulièrement important pour l'obtention de la bonne vitesse de convergence des estimateurs de la "Median Covariation Matrix" (voir [KP12] et le Chapitre 6). Les preuves des lemmes techniques sont données dans l'Annexe B.

Au Chapitre 6, on introduit la notion de "Median Covariation Matrix" (MCM), qui est un indicateur de dispersion robuste multivarié lié à la médiane. Un des principaux intérêts de cet opérateur robuste est que si la loi de la variable étudiée est symétrique, il a les mêmes espaces propres que la matrice de variance-covariance (voir [KP12]). On présente ensuite des algorithmes récursifs permettant d'estimer simultanément la médiane géométrique et la MCM. La construction de ces algorithmes consiste à reprendre les algorithmes d'estimation de la médiane, et de les "injecter" dans un algorithme de gradient stochastique et sa version

moyennée pour estimer la MCM. La forte consistance de ces algorithmes est établie avant de donner les vitesses de convergence en moyenne quadratique. Finalement, on présente une application à l'ACP robuste en ligne avec un algorithme itératif permettant d'estimer les principaux vecteurs propres de la MCM. Une étude sur des données réelles, l'audience TV mesurée à un pas de temps fin pour un échantillon de plus de 5000 individus, confirme l'intérêt de cette méthode. Les preuves techniques sont détaillées dans l'Annexe C, à laquelle on ajoute également des compléments sur l'algorithme de Weiszfeld.

Le Chapitre 7 propose un cadre plus général permettant d'obtenir les vitesses de convergence des algorithmes de type Robbins-Monro et de leur version moyennée dans des espaces de Hilbert. Notons que les hypothèses sont proches de celles introduites par [Pel98] pour les algorithmes en dimension finie, mais les preuves ne dépendent pas de la dimension de l'espace. De plus, sous ces hypothèses, on obtient les vitesses de convergence L^p des algorithmes, et ce, sans avoir à introduire d'hypothèse de forte convexité globale (voir [BM13]) ou sans avoir à supposer que le gradient de la fonction que l'on veut minimiser est borné (voir [Bac14]). Les preuves techniques sont données dans l'Annexe D.

Le Chapitre 8 traite de l'ajustement d'une sphère sur un nuage de points 3D répartis autour d'une sphère complète ou tronquée. On considère des variables aléatoires suivant des lois elliptiques tronquées de la forme $X = \mu + rWU_\Omega$, où W est une variable aléatoire à valeurs dans $\mathbb{R}_+$ et U_Ω suit une loi uniforme sur une partie Ω de la sphère unité dans $\mathbb{R}^d$. On estime les paramètre μ (le centre) et r (le rayon) avec un algorithme de type Robbins-Monro projeté et sa version moyennée. Les vitesses de convergence en moyenne quadratique ainsi que la normalité asymptotique de l'estimateur moyenné sont établies. Finalement, on montre l'efficacité de cette méthode à travers des simulations. Remarquons que cette partie représente une ouverture sur les algorithmes de gradient stochastiques projetés. Des propriétés de convexité ainsi que les preuves sont données dans l'Annexe E.

Ce travail a donné lieu à plusieurs publications, acceptées ou soumises.

Chapitre 4

Hervé Cardot, Peggy Cénac, Antoine Godichon-Baggioni (2016). Online estimation of the geometric median in Hilbert spaces : non asymptotic confidence balls, Accepté dans *Annals of statistics*, http ://arxiv.org/abs/1501.06930.

Chapitre 5

Antoine Godichon-Baggioni (2015). Estimating the geometric median in Hilbert spaces with stochastic gradient algorithms : Lp and almost sure rates of convergence, Publié dans *Journal of Multivariate Analysis*, doi :10.1016/j.jmva.2015.09.013.

Chapitre 6

Hervé Cardot, Antoine Godichon-Baggioni (2015). Fast Estimation of the Median Covariation Matrix with Application to Online Robust Principal Components Analysis, soumis, http ://arxiv.org/abs/1504.02852.

Chapitre 8

Antoine Godichon-Baggioni, Bruno Portier (2016). An averaged projected Robbins-Monro algorithm for estimating the parameters of a truncated spherical distribution, soumis.

Chapitre 1

Quelques résultats sur les algorithmes stochastiques

L'objectif de cette partie est de donner une version stochastique de quelques algorithmes déterministes usuels (voir [NNY94] et [BV04] parmi d'autres) de recherche de zéro d'une fonction. Plus précisément, on cherche à estimer la solution d'un problème de la forme

$$\Phi(h) := \mathbb{E}\left[\phi(X,h)\right] = 0, \tag{1.1}$$

où X est une variable aléatoire à valeurs dans un espace $\mathcal{X}$ et $\phi : \mathcal{X} \times \mathbb{R}^d \longrightarrow \mathbb{R}^d$. Sous certaines conditions, cela peut revenir à minimiser la fonction

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

où $\nabla_h g(x,h) = \phi(x,h)$ et $\nabla G(h) = \Phi(h)$. Pour estimer cette solution, on s'intéresse aux méthodes de gradient stochastiques.

Ces algorithmes représentent un réel intérêt pour l'estimation à partir de gros échantillons à valeurs dans des espaces de grande dimension. En effet, de manière générale, ils ne demandent pas trop d'efforts de calculs, ils ne nécessitent pas de stocker en mémoire toutes les données, et ils peuvent être facilement mis à jour, ce qui représente un réel intérêt lorsque les données arrivent de manière séquentielle.

1.1 Algorithmes déterministes

1.1.1 Recherche des zéros d'une fonction

On introduit maintenant des algorithmes de gradient déterministes. Soit $\Phi : \mathbb{R}^d \longrightarrow \mathbb{R}^d$ une fonction continue, on cherche m tel que $\Phi(m) = 0$. Pour estimer m, on considère l'algorithme récursif défini pour tout $n \geq 1$ par

$$m_{n+1} = m_n - \gamma_n \Phi(m_n), \tag{1.2}$$

avec $m_1 \in \mathbb{R}^d$. De plus, la suite de pas $(\gamma_n)_{n\geq 1}$ est réelle, positive, et vérifie les conditions usuelles suivantes

$$\sum_{n\geq 1} \gamma_n = +\infty, \qquad \sum_{n\geq 1} \gamma_n^2 < +\infty. \tag{1.3}$$

On donne maintenant deux premières propositions naïves qui permettent de comprendre facilement pourquoi cet algorithme converge bien vers m sous de bonnes conditions. Remarquons qu'il est possible, dans le cas déterministe, d'obtenir une vitesse de convergence exponentielle, mais nous ne donnerons pas ces résultats. L'objectif ici est plus de présenter des résultats déterministes analogues aux résultats classiques dans le cas stochastique.

Proposition 1.1.1. *On suppose qu'il existe un point m qui annule la fonction Φ et*

— *qu'il existe une constante strictement positive C telle que pour tout $h \in \mathbb{R}^d$,*

$$\|\Phi(h)\| \leq C \|h - m\|,$$

— *et qu'il existe une constante strictement positive c telle que pour tout $h \in \mathbb{R}^d$,*

$$\langle \Phi(h), h - m \rangle \geq c \|h - m\|^2.$$

Alors, m est l'unique zéro de la fonction Φ, et

$$\lim_{n\to\infty} \|m_n - m\| = 0$$

Démonstration. On montre d'abord que le point m est l'unique zéro de la fonction Φ. Soit $m' \in \mathbb{R}^d$ tel que $m' \neq m$. Alors,

$$\langle \Phi(m'), m' - m \rangle \geq c \left\|m' - m\right\|^2 > 0,$$

et en particulier $\Phi(m') \neq 0$. Le point m est donc bien l'unique zéro de la fonction Φ. On montre maintenant la convergence de l'algorithme. Pour cela, grâce aux hypothèses,

$$\begin{aligned}\|m_{n+1} - m\|^2 &= \|m_n - m\|^2 - 2\gamma_n \langle \Phi(m_n), m_n - m\rangle + \gamma_n^2 \|\Phi(m_n)\|^2 \\ &\leq \|m_n - m\|^2 - 2c\gamma_n \|m_n - m\|^2 + \gamma_n^2 C^2 \|m_n - m\|^2 \\ &\leq \left(1 - 2c\gamma_n + C^2\gamma_n^2\right) \|m_n - m\|^2 .\end{aligned}$$

Comme la suite $(\gamma_n)_{n\geq 1}$ converge vers 0, il existe un rang n_0 tel que pour tout $n \geq n_0$, on ait $1 - 2c\gamma_n + C^2\gamma_n^2 \leq 1 - c\gamma_n < 1$. Alors,

$$\begin{aligned}\|m_{n+1} - m\|^2 &\leq \prod_{k=1}^{n} \left(1 - 2c\gamma_k + C^2\gamma_k^2\right) \|m_1 - m\|^2 \\ &\leq \left(\prod_{k=n_0}^{n} (1 - c\gamma_k)\right) \left(\prod_{k=1}^{n_0-1} \left(1 - 2c\gamma_k + C^2\gamma_k^2\right)\right) \|m_1 - m\|^2\end{aligned}$$

De plus, comme

$$\begin{aligned}\left(\prod_{k=n_0}^{n} (1 - c\gamma_k)\right) &= \exp\left(\sum_{k=n_0}^{n} \log(1 - c\gamma_k)\right) \\ &\leq \exp\left(-c \sum_{k=n_0}^{n} \gamma_k\right),\end{aligned}$$

on obtient

$$\|m_{n+1} - m\|^2 \leq \exp\left(-c \sum_{k=n_0}^{n} \gamma_k\right) \left(\prod_{k=1}^{n_0-1} \left(1 - 2c\gamma_k + C^2\gamma_k^2\right)\right) \|m_1 - m\|^2 .$$

Enfin, comme $\sum_{n\geq 1} \gamma_n = +\infty$, on obtient le résultat. □

Les hypothèses précédentes sont très restrictives, mais elles permettent de se faire une première idée du fonctionnement de ces algorithmes. On énonce maintenant un résultat avec des hypothèses et un résultat plus usuels.

Proposition 1.1.2. *On suppose qu'il existe un point m qui annule la fonction Φ et*

— *qu'il existe une constante positive C telle que pour tout $h \in \mathbb{R}^d$,*

$$\|\Phi(h)\| \leq C\left(1 + \|h - m\|\right),$$

— *et que pour tout* $h \in \mathbb{R}^d$ *tel que* $h \neq m$,

$$\langle \Phi(h), h - m \rangle > 0.$$

Alors, m *est l'unique zéro de* Φ *et la suite* $(m_n)_{n\geq 1}$ *définie par (1.2) vérifie*

$$\lim_{n\to\infty} \|m_n - m\| = 0.$$

Démonstration. On obtient l'unicité de m de manière analogue au théorème précédent. De plus,

$$\begin{aligned}
\|m_{n+1} - m_n\|^2 &= \|m_n - m\|^2 - 2\gamma_n \langle \Phi(m_n), m_n - m\rangle + \gamma_n^2 \|\Phi(m_n)\|^2 \\
&\leq \|m_n - m\|^2 - 2\gamma_n \langle \Phi(m_n), m_n - m\rangle + \gamma_n^2 C^2 (1 + \|m_n - m\|)^2 \\
&\leq \left(1 + 2C^2\gamma_n^2\right) \|m_n - m\|^2 - 2\gamma_n \langle \Phi(m_n), m_n - m\rangle + 2C^2\gamma_n^2.
\end{aligned}$$

Comme $\sum_{n\geq 1} \gamma_n^2 < +\infty$ et comme $\langle \Phi(m_n), m_n - m\rangle > 0$, remarquons que la suite $\left(\|m_n - m\|^2\right)_{n\geq 1}$ est bornée. De plus, la suite $(V_n)_{n\geq 1}$ définie par

$$V_n := \frac{1}{\prod_{k=1}^n \left(1 + 2C^2\gamma_k^2\right)} \left(\|m_n - m\|^2 + 2\sum_{k=1}^n \gamma_k \langle \Phi(m_k), m_k - m\rangle + 2C^2 \sum_{k=n}^\infty \gamma_k^2 \right),$$

est positive. Enfin cette suite est décroissante. En effet,

$$\begin{aligned}
V_{n+1} &= \frac{1}{\prod_{k=1}^{n+1} \left(1 + 2C^2\gamma_k^2\right)} \left(\|m_{n+1} - m\|^2 + 2\sum_{k=1}^{n+1} \gamma_k \langle \Phi(m_k), m_k - m\rangle + 2C^2 \sum_{k=n+1}^\infty \gamma_k^2 \right) \\
&\leq \frac{1}{\prod_{k=1}^{n+1} \left(1 + 2C^2\gamma_k^2\right)} \left(\|m_n - m\|^2 + 2\sum_{k=1}^{n} \gamma_k \langle \Phi(m_k), m_k - m\rangle + 2C^2 \sum_{k=n}^\infty \gamma_k^2 \right) \\
&= \frac{1}{1 + 2C^2\gamma_{n+1}^2} V_n \\
&\leq V_n.
\end{aligned}$$

La suite $(V_n)_{n\geq 1}$ est donc convergente. En particulier, comme la suite $\left(\prod_{k=1}^n \left(1 + 2C^2\gamma_k^2\right)\right)_{n\geq 1}$ est convergente, la suite $\left(\|m_n - m\|^2\right)_{n\geq 1}$ converge vers une limite finie l et

$$\sum_{n\geq 1} \gamma_n \langle \Phi(m_n), m_n - m\rangle < +\infty.$$

Comme $\sum_{n\geq 1}\gamma_n = +\infty$, on a alors

$$\liminf_n \langle \Phi(m_n), m_n - m\rangle = 0.$$

Rappelons que la suite $(\|m_n - m\|)_{n\geq 1}$ converge vers une limite finie l, et donc, la suite $(m_n)_{n\geq 1}$ est bornée, et on peut extraire (car $\mathbb{R}^d$ est de dimension finie) une sous suite $(m_{n_k})_{k\geq 1}$ convergeant vers une limite m'. Par continuité de Φ, on a alors

$$\langle \Phi(m'), m' - m\rangle = 0,$$

et donc, par hypothèse, si $m' \neq m$, on aurait

$$0 = \langle \Phi(m'), m' - m\rangle > 0.$$

Par conséquent, $m' = m$ et on obtient finalement

$$l = \lim_{n\to\infty} \|m_n - m\|^2 = \lim_{k\to\infty} \|m_{n_k} - m\|^2 = 0.$$

□

Cette preuve permet de mettre en lumière le rôle des hypothèses sur la suite de pas $(\gamma_n)_{n\geq 1}$. Notons aussi que cette preuve n'est pas directement applicable dans le cas des espaces de dimension infinie. En effet, dans ce cas, les boules fermées ne sont pas nécessairement compactes, et on ne peut donc pas extraire une sous-suite convergente.

1.1.2 Recherche des minima d'une fonction

Comme mentionné en préambule, on peut assimiler, dans un certain contexte, la recherche d'un zéro de Φ à la recherche du minimum global d'une fonction. En effet, soit $G : \mathbb{R}^d \longrightarrow \mathbb{R}$ une fonction de classe C^1 et de gradient ∇G, on veut résoudre

$$m := \arg\min_{h\in\mathbb{R}^d} G(h). \tag{1.4}$$

Le point $m \in \mathbb{R}^d$ est alors un zéro du gradient de G, et l'algorithme s'écrit

$$m_{n+1} = m_n - \gamma_n \nabla G(m_n), \tag{1.5}$$

avec $m_1 \in \mathbb{R}^d$ et les mêmes hypothèses que dans la partie précédente sur la suite de pas $(\gamma_n)_{n\geq 1}$. Les Propositions 1.1.1 et 1.1.2 peuvent alors se réécrire en remplaçant la fonction

ϕ par la fonction $\nabla G(.)$. De plus, généralement, on considère une fonction G convexe et on peut se référer aux résultats usuels d'analyse convexe (voir [Roc15] par exemple) pour obtenir l'existence et l'unicité d'un minimum global.

1.2 Algorithmes de gradient stochastiques

1.2.1 Définition et premiers résultats

Dans ce qui suit, on considère une variable aléatoire X à valeurs dans un espace $\mathcal{X}$. L'objectif est d'estimer une solution de l'équation (1.1). De manière générale, on a seulement accès à une échantillon de X et donc seulement accès à plusieurs observations d'une variable aléatoire de moyenne $\Phi(.)$. On ne peut donc pas utiliser directement l'algorithme de gradient déterministe. On se donne maintenant une suite de variables aléatoires indépendantes $(X_n)_{n\geq 1}$ de même loi que X. On introduit l'algorithme de gradient stochastique (voir [RM51]) défini de manière itérative pour tout $n \geq 1$ par

$$m_{n+1} = m_n - \gamma_n \phi\left(X_{n+1}, m_n\right), \tag{1.6}$$

avec $(\gamma_n)_{n\geq 1}$ une suite de réels positifs vérifiant (1.3). On donnera plus de précision sur le point initial m_1 dans les résultats de convergence. Néanmoins, il est usuel de prendre m_1 borné, déterministe, ou admettant un moment d'ordre 2. Introduisons la suite de tribus définies pour tout $n \geq 1$ par $\mathcal{F}_n := \sigma\left(X_1, ..., X_n\right) = \sigma\left(m_1, ..., m_n\right)$, comme la variable aléatoire X_{n+1} est indépendante de $\mathcal{F}_n$, on a alors

$$\mathbb{E}\left[\phi\left(X_{n+1}, m_n\right) | \mathcal{F}_n\right] = \Phi\left(m_n\right).$$

On peut alors écrire l'algorithme défini par (1.6) comme :

$$m_{n+1} = m_n - \gamma_n \Phi\left(m_n\right) + \gamma_n \xi_{n+1}, \tag{1.7}$$

avec $\xi_{n+1} := \Phi\left(m_n\right) - \phi\left(X_{n+1}, m_n\right)$. Notons que (ξ_n) est une suite de différences de martingale adaptée à la filtration $(\mathcal{F}_n)$. On peut alors écrire une version stochastique du Théorème 1.1.1.

Je remercie Anatoli Juditsky pour ses précieuses remarques bibliographiques qui m'ont permis de rétablir la vérité sur la paternité de plusieurs résultats et méthodes de décomposition (voir [DJ92] et [DJ93] pour la décomposition de l'algorithme moyenné par exemple).

Proposition 1.2.1. *On suppose qu'il existe un point $m \in \mathbb{R}^d$ qui annule la fonction Φ et*

— *qu'il existe une constante positive C telle que pour tout $h \in \mathbb{R}^d$,*

$$\|\Phi(h)\| \leq C \|h - m\|,$$

— *que la fonction ϕ est uniformément bornée : il existe une constante positive M telle que pour tout tout $x \in \mathcal{X}$ et $h \in H$,*

$$\|\phi(x, h)\| \leq M,$$

— *et qu'il existe une constante strictement positive c telle que pour tout $h \in \mathbb{R}^d$,*

$$\langle \Phi(h), h - m \rangle \geq c \|h - m\|^2.$$

Alors, m est l'unique zéro de la fonction Φ et la suite $(m_n)_{n\geq 1}$ définie par (1.6) vérifie

$$\lim_{n\to\infty} \mathbb{E}\left[\|m_n - m\|^2\right] = 0.$$

Démonstration. On obtient l'unicité du zéro de la fonction Φ de la même façon que pour le cas déterministe. Montrons maintenant la convergence de l'algorithme. On a

$$\begin{aligned}\|m_{n+1} - m\|^2 &= \|m_n - m\|^2 - 2\gamma_n \langle \phi(X_{n+1}, m_n), m_n - m\rangle + \gamma_n^2 \|\phi(X_{n+1}, m_n)\|^2 \\ &\leq \|m_n - m\|^2 - 2\gamma_n \langle \phi(X_{n+1}, m_n), m_n - m\rangle + M^2\gamma_n^2.\end{aligned}$$

Comme m_n est $\mathcal{F}_n$-mesurable, et par hypothèse,

$$\begin{aligned}\mathbb{E}\left[\|m_{n+1} - m\|^2 |\mathcal{F}_n\right] &\leq \|m_n - m\|^2 - 2\gamma_n \langle \mathbb{E}[\phi(X_{n+1}, m_n)|\mathcal{F}_n], m_n - m\rangle + \gamma_n^2 M^2 \\ &\leq \|m_n - m\|^2 - 2\gamma_n \langle \Phi(m_n), m_n - m\rangle + \gamma_n^2 M^2 \\ &\leq (1 - 2c\gamma_n)\|m_n - m\|^2 + \gamma_n^2 M^2.\end{aligned}$$

On obtient donc la relation de récurrence

$$\mathbb{E}\left[\|m_{n+1} - m\|^2\right] \leq (1 - 2c\gamma_n)\mathbb{E}\left[\|m_n - m\|^2\right] + \gamma_n^2 M^2,$$

et on peut conclure en appliquant un lemme de stabilisation (voir [Duf96], Lemme 4.1.1) ou à l'aide d'une récurrence sur n. □

Afin de pouvoir donner un premier résultat usuel sur la convergence des algorithmes de type Robbins-Monro, on va énoncer un résultat classique, analogue à celui utilisé dans la

preuve du Théorème 1.1.1 pour la suite (V_n).

Théorème 1.2.1 (Théorème de Robbins-Siegmund [RS85]). *Soient $(V_n)_{n\geq 1}$, $(a_n)_{n\geq 1}$, $(b_n)_{n\geq 1}$, $(c_n)_{n\geq 1}$ des suites de variables aléatoires réelles positives telles que*

$$\sum_{n\geq 1} a_n < +\infty \quad p.s, \qquad\qquad \sum_{n\geq 1} b_n < +\infty \quad p.s.$$

Soit $(\mathcal{F}_n)_{n\geq 1}$ une suite de tribus telle que V_n, a_n, b_n, c_n soient $\mathcal{F}_n$-mesurable pour tout $n \geq 1$. Enfin, supposons

$$\mathbb{E}\left[V_{n+1}|\mathcal{F}_n\right] \leq (1 + a_n)\, V_n + b_n - c_n.$$

Alors, la suite $(V_n)_{n\geq 1}$ converge presque sûrement vers une variable aléatoire presque sûrement finie V et

$$\sum_{n\geq 1} c_n < +\infty \quad p.s.$$

On peut maintenant donner des conditions moins fortes pour obtenir la forte consistance de l'algorithme.

Proposition 1.2.2. *On suppose qu'il existe un point $m \in \mathbb{R}^d$ qui annule la fonction Φ et*

— *qu'il existe une constante positive C telle que pour tout $h \in \mathbb{R}^d$,*

$$\mathbb{E}\left[\|\phi(X,h)\|^2\right] \leq C^2\left(1 + \|h-m\|^2\right),$$

— *que pour tout $h \in \mathbb{R}^d$ tel que $h \neq m$,*

$$\langle \Phi(h), h-m\rangle > 0,$$

— *et que la variable m_1 admet une moment d'ordre 2.*

Alors, m est l'unique zéro de la fonction Φ et la suite $(m_n)_{n\geq 1}$ définie par (1.6) vérifie

$$\lim_{n\to\infty} m_n = m \quad p.s.$$

De plus,

$$\lim_{n\to\infty} \mathbb{E}\left[\|m_n - m\|^2\right] = 0.$$

Notons que cette proposition est très usuelle pour les algorithmes de gradient stochastiques, mais reste un résultat faible. En effet, il ne donne pas la vitesse de convergence de l'algorithme, et il ne donne aucune garantie sur le comportement de l'algorithme pour une taille d'échantillon n fixée. Cependant, de la même façon que pour le cas déterministe, ce

résultat permet de mettre en lumière le rôle de la suite de pas $(\gamma_n)_{n\geq 1}$ et les grandes lignes de la preuve représentent une méthode classique pour obtenir la forte consistance des algorithmes de gradient stochastique, et ce, même dans le cas où la dimension de l'espace n'est pas finie.

Démonstration. De la même façon que dans la preuve précédente, comme m_n est $\mathcal{F}_n$-mesurable, on a

$$\begin{aligned}\mathbb{E}\left[\|m_{n+1}-m\|^2|\mathcal{F}_n\right] &\leq \|m_n-m\|^2 - 2\gamma_n\langle\Phi(m_n), m_n-m\rangle + \gamma_n^2\mathbb{E}\left[\|\phi(X_{n+1},m_n)\|^2|\mathcal{F}_n\right]\\ &\leq \left(1+C^2\gamma_n^2\right)\|m_n-m\|^2 - 2\gamma_n\langle\Phi(m_n), m_n-m\rangle + C^2\gamma_n^2. \qquad (1.8)\end{aligned}$$

A l'aide d'une récurrence, comme $\gamma_n\langle\Phi(m_n), m_n-m\rangle > 0$, et comme $\sum_{n\geq 1}\gamma_n^2 < +\infty$, on peut montrer qu'il existe une constante positive M telle que

$$\mathbb{E}\left[\|m_{n+1}-m\|^2\right] \leq \left(\prod_{k=1}^{n}\left(1+2C^2\gamma_k^2\right)\right)\mathbb{E}\left[\|m_1-m\|^2\right] + \left(\prod_{k=1}^{n}\left(1+2C^2\gamma_k^2\right)\right)\sum_{k=1}^{n}2C^2\gamma_k^2 \leq M.$$

De plus, en appliquant le Théorème 1.2.1 et l'inégalité (1.8), comme $\sum_{n\geq 1}\gamma_n^2 < +\infty$ et comme $\langle\Phi(m_n), m_n-m\rangle \geq 0$, la suite $\|m_n-m\|^2$ converge presque sûrement vers une variable aléatoire finie et

$$\sum_{n\geq 1}\gamma_n\langle\Phi(m_n), m_n-m\rangle < +\infty \quad p.s.$$

Donc, pour tout $\omega\in\Omega$, on a

$$\liminf_n \langle\Phi(m_n(\omega)), m_n(\omega)-m\rangle = 0.$$

La fin de la preuve est alors analogue au cas déterministe, et on obtient la convergence en moyenne quadratique par convergence dominée. □

Remarquons que de la même façon que pour le cas déterministe, cette preuve ne peut pas s'appliquer directement si l'on est dans des espaces de dimension infinie.

Application à l'optimisation stochastique :

De la même façon que pour le cas déterministe, on peut voir ce problème comme un problème de minimisation de la forme

$$m := \arg\min_{h\in\mathbb{R}^d} G(h), \qquad (1.9)$$

avec $G : \mathbb{R}^d \longrightarrow \mathbb{R}$ une fonction de la forme

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

où $g : \mathcal{X} \times \mathbb{R}^d \longrightarrow \mathbb{R}$. Supposons de plus que la fonction G est de classe C^1 et que pour presque tout x, $g(x,.)$ l'est aussi, avec

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

où ∇G désigne le gradient de G et $\nabla_h g$ le gradient de g par rapport à la seconde variable. L'algorithme de type Robbins-Monro s'écrit alors

$$m_{n+1} = m_n - \gamma_n \nabla_h g\left(X_{n+1}, m_n\right). \tag{1.10}$$

Dans ce contexte, la Proposition 1.2.2 est vérifiée en remplaçant ϕ par $\nabla_h g$ et Φ par ∇G.

1.2.2 Vitesses de convergence

La littérature sur les vitesses de convergence des algorithmes de type Robbins-Monro est très vaste, et l'on ne donnera que quelques exemples de résultats : des résultats asymptotiques comme la vitesse de convergence presque sûre et la normalité asymptotique (voir [Pel98]), et des résultats non asymptotiques comme la vitesse de convergence en moyenne quadratique (voir [BM13]). Dans ce qui suit, on considère une suite de pas $(\gamma_n)_{n\geq 1}$ de la forme

$$\gamma_n = c_\gamma n^{-\alpha} \quad \text{avec} \quad c_\gamma > 0 \quad \text{et} \quad \alpha \in \left(\frac{1}{2}, 1\right). \tag{1.11}$$

Le cas $\alpha = 1$ existe dans la littérature (voir [Pel98] et [Pel00] par exemple), mais nécessite des informations au préalable sur la fonction dont on cherche le zéro, et plus précisément sur les valeurs propres de sa différentielle en m. On donnera par la suite une méthode permettant d'obtenir la vitesse de convergence optimale, sans avoir à prendre un tel type de pas.

Remarque 1.2.1. *Notons que dans ce qui suit, on considère deux types résultats : asymptotiques (vitesse de convergence presque sûre, normalité asymptotique...), qui ne donnent aucune garantie sur le comportement de l'algorithme pour une taille d'échantillon n fixée, et non-asymptotiques (vitesse de convergence en moyenne quadratique, vitesses L^p, ...).*

Remarque 1.2.2. *Bien que l'on ait fait le choix, dans cette partie, de se concentrer sur les résultats introduits par [Pel98] et [BM13], la littérature est très riche sur les vitesses de convergence des algorithmes de gradient stochastiques (voir [NJLS09] pour les algorithmes de gradient à pas constant, ou [JN+14] pour les algorithmes "Primal dual", par exemple).*

Vitesses de convergence asymptotiques :

Notons que la littérature est large sur les vitesses de convergence presque sûre des algorithmes de gradient stochastiques (voir [Duf97] et [Bot10] parmi d'autres). Dans cette partie, on se concentre sur les résultats introduits par [Pel98]. On considère le problème (1.1), avec $\phi : \mathcal{X} \times \mathbb{R}^d \longrightarrow \mathbb{R}^d$ et m une solution du problème. On souhaite donner la vitesse de convergence de l'algorithme de Robbins-Monro défini par (1.6). Pour cela, on suppose que les hypothèses suivantes sont vérifiées :

(P1) La suite $(m_n)_{n\geq 1}$ converge presque sûrement vers m.

(P2) Il existe une constante $a > 1$ et un voisinage $\mathcal{U}_m$ de m tels que pour tout $h \in \mathcal{U}_m$,

$$\Phi(z) = H(z-m) + O\left(\|z-m\|^a\right),$$

avec H une matrice dont la partie réelle des valeurs propres est strictement positive.

(P3) Notons $\xi_{n+1} := \Phi(m_n) - \phi(X_{n+1}, m_n)$, il existe des constantes positives $M > 0$ et $b > 2$ telles que presque sûrement

$$\sup_{n\geq 1} \mathbb{E}\left[\|\xi_{n+1}\|^b \,|\mathcal{F}_n\right] \mathbb{1}_{\{\|m_n - m\| \leq M\}} < +\infty.$$

De plus, il existe une matrice symétrique définie positive Σ telle que

$$\lim_{n\to\infty} \mathbb{E}\left[\xi_{n+1}\xi_{n+1}^T | \mathcal{F}_n\right] = \Sigma \quad p.s.$$

Remarquons que l'on peut se ramener à la Proposition 1.2.2, par exemple, pour vérifier l'hypothèse **(P1)**. Les théorèmes suivants donnent la vitesse de convergence presque sûre ainsi que la normalité asymptotique des estimateurs.

Théorème 1.2.2 ([Pel98])**.** *Supposons que les hypothèses **(P1)** à **(P3)** sont vérifiées. Alors, pour tout $\delta > 0$,*

$$\|m_n - m\| = o\left(\frac{(\ln n)^\delta}{n^{\alpha/2}}\right) \quad p.s.$$

Théorème 1.2.3 ([Pel98])**.** *Supposons que les hypothèses **(P1)** à **(P3)** sont vérifiées. Alors, on a la convergence en loi*

$$\lim_{n\to\infty} \frac{m_n - m}{\sqrt{\gamma_n}} \sim \mathcal{N}\left(0, \Sigma'\right),$$

avec

$$\Sigma' := \int_0^{+\infty} e^{-sH} \Sigma e^{-sH} ds.$$

Remarque 1.2.3. *Soit $(a_n)_{n\geq 1}$, $(b_n)_{n\geq 1}$ deux suites de variables aléatoires réelles. On note*

$a_n = O(b_n)$ p.s si il existe une variable aléatoire presque sûrement finie K telle que

$$\lim_{n\to\infty} \frac{a_n}{b_n} \leq K \quad p.s.$$

De la même façon, on note $a_n = o(b_n)$ p.s si

$$\lim_{n\to\infty} \frac{a_n}{b_n} = 0 \quad p.s.$$

Enfin, soient X, Y deux variables aléatoires de même loi, on note alors $X \sim Y$.

Heuristique de preuve. **Vitesse de convergence presque sûre :** Rappelons que l'algorithme de type Robbins-Monro peut s'écrire comme

$$m_{n+1} = m_n - \gamma_n \Phi(m_n) + \gamma_n \xi_{n+1},$$

avec $\xi_{n+1} := \Phi(m_n) - \phi(X_{n+1}, m_n)$. Notons que la suite (ξ_n) est une suite de différences de martingale par rapport à la filtration $(\mathcal{F}_n)$. De plus, en linéarisant la fonction Φ grâce à l'hypothèse **(P2)**, on obtient

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

avec $\delta_n := \Phi(m_n) - H(m_n - m)$. Enfin, à l'aide d'une récurrence sur n, on obtient

$$m_n - m = \beta_{n-1}(m_1 - m) + \beta_{n-1} M_n - \beta_{n-1} R_n,$$

avec

$$\beta_n := \prod_{k=1}^{n} (I_d - \gamma_k H), \qquad M_n := \sum_{k=1}^{n-1} \gamma_k \beta_k^{-1} \xi_{k+1},$$

$$\beta_0 := I_d, \qquad R_n := \sum_{k=1}^{n-1} \gamma_k \beta_k^{-1} \delta_k.$$

Notons que $(M_n)_{n\geq 2}$ est une martingale par rapport à la filtration $(\mathcal{F}_n)$. De plus, on montrera que le terme $\beta_{n-1} M_n$ est le terme dominant et le lemme suivant en donne la vitesse de convergence.

Lemme 1.2.1 ([Pel98])**.** *Supposons que les hypothèses **(P1)** à **(P3)** sont vérifiées, alors pour tout $\delta > 0$,*

$$\|\beta_{n-1} M_n\| = O\left(\frac{(\ln n)^\delta}{n^{\alpha/2}}\right) \quad p.s.$$

Pour conclure la preuve, on introduit maintenant la suite des restes $(\Delta_n)_{n\geq 2}$ définie pour tout $n \geq 2$ par

$$\Delta_n = m_n - m - \beta_{n-1} M_n.$$

On a alors

$$\begin{aligned}\Delta_{n+1} &= m_{n+1} - m - \beta_n M_{n+1} \\ &= (I_d - \gamma_n H)(m_n - m) + \gamma_n \xi_{n+1} - \gamma_n \delta_n - (I_d - \gamma_n H)(M_n + \gamma_n \xi_{n+1}) \\ &= (I_d - \gamma_n H)\Delta_n - \gamma_n \delta_n.\end{aligned}$$

En appliquant un lemme de stabilisation (voir [Duf96], Lemme 4.1.1), on obtient

$$\|\Delta_n\| = O(\|\delta_n\|) \quad p.s.$$

De plus, comme la suite $(\|m_n - m\|)_{n\geq 1}$ converge presque sûrement vers 0 d'après l'hypothèse **(P1)**, et grâce à l'hypothèse **(P3)**, on a

$$\|\Delta_n\| = O\left(\|m_n - m\|^a\right) = o(\|m_n - m\|) \quad p.s.$$

En appliquant le lemme précédent, pour tout $\delta > 0$,

$$\begin{aligned}\|m_n - m\| &\leq \|\beta_{n-1} M_n\| + \|\delta_n\| \\ &= o\left(\frac{(\ln n)^\delta}{n^{\alpha/2}}\right) + o(\|m_n - m\|) \quad p.s,\end{aligned}$$

ce qui conclu la preuve pour la vitesse de convergence.

Normalité asymptotique : Rappelons que l'algorithme peut s'écrire

$$m_n - m = \beta_{n-1}(m_1 - m) + \beta_{n-1} M_n + \beta_{n-1} R_n.$$

Le terme $\beta_{n-1}(m_1 - m)$ converge exponentiellement vite vers 0. Avec la vitesse de convergence presque sûre, on peut montrer

$$\|\beta_{n-1} R_n\| = O\left(\|m_n - m\|^a\right) = o\left(\frac{1}{n^{\alpha/2}}\right) \quad p.s.$$

Il ne "reste donc plus qu'à" appliquer un TLC au terme $\frac{\beta_{n-1} M_n}{\sqrt{\gamma_n}}$. □

Remarque 1.2.4. *Un point important est que ces preuves ne sont pas directement applicables en*

dimension infinie. Plus précisément, afin de démontrer le lemme 1.2.1, différentes méthodes sont utilisées dans [Pel98] qui reposent, par exemple, sur l'existence de la trace d'une matrice, ou sur le fait que des sous-espaces propres d'une matrice soient de dimension finie, ce qui n'est pas automatiquement le cas en dimension infinie. Durant ma thèse, une des difficultés a donc été de trouver de nouvelles méthodes de démonstration qui ne dépendaient pas de la dimension de l'espace. Par exemple, une version du lemme 1.2.1 est donnée dans le cadre général des espaces de Hilbert au Chapitre 7.

Vitesse de convergence en moyenne quadratique et premiers pas vers la dimension infinie :

La littérature sur les vitesses de convergence non asymptotiques (voir [Bac14] par exemple) est bien moins vaste que pour celle avec les vitesses asymptotiques. Dans cette partie, on se concentre sur le cadre introduit par [BM13]. Plus précisément, on cherche à estimer la solution du problème (1.9), avec $G : H \longrightarrow \mathbb{R}$, où H est un espace de Hilbert séparable. Pour cela, on considère l'algorithme récursif défini par (1.10) et on suppose qu'il existe $m \in H$ tel que $\nabla G(m) = 0$ et que les hypothèses suivantes sont vérifiées :

(BM1) Il existe une constante positive L telle que pour tout $h, h' \in H$,

$$\mathbb{E}\left[\left\|\nabla_h g\left(X, h\right) - \nabla_h g\left(X, h'\right)\right\|^2\right] \leq L^2 \left\|h - h'\right\|^2.$$

(BM2) La fonction G est fortement convexe : il existe une constante strictement positive μ telle que pour tout $h, h' \in H$,

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

(BM3) Il existe une constante positive σ^2 telle que

$$\mathbb{E}\left[\left\|\nabla_h g\left(X, m\right)\right\|^2\right] \leq \sigma^2.$$

Alors, m est le minimum globale de la fonction G et on a la convergence en moyenne quadratique suivante :

Théorème 1.2.4 ([BM13])**.** *Supposons que les hypothèses **(BM1)** à **(BM3)** sont vérifiées et que l'on a une suite de pas vérifiant (1.11). Alors, il existe une constante positive C telle que pour tout $n \geq 1$,*

$$\mathbb{E}\left[\left\|m_n - m\right\|^2\right] \leq \frac{C}{n^\alpha}.$$

Démonstration. Pour tout $n \geq 1$, comme m_n est $\mathcal{F}_n$-mesurable,

$$\begin{aligned}\mathbb{E}\left[\|m_{n+1}-m\|^2 |\mathcal{F}_n\right] &\leq \|m_n-m\|^2 - 2\gamma_n \langle \mathbb{E}\left[\nabla_h g\left(X_{n+1}, m_n\right)|\mathcal{F}_n\right], m_n - m\rangle \\ &\quad + \gamma_n^2 \mathbb{E}\left[\|\nabla_h g\left(X_{n+1}, m_n\right)\|^2 |\mathcal{F}_n\right] \\ &\leq \|m_n-m\|^2 - 2\gamma_n \langle \nabla G(m_n), m_n - m\rangle + \gamma_n^2 \mathbb{E}\left[\|\nabla_h g\left(X_{n+1}, m_n\right)\|^2 |\mathcal{F}_n\right].\end{aligned}$$

Grâce à l'hypothèse **(BM2)**, on a pour tout $h \in H$,

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

ce qui assure bien, dans un premier temps, que m est le minimum global de la fonction G, et dans un deuxième temps, pour tout $h \in H$,

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

De plus, grâce aux hypothèses **(BM1)** et **(BM3)**, on a pour tout $h \in H$,

$$\begin{aligned}\mathbb{E}\left[\|\nabla_h g\left(X,h\right)\|^2\right] &\leq 2\mathbb{E}\left[\|\nabla_h g\left(X,h\right) - \nabla_h g\left(X,m\right)\|^2\right] + 2\mathbb{E}\left[\|\nabla_h g\left(X,m\right)\|^2\right] \\ &\leq 2L^2\|h-m\|^2 + 2\sigma^2.\end{aligned}$$

Finalement, on obtient

$$\begin{aligned}\mathbb{E}\left[\|m_{n+1}-m\|^2\right] &\leq \mathbb{E}\left[\|m_n-m\|^2\right] - \mu\gamma_n \mathbb{E}\left[\|m_n-m\|^2\right] + 2\gamma_n^2 L^2 \mathbb{E}\left[\|m_n-m\|^2\right] + 2\gamma_n^2\sigma^2 \\ &\leq \left(1 - \mu\gamma_n + \gamma_n^2 L^2\right)\mathbb{E}\left[\|m_n-m\|^2\right] + 2\gamma_n^2\sigma^2,\end{aligned}$$

et on peut conclure la démonstration à l'aide d'une récurrence sur n. □

Remarque 1.2.5. *Le résultat exact donné dans [BM13] donne plus de précisions sur la constante C, précisions que nous ne donnons pas pour simplifier les notations. Ce résultat reste vrai quelle que soit la dimension de l'espace (finie ou infinie), et est assez représentatif des problèmes que l'on peut rencontrer pour obtenir des résultats non asymptotiques. En effet, alors que l'on peut se contenter d'hypothèses locales pour avoir les résultats non asymptotiques, on est souvent obligé d'imposer des hypothèses de majoration uniforme et/ou de forte convexité globale et non plus locale.*

1.2.3 Algorithme projeté

Il peut arriver que le problème (1.1) admette plusieurs solutions, ou que les hypothèses nécessaires à la convergence de l'algorithme ne soient vérifiées que sur un sous-ensemble de l'espace H. Il peut alors être judicieux de projeter cet algorithme sur ce sous-ensemble. Plus précisément, on se donne un convexe fermé $\mathcal{K}$ de H. Le convexe $\mathcal{K}$ peut être connu (voir [BF12] parmis d'autres) ou peut être estimé à l'aide d'une partie de l'échantillon (voir Section 8.5 par exemple). L'algorithme projeté s'écrit alors

$$\widehat{m}_{n+1} := \pi\left(\widehat{m}_n - \gamma_n \phi\left(X_{n+1}, \widehat{m}_n\right)\right), \tag{1.12}$$

où π est la projection euclidienne sur $\mathcal{K}$. Rappelons que $\pi(h) = h$ si $h \in \mathcal{K}$ et $\pi(h) \in \partial\mathcal{K}$ si $h \notin \mathcal{K}$ (où $\partial\mathcal{K}$ est la frontière de $\mathcal{K}$). De plus, pour tout $h, h' \in H$,

$$\left\|\pi(h) - \pi(h')\right\| \leq \left\|h - h'\right\|.$$

De plus, notons que l'algorithme peut aussi s'écrire

$$\widehat{m}_{n+1} = \widehat{m}_n - \gamma_n \Phi\left(\widehat{m}_n\right) + \gamma_n \xi_{n+1} + r_n,$$

avec $\xi_{n+1} := \Phi\left(\widehat{m}_n\right) - \phi\left(X_{n+1}, \widehat{m}_n\right)$. La suite (ξ_n) est une suite de différences de martingale par rapport à la filtration $(\mathcal{F}_n)$, et

$$r_n := \pi\left(\widehat{m}_n - \gamma_n \phi\left(X_{n+1}, \widehat{m}_n\right)\right) - \left(\widehat{m}_n - \gamma_n \phi\left(X_{n+1}, \widehat{m}_n\right)\right).$$

Notons que $r_n = 0$ lorsque $\widehat{m}_n - \gamma_n \phi\left(X_{n+1}, \widehat{m}_n\right) \in \mathcal{K}$, i.e lorsque l'on n'a pas besoin de projeter. On peut donc le voir comme un algorithme de Robbins-Monro "contaminé" et la projection comme une garantie que notre algorithme ne sorte pas de l'ensemble convexe $\mathcal{K}$.

Convergence presque sûre :

On s'intéresse maintenant à la forte consistance de l'algorithme.

Proposition 1.2.3. *On suppose que les hypothèses de la Proposition 1.2.2 sont vérifiées pour tout $h \in \mathcal{K}$. Alors m est l'unique zéro de la fonction Φ sur $\mathcal{K}$ et la suite $\left(\widehat{m}_n\right)_{n\geq 1}$ définie par (1.12) vérifie*

$$\lim_{n\to\infty} \widehat{m}_n = m \quad p.s.$$

De plus,

$$\lim_{n\to\infty} \mathbb{E}\left[\left\|\widehat{m}_n - m\right\|^2\right] = 0.$$

Démonstration. Comme pour tout $h, h' \in \mathbb{R}^d$ on a $\|\pi(h) - \pi(h')\| \leq \|h - h'\|$, et comme $m \in \mathcal{K}$, on a $\pi(m) = m$ et

$$\begin{aligned}\|\widehat{m}_{n+1} - m\|^2 &\leq \|\pi\left(\widehat{m}_n - \gamma_n \phi\left(X_{n+1}, \widehat{m}_n\right)\right) - m\|^2 \\ &\leq \|\pi\left(\widehat{m}_n - \gamma_n \phi\left(X_{n+1}, \widehat{m}_n\right)\right) - \pi(m)\|^2 \\ &\leq \|\widehat{m}_n - m - \gamma_n \phi\left(X_{n+1}, \widehat{m}_n\right)\|^2,\end{aligned}$$

et pour tout $n \geq 1$, par définition de l'algorithme, $\widehat{m}_n \in \mathcal{K}$. On obtient alors

$$\mathbb{E}\left[\|\widehat{m}_{n+1} - m\|^2 |\mathcal{F}_n\right] \leq \|\widehat{m}_n - m\|^2 - 2\gamma_n \langle \Phi(\widehat{m}_n), \widehat{m}_n - m\rangle + \gamma_n^2 \mathbb{E}\left[\|\phi\left(X_{n+1}, \widehat{m}_n\right)\|^2\right],$$

et on peut donc montrer la convergence de l'algorithme de la même façon que pour l'algorithme de Robbins-Monro non projeté. □

Vitesse de convergence asymptotique :

De manière analogue à l'algorithme non-projeté, on suppose que les hypothèses suivantes sont vérifiées :

(P1') La suite $(\widehat{m}_n)_{n\geq 1}$ converge presque sûrement vers m, où m à l'intérieur de $\mathcal{K}$.

(P2') Il existe une constante $a > 1$ et un voisinage $\mathcal{U}_m \subset \mathcal{K}$ de m tels que pour tout $h \in \mathcal{U}_m$,

$$\Phi(z) = H(z - m) + O\left(\|z - m\|^a\right),$$

avec H une matrice dont la partie réelle des valeurs propres est strictement positive.

(P3') Notons $\xi_{n+1} := \Phi\left(\widehat{m}_n\right) - \phi\left(X_{n+1}, \widehat{m}_n\right)$, il existe des constantes positives $M > 0$ et $b > 2$ telles que presque sûrement

$$\sup_{n\geq 1} \mathbb{E}\left[\|\xi_{n+1}\|^b |\mathcal{F}_n\right] \mathbb{1}_{\{\|\widehat{m}_n - m\| \leq M\}} < +\infty.$$

De plus, il existe une matrice symétrique définie positive Σ telle que

$$\lim_{n\to\infty} \mathbb{E}\left[\xi_{n+1}\xi_{n+1}^T |\mathcal{F}_n\right] = \Sigma \quad p.s.$$

Notons que l'hypothèse **(P2')** ne peut être vérifiée que si m est à l'intérieur de $\mathcal{K}$. Les théorèmes suivants donnent la vitesse de convergence presque sûre ainsi que la normalité asymptotique des estimateurs.

Théorème 1.2.5. *Supposons que les hypothèses **(P1')** à **(P3')** sont vérifiées. Alors, pour tout $\delta > 0$,*

$$\|\widehat{m}_n - m\| = o\left(\frac{(\ln n)^{\delta}}{n^{\alpha/2}}\right) \quad p.s.$$

Théorème 1.2.6. *Supposons que les hypothèses **(P1')** à **(P3')** sont vérifiées. Alors, on a la convergence en loi*

$$\lim_{n\to\infty} \frac{\widehat{m}_n - m}{\sqrt{\gamma_n}} \sim \mathcal{N}\left(0, \Sigma'\right),$$

avec

$$\Sigma' := \int_0^{+\infty} e^{-sH}\Sigma e^{-sH} ds.$$

Vitesse de convergence en moyenne quadratique :

On veut maintenant estimer une solution locale du problème défini par (1.9), avec $G : H \longrightarrow \mathbb{R}$, où H est un espace de Hilbert séparable. On suppose que G est convexe sur un sous espace convexe et fermé $\mathcal{K}$ de H et que la fonction $g(x,.)$ est de classe C^1 sur $\mathcal{K}$ pour presque tout $x \in \mathcal{X}$. On cherche

$$m := \arg\min_{h\in\mathcal{K}^o} G(h),$$

avec $\mathcal{K}^o$ l'intérieur de $\mathcal{K}$. Soit π la projection euclidienne sur $\mathcal{K}$, rappelons que l'algorithme projeté s'écrit alors sous la forme

$$\widehat{m}_{n+1} = \pi\left(\widehat{m}_n - \gamma_n \nabla_h g\left(X_{n+1}, \widehat{m}_n\right)\right).$$

On suppose qu'il existe m à l'intérieur de $\mathcal{K}$ tel que $\nabla G(m) = 0$. On a alors la convergence en moyenne quadratique suivante.

Théorème 1.2.7. *Supposons que les hypothèses **(BM1)** et **(BM2)** sont vérifiées seulement pour tout $h, h' \in \mathcal{K}$, que l'hypothèse **(BM3)** est vérifiée et que l'on a une suite de pas de la forme $\gamma_n := c_\gamma n^{-\alpha}$, avec $c_\gamma > 0$ et $\alpha \in (0,1)$. Alors, m est l'unique minimum de la fonction G sur $\mathcal{K}$ et il existe une constante positive C telle que pour tout $n \geq 1$,*

$$\mathbb{E}\left[\|\widehat{m}_n - m\|^2\right] \leq \frac{C}{n^{\alpha}}.$$

1.3 L'algorithme moyenné

1.3.1 Retour sur l'algorithme de Robbins-Monro

Dans les exemples précédents, on a vu que sous certaines hypothèses et en prenant un pas vérifiant (1.11), l'algorithme défini par

$$m_{n+1} = m_n - \gamma_n \phi\left(X_{n+1}, m_n\right)$$

converge avec une vitesse de l'ordre $\frac{1}{n^\alpha}$, ce qui n'est pas la vitesse paramétrique (qui est de l'ordre $\frac{1}{n}$, car $\alpha \neq 1$) pour les algorithmes stochastiques. Une première idée, pour obtenir la vitesse optimale serait donc de prendre un pas de la forme $\frac{c}{n}$. Le théorème suivant donne alors la normalité asymptotique de l'estimateur avec ce type de pas et un bon choix de la constante c.

Théorème 1.3.1 ([Pel98])**.** *Supposons que les hypothèses* ***(P1)*** *à* ***(P3)*** *sont vérifiées. De plus, soit* $\lambda_{\min}$ *la plus petite partie réelle des valeurs propres de* H*. Si on prend un pas de la forme* $\frac{c}{n}$ *avec* $c > \frac{1}{2\lambda_{\min}}$*, on a alors la convergence en loi*

$$\lim_{n\to\infty} \sqrt{n}\left(m_n - m\right) \sim \mathcal{N}\left(0, c\Sigma\right).$$

Remarque 1.3.1. *Un résultat analogue est donné par [Wal77] dans le cadre général des espaces de Hilbert.*

Afin d'obtenir une meilleur covariance, on peut introduire un algorithme de la forme (voir [Pel98])

$$m_{n+1}^A = m_n^A + \frac{A}{n}\phi\left(X_{n+1}, m_n^A\right),$$

avec A une matrice $d \times d$ et inversible telle que la matrice $AH - \frac{1}{2}I_d$ ait des vecteurs propres dont la partie réelle est positive.

Proposition 1.3.1 ([Pel98])**.** *Supposons que les hypothèses* ***(P2)*** *et* ***(P3)*** *sont vérifiées, et que* m_n^A *converge presque sûrement vers* m*. On a alors la convergence en loi*

$$\lim_{n\to\infty} \sqrt{n}\left(m_n^A - m\right) \sim \mathcal{N}\left(0, \Sigma(A)\right),$$

où $\Sigma(A)$ *est la solution de l'équation de Lyapounov*

$$\left(AH - \frac{1}{2}I_d\right)\Sigma(A) + \Sigma(A)\left(H^T A^T - \frac{1}{2}I_d\right) = A\Sigma A^T. \tag{1.13}$$

Le choix optimal de la matrice A pour résoudre l'équation de Lyapounov est $A = H^{-1}$. En prenant $A = H^{-1}$, on obtient donc l'algorithme de Newton

$$m_{n+1}^N = m_n^N + \frac{H^{-1}}{n}\phi\left(X_{n+1}, m_n^N\right).$$

Cependant, la matrice H est généralement inconnue. Une idée pour régler ce problème serait d'avoir un estimateur H_n de H et d'écrire l'algorithme comme

$$m_{n+1}^N = m_n^N - \frac{H_n^{-1}}{n}\phi\left(X_{n+1}, m_n\right).$$

Plusieurs problèmes se posent alors. Cela nécessiterait d'inverser une matrice $d \times d$ à chaque opération, ce qui en terme de temps de calcul devient conséquent si l'on à des données à valeurs dans un espace de grande dimension. Une autre possibilité serait d'estimer la matrice H, puis de l'inverser, et d'injecter cet estimateur dans "l'algorithme de Newton", mais on perd alors le côté itératif de l'algorithme, et cela demanderait, par exemple, de stocker en mémoire toutes les données. Enfin, en pratique, ces algorithmes, avec ce choix de pas, ne sont pas nécessairement plus performants pour une taille d'échantillon n fixée que l'algorithme de Robbins-Monro avec un pas vérifiant (1.11). Enfin, une idée serait de trouver une estimateur récursif de H^{-1}, ce qui n'est pas évident, et ce qui demanderait quand même d'effectuer une opération matricielle à chaque itération. Pour résoudre ce problème, on introduit l'algorithme moyenné.

1.3.2 Définition et comportement asymptotique

On rappelle que l'algorithme de Robbins-Monro est défini de manière itérative pour tout $n \geq 1$ par

$$m_{n+1} = m_n - \gamma_n \phi\left(X_{n+1}, m_n\right).$$

On prend une suite de pas vérifiant (1.11). L'algorithme moyenné introduit par [Rup88] et [PJ92] est défini pour tout $n \geq 1$ par

$$\overline{m}_n = \frac{1}{n}\sum_{k=1}^{n} m_k,$$

ce qui peut s'écrire de manière récursive comme

$$\overline{m}_{n+1} = \overline{m}_n + \frac{1}{n+1}\left(m_{n+1} - \overline{m}_n\right), \tag{1.14}$$

avec $\overline{m}_1 = m_1$. Notons que grâce au lemme de Toeplitz, si l'algorithme de Robbins-Monro converge presque sûrement vers m, alors son moyenné aussi. L'idée de la moyennisation est de "lisser" l'algorithme de Robbins-Monro lorsque celui-ci "tourne autour" de m. Afin de donner les résultats asymptotiques introduit par [Pel00], rappelons que si les hypothèses **(P1)** à **(P3)** sont vérifiées, alors pour tout $\delta > 0$,

$$\|m_n - m\| = o\left(\frac{(\ln n)^\delta}{n^{\alpha/2}}\right) \quad p.s.$$

Théorème 1.3.2 ([Pel00])**.** *Supposons que les hypothèses **(P1)** à **(P3)** sont vérifiées, et que $\alpha > \frac{1}{a}$ (avec a défini dans l'hypothèse **(P2)**). On a alors la vitesse de convergence presque sûre :*

$$\|\overline{m}_n - m\| = O\left(\frac{\sqrt{\ln(\ln n)}}{\sqrt{n}}\right) \quad p.s.$$

Remarquons que si $a \geq 2$, on se ramène alors aux restrictions usuelles sur le pas. De plus, on a la normalité asymptotique suivante :

Théorème 1.3.3 ([Pel00])**.** *Supposons que les hypothèses **(P1)** à **(P3)** sont vérifiées et que $\alpha > \frac{1}{a}$ (avec a défini dans l'hypothèse **(P2)**), on a alors la convergence en loi*

$$\lim_{n\to\infty} \sqrt{n}\,(\overline{m}_n - m) \sim \mathcal{N}\left(0, H^{-1}\Sigma H^{-1}\right).$$

La covariance $H^{-1}\Sigma H^{-1}$ ainsi obtenue est la solution optimale de l'équation de Lyapunov (1.13).

Heuristique de preuve. Rappelons que l'algorithme de type Robbins-Monro peut s'écrire (voir (1.7))

$$m_{n+1} - m = m_n - m - \gamma_n \Phi(m_n) + \gamma_n \xi_{n+1},$$

avec $\xi_{n+1} := \Phi(m_n) - \phi\,(X_{n+1}, m_n)$. La suite (ξ_n) est une suite de différences de martingales adaptée à la filtration $(\mathcal{F}_n)$. En linéarisant la fonction Φ, on obtient

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

avec $\delta_n := \Phi(m_n) - H\,(m_n - m)$. L'inégalité précédente peut aussi s'écrire

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

En sommant ces inégalités, en appliquant la transformée d'Abel et en divisant par $\sqrt{n}$, on

obtient (voir [DJ93], [DJ92])

$$\sqrt{n}H\left(\overline{m}_n - m\right) = \frac{m_1 - m}{\sqrt{n}\gamma_1} - \frac{m_{n+1} - m}{\sqrt{n}\gamma_n} + \frac{1}{\sqrt{n}}\sum_{k=2}^{n}\left(\frac{1}{\gamma_k} - \frac{1}{\gamma_{k-1}}\right)(m_k - m) - \frac{1}{\sqrt{n}}\sum_{k=1}^{n}\delta_k + \frac{1}{\sqrt{n}}\sum_{k=1}^{n}\xi_{k+1} \tag{1.15}$$

On doit donc maintenant donner la vitesse de convergence de chacun de ces termes. Comme pour tout $\delta > 0$, on a

$$\|m_n - m\| = o\left(\frac{(\ln n)^\delta}{n^{\alpha/2}}\right) \quad p.s,$$

et grâce à l'hypothèse **(P2)**

$$\|\delta_n\| = O\left(\|m_n - m\|^a\right) = o\left(\frac{(\ln n)^\delta}{n^{a\alpha/2}}\right) \quad p.s.$$

De plus, comme $\gamma_k^{-1} - \gamma_{k-1}^{-1} \leq 2\alpha c_\gamma^{-1} k^{\alpha-1}$, et $\alpha \in (\frac{1}{a}, 1)$, on obtient

$$\begin{aligned}
&\frac{\|m_1 - m\|}{\sqrt{n}\gamma_1} = o\left(\frac{1}{\sqrt{n}}\right) \quad p.s,\\
&\frac{\|m_{n+1} - m\|}{\sqrt{n}\gamma_n} = o\left(\frac{1}{\sqrt{n}}\right) \quad p.s,\\
&\frac{1}{\sqrt{n}}\left\|\sum_{k=2}^{n}\left(\frac{1}{\gamma_k} - \frac{1}{\gamma_{k-1}}\right)(m_k - m)\right\| = o\left(\frac{1}{\sqrt{n}}\right) \quad p.s,\\
&\frac{1}{\sqrt{n}}\left\|\sum_{k=1}^{n}\delta_k\right\| = o\left(\frac{1}{\sqrt{n}}\right) \quad p.s.
\end{aligned}$$

Enfin, la suite $\left(\sum_{k=1}^{n}\xi_{k+1}\right)_{n\geq 1}$ est une martingale par rapport à la filtration $(\mathcal{F}_n)$, on peut donc appliquer, respectivement, la loi du log itéré et un Théorème Central Limite pour les martingales (voir [Duf97] par exemple) pour obtenir, respectivement, la vitesse de convergence presque sûre et la normalité asymptotique. □

Remarque 1.3.2. *Notons que ces résultats sont basés sur ceux obtenus par [Pel98] pour l'algorithme de type Robbins-Monro, et qui, comme mentionné précédemment, ne sont pas prouvés dans le cas d'espaces de dimension infinie. Les preuves du Théorème précédent ne sont donc pas directement applicables en dimension infinie. Cependant, on donne au Chapitre 5, dans le cas particulier des estimateurs de la médiane, les vitesses de l'algorithme moyenné dans des espaces de Hilbert. De manière analogue, au Chapitre 7, on donne des hypothèses suffisantes pour obtenir la vitesse de convergence*

presque sûre de l'algorithme moyenné, pour des espaces de dimension finie ou non.

De la même façon, afin d'obtenir la normalité asymptotique de l'estimateur, des résultats classiques sur les martingales en dimension finie sont utilisée dans [Pel00]. Cependant, ceux-ci ne sont pas démontrés en dimension infinie. On peut alors, par exemple, considérer le Théorème Central Limite pour les martingales introduits par [Jak88] pour obtenir la normalité asymptotique de l'algorithme moyenné dans le cadre plus général des espaces de Banach.

1.3.3 Vitesse de convergence en moyenne quadratique

On cherche maintenant à estimer la solution du problème défini par (1.9), avec $G : H \longrightarrow \mathbb{R}$ et H un espace de Hilbert séparable. Pour ce faire, on considère la suite d'estimateurs $(m_n)_{n\geq 1}$ définie par (1.10) ainsi que la suite d'estimateurs moyennés $(\overline{m}_n)_{n\geq 1}$ définie par (1.14). On suppose qu'il existe $m \in H$ tel que $\nabla G(m) = 0$, et que les hypothèses **(BM1)** et **(BM2)** ainsi que les hypothèses suivantes sont vérifiées :

(BM3) Il existe une constante positive σ^2 telle que pour tout $h \in H$,

$$\mathbb{E}\left[\|\nabla_h g(X,h)\|^2\right] \leq \sigma^2.$$

(BM4) Pour presque tout $x \in \mathcal{X}$, la fonction $g(x,.)$ est deux fois différentiable et pour tout $h \in H$, on note $\nabla_h^2 g(x,h)$ sa Hessienne en h. De plus, il existe une constante positive M telle que pour tout $h, h' \in H$,

$$\left\|\nabla_h^2 g(x,h) - \nabla_h^2 g(x,h')\right\|_{op} \leq M\left\|h - h'\right\|.$$

(BM5) Il existe une constante positive τ telle que

$$\mathbb{E}\left[\|\nabla_h g(X,m)\|^4\right] \leq \tau^4,$$

et il existe un opérateur auto-adjoint positif Σ tel que

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

Sous ces hypothèses, m est le minimum global de la fonction G et on a la vitesse de convergence en moyenne quadratique suivante.

Théorème 1.3.4 ([BM13])**.** *Supposons que les hypothèses **(BM1)** à **(BM5)** sont vérifiées et que l'on a une suite de pas vérifiant (1.11). Alors, il existe une constante positive C telle que pour tout $n \geq 1$,*

$$\mathbb{E}\left[\|\overline{m}_n - m\|^2\right] \leq \frac{C}{n}.$$

Remarque 1.3.3. *De la même façon que pour l'algorithme de type Robbins-Monro, obtenir des résultats de convergence non-asymptotiques nécessite en général des hypothèses beaucoup plus fortes que pour obtenir des résultats asymptotiques. Dans [BM13], il est nécessaire d'avoir la forte convexité de la fonction que l'on veut minimiser pour obtenir ces résultats, tandis que dans [Bac14], des hypothèses très restrictives sur ses dérivées sont imposées. Aux Chapitres 4 et5, on donne ce type de résultats pour les estimateurs de la médiane, et ce alors que la fonction que l'on veut minimiser ne vérifie pas ce type de conditions. De plus, nous proposons au Chapitre 7, des hypothèses moins restrictives pour obtenir les vitesses de convergence L^p des algorithmes de gradients stochastiques et de leur version moyennée.*

On a donc vu que les algorithmes de gradient stochastiques sont des outils performants pour traiter de gros échantillons à valeurs dans des espaces de grande dimension, et on a présenté différents résultats de la littérature. Afin d'avoir une étude plus approfondie sur ces estimateurs, il est intéressant d'étudier leur comportement lorsqu'un partie des données est contaminée, ce qui est l'objet du Chapitre suivant.

Chapitre 2

Introduction à la notion de robustesse

2.1 Introduction

Avec le développement informatique, il est de plus en plus usuel en statistique de traiter de gros échantillons de données. Malheureusement, l'acquisition d'une importante quantité de données peut s'accompagner d'une contamination de celles-ci, ce qui peut conduire à de mauvaises estimations. Prenons par exemple un échantillon de 20 données dont la dernière est contaminée.

1.1	1.1	1.2	1.2	1.3
1.4	1.5	1.6	1.6	1.7
1.7	1.8	1.8	1.9	2.0
2.1	2.1	2.2	2.2	200

On obtient alors une moyenne empirique $\overline{X}_{20} = 11.575$, alors que si on enlève la donnée contaminée, on obtient $\overline{X}_{19} = 1.66$. Si l'on regarde la médiane, on obtient $m_{20} = 1.7$ et $m_{19} = 1.7$. On peut donc, à travers cet exemple, conjecturer que l'estimateur de la médiane est moins sensible aux données atypiques que l'estimateur de la moyenne. Plus précisément, à travers différents critères, on montrera dans ce que suit que l'estimateur de la médiane est "robuste" au contraire de celui de la moyenne.

Dans tout ce qui suit, on considère deux exemples d'estimateur : la moyenne empirique et un estimateur de la médiane que l'on définira par la suite. On considère une variable aléatoire X à valeurs dans $\mathbb{R}^d$, avec $d \geq 1$ et on se donne des variables aléatoires indépendantes et identiquement distribuées $X_1, ..., X_n, ...$ de même loi que X. Rappelons que la moyenne empirique $\overline{X}_n$ est définie par

$$\overline{X}_n := \frac{1}{n} \sum_{k=1}^{n} X_k.$$

Rappelons maintenant que la médiane géométrique m de X est définie par (voir [Hal48] et [Kem87])

$$m := \arg\min_{h\in\mathbb{R}^d} \mathbb{E}\left[\|X-m\| - \|X\|\right], \tag{2.1}$$

où $\|.\|$ est la norme euclidienne. Un estimateur de la médiane (voir [VZ00] par exemple) peut consister à minimiser la fonction empirique, i.e

$$m_n := \arg\min_{h\in\mathbb{R}^d} \sum_{k=1}^{n} \left(\|X_k-h\| - \|X_k\|\right),$$

ce qui revient à résoudre le problème de Fermat-Weber (voir [WF29]) généré par l'échantillon. Notons que l'on peut approcher cette solution à l'aide de l'algorithme de Weiszfeld ([Wei37b]).

Remarque 2.1.1. *La littérature est très vaste sur les estimateurs robustes et l'on aurait pu se concentrer sur l'un des plus usuels, la fonction de Huber (voir [Hub64] et [HR09]). Du fait des thématiques abordées dans cette thèse, j'ai préféré me concentrer sur la médiane géométrique et faire le parallèle avec la moyenne arithmétique.*

2.1.1 Une première définition de la robustesse

On donne maintenant une idée plus précise de la notion de robustesse. Soit $H = \mathbb{R}^m$ et $\mathcal{M}$ l'ensemble des mesures de probabilités sur H que l'on munit d'une métrique d. Soient $X_1, ..., X_n$ des variables aléatoires indépendantes à valeurs dans H et de même fonction de répartition F. Soit $T_n = T_n(X_1, ..., X_n)$ une suite d'estimateurs. On dit que cette suite d'estimateurs est robuste en F_0 si la suite qui à toute fonction de répartition F associe la fonction de répartition de T_n, notée $\mathcal{L}_F(T_n)$, est equicontinue (voir [HR09]). Plus précisément, on dit que la suite d'estimateurs est robuste si pour tout $\epsilon > 0$, il existe $\delta > 0$ et un rang n_0 tels que pour tout $n \geq n_0$,

$$d(F_0, F) \leq \delta \Longrightarrow d\left(\mathcal{L}_{F_0}(T_n), \mathcal{L}_F(T_n)\right) \leq \epsilon.$$

De plus, en pratique, il est intéressant de savoir dans quelle mesure une perturbation F change la loi $\mathcal{L}_F(T_n)$ d'un estimateur. On suppose maintenant que l'on a la convergence en probabilités de notre estimateur T_n vers $T(F)$ et la convergence en loi

$$\lim_{n\to\infty} \sqrt{n}\left(T_n - T(F)\right) \sim \mathcal{N}\left(0, A(F,T)\right),$$

où $A(F,T)$ est la variance asymptotique de l'estimateur. On peut donc chercher à mesurer l'impact d'une perturbation de F sur ces deux indicateurs. Pour cela, on se donne une

constante $\epsilon > 0$ et un voisinage $V_\epsilon(F_0)$. On peut prendre, par exemple, le voisinage de contamination (voir [HR09]) défini par

$$V_\epsilon(F_0) = \{F, \quad F = (1-\epsilon)F_0 + \epsilon G, \quad G \in \mathcal{M}\},$$

où $\mathcal{M}$ est l'ensemble des fonctions de répartition. D'autres voisinages peuvent être considérés, comme celui de Lévy :

$$L_\epsilon(F_0) = \{F, \quad \forall t \quad , F_0(t-\epsilon) - \epsilon \leq F(t) \leq F_0(t+\epsilon) + \epsilon\}.$$

On peut alors, par exemple, définir le biais maximum $b_1(\epsilon)$ et la variance maximum $v_1(\epsilon)$ par

$$\begin{aligned} b_1(\epsilon) &:= \sup_{F \in V_\epsilon} d(T_F, T(F_0)), \\ v_1(\epsilon) &:= \sup_{F \in V_\epsilon} A(F, T). \end{aligned}$$

A travers cette définition et ces exemples un peu abrupts, on voit que mesurer la robustesse demande d'avoir des connaissances sur le comportement asymptotique des estimateurs. On va donc, dans un premier temps, introduire les M-estimateurs, dont on donnera le comportement asymptotique, avant de donner des "quantificateurs" de la robustesse.

Remarque 2.1.2. *J'ai fait le choix, dans cette partie, de me concentrer sur différents critères usuels en dimension finie plutôt que de faire une synthèse exhaustive des résultats de la littérature sur la robustesse. Cependant, dans le cadre de la dimension infinie, on aurait pu s'intéresser à la généralisation des travaux de [Ham71] proposée par [Cue88], ainsi qu'à des travaux plus récents comme [KSZ12] et [Zäh16].*

2.2 M-estimateurs

Dans cette section, qui s'inspire grandement de [HR09] et [MMY06], on va introduire la notion de M-estimateur. Pour cela, on commence par faire quelques rappels sur un des M-estimateurs les plus usuels, l'estimateur du maximum de vraisemblance, avant de "généraliser" cette définition et de donner l'exemple des M-estimateurs de position.

2.2.1 L'estimateur du Maximum de Vraisemblance

Dans ce qui suit, on considère des variables aléatoires réelles $X_1, ..., X_n$ de même loi que X, où X est à valeurs dans $\mathbb{R}$ et a pour fonction de répartition F_0, et pour densité f_0. On suppose que la fonction de vraisemblance est donnée par

$$L(X_1, ..., X_n, \mu) := \prod_{k=1}^{n} f_0 (X_k - \mu) .$$

L'estimateur du maximum de vraisemblance $\widehat{\mu}_n$ est l'estimateur qui maximise la vraisemblance $L(X_1, ..., X_n, \mu)$, i.e

$$\widehat{\mu}_n := \arg\max_{\mu \in \mathbb{R}} L(X_1, ..., X_n, \mu).$$

Cela s'écrit aussi

$$\widehat{\mu}_n = \arg\min_{\mu \in \mathbb{R}} \sum_{k=1}^{n} - \log (f_0 (X_k - \mu)) . \tag{2.2}$$

2.2.2 M-estimateurs

L'estimateur du maximum de vraisemblance est en réalité un cas particulier de M-estimateur.

Définition 2.2.1. *Un M-estimateur est un estimateur de la forme*

$$\widehat{\mu}_n := \arg\min_{\mu \in \mathbb{R}^d} \sum_{k=1}^{n} g (X_k, \mu) , \tag{2.3}$$

avec $g : \mathcal{X} \times \mathbb{R}^d \longrightarrow \mathbb{R}$.

Donc, si g est de classe C^1 par rapport à la seconde variable, $\widehat{\mu}_n$ est alors solution de l'équation

$$\sum_{k=1}^{n} \nabla_\mu g (X_k, \mu) = 0, \tag{2.4}$$

où $\nabla_\mu(.)$ est le gradient de g par rapport à la deuxième variable.

Remarque 2.2.1. *Au niveau de la "population", i.e asymptotiquement, cela peut s'écrire*

$$\mu_0 := \arg\min_{\mu \in \mathbb{R}^d} \mathbb{E} [g (X, \mu)] ,$$

et si g est de classe C^1, on a

$$\mathbb{E} \left[\nabla_\mu g (X, \mu_0) \right] = 0.$$

On considère maintenant une variable aléatoire X à valeurs dans $\mathbb{R}^d$.

Définition 2.2.2. *Un M-estimateur de position est un estimateur $\widehat{\mu}_n$ de la forme*

$$\widehat{\mu}_n = \arg\min_{\mu \in \mathbb{R}^d} \sum_{k=1}^{n} g(X_k - \mu), \tag{2.5}$$

avec $g : \mathbb{R}^d \longrightarrow \mathbb{R}$.

Si la fonction g est de classe C^1, on peut alors réécrire le problème comme

$$\sum_{k=1}^{n} \nabla g \left(X_i - \widehat{\mu}_n\right) = 0,$$

où $\nabla g(.)$ est le gradient de la fonction g. Remarquons qu'en prenant $g : \mu \longmapsto \|\mu\|^2$, on retrouve la moyenne empirique. En effet, cela revient à résoudre

$$\sum_{k=1}^{n} \left(X_k - \widehat{\mu}_n\right) = 0.$$

Enfin, en prenant $g : \mu \longmapsto \|\mu\|$, on retrouve l'estimateur de la médiane (voir [VZ00] parmi d'autres).

2.3 Comportement asymptotique des M-estimateurs de position

Dans ce qui suit, on considère une suite de M-estimateurs de position $\left(\widehat{\mu}_n\right)_{n\geq 1}$ et on s'intéresse à son comportement lorsque n tend vers l'infini. L'objectif ici est de donner des intuitions permettant de comprendre comment obtenir le comportement asymptotique d'un M-estimateur, sans nécessairement rentrer dans le détail de la façon dont on obtient ces résultats.

2.3.1 Cas unidimensionnel

Dans ce qui suit, on considère une fonction $g : \mathbb{R} \longrightarrow \mathbb{R}$ de classe C^2. De plus, on se donne un échantillon $X_1, ..., X_n, ...$ et on s'intéresse à la suite de M-estimateurs de position $\left(\widehat{\mu}_n\right)_{n\geq 1}$ définie pour tout $n \geq 1$ par

$$\widehat{\mu}_n = \arg\min_{\mu \in \mathbb{R}} \sum_{k=1}^{n} g(X_k - \mu).$$

De plus, on pose

$$\mu_0 = \arg\min_{\mu \in \mathbb{R}} \mathbb{E}\left[g\left(X - \mu\right)\right],$$

et on suppose que $(\widehat{\mu}_n)_{n\geq 1}$ converge en probabilité vers μ_0 et que la fonction g est dérivable. Remarquons qu'en appliquant la loi des grands nombres, si pour tout $\mu \in \mathbb{R}$ la variable aléatoire $g(X-\mu)$ admet un moment d'ordre 1, alors, pour tout $\mu \in \mathbb{R}$, on a la convergence presque sûre (et donc en probabilités)

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} g'(X_k - \mu) = \mathbb{E}\left[g'(X-\mu)\right] \quad p.s.$$

On peut donc supposer, par exemple, que la fonction g est fortement convexe pour obtenir la convergence en probabilité de la suite $(\widehat{\mu}_n)_{n\geq 1}$ vers μ_0. De plus, comme la fonction g est deux fois dérivable, alors, sous certaines conditions, on a la convergence en loi (voir [HR09] et [MMY06])

$$\lim_{n\to\infty} \sqrt{n}\,(\widehat{\mu}_n - \mu_0) \sim \mathcal{N}\left(0, \frac{\mathbb{E}\left[g'(X-\mu_0)^2\right]}{\mathbb{E}\left[g''(X-\mu_0)\right]^2}\right).$$

On donne maintenant une idée de la preuve de ce résultat. Rappelons que $\widehat{\mu}_n$ vérifie

$$\sum_{k=1}^{n} g'(X_k - \widehat{\mu}_n) = 0.$$

A l'aide d'une décomposition de Taylor, on obtient

$$0 = \sum_{k=1}^{n} g'(X_k - \mu_0) - (\widehat{\mu}_n - \mu_0) \sum_{k=1}^{n} g''(X_k - \mu_0) + o(\widehat{\mu}_n - \mu_0).$$

De plus, comme $\mathbb{E}\left[g'(X-\mu_0)\right] = 0$, en divisant par n l'égalité précédente, on obtient

$$\begin{aligned}\mathbb{E}\left[g''(X-\mu_0)\right](\widehat{\mu}_n - \mu_0) = \frac{1}{n}\sum_{k=1}^{n} g'(X_k - \mu_0) - \mathbb{E}\left[g'(X-\mu)\right] \\ - (\widehat{\mu}_n - \mu_0)\frac{1}{n}\sum_{k=1}^{n}\left(g''(X_k-\mu_0) - \mathbb{E}\left[g''(X-\mu_0)\right]\right) + o\left(\frac{\widehat{\mu}_n - \mu_0}{n}\right).\end{aligned}$$

En appliquant un Théorème Central Limite au premier terme sur la droite de l'équation précédente et la loi des grands nombres au deuxième terme, et comme $(\widehat{\mu_n})_{n\geq 1}$ converge en probabilité vers μ_0, on obtient le résultat.

2.3.2 Cas multidimensionnel

On considère maintenant une fonction $g : \mathbb{R}^d \longrightarrow \mathbb{R}$ de classe C^2 et de gradient $\nabla g(.)$, on s'intéresse à la suite de M-estimateurs $(\widehat{\mu}_n)_{n\geq 1}$ définie pour tout $n \geq 1$ par

$$\widehat{\mu}_n = \arg\min_{\mu\in\mathbb{R}^d} \sum_{k=1}^{n} g\left(X_k - \mu\right).$$

De plus, on pose

$$\mu_0 = \arg\min_{\mu\in\mathbb{R}^d} \mathbb{E}\left[g\left(X - \mu\right)\right],$$

et on suppose que la suite $(\widehat{\mu}_n)_{n\geq 1}$ converge en probabilité vers μ_0. De la même façon que pour le cas unidimensionnel, on peut utiliser la loi des grands nombres pour obtenir la convergence de la suite d'estimateurs. De plus, pour tout $\mu \in \mathbb{R}^d$, on note $\nabla^2 g(.)$ la Hessienne de g, et sous certaines conditions, on a la convergence en loi

$$\lim_{n\to\infty} \sqrt{n}\left(\widehat{\mu}_n - \mu_0\right) \sim \mathcal{N}\left(0, \Gamma_{\mu_0}^{-1}\mathbb{E}\left[\nabla g\left(X - \mu_0\right) \otimes \nabla g\left(X - \mu_0\right)\right]\Gamma_{\mu_0}^{-1}\right),$$

avec $\Gamma_{\mu_0} := \mathbb{E}\left[\nabla^2 g\left(X - \mu_0\right)\right]$. De la même façon que pour le cas unidimensionnel, on donne maintenant une idée de la preuve de ce résultat. Dans un premier temps, rappelons que l'estimateur $\widehat{\mu}_n$ vérifie

$$\sum_{k=1}^{n} \nabla g\left(X_k - \widehat{\mu}_n\right) = 0.$$

Grâce à la décomposition de Taylor, on obtient

$$0 = \sum_{k=1}^{n} \nabla g\left(X_k - \mu_0\right) - \sum_{k=1}^{n} \nabla^2 g\left(X_k - \mu_0\right)\left(\widehat{\mu}_n - \mu_0\right) + o\left(\widehat{\mu}_n - \mu_0\right).$$

En divisant par n et comme $\mathbb{E}\left[\nabla g\left(X - \mu_0\right)\right] = 0$, on a

$$\begin{aligned}\mathbb{E}\left[\nabla^2 g\left(X - \mu_0\right)\right]\left(\widehat{\mu}_n - \mu_0\right) = {} & \frac{1}{n}\sum_{k=1}^{n} \nabla g\left(X_k - \mu_0\right) - \mathbb{E}\left[\nabla g\left(X - \mu_0\right)\right] \\ & - \frac{1}{n}\sum_{k=1}^{n}\left(\nabla^2 g\left(X_k - \mu_0\right) - \mathbb{E}\left[\nabla^2 g\left(X - \mu_0\right)\right]\right)\left(\widehat{\mu}_n - \mu_0\right) + o\left(\frac{\widehat{\mu}_n - \mu_0}{n}\right).\end{aligned}$$

On obtient donc le résultat en appliquant un Théorème Central Limite (voir Théorème 2.1.9 dans le livre [Duf97] par exemple, ou voir aussi [Bil13]) au premier terme sur la droite de l'équation précédente et la loi des grands nombres au deuxième terme. Remarquons que l'on peut généraliser ce type de méthodes et résultats à l'ensemble des M-estimateurs (voir

[Gee00]).

2.3.3 Exemples

Estimateur de la moyenne

Rappelons que la moyenne empirique est définie par

$$\overline{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k.$$

Si la variable aléatoire X admet un moment d'ordre 1, la convergence en probabilité et la convergence presque sûre vers $\mathbb{E}[X]$ sont des conséquences de la loi des grands nombres. De plus, si X admet un moment d'ordre 2, on a la convergence en loi

$$\lim_{n\to\infty} \sqrt{n}\left(\overline{X}_n - \mathbb{E}[X]\right) \sim \mathcal{N}\left(0, \Sigma\right).$$

Estimateur de la médiane

Rappelons que la suite de M-estimateurs de la médiane $(m_n)_{n\geq 1}$ est définie pour tout $n \geq 1$ par

$$m_n := \arg\min_{\mu\in\mathbb{R}^d} \sum_{k=1}^{n} \|X_k - \mu\|.$$

On suppose maintenant que les hypothèses suivantes sont vérifiées :

(A1) La variable aléatoire X n'est pas concentrée sur une droite, i.e pour tout $h \in \mathbb{R}^d$, il existe $h' \in \mathbb{R}^d$ tel que $\langle h, h'\rangle = 0$ et

$$\mathrm{Var}\left(\langle X, h'\rangle\right) > 0.$$

(A2) La variable aléatoire X n'est pas concentrée autour de la médiane m, i.e

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

Sous ces hypothèses, la suite d'estimateurs $(m_n)_{n\geq 1}$ converge presque sûrement vers la médiane m, et on peut montrer la convergence en loi (voir [ON85])

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

avec

$$\Sigma := \mathbb{E}\left[\frac{X-m}{\|X-m\|} \otimes \frac{X-m}{\|X-m\|}\right],$$

$$\Gamma_m := \mathbb{E}\left[\frac{1}{\|X-m\|}\left(I_{\mathbb{R}^d} - \frac{X-m}{\|X-m\|} \otimes \frac{X-m}{\|X-m\|}\right)\right].$$

On retrouve ainsi le résultat précédent.

2.4 Comment construire un M-estimateur

On s'est intéressé, dans la partie précédente, au comportement asymptotique des M-estimateurs. Cependant, contrairement au cas de la moyenne, on n'a quasiment jamais de formule explicite des ces estimateurs. Néanmoins, on va voir qu'il est possible de construire "aisément" des M-estimateur de positions réels, ainsi que de construire le M-estimateur de la médiane dans $\mathbb{R}^d$.

2.4.1 Estimateur de position dans $\mathbb{R}$

Dans cette partie, on va introduire des algorithmes permettant d'obtenir des M-estimateurs de positions dans le cas unidimensionnel. On considère une suite de M-estimateurs de position $(\widehat{\mu}_n)_{n\geq 1}$ de la forme (2.5) avec $g : \mathbb{R} \longrightarrow \mathbb{R}$. Supposons de plus que g est de classe C^1, on a alors

$$\sum_{k=1}^{n} g'\left(X_k - \widehat{\mu}_n\right) = 0.$$

Si $\widehat{\mu}_n \in \mathbb{R}\backslash\{X_i, i = 1, ..., n\}$, on pose alors $w_k = \frac{g'(X_k-\widehat{\mu}_n)}{X_k-\widehat{\mu}_n}$, et on peut réécrire l'équation précédente comme

$$\sum_{k=1}^{n} w_k\left(X_k - \widehat{\mu}_n\right) = 0.$$

On obtient donc

$$\widehat{\mu}_n = \frac{\sum_{k=1}^{n} w_k X_k}{\sum_{k=1}^{n} w_k}.$$

L'estimateur $\widehat{\mu}_n$ est donc un point fixe de la fonction $P_{X_1,...,X_n} : \mathbb{R}\backslash\{X_i, i = 1, ..., n\} \longrightarrow \mathbb{R}$ définie pour tout $\mu \in \mathbb{R}\backslash\{X_i, i = 1, ..., n\}$ par

$$P_{X_1,...,X_n}(\mu) := \frac{\sum_{k=1}^{n} \frac{g'(X_k-\mu)}{X_k-\mu} X_k}{\sum_{k=1}^{n} \frac{g'(X_k-\mu)}{X_k-\mu}},$$

et on peut donc bâtir un algorithme de recherche de point fixe de la forme

$$\widehat{\mu}_n^{(k+1)} = P_{X_1,\dots,X_n}\left(\widehat{\mu}_n^{(k)}\right).$$

2.4.2 Cas multidimensionnel

Pour un M-estimateur à valeurs dans un espace multidimensionnel, on ne peut pas introduire la suite de poids $(w_k)_{k\geq 1}$, et on ne peut donc pas obtenir directement de formule aussi explicite. Cependant, on peut utiliser des outils classiques de recherche du zéro d'une fonction tel que la méthode de Newton-Raphson cf Chapitre 1) pour construire cet estimateur. De plus, il existe des exemples classiques de M-estimateurs de positions pour lesquels on peut obtenir un algorithme "explicite" : la moyenne arithmétique et la médiane géométrique.

La moyenne

Rappelons que le M-estimateur de la moyenne est défini par

$$\overline{X}_n = \arg\min_{\mu\in\mathbb{R}^d} \sum_{k=1}^{n} \|X_k - \mu\|^2,$$

ce qui peut s'écrire

$$\sum_{k=1}^{n} (X_k - \widehat{\mu}_n) = 0,$$

et l'on retrouve bien l'estimateur usuel $\overline{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k$.

La médiane

Rappelons que le M-estimateur de la médiane est défini par

$$m_n := \arg\min_{\mu\in\mathbb{R}^d} \sum_{k=1}^{n} \|X_k - \mu\|,$$

ce qui revient à retrouver la solution du problème de Fermat-Weber ([WF29]) généré par l'échantillon. De plus, on peut voir le M-estimateur de la médiane comme la solution de l'équation

$$\sum_{k=1}^{n} \frac{X_k - m_n}{\|X_k - m_n\|} = 0. \tag{2.6}$$

Il peut aussi être vu comme le point fixe de la fonction $P_{X_1,...,X_n} : \mathbb{R}^d \setminus \{X_i, i = 1, ..., n\} \longrightarrow \mathbb{R}^d$, définie pour tout $\mu \in \mathbb{R}^d \setminus \{X_i, i = 1, ..., n\}$ par

$$P_{X_1,...,X_n}(\mu) := \frac{\sum_{k=1}^{n} \frac{X_k}{\|X_k - \mu\|}}{\sum_{k=1}^{n} \frac{1}{\|X_k - \mu\|}},$$

et on retrouve ainsi l'algorithme de Weiszfeld (voir [Wei37b]), i.e

$$m_n^{(k+1)} = P_{X_1,...,X_n}\left(m_n^{(k)}\right).$$

2.5 Fonction d'influence

Dans cette section, on cherche à quantifier l'influence de la présence de données atypiques dans l'échantillon sur le comportement des estimateurs. Pour cela, on va introduire la notion de fonction d'influence.

2.5.1 Définition

Dans ce qui suit, $(\widehat{\mu}_n)_{n\geq 1}$ est une suite d'estimateurs "générée" par une suite de variables aléatoires $X_1, ..., X_n, ...$ indépendantes et identiquement distribuées. Pour toute fonction de répartition F, on notera $\mu_0(F)$ la limite de la suite d'estimateur $(\widehat{\mu}_n)_{n\geq 1}$ lorsque les variables aléatoires $X_1, ..., X_n, ...$ ont pour fonction de répartition F. On va s'intéresser au comportement de l'estimateur lorsqu'une portion de l'échantillon est contaminée. Plus précisément, pour tout $x_0 \in \mathbb{R}^d$, on va s'intéresser au comportement de l'estimateur lorsque X admet une fonction de répartition du type $F_\epsilon := (1-\epsilon) F + \epsilon\delta_{x_0}$. Pour cela, on introduit la notion de fonction d'influence, qui est définie par

$$IF_{\mu_0}(x_0, F) := \lim_{\epsilon \to 0} \frac{\mu_0\left((1-\epsilon) F + \epsilon\delta_{x_0}\right) - \mu_0(F)}{\epsilon} \tag{2.7}$$

avec $\epsilon > 0$. De plus, on a

$$IF_{\mu_0}(x_0, F) = \lim_{\epsilon \to 0} \frac{\partial}{\partial \epsilon} \mu_0\left((1-\epsilon) F + \epsilon\delta_{x_0}\right)$$

En effet, on a

$$\begin{aligned}\lim_{\epsilon\to 0}\frac{\partial}{\partial\epsilon}\mu_0\left((1-\epsilon)F+\epsilon\delta_{x_0}\right) &= \lim_{\epsilon\to 0}\lim_{h\to 0}\frac{\mu_0\left((1-\epsilon-h)F+(\epsilon+h)\delta_{x_0}\right)-\mu_0\left((1-\epsilon)F+\epsilon\delta_{x_0}\right)}{h}\\ &= \lim_{h\to 0}\lim_{\epsilon\to 0}\frac{\mu_0\left((1-\epsilon-h)F+(\epsilon+h)\delta_{x_0}\right)-\mu_0\left((1-\epsilon)F+\epsilon\delta_{x_0}\right)}{h}\\ &= \lim_{h\to 0}\frac{\mu_0\left((1-h)F+h\delta_{x_0}\right)-\mu_0(F)}{h}.\end{aligned}$$

Notons qu'à l'aide de l'égalité (2.7), on peut approcher la limite de l'estimateur contaminé par

$$\mu_0\left((1-\epsilon)F+\epsilon\delta_{x_0}\right)\simeq\mu_0(F)+\epsilon IF_{\mu_0}(x_0,F).$$

2.5.2 Fonction d'influence d'un M-estimateur

L'objectif ici est de donner une formule explicite de la fonction d'influence d'un M-estimateur. Pour cela, on se donne des variables aléatoires $X_1,...,X_n,...$ à valeurs dans un espace $\mathcal{X}$, indépendantes et de même fonction de répartition F, et une suite de M-estimateurs $(\widehat{\mu}_n)_{n\geq 1}$ définis par l'équation (2.3), avec $g:\mathcal{X}\times\mathbb{R}^d\longrightarrow\mathbb{R}$. De plus, on suppose que la fonction $g(x,.)$ est différentiable pour presque tout x, et on note $\nabla_\mu g(x,.)$ son gradient, les estimateurs vérifient alors l'équation (2.4). Enfin, on note

$$\mu_0=\arg\min_{\mu\in\mathbb{R}^d}\mathbb{E}\left[g\left(X,\mu\right)\right].$$

Pour tout $\epsilon>0$ et $x_0\in\mathbb{R}^d$, on note $F_\epsilon=(1-\epsilon)F+\epsilon\delta_{x_0}$. De plus, pour toute fonction de répartition $\tilde{F}$, on note $E_{\tilde{F}}$ l'espérance lorsque la variable aléatoire a pour fonction de répartition $\tilde{F}$. Par définition et en reprenant les notations de la partie précédente, on a

$$\begin{aligned}\mathbb{E}_F\left[g'\left(X,\mu_0\right)\right]&=0,\\ \mathbb{E}_{F_\epsilon}\left[\nabla_\mu g\left(X,\mu_0(F_\epsilon)\right)\right]&=0=(1-\epsilon)E_F\left[\nabla_\mu g\left(X,\mu_0(F_\epsilon)\right)\right]+\epsilon\nabla_\mu g\left(x_0,\mu_0(F_\epsilon)\right).\end{aligned}$$

Si $\mu_0(F_\epsilon)$ converge vers $\mu_0(F)$ lorsque ϵ tend vers 0 et si $\mu_0(F_\epsilon)$ est continument dérivable par rapport à ϵ, on obtient, en dérivant par rapport à ϵ,

$$\begin{aligned}0=&-\mathbb{E}_F\left[\nabla_\mu g\left(X,\mu_0(F_\epsilon)\right)\right]+(1-\epsilon)\frac{\partial}{\partial\epsilon}\mu_0(F_\epsilon)\mathbb{E}_F\left[\frac{\partial}{\partial\mu}\nabla_\mu g\left(X,\mu_0(F_\epsilon)\right)\right]+\nabla_\mu g\left(x_0,\mu_0(F_\epsilon)\right)\\ &+\epsilon\frac{\partial}{\partial\epsilon}\mu_0(F_\epsilon),\end{aligned}$$

et par passage à la limite, on obtient

$$0 = \lim_{\epsilon \to 0} \frac{\partial}{\partial \epsilon} \mu_0(F_\epsilon) \mathbb{E}_F \left[\frac{\partial}{\partial \mu} \nabla_\mu g\left(X, \mu_0(F)\right) \right] + \nabla_\mu g\left(x_0, \mu_0(F)\right).$$

Donc, comme $\lim_{\epsilon \to 0} \frac{\partial}{\partial \epsilon} \mu_0(F_\epsilon) = IC_{\mu_0}(x_0, F)$, on obtient

$$IC_{\mu_0}(x_0, F) = -\mathbb{E}_F \left[\frac{\partial}{\partial \mu} \nabla_\mu g\left(X, \mu_0(F)\right) \right]^{-1} \nabla_\mu g\left(x_0, \mu_0(F)\right). \tag{2.8}$$

2.5.3 Exemples

Moyenne

Soit $x_0 \in \mathbb{R}^d$, pour tout $\epsilon > 0$, on pose $F_\epsilon = (1-\epsilon)F + \epsilon \delta_{x_0}$. Soient $X_{\epsilon,1}, ..., X_{\epsilon,n}$ des variables aléatoires identiquement distribuées et de fonction de répartition F_ϵ, on note $\overline{X}_{\epsilon,n}$ la moyenne empirique associée

$$\overline{X}_{\epsilon,n} := \frac{1}{n} \sum_{k=1}^{n} X_{\epsilon,k}.$$

Sous des conditions usuelles, cet estimateur converge presque sûrement vers

$$\mu_\epsilon = (1-\epsilon)\, \mu_0 + \epsilon x_0.$$

On obtient donc,

$$\begin{aligned} IF_{\mu_0}(x_0, F) &= \lim_{\epsilon \to 0} \frac{(1-\epsilon)\mu_0 + \epsilon x_0 - \mu_0}{\epsilon} \\ &= -\mu_0 + x_0. \end{aligned}$$

Remarquons que la fonction d'influence $IF_{\mu_0}(., F)$ n'est pas bornée et est donc sensible à la contamination.

Médiane

Soit $(m_{n,\epsilon})_{n \geq 1}$ la suite de M-estimateurs de la médiane générée par les observations contaminées $X_{\epsilon,1}, ..., X_{\epsilon,n}, ...$ Ces estimateurs vérifient alors l'équation (2.6) et on obtient, asymptotiquement,

$$(1-\epsilon)\, \mathbb{E}\left[\frac{X - m_\epsilon}{\|X - m_\epsilon\|} \right] + \epsilon \frac{x_0 - m_\epsilon}{\|x_0 - m_\epsilon\|} = 0.$$

Avec des calculs analogues à ceux permettant de trouver l'équation (2.8), on obtient (voir [Ger08])

$$IF_m(x_0) = -\left(\mathbb{E}\left[\frac{1}{\|X-m\|}\left(I_{\mathbb{R}^d} - \frac{(X-m)\otimes(X-m)}{\|X-m\|^2}\right)\right]\right)^{-1} \frac{x_0-m}{\|x_0-m\|}.$$

Notons que la matrice $\left(\mathbb{E}\left[\frac{1}{\|X-m\|}\left(I_{\mathbb{R}^d} - \frac{(X-m)\otimes(X-m)}{\|X-m\|^2}\right)\right]\right)^{-1}$ existe bien lorsque la variable aléatoire X n'est pas concentrée autour d'une droite (voir [Kem87] parmi d'autres).

Notons que contrairement à la moyenne, pour une fonction de répartition F fixée, on peut borner uniformément la fonction d'influence $IC_{\mu_0}(.,F)$. Plus précisément, pour tout $h \in \mathbb{R}^d$,

$$\|IF_m(h)\| \leq \frac{1}{\lambda_{\min}},$$

où $\lambda_{\min}$ est la plus petite valeur propre de la matrice $\mathbb{E}\left[\frac{1}{\|X-m\|}\left(I_{\mathbb{R}^d} - \frac{(X-m)\otimes(X-m)}{\|X-m\|^2}\right)\right]$ (voir [CCZ13] pour plus de détails).

2.6 Biais asymptotique maximum et point de rupture

2.6.1 Définitions

Dans cette partie, l'objectif est d'introduire deux nouveaux quantificateurs de robustesse. Pour cela, on définit le biais asymptotique maximum et le point de rupture d'un estimateur. Dans ce qui suit, on considère une variable aléatoire X à valeurs dans $\mathbb{R}^d$ ayant pour fonction de répartition F. On considère une suite de M-estimateurs $(\widehat{\mu}_n)_{n\geq 1}$ générée par un échantillon $X_1, ..., X_n, ...$ et on notera $\mu_0(F)$ la limite de la suite d'estimateurs $(\widehat{\mu}_n)_{n\geq 1}$ lorsque que pour tout $i \geq 1$, X_i a pour fonction de répartition F. De plus, pour tout $\epsilon > 0$, on considère le voisinage de contamination $V_\epsilon(F)$ défini par

$$V_\epsilon(F) := \{(1-\epsilon)\,F + \epsilon G, \quad G \in \mathcal{G}\},$$

où $\mathcal{G}$ est l'ensemble des fonctions de répartition dans $\mathbb{R}^d$.

Définition 2.6.1. *Pour tout $F_\epsilon \in V_\epsilon(F)$, le biais asymptotique de μ_0 en F_ϵ est défini par*

$$b_{\mu_0}(F_\epsilon) = \mu_0(F_\epsilon) - \mu_0(F).$$

Le biais maximum de μ_0 est défini par

$$BM(\mu_0, \epsilon) = \sup \left\{ \left\| b_{\mu_0}(F_\epsilon) \right\|, \quad F_\epsilon \in V_\epsilon \right\}.$$

On donne maintenant une première définition du point de rupture.

Définition 2.6.2. *Le point de rupture de μ_0 en F, noté $\epsilon^\star(\mu_0, F)$, est le plus grand $\epsilon^\star \in (0,1)$ tel que pour tout $\epsilon < \epsilon^\star$, la fonction qui a toute distribution G associe $\mu_0((1-\epsilon)F + \epsilon G)$ est bornée.*

De manière équivalente, on peut définir le point de rupture par

$$\epsilon^\star(\mu_0, F) = \sup \{ \epsilon \in (0,1), \quad BM(\mu_0, \epsilon) < +\infty \}. \tag{2.9}$$

2.6.2 Exemples

La moyenne

Soit X une variable aléatoire à valeurs dans $\mathbb{R}^d$, de fonction de répartition F et admettant un moment d'ordre 1. L'estimateur de la moyenne converge alors vers $\mu_0(F) = \mathbb{E}[X]$. On considère maintenant une variable aléatoire Y de fonction de répartition G. Pour tout $\epsilon > 0$, on note $F_\epsilon = (1-\epsilon)F + \epsilon G$, et on a

$$\mu_0(F_\epsilon) = (1-\epsilon)\,\mathbb{E}[X] + \epsilon \mathbb{E}[Y].$$

On obtient donc, pour tout $\epsilon > 0$,

$$b_{\mu_0}(F_\epsilon) = -\epsilon \mathbb{E}[X] + \epsilon \mathbb{E}[Y],$$

et en particulier

$$\begin{aligned} BM(\mu_0, \epsilon) &= +\infty, \\ \epsilon^\star(\mu_0, F) &= 0. \end{aligned}$$

En d'autres termes, pour tout $\epsilon > 0$, il est possible de contaminer l'échantillon de telle sorte que l'estimateur de la moyenne diverge vers l'infini.

La médiane

Soit X une variable aléatoire à valeurs dans $\mathbb{R}^d$ et de fonction de répartition F. Rappelons que la médiane géométrique est définie par

$$m(F) := \arg\min_{h\in\mathbb{R}^d} \mathbb{E}\left[\|X-h\| - \|X\|\right]$$

Théorème 2.6.1 ([Kem87],[Ger08])**.** *On suppose que la variable aléatoire X n'est pas concentrée sur une droite. La médiane est alors unique (voir [Kem87]) et a un point de rupture différent de* 0*. Plus précisément,*

$$\epsilon^\star(m,F) = 0.5.$$

De plus, on a

$$\lim_{\epsilon\to 0} BM(m,\epsilon) = 0.$$

Démonstration. **Point de rupture** La médiane ne peut pas avoir un point de rupture plus grand que $\frac{1}{2}$ (voir [MMY06]). On va maintenant montrer par l'absurde que pour tout $\epsilon \in \left(0,\frac{1}{2}\right)$, le biais maximum est fini. S'il existe $\epsilon \in \left(0,\frac{1}{2}\right)$ tel que $BM(m,\epsilon) = +\infty$, alors il existe une suite de fonctions de répartition F_n telles que en notant $F_{n,\epsilon} := (1-\epsilon)F + \epsilon F_n$, on ait

$$\lim_{n\to\infty} \|m\left(F_{n,\epsilon}\right)\| = +\infty,$$

où $m(F_{n,\epsilon})$ est la médiane d'une variable aléatoire ayant $F_{n,\epsilon}$ comme fonction de répartition. De plus, pour tout $h \in \mathbb{R}^d$, on a (voir [Ger08])

$$\begin{aligned} G_{n,\epsilon}(h) :&= \int_{\mathbb{R}^d} \left(\|x-h\| - \|x\|\right) dF_{n,\epsilon}(x) \\ &= (1-\epsilon)\int_{\mathbb{R}^d} \left(\|x-h\| - \|x\|\right) dF(x) + \epsilon \int_{\mathbb{R}^d} \left(\|x-h\| - \|x\|\right) dF_n(x) \\ &\geq (1-\epsilon)\int_{\mathbb{R}^d} \left(\|x-h\| - \|x\|\right) dF(x) - \epsilon\|h\|. \end{aligned}$$

De plus, on a

$$\lim_{\|h\|\to\infty} \frac{\int_{\mathbb{R}^d} \left(\|x-h\| - \|x\|\right) dF(x)}{\|h\|} = 1.$$

On obtient donc, comme $\lim_{n\to\infty} \|m\left(F_{n,\epsilon}\right)\| = +\infty$,

$$\liminf \frac{G_{n,\epsilon}\left(m\left(F_{n,\epsilon}\right)\right)}{\|m\left(F_{n,\epsilon}\right)\|} \geq 1-2\epsilon.$$

De plus, comme $m(F_{n,\epsilon})$ est le minimiseur de la fonction $G_{n,\epsilon}$, on obtient

$$G_{n,\epsilon}(m(F_{n,\epsilon})) \leq G_{n,\epsilon}(0) = 0,$$

et en particulier

$$1 - 2\epsilon \leq \liminf \frac{G_{n,\epsilon}\left(m\left(F_{n,\epsilon}\right)\right)}{\|m\left(F_{n,\epsilon}\right)\|} \leq 0,$$

On a donc $\epsilon > 1/2$, d'où la contradiction.

Convergence du biais maximum Supposons que $\lim_{\epsilon \to 0} BM(m, \epsilon) \neq 0$. Il existe alors une constante strictement positive δ, une suite $(\epsilon_n)_{n\geq 1}$ convergeant vers 0, et une suite de fonctions de répartition $(F_n)_{n\geq 1}$ telles que pour tout $n \geq 1$,

$$\|m(F) - m\left(F_{n,\epsilon}\right)\| \geq \delta,$$

avec $F_{n,\epsilon} := (1 - \epsilon_n)F + \epsilon_n F_n$. De plus, prenons $\epsilon_n < \frac{1}{2}$, et comme le point de rupture est égal à $\frac{1}{2}$, la suite $\left(\|m\left(F_{n,\epsilon}\right) - m(F)\|\right)_{n\geq 1}$ est bornée. On peut donc extraire une sous-suite $\left(m\left(F_{n_k,\epsilon}\right)\right)_{k\geq 1}$ convergeant vers un point $m'(F) \neq m(F)$. On a alors

$$\lim_{k\to\infty} G_{n_k,\epsilon}\left(m\left(F_{n_k,\epsilon}\right)\right) = G_0\left(m'(F)\right).$$

De plus, par unicité de la médiane, $G_0(m') > G_0(m(F))$. Enfin, par définition de $m\left(F_{n_k,\epsilon}\right)$, on a

$$G_0(m(F)) \geq \lim_{n\to\infty} G_{n,\epsilon}\left(m\left(F_{n_k,\epsilon}\right)\right) = G_0(m'(F)),$$

d'où la contradiction. □

Chapitre 3

Synthèse des principaux résultats

3.1 Estimation récursive de la médiane géométrique dans les espaces de Hilbert : boules de confiance non asymptotiques

Dans cette partie, on rappelle la définition de la médiane géométrique ainsi que les estimateurs récursifs introduits par [CCZ13]. On donne ici la vitesse de convergence en moyenne quadratique de l'algorithme de type Robbins-Monro, ce qui nous permettra ensuite de donner des boules de confiance non asymptotiques pour cet algorithme, ainsi que pour sa version moyennée.

3.1.1 Définitions et hypothèses

On considère une variable aléatoire X à valeurs dans un espace de Hilbert séparable H, pas nécessairement de dimension finie. La médiane géométrique m de X est définie par

$$m := \arg\min_{h \in H} \mathbb{E}\left[\|X - h\| - \|X\|\right]. \tag{3.1}$$

Le terme $\|X\|$ permet de ne pas avoir à faire d'hypothèses sur l'existence des moments de X, puisque $\mathbb{E}\left[|\|X - h\| - \|X\||\right] \leq \|h\|$. On fait maintenant les hypothèses suivantes :

(A1) La variable X n'est pas concentrée autour d'une droite : pour tout $h \in H$, il existe $h' \in H$ tel que $\langle h, h' \rangle = 0$ et

$$\mathrm{Var}\left(\langle X, h' \rangle\right) > 0.$$

(A2) La variable X n'est pas concentrée autour de points isolés : il existe une constante C

telle que pour tout $h \in H$,

$$\mathbb{E}\left[\frac{1}{\|X-h\|}\right] \leq C, \qquad \mathbb{E}\left[\frac{1}{\|X-h\|^2}\right] \leq C.$$

Sous l'hypothèse **(A1)**, la médiane est bien définie et est unique ([Kem87]). De plus, on notera $G : H \longrightarrow \mathbb{R}$ la fonction que l'on veut minimiser, la médiane m est l'unique solution de l'équation

$$\nabla G(h) = -\mathbb{E}\left[\frac{X-h}{\|X-h\|}\right] = 0.$$

On rappelle maintenant l'algorithme de type Robbins-Monro introduit par [CCZ13], défini de manière itérative pour tout $n \geq 1$ par

$$Z_{n+1} = Z_n + \gamma_n \frac{X_{n+1} - Z_n}{\|X_{n+1} - Z_n\|}, \tag{3.2}$$

avec Z_1 choisi borné. Par exemple, on peut prendre $Z_1 = X_1 \mathbb{1}_{\{\|X_1\| \leq M\}}$, avec $M > 0$. De plus, la suite de pas $(\gamma_n)_{n\geq 1}$ de la forme $\gamma_n := c_\gamma n^{-\alpha}$ avec $c_\gamma > 0$ et $\alpha \in \left(\frac{1}{2}, 1\right)$. Il est possible de prendre $\alpha = 1$, mais pour ce faire, on doit faire un bon choix de la constante c_γ (voir Section 1.3.1). Pour contourner ce problème et pour obtenir une convergence optimale, on introduit l'algorithme moyenné (voir Section 1.3) défini par

$$\overline{Z}_n = \frac{1}{n}\sum_{k=1}^{n} Z_k,$$

ce qui peut aussi s'écrire de façon itérative : pour tout $n \geq 1$

$$\overline{Z}_{n+1} = \overline{Z}_n + \frac{1}{n+1}\left(Z_{n+1} - \overline{Z}_n\right), \tag{3.3}$$

avec $\overline{Z}_1 = Z_1$.

3.1.2 Vitesses de convergence de l'algorithme de type Robbins-Monro

Sous ces mêmes hypothèses, il a été établi dans [CCZ13] qu'il existait une suite croissante d'évènements $(\Omega_N)_{N\geq 1}$ et des constantes C_N telles que pour tout $n \geq N$,

$$\mathbb{E}\left[\|Z_n - m\|^2 \mathbb{1}_{\Omega_N}\right] \leq C_N \frac{\ln n}{n^\alpha}.$$

Cet algorithme est implémenté dans le package "Gmedian" pour le langage "R".

Malheureusement, ce type de résultat est insuffisant pour obtenir des boules de confiance non asymptotiques. En effet, on n'a aucune information sur le comportement des constantes C_N et sur la vitesse à laquelle la suite d'évènements $(\Omega_N)_{N\geq 1}$ converge. Un premier résultat de cette thèse à été d'établir une meilleure majoration de l'erreur quadratique moyenne ([CCGB15]).

Théorème 3.1.1. *Supposons que les hypothèses **(A1)** et **(A2)** sont vérifiées. Alors, pour tout $\alpha < \beta < 3\alpha - 1$, on a les vitesses de convergence suivantes :*

$$\mathbb{E}\left[\|Z_n - m\|^2\right] = O\left(\frac{1}{n^\alpha}\right),$$
$$\mathbb{E}\left[\|Z_n - m\|^4\right] = O\left(\frac{1}{n^\beta}\right).$$

Il y a donc, par rapport à [CCZ13], une nette amélioration sur la vitesse de convergence en moyenne quadratique de l'algorithme, car le conditionnement sur les évènements Ω_N a disparu ainsi que le terme en $\ln n$. Plus précisément, la vitesse de convergence en moyenne quadratique obtenue est la vitesse optimale.

Proposition 3.1.1. *Supposons que les hypothèses **(A1)** et **(A2)** sont vérifiées. Il existe alors une constante strictement positive c telle que pour tout $n \geq 1$,*

$$\mathbb{E}\left[\|Z_n - m\|^2\right] \geq \frac{c}{n^\alpha}.$$

3.1.3 Boules de confiance non asymptotiques

Boules de confiance pour l'algorithme de type Robbins-Monro

L'objectif est d'obtenir une majoration, si possible fine, de $\mathbb{P}\left[\|Z_n - m\| \geq t\right]$, pour $t > 0$. On pourrait obtenir une première majoration grossière à l'aide de l'inégalité de Markov et du Théorème 3.1.1. On se propose de donner des intervalles plus précis. Pour cela, on introduit une nouvelle inégalité exponentielle pour des termes qui sont "presque" des martingales, ingalité analogue à celle du Théorème 3.1 dans [Pin94]. On obtient alors les boules de confiance suivantes.

Théorème 3.1.2. *Supposons que les hypothèses **(A1)** et **(A2)** sont vérifiées. Il existe alors une constante positive C telle que pour tout $\delta \in (0,1)$, il existe un rang n_δ tel que pour tout $n \geq n_\delta$,*

$$\mathbb{P}\left[\|Z_n - m\| \leq \frac{C}{n^{\alpha/2}} \ln\left(\frac{4}{\delta}\right)\right] \geq 1 - \delta.$$

On parle de boules de confiance non asymptotiques car n_δ est déterministe. De plus, en reprenant et en affinant tous les calculs, ce qui est particulièrement exhaustif, il est possible de donner une majoration de n_δ.

La constante C dépend de la constante qui apparait dans l'hypothèse **(A2)** et des valeurs propres de la Hessienne, tandis que le rang n_δ dépend de δ et des termes de restes, et est estimable. Remarquons de plus que si l'on avait directement appliqué l'inégalité de Markov et le Théorème 3.1.1, on aurait obtenu un résultat de la forme

$$\mathbb{P}\left[\|Z_n - m\| \leq \frac{C}{n^{\alpha/2}}\frac{1}{\delta}\right] \geq 1 - \delta,$$

et donc moins précis.

Boules de confiance pour l'algorithme moyenné

Dans [CCZ13], la normalité asymptotique de l'estimateur a été établie, c'est à dire

$$\lim_{n\to\infty} \sqrt{n}\left(\overline{Z}_n - m\right) \sim \mathcal{N}\left(0, \Gamma_m^{-1}\Sigma\Gamma_m^{-1}\right),$$

avec Γ_m la hessienne de G en m et

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

Cependant, en pratique, ce résultat est difficilement exploitable du fait de la difficulté pour obtenir une estimation de la covariance en grande dimension. On propose donc, à l'aide d'une inégalité exponentielle pour les termes de martingales ([Pin94]), de donner des boules de confiance non asymptotiques.

Théorème 3.1.3. *Supposons que les hypothèses **(A1)** et **(A2)** sont vérifiées. Pour tout $\delta \in (0,1)$, il existe un rang n_δ tel que pour tout $n \geq n_\delta$,*

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

où $\lambda_{\min}$ est la plus petite valeur propre de Γ_m.

On parle de boules de confiance non asymptotiques car n_δ est déterministe. De plus, en reprenant et en affinant tous les calculs, ce qui est particulièrement exhaustif, il est possible de donner une majoration de n_δ.

3.2 Estimation de la médiane géométrique dans les espaces de Hilbert à l'aide d'algorithmes de gradient stochastiques : vitesses de convergence L^p et presque sûre

L'objectif ici est de donner une étude non asymptotique plus poussée des algorithmes introduits précédemment, en donnant leurs vitesses de convergence L^p. Cela permettra ensuite d'obtenir leurs vitesses de convergence presque sûre. Enfin, ces résultats sont particulièrement utiles, par la suite, pour établir la convergence des estimateurs récursifs de la "Median Covariation Matrix".

3.2.1 Décomposition de l'algorithme de type Robbins-Monro

Dans ce qui suit, on suppose que les hypothèses **(A1)** et **(A2)** sont vérifiées. L'algorithme défini par (3.2) peut s'écrire

$$Z_{n+1} = Z_n - \gamma_n \Phi(Z_n) + \gamma_n \xi_{n+1}, \tag{3.4}$$

avec $\xi_{n+1} := \Phi(Z_n) + \frac{X_{n+1}-Z_n}{\|X_{n+1}-Z_n\|}$. Introduisons la suite de tribus $(\mathcal{F}_n)_{n\geq 1}$ définie pour tout $n \geq 1$ par $\mathcal{F}_n := \sigma(X_1, ..., X_n)$. Alors, la suite (ξ_n) est une suite de différences de martingales adaptée à la filtration $(\mathcal{F}_n)$ (voir aussi l'équation (1.7) au Chapitre 1). Finalement, en linéarisant le gradient, on obtient la décomposition suivante :

$$Z_{n+1} - m = (I_H - \gamma_n \Gamma_m)(Z_n - m) + \gamma_n \xi_{n+1} - \gamma_n \delta_n, \tag{3.5}$$

avec Γ_m la hessienne de G en m, et $\delta_n := \Phi(Z_n) - \Gamma_m(Z_n - m)$ est le reste de la décomposition de Taylor du gradient.

3.2.2 Vitesses de convergence L^p des algorithmes

Vitesses de convergence L^p de l'algorithme de type Robbins-Monro

On a obtenu précédemment la vitesse de convergence en moyenne quadratique optimale de l'algorithme de type Robbins-Monro. Afin, par exemple, d'obtenir la vitesse de convergence en moyenne quadratique de l'algorithme moyenné, on donne maintenant les vitesses de convergence L^p de l'algorithme de type Robbins-Monro. Pour cela, on prend une suite de pas vérifiant (1.11).

Théorème 3.2.1. *Supposons que les hypothèses* ***(A1)*** *et* ***(A2)*** *sont vérifiées. Pour tout entier p, il existe une constante positive K_p telle que pour tout $n \geq 1$,*

$$\mathbb{E}\left[\|Z_n - m\|^{2p}\right] \leq \frac{K_p}{n^{p\alpha}}. \tag{3.6}$$

La preuve repose sur une récurrence sur $p \geq 1$. On suppose que le Théorème 3.2.1 est vérifié pour tout entier $k \leq p-1$ et on va prouver à l'aide d'une récurrence sur $n \geq 1$ que pour tout $\beta \in \left(\alpha, \frac{p+2}{p}\alpha - \frac{1}{p}\right)$, il existe des constantes positives $C_p, C_{p,\beta}$ telle que pour tout $n \geq 1$,

$$\mathbb{E}\left[\|Z_n - m\|^{2p}\right] \leq \frac{C_p}{n^{p\alpha}},$$
$$\mathbb{E}\left[\|Z_n - m\|^{2p+2}\right] \leq \frac{C_{p,\beta}}{n^{p\beta}}.$$

On procède donc par une double récurrence. Pour ce faire, on a besoin de majorer le moment d'ordre $2p$. Grâce à la décomposition (3.5), on obtient le lemme suivant.

Lemme 3.2.1. *Supposons que les hypothèses* ***(A1)*** *et* ***(A2)*** *sont vérifiées. Soit $p \geq 1$, supposons que pour tout $k \leq p-1$ l'inégalité (3.6) est vérifiée. Alors, il existe des constantes positives c_0, C_1, C_2 et un rang n_α tel que pour tout $n \geq n_\alpha$,*

$$\mathbb{E}\left[\|Z_{n+1} - m\|^{2p}\right] \leq (1 - c_0\gamma_n)\,\mathbb{E}\left[\|Z_n - m\|^{2p}\right] + \frac{C_1}{n^{(p+1)\alpha}} + C_2\gamma_n\mathbb{E}\left[\|Z_n - m\|^{2p+2}\right].$$

On doit donc majorer le moment d'ordre $2p+2$ qui apparait dans la majoration précédente. Grâce à la décomposition (3.4), on a :

Lemme 3.2.2. *Supposons que les hypothèses* ***(A1)*** *et* ***(A2)*** *sont vérifiées. Soit $p \geq 1$, supposons que pour tout $k \leq p-1$ l'inégalité (3.6) est vérifiée. Alors il existe des constantes positives C_1', C_2' et un rang n_α tel que pour tout $n \geq n_\alpha$,*

$$\mathbb{E}\left[\|Z_{n+1} - m\|^{2p+2}\right] \leq \left(1 - \frac{2}{n}\right)^{p+1}\mathbb{E}\left[\|Z_n - m\|^{2p+2}\right] + \frac{C_1'}{n^{(p+2)\alpha}} + C_2'\gamma_n^2\mathbb{E}\left[\|Z_n - m\|^{2p}\right].$$

Vitesses de convergence L^p de l'algorithme moyenné

On s'intéresse maintenant à l'algorithme moyenné défini par (3.3). Avec l'aide du Théorème 3.2.1, on obtient les vitesses de convergence L^p suivantes :

Théorème 3.2.2. *Supposons que les hypothèses* ***(A1)*** *et* ***(A2)*** *sont vérifiées. Pour tout entier $p \geq 1$, il existe une constante positive K_p' telle que pour tout $n \geq 1$, la suite d'estimateurs définie par (3.3)*

vérifie

$$\mathbb{E}\left[\left\|\overline{Z}_n - m\right\|^{2p}\right] \leq \frac{K'_p}{n^p}.$$

Remarque 3.2.1. *Dans la littérature, il n'est souvent donné que la vitesse de convergence en moyenne quadratique, mais avoir les vitesse L^p est crucial pour montrer la convergence des estimateurs de la "Median Covariation Matrix". Notons enfin que nous avons obtenus ce résultat sans avoir à faire d'hypothèse supplémentaire par rapport à [CCZ13] et [CCGB15].*

Finalement, la proposition suivante assure que la vitesse de convergence en moyenne quadratique ainsi obtenue est optimale.

Proposition 3.2.1. *Supposons que les hypothèses **(A1)** et **(A2)** sont vérifiées. Il existe une constante strictement positive c telle que pour tout $n \geq 1$,*

$$\mathbb{E}\left[\left\|\overline{Z}_n - m\right\|^{2}\right] \geq \frac{c}{n}.$$

De plus, en appliquant l'inégalité de Hölder, la dernière proposition assure que les vitesses L^p obtenues sont elles aussi optimales.

Remarque 3.2.2. *Notons que les hypothèses introduites dans le Chapitre 1 ainsi que dans [Bac14] ne sont pas vérifiées dans ce contexte. Plus précisément, nous avons réussi à obtenir ce type de résultats avec des conditions moins restrictives que celles qui existaient dans la littérature.*

3.2.3 Vitesses de convergence presque sûre des algorithmes

En appliquant le lemme de Borel-Cantelli et le Théorème 3.2.1, on obtient une majoration de la vitesse de convergence presque sûre suivante :

Théorème 3.2.3. *Supposons que les hypothèses **(A1)** et **(A2)** sont vérifiées. Pour tout $\beta < \alpha$,*

$$\|Z_n - m\| = o\left(\frac{1}{n^{\beta/2}}\right) \quad p.s.$$

En appliquant le résultat précédent, on obtient une majoration de la vitesse de convergence presque sûre de l'algorithme moyenné.

Corollaire 3.2.1. *Supposons que les hypothèses **(A1)** et **(A2)** sont vérifiées. Pour tout $\delta > 0$,*

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

Remarque 3.2.3. *Nous avons obtenu, pour l'algorithme moyenné, la même vitesse de convergence presque sûre que dans [Pel00] et au Chapitre 1, et ce pour des données à valeurs dans un espace de Hilbert qui n'est pas nécessairement de dimension finie.*

3.3 Estimation rapide de la "Median Covariation Matrix" et application à l'Analyse des Composantes Principales en ligne

L'Analyse des Composantes Principales (ACP) est très utile en statistique pour réduire la dimension lorsque l'on traite de grands échantillons à valeurs dans des espaces de grande dimension. Dans ce contexte, la détection de données atypiques peut être difficile, et les composantes principales issues de l'analyse spectrale de la matrice de covariance peuvent être très sensibles à ces données atypiques.

On s'intéresse à une nouvelle méthode robuste pour l'ACP, basée sur l'analyse spectrale de la "Median Covariation Matrix" (MCM). Cette matrice (ou opérateur) peut être vue comme une médiane géométrique dans l'espace des matrices carrées (ou opérateurs) équipé de la norme de Frobenius (que l'on définit ci-après), et est donc, de la même façon que la médiane (voir [Kem87] et [Ger08] parmi d'autre), un indicateur de dispersion robuste. L'analyse spectrale de la MCM représente un réel intérêt du fait que sous certaines conditions, elle a les mêmes espaces propres que la variance (voir [KP12]).

De la même façon que pour la médiane, afin d'estimer la MCM, on introduit deux algorithmes récursifs : un algorithme de gradient stochastique et sa version moyennée. De plus, on introduit un algorithme récursif permettant d'estimer les principaux vecteurs propres de la MCM.

3.3.1 Définition et hypothèses

Dans ce qui suit, on suppose que les hypothèses **(A1)** et **(A2)** sont vérifiées. On considère maintenant l'espace des opérateurs linéaires de H dans H, noté $\mathcal{S}(H)$. Soit $\{e_i, i \in I\}$ une base orthonormée de H, l'espace $\mathcal{S}(H)$ muni du produit scalaire

$$\forall A, B \in \mathcal{S}(H), \qquad \langle A, B \rangle_F := \sum_{i \in I} \langle A(e_i), B(e_i) \rangle \, ,,$$

et de la norme associée $\|.\|_F$ (norme de Frobenius) est aussi un espace de Hilbert séparable. La Median Covariation Matrix Γ_m est alors définie par

$$\Gamma_m := \arg\min_{V \in \mathcal{S}(H)} \mathbb{E}\left[\|(X-m)\otimes(X-m) - V\|_F - \|(X-m)\otimes(X-m)\|_F\right], \tag{3.7}$$

où pour tout $h, h' \in H$, $h \otimes h' = \langle h, . \rangle h'$ et m est la médiane géométrique de X. On peut donc voir la Median Covariation Matrix Γ_m comme la médiane géométrique de la variable aléatoire $(X-m)\otimes(X-m)$. De manière analogue à la médiane, afin d'assurer l'existence et l'unicité de la MCM, on suppose maintenant que les hypothèses suivantes sont vérifiées :

(A3) Il existe des vecteurs $V_1, V_2 \in \mathcal{S}(H)$ linéairement indépendants telles que

$$\forall i \in \{1,2\}, \qquad \mathrm{Var}\left(\langle V_i, (X-m)\otimes(X-m)\rangle_F\right) > 0.$$

(A4) Il existe une constante positive C telle que pour tout $V \in \mathcal{S}(H)$ et $h \in H$,

$$(a): \quad \mathbb{E}\left[\|(X-h)\otimes(X-h) - V\|_F^{-1}\right] \leq C$$
$$(b): \quad \mathbb{E}\left[\|(X-h)\otimes(X-h) - V\|_F^{-2}\right] \leq C$$

3.3.2 Les algorithmes

Cas où la médiane m est connue

Si la médiane m est connue, comme la MCM est la médiane géométrique de la variable aléatoire $(X-m)\otimes(X-m)$, on peut l'estimer avec l'algorithme de gradient stochastique et sa version moyennée définis récursivement par :

$$W_{n+1} = W_n + \gamma_n \frac{(X_{n+1}-m)\otimes(X_{n+1}-m) - W_n}{\|(X_{n+1}-m)\otimes(X_{n+1}-m) - W_n\|_F},$$
$$\overline{W}_{n+1} = \overline{W}_n - \frac{1}{n+1}\left(\overline{W}_n - W_{n+1}\right),$$

avec $W_1 = \overline{W}_1$ borné et $(\gamma_n)_{n\geq 1}$ une suite de pasvérifiant (1.11). On peut alors se référer à [CCZ13], [CCGB15] où [GB15] pour la convergence de cet algorithme.

Cas où la médiane est inconnue

En général, on ne connait pas la médiane géométrique m et on ne peut donc pas estimer directement la Median Covariation Matrix Γ_m à l'aide de l'algorithme de Robbins-Monro et de son moyenné. Cependant, il est possible d'estimer simultanément m et Γ_m à l'aide

de deux algorithmes de gradient stochastiques et de leur versions moyennées qui évoluent simultanément. Plus précisément, pour tout $n \geq 1$, on peut considérer

$$\begin{aligned}
m_{n+1} &= m_n + \gamma_n^{(m)} \frac{X_{n+1} - m}{\|X_{n+1} - m\|}, \\
\overline{m}_{n+1} &= \overline{m}_n - \frac{1}{n+1} \left(m_{n+1} - \overline{m}_n \right), \\
V_{n+1} &= V_n + \gamma_n \frac{(X_{n+1} - \overline{m}_n) \otimes (X_{n+1} - \overline{m}_n) - V_n}{\|(X_{n+1} - \overline{m}_n) \otimes (X_{n+1} - \overline{m}_n) - V_n\|_F}, \\
\overline{V}_{n+1} &= \overline{V}_n - \frac{1}{n+1} \left(\overline{V}_n - V_{n+1} \right),
\end{aligned}$$

où $\overline{m}_1 = m_1, \overline{V}_1 = V_1$ sont bornés et les suites de pas $\left(\gamma_n^{(m)}\right)_{n\geq 1}, (\gamma_n)_{n\geq 1}$ sont de la forme $\gamma_n^{(m)} := c_m n^{-\alpha_m}$ et $\gamma_n := c_\gamma n^{-\alpha}$, avec $c_m, c_\gamma > 0$ et $\alpha_m, \alpha \in \left(\frac{1}{2}, 1\right)$. L'objectif est donc d'étudier le comportement de l'estimateur $\overline{V}_n$.

3.3.3 Résultats de convergence

Le premier théorème donne la forte consistance des algorithmes.

Théorème 3.3.1. *Supposons que les hypothèses **(A1)** à **(A4a)** sont vérifiées. Alors,*

$$\begin{aligned}
\lim_{n\to\infty} \|V_n - \Gamma_m\|_F &= 0 \quad p.s, \\
\lim_{n\to\infty} \left\|\overline{V}_n - \Gamma_m\right\|_F &= 0 \quad p.s
\end{aligned}$$

On donne maintenant la vitesse de convergence en moyenne quadratique ainsi que la vitesse L^4 de l'algorithme de gradient stochastique.

Théorème 3.3.2. *Supposons que les hypothèses **(A1)** à **(A4b)** sont vérifiées. Il existe une constante C' et pour tout $\beta \in (\alpha, 2\alpha)$, il existe une constante C_β telles que pour tout $n \geq 1$,*

$$\begin{aligned}
\mathbb{E}\left[\|V_n - \Gamma_m\|_F^2\right] &\leq \frac{C'}{n^\alpha}, \\
\mathbb{E}\left[\|V_n - \Gamma_m\|_F^4\right] &\leq \frac{C_\beta}{n^\beta}.
\end{aligned}$$

Finalement, le théorème suivant donne la vitesse de convergence en moyenne quadratique de l'algorithme moyenné.

Cet algorithme est implémenté dans le package "Gmedian" pour le langage "R".

Théorème 3.3.3. *Supposons que les hypothèses **(A1)** à **(A4b)** sont vérifiées. Il existe une constante positive C'' telle que pour tout $n \geq 1$,*

$$\mathbb{E}\left[\left\|\overline{V}_n - \Gamma_m\right\|_F^2\right] \leq \frac{C''}{n}.$$

Afin d'avoir une étude plus approfondie, il serait intéressant d'établir la normalité asymptotique de l'algorithme moyenné, ainsi que de donner les vitesses de convergence L^p des algorithmes.

3.4 Vitesse de convergence des algorithmes de Robbins-Monro et de leur moyenné. Application à la statistique robuste

On se concentre maintenant sur le problème (1.9), avec $G(h) := \mathbb{E}\left[g\left(X, h\right)\right]$, où H est un espace de Hilbert séparable. De plus, la fonction $G : H \longrightarrow \mathbb{R}$ est convexe. Nous avons vu précédemment les cas de la médiane (voir aussi les Chapitres 4 et 5) et de la "Median Covariation Matrix" (voir aussi Chapitre 6). Au Chapitre 2, nous avons vu qu'une méthode usuelle pour estimer la solution de ce type de problème, en se donnant un échantillon $X_1, ..., X_n, ...$, est de considérer le problème empirique généré par l'échantillon, i.e de considérer le M-estimateur

$$\widehat{m}_n := \arg\min_{h \in H} \frac{1}{n} \sum_{k=1}^{n} g\left(X_k, h\right),$$

et de construire $\widehat{m}_n$ à l'aide de méthodes d'optimisation déterministes usuelles. Cependant, ces méthodes peuvent être très couteuses en temps de calcul, et nécessitent de stocker en mémoire toutes les données, ce qui n'est pas toujours possible pour de gros échantillons à valeurs dans des espaces de grande dimension.

Dans ce contexte, nous avons vu au Chapitre 1 que les algorithmes de gradient stochastiques et leur version moyennée, définis par (1.10) et (1.14) sont de sérieux candidats pour contourner ce genre de problème. Dans cette partie, en s'inspirant de [CCGB15], [GB15] et [CGB15], on établit les vitesses presque sûre de ces algorithmes ainsi que leurs vitesses L^p, et ce, avec des hypothèses moins restrictives que dans la littérature (voir ([BM13]) ou ([Bac14]) parmi d'autres).

3.4.1 Hypothèses

Dans ce qui suit, on suppose que les hypothèses suivantes sont vérifiées :

(A1) La fonction g est Fréchet-différentiable par rapport à la seconde variable. De plus,

G est différentiable, et en notant Φ son gradient, il existe $m \in H$ tel que

$$\Phi(m) := \nabla G(m) = 0.$$

(A2) La fonction G est deux fois continûment différentiable et pour toute constante positive A, il existe une constante positive C_A telle que pour tout $h \in \mathcal{B}(m, A)$,

$$\|\Gamma_h\|_{op} \leq C_A,$$

où Γ_h est la Hessienne de la fonction G en h et $\|.\|_{op}$ est la norme spectrale usuelle pour les opérateurs linéaires.

(A3) Il existe une constante strictement positive ϵ telle que pour tout $h \in \mathcal{B}(m, \epsilon)$, la Hessienne Γ_h set diagonalisable. De plus, on note $\lambda_{\min}$ la limite inférieure des valeurs propres de Γ_m, alors $\lambda_{\min} > 0$. Finalement, pour tout $h \in \mathcal{B}(m, \epsilon)$, et pour toute valeur propre λ_h de Γ_h, on a $\lambda_h \geq \frac{\lambda_{\min}}{2} > 0$.

(A4) Il existe des constantes strictement positives ϵ, C_ϵ telles que pour tout $h \in \mathcal{B}(m, \epsilon)$,

$$\|\Phi(h) - \Gamma_m(h - m)\| \leq C_\epsilon \|h - m\|^2.$$

(A5) Soit $f : \mathcal{X} \times H \longrightarrow \mathbb{R}_+$ et soit C une constante positive telle que pour presque tout $x \in \mathcal{X}$ et pour tout $h \in H$, $\|\nabla_h g(x, h)\| \leq f(x, h) + C\|h - m\|$ presque sûrement, et

(a) Il existe une constante positive L_1 telle que pour tout $h \in H$,

$$\mathbb{E}\left[f(X, h)^2\right] \leq L_1.$$

(b) Pour tout entier positif p, il existe une constante positive L_p telle que pour tout $h \in H$,

$$\mathbb{E}\left[f(X, h)^{2p}\right] \leq L_p.$$

L'hypothèse **(A1)** est cruciale pour introduire un algorithme de gradient stochastique, tandis que les hypothèses **(A2)** et **(A3)** assurent la forte convexité locale de la fonction G (et non globale comme dans [BM13]), tout en ayant un contrôle sur la "perte" de forte convexité. Notons que ces hypothèses sont aussi présentes dans [Bac14], mais de manière plus restrictive. L'hypothèse **(A4)** permet de borner localement le terme de reste dans la décomposition de Taylor du gradient de G. Finalement, l'hypothèse **(A5)** permet, au lieu de borner uniformément le gradient de g (voir [Bac14]), de séparer la majoration en deux parties : une dont l'espérance est bornée uniformément, et une autre qui dépend de l'erreur d'estimation.

Remarque 3.4.1. *On a ici des hypothèses assez proches de celles introduites au Chapitre 1 et par*

[Pel00] pour obtenir les vitesses de convergence presque sûre des algorithmes mais avec des résultats valables pour des espaces de dimension infinie. De plus, ces hypothèses sont clairement moins restrictives que celles introduites par [BM13] ou [Bac14] pour l'obtention des vitesses de convergence en moyenne quadratique.

Remarque 3.4.2. *Soient H, H' deux espaces vectoriels normés, et $\mathcal{U}$ un ouvert de H. Soit $a \in \mathcal{U}$, une fonction $f : \mathcal{U} \longrightarrow H'$ est Fréchet-différentiable en a si il existe une application linéaire $Df : H \longrightarrow H'$ telle que*

$$\lim_{h \to 0} \frac{f(a+h) - f(a) - Df(h)}{\|h\|} = 0.$$

3.4.2 Vitesses de convergence

On donne maintenant les vitesses de convergence des algorithmes définis par (1.6) et (1.14).

Théorème 3.4.1. *On suppose que les hypothèses **(A1)** à **(A5a)** sont vérifiées. Alors, pour tout $\delta, \delta' > 0$,*

$$\|Z_n - m\|^2 = o\left(\frac{(\ln n)^\delta}{n^\alpha}\right) \quad p.s,$$
$$\left\|\overline{Z}_n - m\right\|^2 = o\left(\frac{(\ln n)^{1+\delta'}}{n}\right), \quad p.s.$$

Remarque 3.4.3. *Notons que des résultats analogues sont donné dans [Pel98] et [Pel00], mais seulement dans le cas d'espace de dimension finie, et les preuves ne peuvent pas être directement appliquées en dimension infinie. En effet, elles reposent sur le fait que les sous-espaces vectoriels de la Hessienne G soient de dimension finie, et que la trace d'une matrice existe, ce qui n'est pas toujours le cas en dimension infinie.*

En s'inspirant de [CCGB15], [GB15] et [CGB15], on obtient les vitesses L^p de l'algorithme de gradient stochastique.

Théorème 3.4.2. *On suppose que les hypothèses **(A1)** à **(A5b)** sont vérifiées. Alors, pour tout entier positif p, il existe une constante positive K_p telle que pour tout $n \geq 1$,*

$$\mathbb{E}\left[\|Z_n - m\|^{2p}\right] \leq \frac{K_p}{n^{p\alpha}}.$$

Finalement, on obtient les vitesses de convergence L^p de l'algorithme moyenné.

Théorème 3.4.3. *On suppose que les hypothèses **(A1)** à **(A5b)** sont vérifiées. Alors, pour tout entier*

positif p, il existe une constante positive K'_p telle que pour tout $n \geq 1$,

$$\mathbb{E}\left[\left\|\overline{Z}_n - m\right\|^{2p}\right] \leq \frac{K'_p}{n^p}.$$

Remarque 3.4.4. *Avec une restriction sur la suite de pas (γ_n), on peut remplacer l'hypothèse **(A4)** par une hypothèse plus usuelle. En effet, soit $\beta \in (1,2]$, on obtiendrait la même vitesse de convergence en moyenne quadratique et presque sûre pour l'algorithme de Robbins-Monro en remplaçant l'hypothèse **(A4)** par*

$$\|\Phi(h) - \Gamma_m (h-m)\| \leq \|h-m\|^\beta$$

pour tout $h \in \mathcal{B}(m,\epsilon)$. De plus, on aurait les mêmes vitesses pour l'algorithme moyenné en prenant une suite de pas de la forme $\gamma_n := c_\gamma n^{-\alpha}$ avec $\alpha \in \left(\frac{1}{\beta}, 1\right)$.

3.4.3 Applications

Application aux quantiles géométriques

Soit H un espace de Hilbert séparable et X une variable aléatoire à valeurs dans H. Le quantile géométrique m^v correspondant à la direction v, où $v \in H$ et $\|v\| \leq 1$, est défini par

$$m^v := \arg\min_{h \in H} \mathbb{E}\left[\|X-h\| - \|X\|\right] - \langle h, v\rangle . \tag{3.8}$$

Notons que si $v = 0$, on retrouve alors la médiane géométrique (voir la Section 3.1.1). On note G_v la fonction que l'on veut minimiser, et elle est définie pour tout $h \in H$ par $G_v(h) := \mathbb{E}\left[\|X-h\| + \langle X-h, v\rangle\right]$. On suppose maintenant que les hypothèses de la Section 3.1.1 sont vérifiées. Alors m^v est bien défini, unique (voir [Cha96]) et est solution de l'équation

$$\Phi_v(h) := \nabla G_v(h) = -\mathbb{E}\left[\frac{X-h}{\|X-h\|}\right] - v = 0.$$

Les hypothèses **(A1)** et **(A5b)** sont alors vérifiées, et les algorithmes définis pour tout $n \geq 1$ par

$$\begin{aligned} m^v_{n+1} &= m^v_n + \gamma_n \left(\frac{X_{n+1} - m^v_n}{\|X_{n+1} - m^v_n\|} + v\right), \\ \overline{m}_{n+1} &= \overline{m}_n + \frac{1}{n+1}\left(m_{n+1} - \overline{m}_n\right), \end{aligned}$$

vérifient les théorèmes précédents.

Application à la régression logistique robuste

Soit $d \geq 1$ et $H = \mathbb{R}^d$. Soit (X, Y) une paire de variables aléatoires à valeurs dans $H \times \{-1, 1\}$, on veut minimiser la fonction G_r définie pour tout $h \in \mathbb{R}^d$ par (voir [Bac14])

$$G_r(h) := \mathbb{E}\left[\log\left(\cosh\left(Y - \langle X, h\rangle\right)\right)\right]. \tag{3.9}$$

Afin d'assurer l'existence et l'unicité de la solution, on suppose maintenant que les hypothèses suivantes sont vérifiées :

(B1) Il existe $m^r \in \mathbb{R}^d$ tel que $\nabla G_r\left(m^r\right) = 0$.

(B2) La Hessienne de la fonction G_r en m^r est définie positive.

(B3) Pour tout entier positif p, la variable aléatoire X admet un moment d'ordre p.

L'hypothèse **(B1)** assure l'existence d'une solution, tandis que l'hypothèse **(B2)** assure son unicité. Finalement, l'hypothèse **(B3)** permet de vérifier **(A5)**. Les hypothèses **(A1)** à **(A5b)** sont alors vérifiées et les algorithmes définis pour tout $n \geq 1$ par

$$\begin{aligned} m^r_{n+1} &= m^r_n + \gamma_n \frac{\sinh\left(Y_{n+1} - \langle X_{n+1}, m^r_n\rangle\right)}{\cosh\left(Y_{n+1} - \langle X_{n+1}, m^r_n\rangle\right)} X_{n+1}, \\ \overline{m}^r_{n+1} &= \overline{m}^r_n + \frac{1}{n+1}\left(m^r_{n+1} - \overline{m}^r_n\right) \end{aligned}$$

vérifient alors les théorèmes précédents.

3.5 Un algorithme de Robbins-Monro projeté et sa version moyennée pour l'estimation des paramètres d'une distribution sphérique tronquée

Dans cette partie, on propose un algorithme permettant d'ajuster une sphère à un nuage de points 3D distribué autour d'une sphère complète ou tronquée (voir [BP14]). Pour ce faire, on suppose que les observations sont des réalisations indépendantes et identiquement distribuées d'un vecteur aléatoire X défini comme

$$X := \mu + rWU_\Omega, \tag{3.10}$$

où W est une variable aléatoire réelle positive telle que $\mathbb{E}[W] = 1$ et U_Ω est uniformément distribuée sur une partie Ω de la sphère unité de $\mathbb{R}^d$, avec $d \geq 2$. Les paramètres $\mu \in \mathbb{R}^d$ et $r > 0$ sont respectivement le centre et le rayon de la sphère que l'on veut ajuster sur le nuage de points.

Afin d'estimer ces paramètres, on introduit un algorithme de gradient stochastique projeté et sa version moyennée. On établit des résultats asymptotiques tels que la forte consistance de ces algorithmes ainsi que la normalité asymptotique du moyenné. De plus, quelques résultats non-asymptotiques sont donnés, telles que les vitesses de convergence en moyenne quadratique.

3.5.1 Cadre de travail

Hypothèses

Dans ce qui suit, on suppose que les variables aléatoires W et U_Ω définies en introduction sont indépendantes. On s'intéresse à l'estimation des paramètres μ et $r^\star := r\mathbb{E}[W]$, et on note $\theta := (\mu, r^\star)$. On suppose à partir de maintenant que les hypothèses suivantes sont vérifiées :

(H1) La variable aléatoire X n'est pas concentrée autour du centre μ :

$$\mathbb{E}\left[\frac{1}{\|X-\mu\|^2}\right] < +\infty.$$

(H2) La variable aléatoire X admet un moment du second ordre :

$$\mathbb{E}\left[\|X-\mu\|^2\right] < +\infty.$$

Ces hypothèses assurent que la variable aléatoire X n'est pas concentrée autour du centre mais bien autour de la sphère, et ce, sans trop de dispersion.

Quelques propriétés

Afin d'introduire deux algorithmes qui permettent d'estimer le paramètre θ, on commence par remarquer que le paramètre θ peut être vu comme un minimiseur local d'une fonction. En effet, soit $G : \mathbb{R}^d \times \mathbb{R}_+^\star \longrightarrow \mathbb{R}_+$ définie pour tout $y = (z, a) \in \mathbb{R}^d \times \mathbb{R}_+^*$ par

$$G(y) := \frac{1}{2}\mathbb{E}\left[(\|X-z\| - a)^2\right]. \tag{3.11}$$

La fonction G est Fréchet-différentiable et on note $\Phi(.)$ son gradient, qui est défini pour tout $y = (z, a) \in \mathbb{R}^d \times \mathbb{R}_+^*$ par

$$\Phi(y) := \nabla G(y) = \begin{pmatrix} z - \mathbb{E}[X] - a\mathbb{E}\left[\frac{z-X}{\|z-X\|}\right] \\ a - \mathbb{E}\left[\|z-X\|\right] \end{pmatrix}. \tag{3.12}$$

Remarquons que l'on a en particulier $\Phi(\theta) = 0$. Une idée serait donc d'introduire un algorithme de gradient stochastique pour estimer θ. Cependant, la fonction que l'on veut minimiser n'est pas convexe sur $\mathbb{R}^d \times \mathbb{R}_+$ et l'algorithme de Robbins-Monro ne converge donc pas nécessairement. On va donc chercher à le projeter sur un sous-espace avec de bonnes propriétés. On suppose à partir de maintenant que l'hypothèse suivante est vérifiée :

(H3) Il existe des constantes positives R_μ et R_r telles que pour tout $y = (z, a) \in \overline{\mathcal{B}(\mu, R_\mu)} \times \overline{\mathcal{B}(r^\star, R_r)}$,

$$\sup_{z \in \overline{\mathcal{B}(\mu, R_\mu)}} \lambda_{\max}(\Gamma(z)) < \frac{1 - \|\mathbb{E}[U_\Omega]\|^2 / A}{r^\star + \frac{3}{2} R_r},$$

avec A telle que $\|\mathbb{E}[U_\Omega]\|^2 < A < 1$, et $\lambda_{\max}(M)$ donne la plus grande valeur propre de la matrice M et

$$\Gamma(z) := \mathbb{E}\left[\frac{1}{\|X - z\|}\left(I_d - \frac{(X-z)(X-z)^T}{\|X-z\|^2}\right)\right].$$

Notons que cette hypothèse est vérifiée dès que la sphère n'est pas trop tronquée et que la variable aléatoire n'est pas trop dispersée autour de la sphère. En effet, dans le cas de la sphère complète, on a

$$\begin{aligned}\Gamma(\theta) &= \mathbb{E}\left[\frac{1}{W}\right](I_d - \mathbb{E}[U \otimes U]) \\ &= \frac{2}{3}\mathbb{E}\left[\frac{1}{W}\right] I_d.\end{aligned}$$

La principale conséquence de cette hypothèse est la proposition suivante, qui est cruciale pour introduire un algorithme de gradient projeté et assurer sa convergence.

Proposition 3.5.1. *Supposons que les hypothèses **(H1)** à **(H3)** sont vérifiées. Alors, il existe une constante positive c telle que pour tout $y \in \overline{\mathcal{B}(\mu, R_\mu)} \times \overline{\mathcal{B}(r^\star, R_r)}$,*

$$\langle \Phi(y), y - \theta \rangle \geq c \|y - \theta\|^2.$$

3.5.2 Les algorithmes

Afin d'introduire un algorithme projeté, dans ce qui suit, on considère un sous ensemble $\mathcal{K}$ de $\overline{\mathcal{B}(\mu, R_\mu)} \times \overline{\mathcal{B}(r^\star, R_r)}$ compact et convexe, et tel que $\theta \notin \partial\mathcal{K}$. On verra en simulation (voir Section 8.5) comment construire un tel compact. De plus, on se donne une projection π

sur $\mathcal{K}$ telle que pour tout $y \in \mathcal{K}$, on ait $\pi(y) = y$ et pour tout $y \notin \mathcal{K}$, on ait $\pi(y) \in \partial\mathcal{K}$. Enfin, la projection π est 1-lippschitzienne, i.e pour tout $y, y' \in \mathbb{R}^d \times \mathbb{R}_+^\star$,

$$\left\|\pi(y) - \pi(y')\right\| \leq \left\|y - y'\right\|.$$

Notons que l'on peut prendre, par exemple, la projection euclidienne sur $\mathcal{K}$. On considère maintenant des variables aléatoires $X_1, ..., X_n, ...$ indépendantes et de même loi que X et on introduit l'algorithme de Robbins-Monro projeté (PRM) (voir Section 1.2.3), défini récursivement pour tout $n \geq 1$ par

$$\widehat{\theta}_{n+1} = \pi\left(\widehat{\theta}_n - \gamma_n \nabla_y g\left(X_{n+1}, \widehat{\theta}_n\right)\right), \tag{3.13}$$

où pour tout $y = (z, a) \in \mathbb{R}^d \times \mathbb{R}_+^\star$ et $x \in \mathbb{R}^d$,

$$\nabla_y g\left(x, y\right) := \begin{pmatrix} z - x - a\frac{z-x}{\|z-x\|} \\ a - \|z - x\| \end{pmatrix}.$$

De plus, on prend $\widehat{\theta}_1 \in \mathcal{K}$ et une suite de pas $(\gamma_n)_{n\geq 1}$ vérifiant (1.11). Afin d'améliorer la convergence, on peut maintenant introduire l'algorithme moyenné défini récursivement pour tout $n \geq 1$ par

$$\overline{\theta}_{n+1} = \overline{\theta}_n + \frac{1}{n+1}\left(\widehat{\theta}_{n+1} - \overline{\theta}_n\right), \tag{3.14}$$

avec $\overline{\theta}_1 = \widehat{\theta}_1$.

3.5.3 Vitesses de convergence

On donne maintenant donner la vitesse de convergence en moyenne quadratique et les vitesses L^p (sous conditions) de l'algorithme projeté ainsi qu'une majoration de la probabilité avec laquelle le vecteur aléatoire $\widehat{\theta}_n - \gamma_n \nabla_y g\left(X_{n+1}, \widehat{\theta}_n\right)$ sort du compact $\mathcal{K}$.

Théorème 3.5.1. *Supposons que les hypothèses* ***(H1)*** *à* ***(H3)*** *sont vérifiées. Alors, il existe une constante positive C_1 telle que pour tout $n \geq 1$,*

$$\mathbb{E}\left[\left\|\widehat{\theta}_n - \theta\right\|^2\right] \leq \frac{C_1}{n^\alpha}.$$

De plus, soit p un entier positif, supposons que $\mathbb{E}\left[\|X - \mu\|^{2p}\right] < +\infty$, alors il existe une constante C_p telle que pour tout $n \geq 1$,

$$\mathbb{E}\left[\left\|\widehat{\theta}_n - \theta\right\|^{2p}\right] \leq \frac{C_p}{n^{p\alpha}},$$

et pour tout $n \geq 1$,

$$\mathbb{P}\left[\widehat{\theta}_n - \gamma_n \nabla_y g\left(X_{n+1}, \widehat{\theta}_n\right) \notin \mathcal{K}\right] \leq \frac{C_p}{d_{\min}^{2p} n^{p\alpha}},$$

où $d_{\min} := \inf_{y \in \partial\mathcal{K}} \{\|y - \theta\|\} > 0$.

Afin de donner la vitesse de convergence de l'algorithme moyenné, on introduit maintenant une dernière hypothèse :

(H4) La Hessienne de la fonction G en θ, notée Γ_θ et définie par

$$\Gamma_\theta := \begin{pmatrix} I_d - r^\star \mathbb{E}\left[\frac{1}{\|X-\mu\|}\left(I_d - \frac{(X-\mu)\otimes(X-\mu)}{\|X-\mu\|^2}\right)\right] & \mathbb{E}\left[\frac{X-\mu}{\|X-\mu\|}\right] \\ \mathbb{E}\left[\frac{X-\mu}{\|X-\mu\|}\right]^T & 1 \end{pmatrix},$$

est définie positive.

Le théorème suivant donne alors la convergence en moyenne quadratique de l'algorithme moyenné.

Théorème 3.5.2. *Supposons que les hypothèses* ***(H1)*** *à* ***(H4)*** *sont vérifiées et que* $\mathbb{E}\left[\|X-\mu\|^{12}\right] < +\infty$. *Alors, il existe une constante positive* C *telle que pour tout* $n \geq 1$,

$$\mathbb{E}\left[\left\|\widehat{\theta}_n - \theta\right\|^2\right] \leq \frac{C}{n}.$$

Notons qu'à notre connaissance, les résultats sur les algorithmes projetés et moyennés sont rares dans la littérature. Enfin, on peut donner la normalité asymptotique de l'estimateur moyenné.

Théorème 3.5.3. *Supposons que les hypothèses* ***(H1)*** *à* ***(H4)*** *sont vérifiées et que* $\mathbb{E}\left[\|X-\mu\|^{12}\right] < +\infty$. *Alors,*

$$\lim_{n\to\infty} \sqrt{n}\left(\overline{\theta}_n - \theta\right) \sim \mathcal{N}\left(0, \Gamma_\theta^{-1}\Sigma\Gamma_\theta^{-1}\right),$$

avec

$$\Sigma := \mathbb{E}\left[\begin{pmatrix} \mu - X - r^\star \frac{(\mu-X)}{\|\mu-X\|} \\ r^\star - \|\mu - X\| \end{pmatrix}\begin{pmatrix} \mu - X - r^\star \frac{(\mu-X)}{\|\mu-X\|} \\ r^\star - \|\mu - X\| \end{pmatrix}^T\right].$$

En particulier, on a

$$\lim_{n\to\infty} \sqrt{n}\Sigma^{-1/2}\Gamma_\theta\left(\overline{\theta}_n - \theta\right) \sim \mathcal{N}\left(0, I_{d+1}\right).$$

Notons que contrairement à [BP14], on peut obtenir pour les vitesses de convergence des algorithmes, et ce, même dans le cas de la sphère tronquée.

Première partie

Estimation récursive de la médiane géométrique dans les espaces de Hilbert

Chapitre 4

Online estimation of the geometric median in Hilbert spaces : non asymptotic confidence balls

Résumé

On a vu au Chapitre 1 que les algorithmes de gradient stochastique et leur version moyennée sont des outils efficaces pour traiter de gros échantillons à valeurs dans des espaces de grandes dimensions. De plus, on a vu au Chapitre 2 que contrairement à la moyenne, la médiane géométrique est un indicateur de position robuste. Dans ce contexte, on s'intéresse aux estimateurs de la médiane développé par [CCZ13]. Des résultats de convergence ont déjà été établis, tels que la normalité asymptotique, mais ne donnent aucune garantie sur le comportement des algorithmes pour une taille d'échantillon n fixée. Ce travail vise à étudier plus précisément le comportement non asymptotique de cet algorithme non linéaire en donnant des boules de confiance non-asymptotiques dans des espaces de Hilbert.

Pour ce faire, on se concentre dans un premiers temps sur les vitesses de convergence de l'algorithme de type Robbins-Monro. Plus précisément, on en donne la vitesse de convergence en moyenne quadratique ainsi qu'une majoration de la vitesse L^4 (Théorème 4.3.1). Dans un deuxième temps, on introduit une nouvelle inégalité exponentielle pour des termes qui sont presque des martingales (Proposition 4.4.1). Cela permet alors d'obtenir des boules de confiances non asymptotiques pour l'algorithme de gradient stochastique (Théorème 4.4.1) ainsi que pour son moyenné (Théorème 4.4.2).

This Chapter is based on a work with Hervé Cardot and Peggy Cénac accepted in the Annals of Statistics ([CCGB15]).

Abstract

Estimation procedures based on recursive algorithms are interesting and powerful techniques that are able to deal rapidly with very large samples of high dimensional data. The collected data may be contaminated by noise so that robust location indicators, such as the geometric median, may be preferred to the mean. In this context, an estimator of the geometric median based on a fast and efficient averaged non linear stochastic gradient algorithm has been developed by [CCZ13]. This work aims at studying more precisely the non asymptotic behavior of this non linear algorithm by giving non asymptotic confidence balls in general separable Hilbert spaces. This new result is based on the derivation of improved L^2 rates of convergence as well as an exponential inequality for the nearly martingale terms of the recursive non linear Robbins-Monro algorithm.

4.1 Introduction

Dealing with large samples of observations taking values in high dimensional spaces, such as functional spaces, is not unusual nowadays. In this context, simple estimators of location such as the arithmetic mean can be greatly influenced by a small number of outlying values and robust indicators of location may be preferred to the mean. We focus in this work on the estimation of the geometric median, also called L^1-median or spatial median. It is a multivariate generalization of the real median introduced by [Hal48] that can be defined in general metric spaces.

Let H be a separable Hilbert space, we denote by $\langle .,. \rangle$ its inner product and by $\|.\|$ the associated norm. Let X be a random variable taking values in H, the geometric median m of X is defined by :

$$m := \arg\min_{h \in H} \mathbb{E}\left[\|X - h\| - \|X\|\right]. \tag{4.1}$$

Many properties of this median in the the general setting of separable Banach spaces, such as existence and uniqueness, as well as robustness are given in [Kem87] (see also the review [Sma90]). Recently, this median has received much attention in the literature. For example, [Min14] suggests to consider, in various statistical contexts, the geometric median of independent estimators in order to obtain much tighter concentration bounds. In functional data analysis, [KP12] consider resistant estimators of the covariance operator based on the geometric median in order to derive a robust test of equality of the second-order structure for two samples. The geometric median is also chosen to be the central location indicator in various types of robust functional principal components analysis (see [LMS$^+$99], [Ger08] and [BBT$^+$11]). The posterior geometric median of estimators has also been used in a robust bayesian context by [MSLD14]. Finally, a general definition of the geometric median on manifolds is given in [FVJ09] and [ADPY12] with signal processing issues in mind.

Consider a sequence of i.i.d copies $X_1, X_2, \ldots, X_n, \ldots$ of X. A natural estimator $\widehat{m}_n$ of m, based on $X_1, \ldots, X_n$, is obtained by minimizing the empirical risk

$$\widehat{m}_n := \arg\min_{h \in H} \sum_{i=1}^{n} \left[\|X_i - h\| - \|X_i\|\right]. \tag{4.2}$$

Convergence properties of the empirical estimator $\widehat{m}_n$ are reviewed in [MNO10] when the dimension of H is finite whereas the recent work of [CC14] proposes a deep asymptotic study for random variables taking values in separable Banach spaces. Given a sample $X_1, \ldots, X_n$ the computation of $\widehat{m}_n$ generally relies on a variant of the Weiszfeld's algorithm (see *e.g.* [Wei37b] and [Kuh73]) introduced by [VZ00]. This iterative algorithm is relatively fast (see

[BS15] for an improved version) but it is not adapted to handle very large datasets of high-dimensional data since it requires to store all the data in memory.

However huge datasets are not unusual anymore with the development of automatic sensors and smart meters. In this context, [CCZ13] have developed a much faster algorithm, which thanks to its recursive nature does not require to store all the data and can be updated automatically when the data arrive sequentially. The estimation procedure is based on the simple following recursive scheme,

$$Z_{n+1} = Z_n + \gamma_n \frac{X_{n+1} - Z_n}{\|X_{n+1} - Z_n\|} \tag{4.3}$$

where the sequence of steps (γ_n) controls the convergence of the algorithm and satisfy the usual conditions for the convergence of Robbins Monro algorithms (see Section 4.3). The averaged version of the algorithm is defined as follows

$$\overline{Z}_{n+1} = \overline{Z}_n + \frac{1}{n+1}\left(Z_{n+1} - \overline{Z}_n\right), \tag{4.4}$$

with $\overline{Z}_0 = 0$, so that $\overline{Z}_n = \frac{1}{n}\sum_{i=1}^{n} Z_i$. The averaging step described in (4.4), and first studied in [PJ92], allows a considerable improvement of the convergence compared to the initial Robbins-Monro algorithm described in (4.3). It is shown in [CCZ13] that the recursive averaged estimator $\overline{Z}_n$ and the empirical estimator $\widehat{m}_n$ have the same Gaussian limiting distribution.

However the asymptotic normality shown in [CCZ13] does not give any clue of how far the distribution of the estimator is from its asymptotic law for any fixed sample size n. The aim of this work is to give new arguments in favor of the averaged stochastic estimator of the geometric median by providing a sharp control of its deviations around the true median, for finite samples. Indeed the obtention of finite sample guarantees with high probability is always desirable for the statisticians who have to study real data, since the samples under study will always have a finite sample size. Nice arguments for considering non asymptotic properties of estimators are given for example in [RV10]. The obtention of such results generally requires much more mathematical efforts compared to more classical weak convergence results as well as more restrictive conditions on the existence of all the moments of the variable (see for example [Woo72] or [TY14]). Note also that, as far as we know, there are only very few results in the literature on non asymptotic bounds for non linear recursive algorithms (see however [BDF13] for recursive PCA or [BM13]).

The construction of our non asymptotic confidence balls (see Theorem 4.4.1 and Theo-

rem 4.4.2) rely on the obtention of the optimal rate of convergence in quadratic mean (see Theorem 4.3.1) of the Robbins-Monro algorithm used for estimating the geometric median as well as new exponential inequalities for "near" martingale sequences in Hilbert spaces (see Proposition 4.4.1), similar to the seminal result of [Pin94] for martingales. These properties do not require any additional conditions on the moments of the data to hold The proof of Theorem 4.3.1 is based on a new approach which consists in obtaining first, relations between the L^2 and the L^4 estimation errors and then make an induction using these relations to get the optimal rate of convergence in quadratic mean of Robbins-Monro algorithms. This new approach may give keys to obtain non asymptotic results when the objective function only possesses locally strong convexity properties.

The paper is organized as follows. Section 4.2 recalls some convexity properties of the geometric median as well as the basic assumptions ensuring the uniqueness of the geometric median. In Section 4.3, the rates of convergence of the stochastic gradient algorithm are derived in quadratic mean as well as in L^4. In Section 4.4, an exponential inequality is derived borrowing ideas from [TY14]. It enables us to build non asymptotic confidence balls for the Robbins-Monro algorithm as well as its averaged version. The most innovative part of the proofs is given in Section 4.5 whereas the other technical details are gathered in a supplementary file.

4.2 Assumptions on the median and convexity properties

Let us first state basic assumptions on the median.

(A1) The random variable X is not concentrated on a straight line : for all $h \in H$, there exists $h' \in H$ such that $\langle h, h' \rangle = 0$ and

$$\mathrm{Var}\left(\langle h', X \rangle\right) > 0.$$

(A2) X is not concentrated around single points : there is a constant $C > 0$ such that for all $h \in H$:

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

Assumption **(A1)** ensures that the median m is uniquely defined ([Kem87]). Assumption **(A2)** is closely related to small ball probabilities and to the dimension of H. It was proved in [Cha92] that when $H = \mathbb{R}^d$, assumption **(A2)** is satisfied when $d \geq 2$ under classical assumptions on the density of X. A detailed discussion on assumption **(A2)** and its connection with small balls probabilities can be found in [CCZ13].

We now recall some results about convexity and robustness of the geometric median. We denote by $G : H \longrightarrow \mathbb{R}$ the convex function we would like to minimize, defined for all $h \in H$ by

$$G(h) := \mathbb{E}\left[\|X - h\| - \|X\|\right]. \tag{4.5}$$

This function is Fréchet differentiable on H, we denote by Φ its Fréchet derivative, and for all $h \in H$:

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

Under previous assumptions, m is the unique zero of Φ.

Let us define $U_{n+1} := -\frac{X_{n+1}-Z_n}{\|X_{n+1}-Z_n\|}$ and let us introduce the sequence of σ-algebra $\mathcal{F}_n := \sigma(Z_1, ..., Z_n) = \sigma(X_1, ..., X_n)$. For all integer $n \geq 1$,

$$\mathbb{E}\left[U_{n+1}|\mathcal{F}_n\right] = \Phi(Z_n). \tag{4.6}$$

The sequence $(\xi_n)_n$ defined by $\xi_{n+1} := \Phi(Z_n) - U_{n+1}$ is a martingale difference sequence with respect to the filtration $(\mathcal{F}_n)$. Moreover, we have for all n, $\|\xi_{n+1}\| \leq 2$ and

$$\mathbb{E}\left[\|\xi_{n+1}\|^2|\mathcal{F}_n\right] \leq 1 - \|\Phi(Z_n)\|^2 \leq 1. \tag{4.7}$$

Algorithm (4.3) can be written as a Robbins-Monro or a stochastic gradient algorithm :

$$Z_{n+1} - m = Z_n - m - \gamma_n \Phi(Z_n) + \gamma_n \xi_{n+1}. \tag{4.8}$$

We now consider the Hessian of G, which is denoted by $\Gamma_h : H \longrightarrow H$. It satisfies (see [Ger08])

$$\Gamma_h = \mathbb{E}\left[\frac{1}{\|X-h\|}\left(I_H - \frac{(X-h)\otimes(X-h)}{\|X-h\|^2}\right)\right],$$

where I_H is the identity operator in H and $u \otimes v(h) = \langle u, h\rangle v$ for all $u, v, h \in H$. The following (local) strong convexity properties will be useful (see [CCZ13] for proofs).

Proposition 4.2.1 ([CCZ13]). *Under assumptions **(A1)** and **(A2)**, for any real number $A > 0$, there is a positive constant c_A such that for all $h \in H$ with $\|h\| \leq A$, and for all $h' \in H$:*

$$c_A\|h'\|^2 \leq \langle h', \Gamma_h h'\rangle \leq C\|h'\|^2.$$

As a particular case, there is a positive constant c_m such that for all $h' \in H$:

$$c_m\|h'\|^2 \leq \langle h', \Gamma_m h'\rangle \leq C\|h'\|^2. \tag{4.9}$$

The following corollary recall some properties of the spectrum of the Hessian of G, in particular on the spectrum of Γ_m.

Corollary 4.2.1. *Under assumptions **(A1)** and **(A2)**, for all $h \in H$, there is an increasing sequence of non-negative eigenvalues $(\lambda_{j,h})$ and an orthonormal basis $(v_{j,h})$ of eigenvectors of Γ_h such that*

$$\begin{aligned} \Gamma_h v_{j,h} &= \lambda_{j,h} v_{j,h}, \\ \sigma(\Gamma_h) &= \{\lambda_{j,h}, j \in \mathbb{N}\}, \\ \lambda_{j,h} &\leq C. \end{aligned}$$

Moreover, if $\|h\| \leq A$, for all $j \in \mathbb{N}$ we have $c_A \leq \lambda_{j,h} \leq C$.

As a particular case, the eigenvalues $\lambda_{j,m}$ of Γ_m satisfy, $c_m \leq \lambda_{j,m} \leq C$, for all $j \in \mathbb{N}$.

The bounds are an immediate consequence of Proposition 4.2.1. Remark that with these different convexity properties of the geometric median, we are close to the framework of [Bac14]. The difference comes from the fact that G does not satisfy the generalized self-concordance assumption which is central in the latter work.

4.3 Rates of convergence of the Robbins-Monro algorithms

If the sequence $(\gamma_n)_n$ of stepsizes fulfills the classical following assumptions :

$$\sum_{n\geq 1} \gamma_n^2 < \infty \quad \text{and} \quad \sum_{n\geq 1} \gamma_n = \infty,$$

and **(A1)** and **(A2)** hold, the recursive estimator Z_n is strongly consistent (see [CCZ13], Theorem 3.1). The first condition on the stepsizes ensures that the recursive algorithm converges towards some value in H whereas the second condition forces the algorithm to converge to m, the unique minimizer of G.

From now on, Z_1 is chosen so that it is bounded (consider for example $Z_1 = X_1 \mathbb{1}_{\{\|X\| \leq M'\}}$ for some non negative constant M'). Consequently, there is a positive constant M such that for all $n \geq 1$:

$$\mathbb{E}\left[\|Z_n - m\|^2\right] \leq M.$$

Let us consider now sequences $(\gamma_n)_n$ of the form $\gamma_n = c_\gamma n^{-\alpha}$ where c_γ is a positive constant, and $\alpha \in (1/2, 1)$. Note that considering $\alpha = 1$ would be possible, with a suitable constant c_γ which is unknown in practice, in order to obtain the optimal parametric rate of convergence. The algorithm can very sensitive to the values c_γ. That is why we prefer to introduce an averaging step with $\alpha < 1$, which is in practice and theoretically more efficient,

since it has the same asymptotic variance as the empirical risk minimizer ([CCZ13], Theorem 3.4).

In order to get confidence balls for the median, the following additional assumption is supposed to hold.

(A3) There is a positive constant C such that for all $h \in H$:

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

This assumption ensures that the remainder term in the Taylor approximation to the gradient is bounded. Note that this assumption is also required to get the asymptotic normality in [CCZ13]. It is also assumed in [CC14] for deriving the asymptotic normality of the empirical median estimator. Remark that for the sake of simplicity, we have considered the same constant C in **(A2)** and **(A3)**. As in **(A2)**, Assumption **(A3)** is closely related to small ball probabilities and when $H = \mathbb{R}^d$, this assumption is satisfied when $d \geq 3$ under weak conditions.

We state now the first new and important result on the rates of convergence in quadratic mean of the Robbins Monro algorithm. A comparison with Proposition 3.2 in [CCZ13] reveals that the term $\log n$ has disappeared as well as the constant C_N that was related to a sequence $(\Omega_N)_N$ of events whose probability was tending to one. This is a significant improvement which is crucial to get a deep study of the estimators and to get non asymptotic results.

Theorem 4.3.1. *Assuming **(A1)-(A3)** hold, the algorithm (Z_n) defined by (4.3), with $\gamma_n = c_\gamma n^{-\alpha}$, converges in quadratic mean, for all $\alpha \in (1/2, 1)$ and for all $\alpha < \beta < 3\alpha - 1$, with the following rate :*

$$\mathbb{E}\left[\|Z_n - m\|^2\right] = O\left(\frac{1}{n^\alpha}\right), \tag{4.10}$$

$$\mathbb{E}\left[\|Z_n - m\|^4\right] = O\left(\frac{1}{n^\beta}\right). \tag{4.11}$$

Upper bounds for the rates of convergence at order four are also given because they will be useful in several proofs. Remark that obtaining better rates of convergence at the order four would also be possible at the expense of longer proofs, and since it is not necessary here, it is not given.

The proof of this theorem relies on a new approach which consists in an induction on n using two decompositions of the algorithm which enables us to obtain an upper bound of the quadratic mean error and the L^4 error. Note that this approach can be used in several cases when the function we would like to minimize is only locally strongly convex.

Lemma 4.3.1. *Assuming **(A1)-(A3)** hold, there are positive constants C_1, C_2, C_3, C_4 such that for all $n \geq 1$:*

$$\mathbb{E}\left[\|Z_n - m\|^2\right] \leq C_1 e^{-C_4 n^{1-\alpha}} + \frac{C_2}{n^\alpha} + C_3 \sup_{n/2-1 \leq k \leq n} \mathbb{E}\left[\|Z_k - m\|^4\right]. \tag{4.12}$$

The proof of Lemma 4.3.1 is given in Section 4.5. In order to get a rate of convergence of the last term in previous inequality, we use a second decomposition (see equation (4.8)), to get a bound of the 4-th moment.

Lemma 4.3.2. *Assuming the three assumptions **(A1)** to **(A3)**, for all $\alpha \in (1/2, 1)$, there are a rank n_α and positive constants C_1', C_2' such that for all $n \geq n_\alpha$:*

$$\mathbb{E}\left[\|Z_{n+1} - m\|^4\right] \leq \left(1 - \frac{1}{n}\right)^2 \mathbb{E}\left[\|Z_n - m\|^4\right] + \frac{C_1'}{n^{3\alpha}} + C_2' \frac{1}{n^{2\alpha}} \mathbb{E}\left[\|Z_n - m\|^2\right]. \tag{4.13}$$

The proof of Lemma 4.3.2 is given in Section 4.5. The next result gives the exact rate of convergence in quadratic mean and states that it is not possible to get the parametric rates of convergence with the Robbins Monro algorithm when $\alpha \in (1/2, 1)$.

Proposition 4.3.1. *Assume **(A1)-(A3)** hold, for all $\alpha \in (1/2, 1)$, there is a positive constant C' such that for all $n \geq 1$,*

$$\mathbb{E}\left[\|Z_n - m\|^2\right] \geq \frac{C'}{n^\alpha}.$$

The proof of Proposition 4.3.1 is given in the supplementary file.

4.4 Non asymptotic confidence balls

4.4.1 Non asymptotic confidence balls for the Robbins-Monro algorithm

The aim is now to derive an upper bound for $\mathbb{P}\left[\|Z_n - m\| \geq t\right]$, for $t > 0$. A simple and first result can be obtained by applying Markov's inequality and Theorem 4.3.1. We give below a sharper bound that relies on exponential inequalities that are close to the ones given in Theorem 3.1 in [Pin94]. The following theorem gives non asymptotic confidence balls for the Robbins-Monro algorithm.

Theorem 4.4.1. *Assume that **(A1)-(A3)** hold. There is a positive constant C such that for all $\delta \in$*

This result is considered as non asymptotic since the constant C can be calculated and since the rank n_δ can be bounded.

$(0,1)$, *there is a rank* n_δ *such that for all* $n \geq n_\delta$,

$$\mathbb{P}\left[\|Z_n - m\| \leq \frac{C}{n^{\alpha/2}} \ln\left(\frac{4}{\delta}\right)\right] \geq 1 - \delta.$$

The proof is given in a supplementary file. This result is obtained via the study of a linearized version of the gradient (4.8),

$$Z_{n+1} - m = Z_n - m - \gamma_n \Gamma_m (Z_n - m) + \gamma_n \xi_{n+1} - \gamma_n \delta_n, \tag{4.14}$$

where $\delta_n := \Phi(Z_n) - \Gamma_m(Z_n - m)$. Introducing for all $n \geq 1$, the following operators :

$$\begin{aligned}
\alpha_n &:= I_H - \gamma_n \Gamma_m, \\
\beta_n &:= \prod_{k=1}^{n} \alpha_k = \prod_{k=1}^{n} \left(I_H - \gamma_k \Gamma_k\right), \\
\beta_0 &:= I_H,
\end{aligned}$$

by induction, (4.14) yields

$$Z_n - m = \beta_{n-1}(Z_1 - m) + \beta_{n-1} M_n - \beta_{n-1} R_n, \tag{4.15}$$

with $R_n := \sum_{k=1}^{n-1} \gamma_k \beta_k^{-1} \delta_k$ and $M_n := \sum_{k=1}^{n-1} \gamma_k \beta_k^{-1} \xi_{k+1}$. Note that (M_n) is a martingale sequence adapted to the filtration $(\mathcal{F}_n)$. Moreover,

$$\begin{aligned}
\mathbb{P}\left[\|Z_n - m\| \geq t\right] &\leq \mathbb{P}\left[\|\beta_{n-1} M_n\| \geq \frac{t}{2}\right] + \mathbb{P}\left[\|\beta_{n-1} R_n\| \geq \frac{t}{4}\right] + \mathbb{P}\left[\|\beta_{n-1}(Z_1 - m)\| \geq \frac{t}{4}\right] \\
&\leq \mathbb{P}\left[\|\beta_{n-1} M_n\| \geq \frac{t}{2}\right] + 4\frac{\mathbb{E}\left[\|\beta_{n-1} R_n\|\right]}{t} + 16\frac{\mathbb{E}\left[\|\beta_{n-1}(Z_1 - m))\|^2\right]}{t^2}.
\end{aligned} \tag{4.16}$$

Then, we must get upper bounds for each term on the right-hand side of previous inequality. As explained in Remark 4.4.1 below, it is not possible to directly apply Theorem 3.1 of [Pin94] to the quasi martingale term but the following proposition gives an analogous exponential inequality in the case where we do not have exactly a sequence of martingale differences.

Proposition 4.4.1. *Let* $(\beta_{n,k})_{(k,n)\in\mathbb{N}\times\mathbb{N}}$ *be a sequence of linear operators on* H *and* (ξ_n) *be a sequence of* H*-valued martingale differences adapted to a filtration* $(\mathcal{F}_n)$*. Moreover, let* (γ_n) *be a sequence of*

positive real numbers. Then, for all $r > 0$ and for all $n \geq 1$,

$$\mathbb{P}\left[\left\|\sum_{k=1}^{n-1}\gamma_k\beta_{n-1,k}\xi_{k+1}\right\| \geq r\right] \leq 2e^{-r}\left\|\prod_{j=2}^{n}\left(1+\mathbb{E}\left[e^{\left\|\gamma_{j-1}\beta_{n-1,j-1}\xi_j\right\|}-1-\left\|\gamma_{j-1}\beta_{n-1,j-1}\xi_j\right\|\,\middle|\mathcal{F}_{j-1}\right]\right)\right\|$$
$$\leq 2\exp\left(-r+\left\|\sum_{j=2}^{n}\mathbb{E}\left[e^{\left\|\gamma_{j-1}\beta_{n-1,j-1}\xi_j\right\|}-1-\left\|\gamma_{j-1}\beta_{n-1,j-1}\xi_j\right\|\,\middle|\mathcal{F}_{j-1}\right]\right\|\right).$$

The proof of Proposition 4.4.1 is postponed in the supplementary file. As in [TY14], it enables to give a sharp upper bound for $\mathbb{P}\left[\left\|\sum_{k=1}^{n-1}\gamma_k\beta_{n-1,k}\xi_{k+1}\right\| \geq t\right]$.

Corollary 4.4.1. *Let $(\beta_{n,k})$ be sequence of linear operators on H, (ξ_n) be a sequence of H-valued martingale differences adapted to a filtration $(\mathcal{F}_n)$ and (γ_n) be a sequence of positive real numbers. Let (N_n) and (σ_n^2) be two deterministic sequences such that*

$$N_n \geq \sup_{k\leq n-1}\|\gamma_k\beta_{n-1,k}\xi_{k+1}\| \quad a.s. \quad \text{and} \quad \sigma_n^2 \geq \sum_{k=1}^{n-1}\mathbb{E}\left[\|\gamma_k\beta_{n-1,k}\xi_{k+1}\|\,\middle|\mathcal{F}_n\right].$$

For all $t > 0$ and all $n \geq 1$,

$$\mathbb{P}\left[\left\|\sum_{k=1}^{n-1}\gamma_k\beta_{n-1,k}\xi_{k+1}\right\| \geq t\right] \leq 2\exp\left(-\frac{t^2}{2(\sigma_n^2+tN_n/3)}\right).$$

In our context, Corollary 4.4.1 can be written as follows :

Corollary 4.4.2. *Let $(N_n)_{n\geq 1}$ and $\left(\sigma_n^2\right)_{n\geq 1}$ be two deterministic sequences such that*

$$N_n \geq \sup_{k\leq n-1}\left\|\gamma_k\beta_{n-1}\beta_k^{-1}\xi_{k+1}\right\| \quad a.s. \quad \text{and} \quad \sigma_n^2 \geq \sum_{k=1}^{n-1}\mathbb{E}\left[\left\|\gamma_k\beta_{n-1}\beta_k^{-1}\xi_{k+1}\right\|\,|\mathcal{F}_n\right].$$

Then, for all $t > 0$ and for all $n \geq 1$,

$$\mathbb{P}\left[\left\|\sum_{k=1}^{n-1}\gamma_k\beta_{n-1}\beta_k^{-1}\xi_{k+1}\right\| \geq t\right] \leq 2\exp\left(-\frac{t^2}{2(\sigma_n^2+tN_n/3)}\right).$$

Remark 4.4.1. *Note that $(\beta_{n-1}M_n)$ is not a martingale sequence. Then, a first idea could be to apply Theorem 3.1 in [Pin94] to the martingale term $M_n = \sum_{k=1}^{n-1}\beta_k^{-1}\gamma_k\xi_{k+1}$ but this does not work. Indeed, although there is a positive constant M such that $\|\beta_{n-1}M_n\| \leq M$ for all $n \geq 1$, the sequence $\|\beta_{n-1}\|\,\|M_n\|$ may not be convergent ($\|\beta_{n-1}\|$ denotes the usual spectral norm of operator*

β_{n-1}). Then, it is possible to exhibit sequences (ξ_n) such that for all $t > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left[\|\beta_{n-1}\| \, \|M_n\| \geq t\right] = 1,$$
$$\lim_{n\to\infty} \mathbb{P}\left[\|\beta_{n-1} M_n\| \geq t\right] = 0.$$

Indeed, let $\lambda_{\min}$ and $\lambda_{\max}$ be the $\liminf$ *and* $\limsup$ *of the eigenvalues of the hessian Γ_m and suppose that $\lambda_{\min} < \lambda_{\max}$ and suppose $\gamma_n \lambda_{\max} \leq 1$ for all $n \geq 1$. Then $\|\beta_{n-1}\| = \prod_{k=1}^{n-1} (1 - \lambda_{\min}\gamma_k)$. Moreover, there exists a sequence $(h_n)_{n\geq 1}$ such that $\|h_n\| = 1$ for all $n \geq 1$, and a positive constant λ such that $\lambda_{\min} < \lambda \leq \lambda_{\max}$, and*

$$\left\| \sum_{k=1}^{n-1} \gamma_k \beta_k^{-1} h_k \right\| = \sum_{k=1}^{n-1} \gamma_k \prod_{j=1}^{k} \left(1 - \lambda\gamma_j\right)^{-1}.$$

Thus,

$$\|\beta_{n-1}\| \left\| \sum_{k=1}^{n-1} \gamma_k \beta_k^{-1} h_k \right\| \xrightarrow[n\to\infty]{} +\infty.$$

4.4.2 Non asymptotic confidence balls for the averaged algorithm :

The following theorem, which is one of the most important result of this paper, provides non asymptotic confidence balls for the averaged algorithm.

Theorem 4.4.2. *Assume that **(A1)-(A3)** hold. For all $\delta \in (0, 1)$, there is a rank n_δ such that for all $n \geq n_\delta$,*

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

The proof heavily relies on the following decomposition, which is obtained, as in [CCZ13] and [Pel00], using decomposition (4.14). Indeed, summing and applying Abel's transform, we get :

$$\Gamma_m\left(\overline{Z}_n - m\right) = \frac{Z_1 - m}{\gamma_1 n} - \frac{Z_{n+1} - m}{\gamma_n n} + \frac{1}{n}\sum_{k=2}^{n} \left[\frac{1}{\gamma_k} - \frac{1}{\gamma_{k-1}}\right](Z_k - m) - \frac{1}{n}\sum_{k=1}^{n} \delta_k + \frac{1}{n}\sum_{k=1}^{n} \xi_{k+1}. \tag{4.17}$$

Noting that $\sum_{k=1}^{n} \xi_{k+1}$ is a martingale term adapted to the filtration $(\mathcal{F}_n)$, the proof of Theorem 4.4.2 relies on the application of Pinelis-Bernstein's Lemma (see [TY14], Appendix A) to this term and on the fact that, thanks to Theorem 4.3.1, it can be shown that the other terms at the right-hand side of (4.17) are negligible.

This result is considered as non asymptotic since the rank n_δ can be bounded.

Remark 4.4.2. *We can also have a more precise form of the rank n_δ (see the Proof of Theorem 4.4.2) :*

$$n_\delta := \max\left\{ \left(\frac{6C_1'}{\delta \ln\left(\frac{4}{\delta}\right)}\right)^{\frac{1}{1/2-\alpha/2}}, \left(\frac{6C_2'}{\delta \ln\left(\frac{4}{\delta}\right)}\right)^{\frac{1}{\alpha-1/2}}, \left(\frac{6C_3'}{\delta \ln\left(\frac{4}{\delta}\right)}\right)^{\frac{1}{2}} \right\}, \tag{4.18}$$

where C_1', C_2' and C_3' are constants. We can remark that the first two terms are the leading ones and if the rate α is chosen equal to $2/3$, they are of the same order that is $n_\delta = O\left(\left(\frac{-1}{\delta \ln \delta}\right)^6\right)$.

Remark 4.4.3. *We can make an informal comparison of previous result with the central limit theorem stated in ([CCZ13], Theorem 3.4), even if the latter result is only of asymptotic nature. Under assumptions **(A1)-(A3)**, it has been shown that*

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

with,

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

This implies, with the continuity of the norm in H, that for all $t > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left[\left\|\sqrt{n}\left(\overline{Z}_n - m\right)\right\| \geq t\right] = \mathbb{P}\left[\|V\| \geq t\right],$$

where V is a centered H-valued Gaussian random vector with covariance operator $\Delta_V = \Gamma_m^{-1}\Sigma\Gamma_m^{-1}$. Operator Δ_V is self-adjoint and non negative, so that it admits a spectral decomposition $\Delta_V = \sum_{j\geq 1} \eta_j v_j \otimes v_j$, where $\eta_1 \geq \eta_2 \geq \ldots \geq 0$ is the sequence of ordered eigenvalues associated to the orthonormal eigenvectors v_j, $j \geq 1$. Using the Karhunen-Loève's expansion of V, we directly get that

$$\|V\|^2 = \sum_{j\geq 1} \eta_j^2 V_j^2$$

where $V_1, V_2, \ldots$ are i.i.d. centered Gaussian variables with unit variance. Thus the distribution of $\|V\|^2$ is a mixture of independent Chi-square random variables with one degree of freedom. Computing the quantiles of $\|V\|$ to build confidence balls would require to know, or to estimate, all the (leading) eigenvalues of the rather complicated operator Δ_V and this is not such an easy task. Indeed, it would be necessary to project on a finite dimensional space to get the inverse of the Hessian before extracting the leading eigenvectors of the covariance. Finally, the last problem of using the central limit theorem to get confidence balls is that, as far as we know, we do not know its rate of convergence.

On the other hand, the use of the confidence balls given in Theorem 4.4.2 only requires the know-

ledge of $\lambda_{\min}$. This eigenvalue is not difficult to estimate since it can also be written as

$$\lambda_{\min} = \mathbb{E}\left[\frac{1}{\|X-m\|}\right] - \lambda_{\max}\left(\mathbb{E}\left[\frac{1}{\|X-m\|^3}(X-m)\otimes(X-m)\right]\right),$$

where $\lambda_{\max}(A)$ denotes the largest eigenvalue of operator A.

Remark 4.4.4. *Under previous assumptions, with analogous calculus to the ones in the proof of Theorem 4.4.2 and applying Theorem 4.3.1, it can be shown that there is a positive constant C' such that for all $n \geq 1$,*

$$\mathbb{E}\left[\left\|\overline{Z}_n - m\right\|\right] \leq \frac{C'}{\sqrt{n}}.$$

Moreover, assuming the additional condition $\alpha > 2/3$, it can be shown that there is a positive constant C'' such that

$$\mathbb{E}\left[\|\overline{Z}_n - m\|^2\right] \leq \frac{C''}{n}.$$

The averaged algorithm converges at the parametric rate of convergence in quadratic mean.

4.5 Proofs

4.5.1 Proof of Theorem 4.3.1

As explained in Section 4.3, the proof of Theorem 4.3.1 is based on Lemma 4.3.1, which allows to obtain an upper bound of the quadratic mean error, and on Lemma 4.3.2, which gives an upper bound of the L^4 error. We first prove Lemma 4.3.1. In order to do so, we have to introduce a new technical lemma which gives a bound of the rest in the Taylor's expansion of the gradient. This will enable us to bound the rest term $\beta_{n-1}R_n$ in decomposition (4.15).

Lemma 4.5.1. *Assuming assumption (A3), there is a constant C_m such that for all $n \geq 1$:*

$$\|\delta_n\| \leq C_m \|Z_n - m\|^2, \tag{4.19}$$

where $\delta_n := \Phi(Z_n) - \Gamma_m(Z_n - m)$ is the second order term in the Taylor's decomposition of $\Phi(Z_n)$.

The proof is given in a supplementary file. We can now prove Lemma 4.3.1.

Proof of Lemma 4.3.1 : Let us study the asymptotic behavior of the sequence of operators (β_n). Since Γ_m admits a spectral decomposition, we have $\|\alpha_k\| \leq \sup_j |1 - \gamma_k \lambda_j|$ where (λ_j) is the

sequence of eigenvalues of Γ_m. Since for all $j \geq 1$ we have $0 < c_m \leq \lambda_j \leq C$, there is a rank n_0 such that for all $n \geq n_0, \gamma_n C < 1$. In particular, for all $n \geq n_0$ we have $\|\alpha_n\| \leq 1 - \gamma_n c_m$. Thus, there is a positive constant c_1 such that for all $n \geq 1$:

$$\|\beta_{n-1}\| \leq c_1 \exp\left(-\lambda_{\min} \sum_{k=1}^{n-1} \gamma_k\right) \leq c_1 \exp\left(-c_m \sum_{k=1}^{n-1} \gamma_k\right), \tag{4.20}$$

where $\lambda_{\min} > 0$ is the smallest eigenvalue of Γ_m. Similarly, there is a positive constant c_2 such that for all integer n and for all integer $k \leq n-1$:

$$\left\|\beta_{n-1}\beta_k^{-1}\right\| \leq c_2 \exp\left(-c_m \sum_{j=k+1}^{n-1} \gamma_j\right). \tag{4.21}$$

Moreover, for all $n > n_0, k \geq n_0$ such that $k \leq n-1$, (see [CCZ13] for more details),

$$\left\|\beta_{n-1}\beta_k^{-1}\right\| \leq \exp\left(-c_m \sum_{j=k+1}^{n-1} \gamma_j\right). \tag{4.22}$$

Using decomposition (4.15) again, we get

$$\mathbb{E}\left[\|Z_n - m\|^2\right] \leq 3\mathbb{E}\left[\|\beta_{n-1}(Z_1 - m)\|^2\right] + 3\mathbb{E}\left[\|\beta_{n-1}M_n\|^2\right] + 3\mathbb{E}\left[\|\beta_{n-1}R_n\|^2\right]. \tag{4.23}$$

We now bound each term on the right-hand side of previous inequality.

Step 1 : The quasi-deterministic term : Using inequality (4.20), with help of an integral test for convergence, for all $n \geq 1$:

$$\begin{aligned}
\mathbb{E}\left[\|\beta_{n-1}(Z_1 - m)\|^2\right] &\leq c_1^2 \exp\left(-2c_m \sum_{k=1}^{n-1} \gamma_k\right) \mathbb{E}\left[\|Z_1 - m\|^2\right] \\
&\leq c_1^2 \left(-2c_m c_\gamma \int_1^n t^{-\alpha} dt\right) \mathbb{E}\left[\|Z_1 - m\|^2\right] \\
&\leq c_1^2 M \exp\left(2\frac{c_m c_\gamma}{1-\alpha}\right) \exp\left(-2\frac{c_m c_\gamma}{1-\alpha} n^{1-\alpha}\right).
\end{aligned}$$

Since $\alpha < 1$, this term converges exponentially to 0.

Step 2 : The martingale term : We have

$$\begin{aligned}\|\beta_{n-1}M_n\|^2 &= \left\|\sum_{k=1}^{n-1}\gamma_k\beta_{n-1}\beta_k^{-1}\xi_{k+1}\right\|^2 \\ &\leq \sum_{k=1}^{n-1}\gamma_k^2\left\|\beta_{n-1}\beta_k^{-1}\right\|^2\|\xi_{k+1}\|^2 + 2\sum_{k=1}^{n-1}\sum_{k'<k}\gamma_k\gamma_{k'}\langle\beta_{n-1}\beta_k^{-1}\xi_{k+1},\beta_{n-1}\beta_{k'}^{-1}\xi_{k'+1}\rangle.\end{aligned}$$

Since (ξ_n) is a sequence of martingale differences, for all $k' < k$ we have $\mathbb{E}\left[\langle\xi_{k+1},\xi_{k'+1}\rangle\right] = 0$. Thus,

$$\mathbb{E}\left[\|\beta_{n-1}M_n\|^2\right] \leq \sum_{k=1}^{n-1}\gamma_k^2\left\|\beta_{n-1}\beta_k^{-1}\right\|^2, \tag{4.24}$$

because for all $k \in \mathbb{N}$, $\mathbb{E}\left[\|\xi_{k+1}\|^2\right] \leq 1$. The term $\|\beta_{n-1}\beta_k^{-1}\|$ converges exponentially to 0 when k is lower enough than n. We denote by $E(.)$ the integer function and we isolate the dominating term. Let us split the sum into two parts :

$$\sum_{k=1}^{n-1}\gamma_k^2\left\|\beta_{n-1}\beta_k^{-1}\right\|^2 = \sum_{k=1}^{E(n/2)-1}\gamma_k^2\|\beta_{n-1}\beta_k^{-1}\|^2 + \sum_{k=E(n/2)}^{n-1}\gamma_k^2\|\beta_{n-1}\beta_k^{-1}\|^2. \tag{4.25}$$

We shall show that the first term on the right-hand side in (4.25) converges exponentially to 0 and that the second term on the right-hand side, which is the dominating one, converges at the rate $\frac{1}{n^\alpha}$. Indeed, we deduce from inequality (4.21) :

$$\sum_{k=1}^{E(n/2)-1}\gamma_k^2\left\|\beta_{n-1}\beta_k^{-1}\right\|^2 \leq c_2\sum_{k=1}^{E(n/2)-1}\gamma_k^2 e^{-2c_m\frac{n}{2}\frac{c_\gamma}{n^\alpha}} \leq c_2 e^{-c_m c_\gamma n^{1-\alpha}}\sum_{k=1}^{E(n/2)-1}\gamma_k^2.$$

Since $\sum\gamma_k^2 < \infty$, we get $\sum_{k=1}^{E(n/2)-1}\gamma_k^2\|\beta_{n-1}\beta_k^{-1}\|^2 = O\left(e^{-c_m c_\gamma n^{1-\alpha}}\right)$.
We now bound the second term on the right-hand side of equality (4.25). Using inequality (4.22), for all $n > 2n_0$:

$$\begin{aligned}\sum_{k=E(n/2)}^{n-1}\gamma_k^2\left\|\beta_{n-1}\beta_k^{-1}\right\|^2 &\leq \sum_{k=E(n/2)}^{n-2}\gamma_k^2 e^{-2c_m\sum_{j=k+1}^{n-1}\gamma_j} + \gamma_{n-1}^2 \\ &\leq c_\gamma\left(\frac{1}{E(n/2)}\right)^\alpha\sum_{k=E(n/2)}^{n-2}\gamma_k e^{-2c_m\sum_{j=k+1}^{n-1}\gamma_j} + \gamma_{n-1}^2 \\ &\leq \frac{2^\alpha c_\gamma}{n^\alpha}\sum_{k=E(n/2)}^{n-2}\gamma_k e^{-2c_m\sum_{j=k+1}^{n-1}\gamma_j} + \gamma_{n-1}^2.\end{aligned}$$

Moreover, for all $n > 2n_0$ and $k \leq n-2$,

$$\sum_{j=k+1}^{n-1} \gamma_j \leq \int_{k+1}^{n} \frac{c_\gamma}{s^\alpha} ds = \frac{c_\gamma}{1-\alpha}\left[n^{1-\alpha} - (k+1)^{1-\alpha}\right],$$

and hence $e^{-2c_m \sum_{j=k+1}^{n-1} \gamma_j} \leq e^{-2c_m \frac{c_\gamma}{1-\alpha}\left[n^{1-\alpha}-(k+1)^{1-\alpha}\right]}$. Since $\frac{1}{k^\alpha} \leq \frac{2}{(k+1)^\alpha}$,

$$\begin{aligned}
\sum_{k=E(n/2)}^{n-2} \gamma_k e^{2c_m \frac{c_\gamma}{1-\alpha}(k+1)^{1-\alpha}} &\leq 2^\alpha c_\gamma \sum_{k=E(n/2)}^{n-2} \frac{1}{(k+1)^\alpha} e^{2c_m \frac{c_\gamma}{1-\alpha}(k+1)^{1-\alpha}} \\
&\leq 2^\alpha c_\gamma \int_{E(n/2)}^{n-1} \frac{1}{(t+1)^\alpha} e^{2c_m \frac{c_\gamma}{1-\alpha}(t+1)^{1-\alpha}} dt \\
&\leq \frac{2^{\alpha-1}}{c_m} e^{2c_m \frac{c_\gamma}{1-\alpha} n^{1-\alpha}}.
\end{aligned}$$

Note that the integral test for convergence is valid because there is a rank $n_0' \in \mathbb{N}$ such that the function $t \longmapsto \frac{1}{(t+1)^\alpha} e^{2c_m \frac{c_\gamma}{1-\alpha}(t+1)^{1-\alpha}}$ is increasing on $[n_0', \infty)$. Let $n_1 := \max\{2n_0+1, n_0'\}$, for all $n \geq n_1$:

$$\sum_{k=E(n/2)}^{n-1} \gamma_k^2 \left\| \beta_{n-1}\beta_k^{-1} \right\|^2 \leq \frac{2^{2\alpha-1} c_\gamma}{c_m} \frac{1}{n^\alpha} + c_\gamma 2^{2\alpha} \frac{1}{n^{2\alpha}}. \tag{4.26}$$

Consequently, there is a positive constant C_2 such that for all $n \geq 1$,

$$3\mathbb{E}\left[\|\beta_{n-1} M_n\|^2\right] \leq C_2 \frac{1}{n^\alpha}. \tag{4.27}$$

Remark 4.5.1. *Note that splitting the sum in equation (4.25) is really crucial to get the good rate of convergence of the martingale term. Remark that a different split was considered in [CCZ13], which leads to a non optimal bound of the form*

$$\mathbb{E}\left[\|\beta_{n-1} M_n\|^2\right] \leq \frac{C_2 \ln n}{n^\alpha}.$$

Step 3 : The rest term : In the same way, we split the sum into two parts :

$$\sum_{k=1}^{n-1} \gamma_k \beta_{n-1}\beta_k^{-1}\delta_k = \sum_{k=1}^{E(n/2)-1} \gamma_k \beta_{n-1}\beta_k^{-1}\delta_k + \sum_{k=E(n/2)}^{n-1} \gamma_k \beta_{n-1}\beta_k^{-1}\delta_k. \tag{4.28}$$

One can check (see the proof of Lemma 4.5.3 for more details) that there is a positive constant M such that for all $n \geq 1$,

$$\mathbb{E}\left[\|Z_n - m\|^4\right] \leq M. \tag{4.29}$$

Moreover, by Lemma 4.5.1, $\|\delta_n\| \leq C_m \|Z_n - m\|^2$. Thus, for all $k, k' \geq 1$, the application of Cauchy-Schwarz's inequality gives us

$$\mathbb{E}\left[\|\delta_k\|\|\delta_{k'}\|\right] \leq C_m^2 \mathbb{E}\left[\|Z_k - m\|^2 \|Z_{k'} - m\|^2\right] \leq C_m^2 \sup_{n\geq 1} \mathbb{E}\left[\|Z_n - m\|^4\right] \leq C_m^2 M.$$

As a particular case, we also have $\mathbb{E}\left[|\langle \delta_k, \delta_{k'} \rangle|\right] \leq C_m^2 M$. Applying this result to the term on the right-hand side in (4.28),

$$\begin{aligned}
\mathbb{E}\left[\left\|\sum_{k=1}^{E(n/2)-1} \gamma_k \beta_{n-1}\beta_k^{-1}\delta_k\right\|^2\right] &\leq C_m^2 M \left[\sum_{k=1}^{E(n/2)-1} \gamma_k \|\beta_{n-1}\beta_k^{-1}\|\right]^2 \\
&\leq c_2 C_m^2 M e^{-2c_m c_\gamma n^{1-\alpha}} \left(\sum_{k=1}^{E(n/2)-1} \gamma_k\right)^2 \\
&\leq C_1' e^{-2c_m c_\gamma n^{1-\alpha}} n^{2-2\alpha}.
\end{aligned}$$

This term converges exponentially to 0. To bound the second term, we use the same idea as for the martingale term. Applying previous inequalities for the terms $\mathbb{E}\left[\|\delta_k\|\|\delta_{k'}\|\right]$ which appear in the double products, we get :

$$\begin{aligned}
\mathbb{E}\left[\left\|\sum_{k=E(n/2)}^{n-1} \gamma_k \beta_{n-1}\beta_k^{-1}\delta_k\right\|^2\right] &\leq C_m^2 \sup_{E(n/2)\leq k\leq n-1} \mathbb{E}\left[\|Z_k - m\|^4\right] \left[\sum_{k=E(n/2)}^{n-1} \gamma_k \|\beta_{n-1}\beta_k^{-1}\|\right]^2 \\
&\leq C_3 \sup_{E(n/2)\leq k\leq n-1} \mathbb{E}\left[\|Z_k - m\|^4\right],
\end{aligned}$$

since $\left[\sum_{k=E(n/2)}^{n-1} \gamma_k \|\beta_{n-1}\beta_k^{-1}\|\right]^2$ is bounded. Indeed, one can check it with similar calculus to the ones in the proof of inequality (4.27). We put together the terms which converge exponentially to 0. □

To prove Lemma 4.3.2, we introduce two technical lemmas. The first one gives a bound of the decomposition in the particular case when $\|Z_n - m\|$ is not too large.

Lemma 4.5.2. *If assumptions* ***(A1)*** *and* ***(A2)*** *holds, there are a rank n_α and a constant c such that for all $n \geq n_\alpha$, $\|Z_n - m\| \leq cn^{1-\alpha}$ yields*

$$\langle \Phi(Z_n), Z_n - m \rangle \geq \frac{1}{c_\gamma n^{1-\alpha}} \|Z_n - m\|^2. \tag{4.30}$$

As a corollary, there is also a deterministic rank n'_α such that for all $n \geq n'_\alpha$, $\|Z_n - m\| \leq cn^{1-\alpha}$

yields

$$\|Z_n - m - \gamma_n \Phi(Z_n)\|^2 \leq \left(1 - \frac{1}{n}\right) \|Z_n - m\|^2. \tag{4.31}$$

Proof of Lemma 4.5.2. We suppose that $\|Z_n - m\| \leq cn^{1-\alpha}$. We must consider two cases. If $\|Z_n - m\| \leq 1$, then we have in particular $\|Z_n\| \leq \|m\| + 1$. Consequently, we get with Corollary 2.2 in [CCZ13] that there is a positive constant c_1 such that $\langle \Phi(Z_n), Z_n - m \rangle \geq c_1 \|Z_n - m\|^2$.

If $\|Z_n - m\| \geq 1$, since $\Phi(Z_n) = \int_0^1 \Gamma_{m+t(Z_n-m)}(Z_n - m)dt$,

$$\langle \Phi(Z_n), Z_n - m \rangle = \int_0^1 \left\langle Z_n - m, \Gamma_{m+t(Z_n-m)}(Z_n - m) \right\rangle dt.$$

Moreover, operators Γ_h are non negative for all $h \in H$. Applying Proposition 2.1 of [CCZ13], and since for all $t \in \left[0, \frac{1}{\|Z_n - m\|}\right]$ we have $\|m + t(Z_n - m)\| \leq \|m\| + 1$, there is a positive constant c_2 such that :

$$\begin{aligned}
\langle \Phi(Z_n), Z_n - m \rangle &= \int_0^1 \left\langle Z_n - m, \Gamma_{m+t(Z_n-m)}(Z_n - m) \right\rangle dt \\
&\geq \int_0^{1/\|Z_n-m\|} \left\langle Z_n - m, \Gamma_{m+t(Z_n-m)}(Z_n - m) \right\rangle dt \\
&\geq \int_0^{1/\|Z_n-m\|} c_2 \|Z_n - m\|^2 dt \\
&\geq \frac{c_2}{cn^{1-\alpha}} \|Z_n - m\|^2.
\end{aligned}$$

We can choose a rank n_α such that for all $n \geq n_\alpha$ we have $c_1 \geq \frac{1}{c_\gamma n^{1-\alpha}}$ which concludes the proof of inequality (4.30) with $c = c_2 c_\gamma$.

We now prove inequality (4.31). For all $n \geq n_\alpha$, $\|Z_n - m\| \leq cn^{1-\alpha}$ yields

$$\begin{aligned}
\|Z_n - m - \gamma_n \Phi(Z_n)\|^2 &\leq \|Z_n - m\|^2 - \frac{2}{c_\gamma n^{1-\alpha}} \frac{c_\gamma}{n^\alpha} \|Z_n - m\|^2 + \gamma_n^2 C^2 \|Z_n - m\|^2 \\
&= \left(1 - \frac{2}{n} + C^2 \frac{c_\gamma^2}{n^{2\alpha}}\right) \|Z_n - m\|^2.
\end{aligned}$$

Thus, we can choose a rank $n'_\alpha \geq n_\alpha$ such that for all $n \geq n'_\alpha$ we have $C^2 \frac{c_\gamma^2}{n^{2\alpha}} \leq \frac{1}{n}$. Note that this is possible since $\alpha > 1/2$. □

The second lemma shows that the probability for $\|Z_n - m\|$ to be large is very small as n

increases.

Lemma 4.5.3. *There is a positive constant C_α such that for all $n \geq 1$,*

$$\mathbb{P}\left[\|Z_n - m\| \geq cn^{1-\alpha}\right] \leq \frac{C_\alpha}{n^{4-\alpha}},$$

where c has been defined in the previous lemma.

The proof is given in the supplementary file.

Proof of Lemma 4.3.2. For all $n \geq 1$,

$$\mathbb{E}\left[\|Z_{n+1} - m\|^4\right] = \mathbb{E}\left[\|Z_{n+1} - m\|^4 \mathbb{1}_{\|Z_n - m\| \geq cn^{1-\alpha}}\right] + \mathbb{E}\left[\|Z_{n+1} - m\|^4 \mathbb{1}_{\|Z_n - m\| < cn^{1-\alpha}}\right], \tag{4.32}$$

with c defined in Lemma 4.5.2. Let us bound the first term in (4.32). Since $\|Z_{n+1} - m\| \leq \|Z_n - m\| + \gamma_n \leq \|Z_1 - m\| + \sum_{k=1}^n \gamma_k$ and since Z_1 is bounded or deterministic, there is a constant C'_α such that for all integer $n \geq 1$,

$$\|Z_n - m\| \leq C'_\alpha n^{1-\alpha}.$$

Consequently,

$$\begin{aligned}
\mathbb{E}\left[\|Z_{n+1} - m\|^4 \mathbb{1}_{\|Z_n - m\| \geq cn^{1-\alpha}}\right] &\leq \mathbb{E}\left[\left(C'_\alpha (n+1)^{1-\alpha}\right)^4 \mathbb{1}_{\|Z_n - m\| \geq cn^{1-\alpha}}\right] \\
&\leq \left(C'_\alpha (n+1)^{1-\alpha}\right)^4 \mathbb{P}\left[\|Z_n - m\| \geq cn^{1-\alpha}\right].
\end{aligned}$$

Thus, applying Lemma 4.5.3, we get

$$\left(C'_\alpha (n+1)^{1-\alpha}\right)^4 \mathbb{P}\left[\|Z_n - m\| \geq cn^{1-\alpha}\right] \leq \frac{C'^4_\alpha C_\alpha (n+1)^{4-4\alpha}}{n^{4-\alpha}} \leq 2^{4-4\alpha} \frac{C'^4_\alpha C_\alpha}{n^{3\alpha}}.$$

We now bound the second term. Suppose that $\|Z_n - m\| \leq cn^{1-\alpha}$. Since $\|\xi_{n+1}\| \leq 2$, using Lemma 4.5.2, there is a rank n_α such that for all $n \geq n_\alpha$,

$$\begin{aligned}
&\|Z_{n+1} - m\|^2 \mathbb{1}_{\|Z_n - m\| < cn^{1-\alpha}} \\
&\quad = \left(\|Z_n - m - \gamma_n \Phi(Z_n)\|^2 + \gamma_n^2 \|\xi_{n+1}\|^2 + 2\gamma_n \langle Z_n - m - \gamma_n \Phi(Z_n), \xi_{n+1}\rangle\right) \mathbb{1}_{\|Z_n - m\| < cn^{1-\alpha}} \\
&\quad \leq \left(\left(1 - \frac{1}{n}\right)\|Z_n - m\|^2 + 4\gamma_n^2 + 2\gamma_n \langle Z_n - m - \gamma_n \Phi(Z_n), \xi_{n+1}\rangle\right) \mathbb{1}_{\|Z_n - m\| < cn^{1-\alpha}}.
\end{aligned}$$

Moreover, since (ξ_{n+1}) is a sequence of martingale differences for the filtration $(\mathcal{F}_n)$,

$$\begin{aligned}\mathbb{E}\left[\langle Z_n - m - \gamma_n \Phi(Z_n), \xi_{n+1}\rangle \mathbb{1}_{\|Z_n - m\| \leq cn^{1-\alpha}} | \mathcal{F}_n\right] &= 0,\\ \mathbb{E}\left[\langle Z_n - m - \gamma_n \Phi(Z_n), \xi_{n+1}\rangle \|Z_n - m\|^2 \mathbb{1}_{\|Z_n - m\| \leq cn^{1-\alpha}} | \mathcal{F}_n\right] &= 0.\end{aligned}$$

Applying Cauchy-Schwarz's inequality,

$$\begin{aligned}\mathbb{E}\left[\|Z_{n+1} - m\|^4 \mathbb{1}_{\|Z_n - m\| \leq cn^{1-\alpha}}\right] &\leq \left(1 - \frac{1}{n}\right)^2 \mathbb{E}\left[\|Z_n - m\|^4 \mathbb{1}_{\|Z_n - m\| \leq cn^{1-\alpha}}\right] + 16\gamma_n^4\\ &\quad + 8\gamma_n^2 \left(1 - \frac{1}{n}\right) \mathbb{E}\left[\|Z_n - m\|^2 \mathbb{1}_{\|Z_n - m\| \leq cn^{1-\alpha}}\right]\\ &\quad + 4\gamma_n^2 \mathbb{E}\left[\langle Z_n - m - \gamma_n \Phi(Z_n), \xi_{n+1}\rangle^2 \mathbb{1}_{\|Z_n - m\| \leq cn^{1-\alpha}}\right]\\ &\leq \left(1 - \frac{1}{n}\right)^2 \mathbb{E}\left[\|Z_n - m\|^4\right] + 16\gamma_n^4 + 8\gamma_n^2 \mathbb{E}\left[\|Z_n - m\|^2\right]\\ &\quad + 4\gamma_n^2 \mathbb{E}\left[\|Z_n - m - \gamma_n \Phi(Z_n)\|^2 \mathbb{E}\left[\|\xi_{n+1}\|^2 | \mathcal{F}_n\right] \mathbb{1}_{\|Z_n - m\| \leq cn^{1-\alpha}}\right].\end{aligned}$$

Finally, since $\mathbb{E}\left[\|\xi_{n+1}\|^2 | \mathcal{F}_n\right] \leq 1$, applying Lemma 4.5.3 we get

$$\begin{aligned}\mathbb{E}\left[\|Z_{n+1} - m\|^4 \mathbb{1}_{\|Z_n - m\| \leq cn^{1-\alpha}}\right] &\leq \left(1 - \frac{1}{n}\right)^2 \mathbb{E}\left[\|Z_n - m\|^4\right] + 16\gamma_n^4 + 8\gamma_n^2 \mathbb{E}\left[\|Z_n - m\|^2\right]\\ &\quad + 4\gamma_n^2 \left(1 - \frac{1}{n}\right) \mathbb{E}\left[\|Z_n - m\|^2\right]\\ &\leq \left(1 - \frac{1}{n}\right)^2 \mathbb{E}\left[\|Z_n - m\|^4\right] + 16\gamma_n^4 + 12\gamma_n^2 \mathbb{E}\left[\|Z_n - m\|^2\right].\end{aligned}$$

Since $\gamma_n^4 = o\left(\frac{1}{n^{3\alpha}}\right)$, there are positive constants C_1', C_2' such that for all $n \geq n_\alpha$,

$$\begin{aligned}\mathbb{E}\left[\|Z_{n+1} - m\|^4\right] &= \mathbb{E}\left[\|Z_{n+1} - m\|^4 \mathbb{1}_{\|Z_n - m\| \geq cn^{1-\alpha}}\right] + \mathbb{E}\left[\|Z_{n+1} - m\|^4 \mathbb{1}_{\|Z_n - m\| \leq cn^{1-\alpha}}\right]\\ &\leq \frac{2^{4-4\alpha} C_\alpha'^4 C_\alpha}{n^{3\alpha}} + \left(1 - \frac{1}{n}\right)^2 \mathbb{E}\left[\|Z_n - m\|^4\right] + 16\gamma_n^4 + 12\gamma_n^2 \mathbb{E}\left[\|Z_n - m\|^2\right]\\ &\leq \left(1 - \frac{1}{n}\right)^2 \mathbb{E}\left[\|Z_n - m\|^4\right] + C_1' \frac{1}{n^{3\alpha}} + C_2' \frac{1}{n^{2\alpha}} \mathbb{E}\left[\|Z_n - m\|^2\right].\end{aligned}$$

□

Proof of Theorem 4.3.1. Let $\beta \in (\alpha, 3\alpha - 1)$, there is a rank $n_\beta \geq n_\alpha$ (n_α is defined in Lemma 4.3.2) such that for all $n \geq n_\beta$ we have $\left(1 - \frac{1}{n}\right)^2 \left(\frac{n+1}{n}\right)^\beta + (C_1' + C_2') 2^{3\alpha} \frac{1}{(n+1)^{3\alpha - \beta}} \leq 1$ (C_1', C_2'

are defined in Lemma 4.3.2). Indeed, since $\beta < 3\alpha - 1 < 2$,

$$\left(1-\frac{1}{n}\right)^2\left(\frac{n+1}{n}\right)^\beta + \left(C_1' + C_2'\right)2^{3\alpha}\frac{1}{(n+1)^{3\alpha-\beta}} = 1 - (2-\beta)\frac{1}{n} + o\left(\frac{1}{n}\right).$$

We now prove by induction that there are positive constants C', C'' such that $2C' \geq C'' \geq C' \geq 1$ and such that for all $n \geq n_\beta$,

$$\mathbb{E}\left[\|Z_n - m\|^2\right] \leq \frac{C'}{n^\alpha}$$
$$\mathbb{E}\left[\|Z_n - m\|^4\right] \leq \frac{C''}{n^\beta}.$$

Let us choose $C' \geq n_\beta \mathbb{E}\left[\|Z_{n_\beta} - m\|^2\right]$ and $C'' \geq n_\beta \mathbb{E}\left[\|Z_{n_\beta} - m\|^4\right]$. This is possible since there is a positive constant M such that for all $n \geq 1$, $\sup\left\{\mathbb{E}\left[\|Z_n - m\|^2\right], \mathbb{E}\left[\|Z_n - m\|^4\right]\right\} \leq M$. Let $n \geq n_\beta$, using Lemma 4.3.2 and by induction,

$$\begin{aligned}\mathbb{E}\left[\|Z_{n+1} - m\|^4\right] &\leq \left(1-\frac{1}{n}\right)^2 \mathbb{E}\left[\|Z_n - m\|^4\|\right] + \frac{C_1'}{n^{3\alpha}} + \frac{C_2'}{n^{2\alpha}}\mathbb{E}\left[\|Z_n - m\|^2\right] \\ &\leq \left(1-\frac{1}{n}\right)^2 \frac{C''}{n^\beta} + \frac{C_1'}{n^{3\alpha}} + \frac{C_2'C'}{n^{3\alpha}}.\end{aligned}$$

Moreover, since $C' \leq C''$ and since $C'' \geq 1$,

$$\mathbb{E}\left[\|Z_{n+1} - m\|^4\right] \leq \left(1-\frac{1}{n}\right)^2 \frac{C''}{n^\beta} + \frac{C_1'C''}{n^{3\alpha}} + \frac{C_2'C''}{n^{3\alpha}}.$$

Factorizing by $\frac{C''}{(n+1)^\beta}$, we get

$$\begin{aligned}\mathbb{E}\left[\|Z_{n+1} - m\|^4\right] &\leq \left(1-\frac{1}{n}\right)^2\left(1+\frac{1}{n}\right)^\beta \frac{C''}{(n+1)^\beta} + \left(C_1' + C_2'\right)\left(1+\frac{1}{n}\right)^{3\alpha}\frac{1}{(n+1)^{3\alpha-\beta}}\frac{C''}{(n+1)^\beta} \\ &\leq \left(\left(1-\frac{1}{n}\right)^2\left(1+\frac{1}{n}\right)^\beta + \left(C_1' + C_2'\right)2^{3\alpha}\frac{1}{(n+1)^{3\alpha-\beta}}\right)\frac{C''}{(n+1)^\beta}.\end{aligned}$$

By definition of n_β,

$$\mathbb{E}\left[\|Z_{n+1} - m\|^4\right] \leq \frac{C''}{(n+1)^\beta}. \tag{4.33}$$

We now prove that $\mathbb{E}\left[\|Z_{n+1} - m\|^2\right] \leq \frac{C'}{(n+1)^\alpha}$. Since $C'' \leq 2C'$, by Lemma 4.3.1 and by

induction, there is a constant $C''' > 0$ such that

$$\begin{aligned}\mathbb{E}\left[\|Z_{n+1}-m\|^2\right] &\leq \frac{C'''}{(n+1)^\alpha} + C_3 \sup_{n/2+1\leq k\leq n+1} \mathbb{E}\left[\|Z_k-m\|^4\right] \\ &\leq \frac{C'''}{(n+1)^\alpha} + 2^{\beta+1}C_3\frac{1}{(n+1)^{\beta-\alpha}}\frac{C'}{(n+1)^\alpha}\end{aligned}$$

To get $\mathbb{E}\left[\|Z_{n+1}-m\|^2\right] \leq \frac{C'}{(n+1)^\alpha}$, we only need to take $C' \geq C''' + 2^{\beta+1}C_3\frac{1}{(n+1)^{\beta-\alpha}}$, which concludes the induction.

The proof is complete for all $n \geq 1$ by taking $C' \geq \max_{n\leq n_\beta}\left\{n^\alpha \mathbb{E}\left[\|Z_n-m\|^2\right]\right\}$ and $C'' \geq \max_{n\leq n_\beta}\left\{n^\beta \mathbb{E}\left[\|Z_n-m\|^4\right]\right\}$. □

4.5.2 Proof of Theorem 4.4.2

Let us recall the decomposition of the averaged algorithm

$$\Gamma_m\left(\overline{Z}_n - m\right) = \frac{Z_1-m}{n\gamma_1} - \frac{Z_{n+1}-m}{n\gamma_n} + \sum_{k=2}^{n}\left(\frac{1}{\gamma_k}-\frac{1}{\gamma_{k-1}}\right)(Z_k-m) - \frac{1}{n}\sum_{k=1}^{n}\delta_k + \frac{1}{n}\sum_{k=1}^{n}\xi_{k+1}.$$

We now bound each term on the right-hand side of previous inequality. Note that since $\mathbb{E}\left[\|Z_n-m\|^2\right] \leq \frac{C'}{n^\alpha}$, applying Cauchy-Schwarz's inequality, we have $\mathbb{E}\left[\|Z_n-m\|\right] \leq \sqrt{\frac{C'}{n^\alpha}}$. Then,

$$\mathbb{E}\left[\left\|\frac{Z_{n+1}-m}{n\gamma_n}\right\|^2\right] \leq \frac{n^{2\alpha}}{c_\gamma n^2}\mathbb{E}\left[\|Z_{n+1}-m\|^2\right] \leq \frac{2^\alpha C'}{c_\gamma}\frac{1}{n^{2-\alpha}}.$$

Since $\alpha < 1$, remark that $\frac{2-\alpha}{2} > \frac{1}{2}$. Moreover, since $\gamma_k^{-1}-\gamma_{k-1}^{-1} \leq 2\alpha c_\gamma^{-1}k^{\alpha-1}$, there is a positive constant C_1 such that :

$$\begin{aligned}\mathbb{E}\left[\left\|\frac{1}{n}\sum_{k=2}^{n}(Z_k-m)\left(\gamma_k^{-1}-\gamma_{k-1}^{-1}\right)\right\|\right] &\leq \frac{2\alpha c_\gamma^{-1}}{n}\sum_{k=2}^{n}\mathbb{E}\left[\|Z_k-m\|\right]k^{\alpha-1} \\ &\leq \frac{2\alpha c_\gamma^{-1}\sqrt{C'}}{n}\sum_{k=2}^{n/2-1}k^{\alpha/2-1} \leq \frac{C_1}{n^{1-\alpha/2}}.\end{aligned}$$

Note also that since $\alpha < 1$, we have $1-\alpha/2 \geq 1/2$. Moreover, since $\|\delta_n\| \leq C_m\|T_n\|^2$, there is a positive constant C_2 such that

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{k=1}^{n}\delta_k\right\|\right] \leq \frac{C_m}{n}\sum_{k=1}^{n}\mathbb{E}\left[\|Z_k-m\|^2\right] \leq \frac{C_mC'}{n}\sum_{k=1}^{n}k^{-\alpha} \leq C_2\frac{1}{n^\alpha}.$$

Finally, there is a positive constant C_3 such that $\mathbb{E}\left[\left\|\frac{Z_1-m}{\gamma_1 n}\right\|\right] \leq \frac{C_3}{n}$.

We now study the martingale term. Let M be a constant and (σ_n) be a sequence of positive real numbers defined by $M := 2 \geq \sup_i \|\xi_i\|$ and $\sigma_n^2 := n \geq \sum_{k=1}^n \mathbb{E}\left[\|\xi_k\|^2 | \mathcal{F}_{k-1}\right]$. Applying Pinelis-Bernstein's Lemma, we have for all $t > 0$,

$$\mathbb{P}\left(\sup_{1\leq k\leq n}\left\|\sum_{j=1}^{k}\xi_{j+1}\right\| \geq t\right) \leq 2\exp\left[-\frac{t^2}{2\left(\sigma_n^2 + Mt/3\right)}\right].$$

Consequently,

$$\mathbb{P}\left(\frac{\left\|\sum_{k=1}^n \xi_{k+1}\right\|}{n} \geq t\right) \leq \mathbb{P}\left(\sup_{1\leq k\leq n}\left\|\sum_{j=1}^{k}\xi_{j+1}\right\| \geq tn\right) \leq 2\exp\left[-\frac{t^2}{2\left(\sigma_n'^2 + N_n' t/3\right)}\right],$$

with $\sigma_n'^2 := 1/n$ and $N_n' := 2/n$. As in the proof of Theorem 4.1, there are positive constants C_1', C_2', C_3' such that for all $t > 0$,

$$\mathbb{P}\left[\left\|\Gamma_m\left(\overline{Z}_n - m\right)\right\| \geq t\right] \leq 2\exp\left[-\frac{(t/2)^2}{2\left(\sigma_n'^2 + N_n' t/6\right)}\right] + \frac{C_1'}{n^{1-\alpha/2}} + \frac{C_2'}{n^\alpha} + \frac{C_3'}{n} =: g(t,n).$$

We search values of t such that $g(t,n) \leq \delta$ and we must solve the following system of inequalities,

$$2\exp\left[-\frac{(t/2)^2}{2(\sigma_n'^2 + N_n t/6)}\right] \leq \delta/2, \quad \frac{C_1'}{tn^{1-\alpha/2}} \leq \delta/6, \quad \frac{C_2'}{tn^\alpha} \leq \delta/6, \quad \frac{C_3'}{tn} \leq \delta/6.$$

We get that t must satisfy (see [TY14] , Appendix A, for the martingale term) :

$$t \geq 4\left(\frac{N_n'}{3} + \sigma_n'\right)\ln\left(\frac{4}{\delta}\right), \quad t \geq \frac{6C_1'}{\delta}\frac{1}{n^{1-\alpha/2}}, \quad t \geq \frac{6C_2'}{\delta}\frac{1}{n^\alpha}, \quad t \geq \frac{6C_3'}{\delta}\frac{1}{n}.$$

Since $\left(\frac{N_n'}{3} + \sigma_n'\right) = \frac{2}{3n} + \frac{1}{\sqrt{n}}$, the other terms are negligible for n large enough and we can choose

$$n_\delta := \max\left\{\left(\frac{6C_1'}{\delta\ln\left(\frac{4}{\delta}\right)}\right)^{\frac{1}{1/2-\alpha/2}}, \left(\frac{6C_2'}{\delta\ln\left(\frac{4}{\delta}\right)}\right)^{\frac{1}{\alpha-1/2}}, \left(\frac{6C_3'}{\delta\ln\left(\frac{4}{\delta}\right)}\right)^{\frac{1}{2}}\right\}. \tag{4.34}$$

Annexe A

Online estimation of the geometric median in Hilbert spaces : non asymptotic confidence balls. Appendix

Résumé

Dans cette partie, on rappelle les décompositions des algorithmes avant de donner les preuves des propositions et lemmes techniques. Plus précisément, on donne la preuve du Lemme 4.5.1 (lemme qui permet de borner le terme de reste dans la décomposition de Taylor du gradient) ainsi que la preuve de la Proposition 4.3.1 (qui assure que la vitesse obtenue est bien la vitesse optimale). Finalement, on donne la preuve de la Proposition 4.4.1, qui est analogue à celle du Théorème 4.1 dans [Pin94] avant de l'utiliser dans la preuve du Théorème 4.4.1.

A.1 Decomposition of the Robbins-Monro algorithm

First, let us recall some decompositions of the Robbins-Monro algorithm :

$$Z_{n+1} - m = Z_n - m - \gamma_n \Phi(Z_n) + \gamma_n \xi_{n+1}, \tag{A.1}$$

with $\Phi(h) = -\mathbb{E}\left[\frac{X-h}{\|X-h\|}\right]$ for all $h \in H$ and $\xi_{n+1} := \Phi(Z_n) + \frac{X_{n+1}-Z_n}{\|X_{n+1}-Z_n\|}$. Note that (ξ_n) is a sequence of martingale differences adapted to the filtration $(\mathcal{F}_n)$. Moreover, linearizing the gradient,

$$Z_{n+1} - m = Z_n - m - \gamma_n \Gamma_m (Z_n - m) + \gamma_n \xi_{n+1} - \gamma_n \delta_n, \tag{A.2}$$

with Γ_m the Hessian of the functional G at m and $\delta_n := \Phi(Z_n) - \Gamma_m (Z_n - m)$ is the remainder term in the Taylor's expansion of the gradient. Finally, by induction, we have

$$Z_n - m = \beta_{n-1}(Z_1 - m) + \beta_{n-1} M_n + \beta_{n-1} R_n, \tag{A.3}$$

with

$$\beta_n := \prod_{k=1}^{n} (I_H - \gamma_k \Gamma_m), \qquad M_n := \sum_{k=1}^{n-1} \gamma_k \beta_k^{-1} \xi_{k+1},$$

$$\beta_0 = I_H, \qquad R_n := \sum_{k=1}^{n-1} \gamma_k \beta_k^{-1} \delta_k.$$

Note that equations (A.1), (A.2) and (A.3) correspond respectively to equations (4.8), (4.14) and (4.15) in the main document.

A.2 Proof of technical lemma and of Proposition 4.3.1

Proof of Lemma 4.5.1. Using Taylor's expansion with integral rest,

$$\Phi(Z_n) = \int_0^1 \Gamma_{m+t(Z_n-m)}(Z_n - m)dt. \tag{A.4}$$

and $\delta_n = \int_0^1 \left(\Gamma_{m+t(Z_n-m)} - \Gamma_m\right)(Z_n - m)dt$. For all $h, h' \in H$, we denote by $\varphi_{h,h'}$ the function defined as follows :

$$\begin{aligned} \varphi_{h,h'} : [0,1] &\longrightarrow H \\ t &\longmapsto \varphi_{h,h'}(t) := \Gamma_{m+th}(h'). \end{aligned}$$

Let $U_h : [0,1] \longrightarrow \mathbb{R}_+$ and $V_{h,h'} : [0,1] \longrightarrow H$ be two random functions defined for all $t \in [0,1]$ by

$$U_h(t) := \frac{1}{\|X - m - th\|},$$
$$V_{h,h'}(t) := h' - \frac{\langle X - m - th, h' \rangle (X - m - th)}{\|X - m - th\|^2}.$$

Let $V'_{h,h'}(t) = \frac{d}{dt} V_{h,h'}(t) = \lim_{t' \to 0} \frac{v_{h,h'}(t+t') - v_{h,h'}(t)}{t'}$ and $U'_{h,}(t) = \frac{d}{dt} U_h(t) = \lim_{t' \to 0} \frac{u_{h,h'}(t+t') - u_{h,h'}(t)}{t}$. Let $\varphi'_{h,h'}(t) = \frac{d}{dt} \varphi_{h,h'}(t)$, by dominated convergence, $\varphi_{h,h'}$ is differentiable on $[0,1]$ and $\|\varphi'_{h,h'}(t)\| \leq \mathbb{E}\left[|U'_h(t)| \|V_{h,h'}(t)\| + |U_h(t)| \|V'_{h,h'}(t)\|\right]$. Using Cauchy-Schwarz inequality,

$$|U_h(t)| = \frac{1}{\|X - m - th\|},$$
$$|U'_h(t)| \leq \frac{\|h\|}{\|X - m - th\|^2},$$
$$\|V_{h,h'}(t)\| \leq 2\|h'\|,$$
$$\|V'_{h,h'}(t)\| \leq \frac{4\|h\| \|h'\|}{\|X - m - th\|}.$$

Finally, using assumption (A3),

$$\|\varphi_{h,h'}(t)\| \leq 6\|h\| \|h'\| \mathbb{E}\left[\frac{1}{\|X - m - th\|^2}\right]$$
$$\leq 6\|h\| \|h'\| C.$$

We obtain for all $h \in H$

$$\|\Phi(m+h) - \Gamma_m(h)\| \leq \int_0^1 \|\Gamma_{m+th}(h) - \Gamma_m(h)\| \, dt$$
$$\leq \int_0^1 \|\varphi_{h,h}(t) - \varphi_{h,h}(0)\| dt$$
$$\leq \int_0^1 \sup_{t' \in [0,t]} \|\varphi'_{h,h}(t')\| dt$$
$$\leq 6C\|h\|^2.$$

Taking $h = Z_n - m$, for all $n \geq 1$:

$$\|\delta_n\| \leq C_m \|Z_n - m\|^2,$$

with $C_m = 6C$. □

Proof of Lemma 4.5.3. In order to use Markov's inequality, we prove by induction that for all integer $p \geq 1$, there is a positive constant M_p such that for all n :

$$\mathbb{E}\left[\|Z_n - m\|^{2p}\right] \leq M_p.$$

[CCZ13] have proved previous inequality in the particular case $p = 1$. Decomposition (A.1) yields

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

Moreover, $\langle \xi_{n+1}, \Phi(Z_n)\rangle = -\langle U_{n+1}, \Phi(Z_n)\rangle + \|\Phi(Z_n)\|^2$. Since $\|\Phi(Z_n)\| \leq 1$, $\|\xi_{n+1}\| \leq 2$ and $\langle \Phi(Z_n), Z_n - m\rangle \geq 0$, applying Cauchy-Schwarz's inequality, for all $n \geq 1$

$$\|Z_{n+1} - m\|^2 \leq \|Z_n - m\|^2 + 6\gamma_n^2 + 2\gamma_n\langle \xi_{n+1}, Z_n - m\rangle. \tag{A.5}$$

Using this inequality,

$$\begin{aligned}\|Z_{n+1} - m\|^{2p} &\leq \left(\|Z_n - m\|^2 + 6\gamma_n^2 + 2\gamma_n\langle \xi_{n+1}, Z_n - m - \gamma_n\Phi(Z_n)\rangle\right)^p \\ &= \sum_{k=0}^{p}\binom{p}{k}\left(2\gamma_n\langle \xi_{n+1}, Z_n - m\rangle\right)^k\left(\|Z_n - m\|^2 + 6\gamma_n^2\right)^{p-k} \\ &= \left(\|Z_n - m\|^2 + 6\gamma_n^2\right)^p + 2p\gamma_n\langle \xi_{n+1}, Z_n - m\rangle\left(\|Z_n - m\|^2 + 6\gamma_n^2\right)^{p-1} \\ &+ \sum_{k=2}^{p}\binom{p}{k}\left(2\gamma_n\langle \xi_{n+1}, Z_n - m\rangle\right)^k\left(\|Z_n - m\|^2 + 6\gamma_n^2\right)^{p-k}.\end{aligned} \tag{A.6}$$

We now bound the three terms in (A.6). First, using induction assumptions,

$$\begin{aligned}\mathbb{E}\left[\left(\|Z_n - m\|^2 + 6\gamma_n^2\right)^p\right] &= \mathbb{E}\left[\|Z_n - m\|^{2p} + \sum_{k=0}^{p-1}\binom{p}{k}\|Z_n - m\|^{2k}(6\gamma_n^2)^{p-k}\right] \\ &= \mathbb{E}\left[\|Z_n - m\|^{2p}\right] + \sum_{k=0}^{p-1}\binom{p}{k}\mathbb{E}\left[\|Z_n - m\|^{2k}\right](6\gamma_n^2)^{p-k} \\ &\leq \mathbb{E}\left[\|Z_n - m\|^{2p}\right] + \sum_{k=0}^{p-1}\binom{p}{k}M_k(6\gamma_n^2)^{p-k}.\end{aligned}$$

Since for all $k \leq p-1$ we have $\left(\gamma_n^2\right)^{p-k} = o\left(\gamma_n^2\right)$, there is a positive constant C_p such that for

all $n \geq 1$,

$$\mathbb{E}\left[\left(\|Z_n - m\|^2 + 6\gamma_n^2\right)^p\right] \leq \mathbb{E}\left[\|Z_n - m\|^{2p}\right] + C_p\gamma_n^2. \tag{A.7}$$

Remark that C_p does not depend on n. Let us now deal with the second term in (A.6). Since (ξ_n) is a sequence of martingale differences adapted to the filtration $(\mathcal{F}_n)$ and since Z_n is $\mathcal{F}_n$-measurable, for all $n \geq 1$,

$$\mathbb{E}\left[2\gamma_n\langle\xi_{n+1}, Z_n - m\rangle\left(\|Z_n - m\|^2 + 6\gamma_n^2\right)^{p-1}|\mathcal{F}_n\right] = 0.$$

It remains to bound the last term in (A.6). Applying Cauchy Schwarz's inequalities, for all $n \geq 1$, we get

$$\begin{aligned}
&\sum_{k=2}^{p}\binom{p}{k}\mathbb{E}\left[\left(2\gamma_n\langle\xi_{n+1}, Z_n - m\rangle\right)^k\left(\|Z_n - m\|^2 + 6\gamma_n^2\right)^{p-k}\right] \\
&\quad= \sum_{k=2}^{p}\binom{p}{k}\mathbb{E}\left[\left(2\gamma_n\langle\xi_{n+1}, Z_n - m\rangle\right)^k\sum_{j=0}^{p-k}\binom{p-k}{j}\left(\|Z_n - m\|^2\right)^{p-k-j}\left(6\gamma_n^2\right)^j\right] \\
&\quad= \sum_{k=2}^{p}\sum_{j=0}^{p-k}\binom{p-k}{j}\binom{p}{k}2^{k+j}3^j\gamma_n^{k+2j}\mathbb{E}\left[\left(\langle\xi_{n+1}, Z_n - m\rangle\right)^k\|Z_n - m\|^{2(p-k-j)}\right] \\
&\quad\leq \sum_{k=2}^{p}\sum_{j=0}^{p-k}\binom{p-k}{j}\binom{p}{k}2^{k+j}3^j\gamma_n^{k+2j}\mathbb{E}\left[\|\xi_{n+1}\|^k\|Z_n - m\|^{2p-k-2j}\right].
\end{aligned}$$

For all $n \geq 1$, since $\|\xi_{n+1}\| \leq 2$,

$$\begin{aligned}
&\sum_{k=2}^{p}\sum_{j=0}^{p-k}\binom{p-k}{j}\binom{p}{k}2^{k+j}3^j\gamma_n^{k+2j}\mathbb{E}\left[\|\xi_{n+1}\|^k\|Z_n - m\|^{2p-k-2j}\right] \\
&\qquad\leq \sum_{k=2}^{p}\sum_{j=0}^{p-k}\binom{p-k}{j}\binom{p}{k}2^{2k+j}3^j\gamma_n^{k+2j}\mathbb{E}\left[\|Z_n - m\|^{2p-k-2j}\right].
\end{aligned}$$

Finally, using Cauchy-Schwarz's inequality and by induction, we get

$$\begin{aligned}
&\sum_{k=2}^{p}\sum_{k=2}^{p}\sum_{j=0}^{p-k}\binom{p-k}{j}\binom{p}{k}2^{2k+j}3^j\gamma_n^{k+2j}\mathbb{E}\left[\|Z_n - m\|^{2p-k-2j}\right] \\
&\quad\leq \sum_{k=2}^{p}\sum_{j=0}^{p-k}\binom{p-k}{j}\binom{p}{k}2^{2k+j}3^j\gamma_n^{k+2j}\sqrt{\mathbb{E}\left[\|Z_n - m\|^{2(p-1-j)}\right]}\sqrt{\mathbb{E}\left[\|Z_n - m\|^{2(p-k-j+1)}\right]} \\
&\quad\leq \sum_{k=2}^{p}\sum_{j=0}^{p-k}\binom{p-k}{j}\binom{p}{k}2^{2k+j}3^j\gamma_n^{k+2j}\sqrt{M_{p-1-j}}\sqrt{M_{p-k-j+1}}.
\end{aligned}$$

Moreover, for all $k \geq 2$ and $j \geq 0$, $\gamma_n^{2j+k} = O(\gamma_n^2)$, so there is a constant C'_p such that for all $n \geq 1$:

$$\sum_{k=2}^{p} \binom{p}{k} \mathbb{E}\left[\left(2\gamma_n \langle \xi_{n+1}, Z_n - m \rangle\right)^k \left(\|Z_n - m\|^2 + 6\gamma_n^2 \right)^{p-k} \right] \leq C'_p \gamma_n^2$$

Remark that C'_p does not depend on n. Since Z_1 is chosen bounded or deterministic, we get by induction

$$\begin{aligned}
\mathbb{E}\left[\|Z_{n+1} - m\|^{2p} \right] &\leq \mathbb{E}\left[\|Z_n - m\|^{2p} \right] + \left(C_p + C'_p \right) \gamma_n^2 \\
&\leq \mathbb{E}\left[\|Z_1 - m\|^{2p} \right] + \left(C_p + C'_p \right) \sum_{k=1}^{n} \gamma_k^2 \\
&\leq \mathbb{E}\left[\|Z_1 - m\|^{2p} \right] + \left(C_p + C'_p \right) \sum_{k=1}^{\infty} \gamma_k^2 \\
&\leq M_p,
\end{aligned}$$

which concludes the induction.

Applying Markov's inequality, for all integer $p \geq 1$:

$$\mathbb{P}\left[\|Z_n - m\| \geq c n^{1-\alpha} \right] \leq \frac{\mathbb{E}\left[\|Z_n - m\|^{2p} \right]}{(c n^{1-\alpha})^{2p}} \leq \frac{M_p}{(c n^{1-\alpha})^{2p}}.$$

To get the result, we take $p \geq \frac{4-\alpha}{2(1-\alpha)}$. This is possible since $\alpha \neq 1$.

□

Proof of Proposition 4.3.1. A lower bound for $\|Z_n - m - \Phi(Z_n)\|$ is obtained by using decomposition (A.1). Using Corollary 2.1, for all $h \in H$,

$$\begin{aligned}
\|\Phi(m+h)\| &\leq \left\| \int_0^1 \Gamma_{m+th}(h) dt \right\| \\
&\leq \int_0^1 \|\Gamma_{m+th}(h)\| \, dt \\
&\leq C\|h\|.
\end{aligned}$$

So, there is a rank n_0 such that for all $n \geq n_0$,

$$\begin{aligned}
\|h - \gamma_n \Phi(m+h)\| &\geq \left| \|h\| - \gamma_n \|\Phi(m+h)\| \right| \\
&\geq \|h\| - C\gamma_n \|h\|.
\end{aligned}$$

In a particular case, for all $n \geq n_0$,

$$\|Z_n - m - \gamma_n \Phi(Z_n)\| \geq (1 - C\gamma_n) \|Z_n - m\|.$$

Since $\lim_{n\to\infty} \mathbb{E}\left[\|Z_n - m\|^2\right] = 0$, there is a rank n_0' such that for all $n \geq n_0'$,

$$\mathbb{E}\left[\|\xi_{n+1}\|^2\right] = 1 - \mathbb{E}\left[\|\Phi(Z_n)\|^2\right] \geq 1 - C^2\mathbb{E}\left[\|Z_n - m\|^2\right].$$

Finally, since (ξ_{n+1}) is a sequence of martingale differences adapted to the filtration $(\mathcal{F}_n)$, there is a rank $n_1 \geq n_0'$ such that for all $n \geq n_1$,

$$\begin{aligned}\mathbb{E}\left[\|Z_{n+1} - m\|^2\right] &\geq (1 - C\gamma_n)^2 \mathbb{E}\left[\|Z_n - m\|^2\right] + \gamma_n^2 \left(1 - 2C^2\mathbb{E}\left[\|Z_n - m\|^2\right]\right) \\ &\geq (1 - 2C\gamma_n) \mathbb{E}\left[\|Z_n - m\|^2\right] + \gamma_n^2.\end{aligned}$$

We can prove by induction that there is a positive constant C_0 such that for all $n \geq n_1$,

$$\mathbb{E}\left[\|Z_n - m\|^2\right] \geq \frac{C_0}{n^\alpha}.$$

To conclude the proof, we just have to take $C' := \min\left\{\min_{1\leq n\leq n_1}\left\{\mathbb{E}\left[\|Z_n - m\|^2\right] n^\alpha\right\}, C_0\right\}$. □

A.3 Proofs of Proposition 4.4.1 and Theorem 4.4.1

Proof of Proposition 4.4.1. As in [Pin94], For all integer $j, n \geq 1$ such that $2 \leq j \leq n$, let us define

$$\begin{aligned}f_{j,n} &:= \sum_{k=1}^{j-1} \gamma_k \beta_{n-1} \beta_k^{-1} \xi_{k+1}, \\ d_{j,n} &:= f_{j,n} - f_{j-1,n} = \beta_{n-1}\beta_{j-1}^{-1}\gamma_{j-1}\xi_j, \\ e_{j,n} &:= \mathbb{E}\left[e^{\|d_{j,n}\|} - 1 - \|d_{j,n}\| | \mathcal{F}_{j-1}\right],\end{aligned}$$

with $f_{0,n} = 0$. Remark that for all $k \leq n - 1$,

$$\mathbb{E}\left[\beta_{n-1}\beta_k^{-1}\xi_{k+1} | \mathcal{F}_k\right] = 0.$$

We cannot apply directly Theorem 3.1 of [Pin94] because the sequence $\left(\beta_{n-1}\beta_k^{-1}\xi_{k+1}\right)$ is not properly a martingale differences sequence. As in [Pin94], for all $t \in [0, 1]$, let us define

$u(t) := \|x + tv\|$, with $x, v \in H$. We have for all $t \in [0,1]$, $u'(t) \leq \|v\|$ and $\left(u^2(t)\right)'' \leq 2\|v\|^2$. Moreover, since for all $u \in \mathbb{R}$, $\cosh u \geq \sinh u$, we also get

$$(\cosh u)''(t) \leq \|v\|^2 \cosh u$$

Let $\varphi(t) := \mathbb{E}\left[\cosh\left(\|f_{j-1,n} + td_{j,n}\|\right)|\mathbb{F}_j\right]$,

$$\begin{aligned}\varphi''(t) &\leq \mathbb{E}\left[\|d_{j,n}\|^2 \cosh\left(\|f_{j-1,n} + td_{j,n}\|\right)|\mathcal{F}_{j-1}\right] \\ &\leq \mathbb{E}\left[\|d_{j,n}\|^2 e^{t\|d_{j,n)}\|} \cosh\left(\|f_{j-1,n}\|\right)|\mathcal{F}_{j-1}\right].\end{aligned}$$

Moreover, since (ξ_n) is a sequence of martingale differences adapted to the filtration $(\mathcal{F}_n)$, for all $j \geq 1$, $\mathbb{E}\left[d_{j,n}|\mathcal{F}_{j-1}\right] = 0$ and $\varphi'(0) = 0$. We get for all $j \geq 1$ such that $j \leq n$,

$$\begin{aligned}\mathbb{E}\left[\cosh\left(\|f_{j,n}\|\right)|\mathcal{F}_{j-1}\right] &= \varphi(1) \\ &= \varphi(0) + \int_0^1 (1-t)\varphi''(t)dt \\ &\leq (1 + e_{j,n}) \cosh\left(\|f_{j-1,n}\|\right).\end{aligned}$$

Let $G_1 := 1$ and for all $2 \leq j \leq n$, let $G_j := \frac{\cosh(\|f_{j,n}\|)}{\prod_{i=2}^{j}(1+e_{i,n})}$. Using previous inequality, since $e_{j+1,n}$ is $\mathcal{F}_j$-measurable,

$$\begin{aligned}\mathbb{E}\left[G_{j+1}|\mathcal{F}_j\right] &= \mathbb{E}\left[\frac{\cosh\left(\left\|f_{j+1,n}\right\|\right)}{\prod_{i=2}^{j+1}(1+e_{i,n})}\Big|\mathcal{F}_j\right] \\ &= \frac{\mathbb{E}\left[\cosh\left(\left\|f_{j+1,n}\right\|\right)|\mathcal{F}_j\right]}{\prod_{i=2}^{j+1}(1+e_{i,n})} \\ &\leq \frac{(1+e_{j+1,n})\cosh\left(\left\|f_{j,n}\right\|\right)}{\prod_{i=2}^{j+1}(1+e_{i,n})} \\ &= G_j.\end{aligned}$$

By induction, $\mathbb{E}[G_n] \leq \mathbb{E}[G_1] \leq 1$. Finally,

$$\begin{aligned}
\mathbb{P}[\|f_{n,n}\| \geq r] &\leq \mathbb{P}\left[G_n \geq \frac{\cosh r}{\left\|\prod_{j=2}^{n}(1+e_{j,n})\right\|}\right] \\
&\leq \mathbb{P}\left[G_n \geq \frac{1}{2}\frac{\exp(r)}{\left\|\prod_{j=2}^{n}(1+e_{j,n})\right\|}\right] \\
&\leq 2\mathbb{E}[G_n]\, e^{-r}\left\|\prod_{j=2}^{n}(1+e_{j,n})\right\| \\
&\leq 2e^{-r}\left\|\prod_{j=2}^{n}(1+e_{j,n})\right\|.
\end{aligned}$$

□

Proof of Theorem 4.4.1. Using Theorem 4.3.1, one can check that $\mathbb{E}[\|\beta_{n-1}R_n\|] = O\left(\frac{1}{n^\alpha}\right)$. Indeed, applying Lemma 4.5.1,

$$\begin{aligned}
\mathbb{E}[\|\beta_{n-1}R_n\|] &\leq \sum_{k=1}^{n-1}\gamma_k\|\beta_{n-1}\beta_k^{-1}\|\mathbb{E}[\|\delta_k\|] \\
&\leq C_m\sum_{k=1}^{n-1}\gamma_k\|\beta_{n-1}\beta_k^{-1}\|\mathbb{E}\left[\|Z_k-m\|^2\right].
\end{aligned}$$

Moreover, with calculus similar to the ones for the upper bound of the martingale term in the proof of Lemma 4.3.1 and applying Theorem 4.3.1,

$$\begin{aligned}
C_m\sum_{k=1}^{n-1}\gamma_k\|\beta_{n-1}\beta_k^{-1}\|\mathbb{E}\left[\|Z_k-m\|^2\right] &\leq C_m C' c_\gamma\sum_{k=1}^{n-1}\frac{1}{k^{2\alpha}}\|\beta_{n-1}\beta_k^{-1}\| \\
&= O\left(\frac{1}{n^\alpha}\right).
\end{aligned}$$

Finally, the deterministic term $\beta_{n-1}(Z_1-m)$ converges exponentially to 0. So, there are positive constants C_1, C_1', C_2 such that

$$\mathbb{P}[\|Z_n-m\| \geq t] \leq \mathbb{P}\left[\|\beta_{n-1}M_n\| \geq \frac{t}{2}\right] + \frac{C_1 e^{-C_1' n^{1-\alpha}}}{t^2} + \frac{C_2}{n^\alpha}\frac{1}{t} \tag{A.8}$$

We now give a "good" choice of sequences (N_n) and (σ_n^2) to apply Corollary 4.4.2.

Step 1 : Choice of N_n : Since $\left\|\beta_{n-1}\beta_k^{-1}\right\| \leq e^{-\lambda_{\min}\sum_{j=k+1}^{n}\gamma_j}$ and since $\|\xi_{n+1}\| \leq 2$, we have $\|\beta_{n-1}\beta_k^{-1}\xi_{k+1}\| \leq 2c_2\gamma_k e^{-\lambda_{\min}\sum_{j=k+1}^{n-1}\gamma_j}$ if $k \neq n-1$, where $\lambda_{\min}$ is the smallest eigenvalue of Γ_m. With calculus analogous to the ones for the bound of the martingale term in the proof of Lemma 4.3.1, one can check that if $k \leq n/2$,

$$\|\beta_{n-1}\beta_k^{-1}\gamma_k\xi_{k+1}\| \leq 2c_2 e^{-2\lambda_{\min}c_\gamma n^{1-\alpha}}\gamma_1.$$

Moreover, if $k \geq n/2$ and $k \neq n-1$,

$$2c_2\gamma_k e^{-\lambda_{\min}\sum_{j=k+1}^{n-1}\gamma_j} \leq 2c_2\gamma_k \leq 2c_2 2^\alpha c_\gamma \frac{1}{n^\alpha}.$$

Finally, if $k = n-1$,

$$\|\beta_{n-1}\beta_k^{-1}\gamma_{n-1}\xi_n\| \leq c_\gamma 2^\alpha \frac{1}{n^\alpha}.$$

Let $C_N := \max\left\{\sup_{n\geq 1}\left\{e^{-2\lambda_{\min}c_\gamma n^{1-\alpha}}n^\alpha\right\}, 2c_2, 1\right\}$, thus for all $n \geq 1$, $\sup_{k\leq n-1}\left\{\|\beta_{n-1}\beta_k^1\gamma_k\xi_{k+1}\|\right\} \leq \frac{C_N}{n^\alpha}$. So we take

$$N_n = \frac{C_N}{n^\alpha}. \tag{A.9}$$

Step 2 : Choice of σ_n^2 : In the same way, for n enough large, we have $\sum_{k=1}^{n-1}\mathbb{E}\left[\left\|\beta_{n-1}\beta_k^{-1}\gamma_k\xi_{k+1}\right\|^2|\mathcal{F}_k\right] \leq \frac{2^{\alpha+1}c_\gamma}{c_m}\frac{1}{n^\alpha}$. Indeed, we just have to divide the sum into two parts, the first one converges exponentially to 0, and is lower than the second one from a certain rank. For n large enough, we can take

$$\sigma_n^2 = c_\gamma \frac{2^{1+\alpha}}{c_m}\frac{1}{n^\alpha}. \tag{A.10}$$

Using inequality (A.8) and Corollary 4.4.2,

$$\mathbb{P}\left[\|Z_n - m\| \geq t\right] \leq 2\exp\left(-\frac{(t/2)^2}{2(\sigma_n^2 + N_n(t/2)/3)}\right) + \frac{C_1 e^{-C_1' n^{1-\alpha}}}{t^2} + \frac{C_2}{n^\alpha}\frac{1}{t} =: f(t,n).$$

We look for values of t for which $f(t,n) \leq \delta$. We search to solve :

$$\begin{aligned} 2\exp\left(-\frac{(t/2)^2}{2(\sigma_n^2 + N_n t/6)}\right) &\leq \delta/2, \\ \frac{C_1 e^{-C_1' n^{1-\alpha}}}{t^2} &\leq \delta/4, \\ \frac{C_2}{n^\alpha}\frac{1}{t} &\leq \delta/4. \end{aligned}$$

We get (see [TY14] , Appendix A, for the exponential term) :

$$\begin{aligned} t &\geq 4\left(\frac{N_n}{3} + \sigma_n\right)\ln\frac{4}{\delta}, \\ t &\geq 2\sqrt{\frac{C_1 e^{-C_1' n^{1-\alpha}}}{\delta}}, \\ t &\geq 4\frac{C_2}{n^\alpha}\frac{1}{\delta}. \end{aligned}$$

Let us take a rank n_δ such that for all $n \geq n_\delta$, with (A.10),

$$\begin{aligned} 4\left(\frac{N_n}{3} + \sigma_n\right)\ln\frac{4}{\delta} &\geq 2\sqrt{\frac{C_1 e^{-C_1' n^{1-\alpha}}}{\delta}}, \\ 4\left(\frac{N_n}{3} + \sigma_n\right)\ln\frac{4}{\delta} &\geq 4\frac{C_2}{n^\alpha}\frac{1}{\delta}. \end{aligned}$$

Thus, for all $n \geq n_\delta$, with probability at least $1-\delta$:

$$\|Z_n - m\| \leq 4\left(\frac{N_n}{3} + \sigma_n\right)\ln\frac{4}{\delta}. \tag{A.11}$$

□

Chapitre 5

Estimating the geometric median in Hilbert spaces with stochastic gradient algorithms : L^p and almost sure rates of convergence

Résumé

Au Chapitre 4, nous avons donné des boules de confiance non asymptotiques pour les estimateurs de la médiane introduits par [CCZ13]. Cependant, ces résultats ne nous permettaient pas de pouvoir établir la convergence des estimateurs de la "Median Covariation Matrix", par exemple. Ce travail a pour but d'étudier plus précisément le comportement asymptotique des estimateurs récursifs de la médiane géométrique définis par (4.3) et (4.4). Pour cela, on donne les vitesses L^p de l'algorithme de gradient stochastique (Théorème 5.4.1) ainsi que celles de l'algorithme moyenné (Théorème 5.4.2). De plus on montre que ces dernières vitesses sont optimales (Proposition 5.4.1). Enfin, on donne les vitesses de convergences presque sûre de l'algorithme de Robbins-Monro (Théorème 5.5.1) et de son moyenné (Corollaire 5.5.1).

This Chapter is based on a work accepted in Journal of Multivariate Analysis ([GB15]).

Abstract

The geometric median, also called L^1-median, is often used in robust statistics. Moreover, it is more and more usual to deal with large samples taking values in high dimensional spaces. In this context, a fast recursive estimator has been introduced by [CCZ13]. This work aims at studying more precisely the asymptotic behavior of the estimators of the geometric median based on such non linear stochastic gradient algorithms. The L^p rates of convergence as well as almost sure rates of convergence of these estimators are derived in general separable Hilbert spaces. Moreover, the optimal rates of convergence in quadratic mean of the averaged algorithm are also given.

5.1 Introduction

The geometric median, also called L^1-median, is a generalization of the real median introduced by [Hal48]. In the multivariate case, it is closely related to the Fermat-Webber's problem (see [WF29]), which consists in finding a point minimizing the sum of distances from given points. This is a well known convex optimization problem. The literature is very wide on the estimation of the solution of this problem. One of the most usual method is to use Weiszfeld's algorithm (see [Kuh73]), or more recently, to use the algorithm proposed by [BS15].

In the more general context of Banach spaces, [Kem87] gives many properties on the median, such as its existence, its uniqueness, and maybe the most important, its robustness. Because of this last property, the median is often used in robust statistics. For example, [Min14] considers it in order to get much tighter concentration bounds for aggregation of estimators. [CCM12] propose a recursive algorithm using the median for clustering, which is few sensitive to outliers than the k-means. One can also see [CC14], [Cue14], [BBT$^+$11] or [Ger08] among others for other examples.

In this context, several estimators of the median are proposed in the literature. In the multivariate case, one of the most usual method is to consider the Fermat-Webber's problem generated by the sample, and to solve it using Weiszfeld's algorithm (see [VZ00] and [MNO10] for example). This method is fast, but can encounter many difficulties when we deal with a large sample taking values in relatively high dimensional spaces. Indeed, since it requires to store all the data, it can be difficult or impossible to perform the algorithm.

Dealing with high dimensional of functional data is more and more usual. There exists a large recent literature on functional data analysis (see [BSGV14], [FV06] or [SR05] for example), but few of them speak about robustness (see [Cad01] and [Cue14]).

In this large sample and high dimensional context, recursive algorithms have been introduced by [CCZ13]; a stochastic gradient algorithm, or Robbins-Monro algorithm (see [RM51], [BDM01], [Duf97], [BMP90], [KY03] among others), and its averaged version (see [PJ92]). It enables us to estimate the median in Hilbert spaces, whose dimension is not necessarily finite, such as functional spaces. The advantage of these algorithms is that they can treat all the data, can be simply updated, and do not require too much computational efforts. Moreover, it has been proven in [CCZ13] that the averaged version and the estimator proposed by [VZ00] have the same asymptotic distribution. Other properties were given, such as the strong consistency of these algorithms. Moreover, the optimal rate of convergence in quadratic mean of the Robbins-monro algorithm as well as non asymptotic confidence balls for both algorithms are given in [CCGB15].

The aim of this work is to give new asymptotic convergence properties in order to have a deeper knowledge of the asymptotic behaviour of these algorithms. Optimal L^p rates of convergence for the Robbins-Monro algorithm are given. This enables, in a first time, to get the optimal rate of convergence in quadratic mean of the averaged algorithm. In a second time, it enables us to get the L^p rates of convergence. In a third time, thanks to these results, applying Borel-Cantelli's Lemma, we give an almost sure rate of convergence of the Robbins-Monro algorithm. Finally, applying a law of large numbers for martingales (see [Duf97] for example), we give an almost sure rate of convergence of the averaged algorithm.

The paper is organized as follows. In Section 5.2, we recall the definition of the median and some important convexity properties. The Robbins-Monro algorithm and its averaged version are defined in Section 5.3. After recalling the rate of convergence in quadratic mean of the Robbins-Monro algorithm given by [CCGB15], we give the L^p-rates of convergence of the stochastic gradient algorithm as well as the optimal rate of convergence in quadratic mean of the averaged algorithm in Section 5.4. Finally, almost sure rates of convergence of the algorithms are given in Section 5.5. The lemma that help understanding the structure of the proofs are given all along the text, but the proofs are postponed in an Section 5.6 and in a supplementary file.

5.2 Definitions and convexity properties

Let H be a separable Hilbert space, we denote by $\langle .,. \rangle$ its inner product and by $\|.\|$ the associated norm. Let X be a random variable taking values in H, the geometric median m of X is defined by

$$m := \arg\min_{h \in H} \mathbb{E}\left[\|X-h\| - \|X\|\right]. \tag{5.1}$$

We suppose from now that the following assumptions are fulfilled :

- **(A1)** X is not concentrated on a straight line : for all $h \in H$, there is $h' \in H$ such that $\langle h, h' \rangle = 0$ and $\mathrm{Var}\left(\langle X, h' \rangle\right) > 0$.
- **(A2)** X is not concentrated around single points : there is a positive constant C such that for all $h \in H$,

$$\mathbb{E}\left[\frac{1}{\|X-h\|}\right] \leq C, \qquad \mathbb{E}\left[\frac{1}{\|X-h\|^2}\right] \leq C.$$

Remark that since $\mathbb{E}\left[\frac{1}{\|X-h\|^2}\right] \leq C$, as a particular case, $\mathbb{E}\left[\frac{1}{\|X-h\|}\right] \leq \sqrt{C}$. Note that for the sake of simplicity, even if it means supposing $C \geq 1$, we take C instead of $\sqrt{C}$. Assumption **(A1)** ensures that the median m is uniquely defined [Kem87]. Assumption **(A2)** is

not restrictive whenever $d \geq 3$, where d is the dimension of H, not necessarily finite (see [CCZ13] and [Cha92] for more details). Note that many convergence results can be found without Assumption **(A2)** if we deal with data taking values in compact sets (see [ADPY12] or [Yan10] for example).

Let G be the function we would like to minimize. It is defined for all $h \in H$ by

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

This function is convex and many convexity properties are given in [Cha92], [Ger08], [CCZ13] and [CCGB15]. We recall two important ones :

(P1) G is Fréchet-differentiable and its gradient is given for all $h \in H$ by

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

The median m is the unique zero of Φ.

(P2) G is twice differentiable and for all $h \in H$, Γ_h stands for the Hessian of G at h. Thus, H admits an orthonormal basis composed of eigenvectors of Γ_h, and let $(\lambda_{i,h})$ be the eigenvalues of Γ_h, we have $0 \leq \lambda_{i,h} \leq C$.

Moreover, for all positive constant A, there is a positive constant c_A such that for all $h \in \mathcal{B}(0, A)$, $c_A \leq \lambda_{i,h} \leq C$.

As a particular case, let $\lambda_{\min}$ be the smallest eigenvalue of Γ_m, there is a positive constant c_m such that $0 < c_m < \lambda_{\min} \leq C$.

5.3 The algorithms

Let $X_1, ..., X_n, ...$ be independent random variables with the same law as X. We recall the algorithm for estimation of the geometric median, defined as follows :

$$Z_{n+1} = Z_n + \gamma_n \frac{X_{n+1} - Z_n}{\|X_{n+1} - Z_n\|}, \tag{5.2}$$

where the initialization Z_1 is chosen bounded ($Z_1 = X_1 \mathbb{1}_{\{\|X_1\| \leq M\}}$ for example) or deterministic. The sequence (γ_n) of steps is positive and verifies the following usual conditions

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

The averaged version of the algorithm (see [PJ92], [CCZ13]) is given iteratively by

$$\overline{Z}_{n+1} = \overline{Z}_n - \frac{1}{n+1}\left(\overline{Z}_n - Z_{n+1}\right), \tag{5.3}$$

where $\overline{Z}_1 = Z_1$. This can be written as $\overline{Z}_n = \frac{1}{n}\sum_{k=1}^{n} Z_k$.

The algorithm defined by (5.2) is a stochastic gradient or Robbins-Monro algorithm. Indeed, it can be written as follows :

$$Z_{n+1} = Z_n - \gamma_n \Phi(Z_n) + \gamma_n \xi_{n+1}, \tag{5.4}$$

where $\xi_{n+1} := \Phi(Z_n) + \frac{X_{n+1}-Z_n}{\|X_{n+1}-Z_n\|}$. Let $\mathcal{F}_n$ be the σ-algebra defined by $\mathcal{F}_n := \sigma(X_1, ..., X_n) = \sigma(Z_1, ..., Z_n)$. Thus, (ξ_n) is a sequence of martingale differences adapted to the filtration $(\mathcal{F}_n)$. Indeed, for all $n \geq 1$, we have almost surely $\mathbb{E}[\xi_{n+1}|\mathcal{F}_n] = 0$. Linearizing the gradient,

$$Z_{n+1} - m = (I_H - \gamma_n \Gamma_m)(Z_n - m) + \gamma_n \xi_{n+1} - \gamma_n \delta_n, \tag{5.5}$$

with $\delta_n := \Phi(Z_n) - \Gamma_m(Z_n - m)$. Note that there is a positive deterministic constant C_m such that for all $n \geq 1$ (see [CCGB15]), almost surely,

$$\|\delta_n\| \leq C_m \|Z_n - m\|^2. \tag{5.6}$$

Moreover, since $\Phi(Z_n) = \int_0^1 \Gamma_{m+t(Z_n-m)}(Z_n - m)dt$, applying convexity property **(P2)**, one can check that almost surely

$$\|\delta_n\| \leq 2C \|Z_n - m\|. \tag{5.7}$$

5.4 L^p rates convergence of the algorithms

We now consider a step sequence (γ_n) of the form $\gamma_n = c_\gamma n^{-\alpha}$ with $c_\gamma > 0$ and $\alpha \in (1/2, 1)$. The optimal rate of convergence in quadratic mean of the Robbins-Monro algorithm is given in [CCGB15]. Indeed, it was proven that there are positive constants c', C' such that for all $n \geq 1$,

$$\frac{c'}{n^\alpha} \leq \mathbb{E}\left[\|Z_n - m\|^2\right] \leq \frac{C'}{n^\alpha}. \tag{5.8}$$

Moreover, the L^p rates of convergence were not given, but it was proven that the p-th moments are bounded for all integer p : there exists a positive constant M_p such that for all

$n \geq 1$,

$$\mathbb{E}\left[\|Z_n - m\|^{2p}\right] \leq M_p. \tag{5.9}$$

5.4.1 L^p rates of convergence of the Robbins-Monro algorithm

Theorem 5.4.1. *Assume **(A1)** and **(A2)** hold. For all $p \geq 1$, there is a positive constant K_p such that for all $n \geq 1$,*

$$\mathbb{E}\left[\|Z_n - m\|^{2p}\right] \leq \frac{K_p}{n^{p\alpha}}. \tag{5.10}$$

As a corollary, applying Cauchy-Schwarz's inequality, for all $p \geq 1$ and for all $n \geq 1$,

$$\mathbb{E}\left[\|Z_n - m\|^{p}\right] \leq \frac{\sqrt{K_p}}{n^{\frac{p\alpha}{2}}}. \tag{5.11}$$

The proof is given in Section 5.6. Since it was proven (see [CCGB15]) that the rate for $p = 1$ is the optimal one, one can check, applying Hölder's inequality, that the given ones for $p \geq 2$ are also optimal. In order to prove this theorem with a strong induction on p and n, we have to introduce two technical lemma. The first one gives an upper bound for $\mathbb{E}\left[\|Z_{n+1} - m\|^{2p}\right]$ when inequality (5.10) is verified for all integer from 0 to $p-1$, i.e when the strong induction assumptions are verified.

Lemma 5.4.1. *Assume **(A1)** and **(A2)** hold, let $p \geq 2$, if inequality (5.10) is verified for all integer fom 0 to $p-1$, there are a rank n_α and non-negative constants c_0, C_1, C_2 such that for all $n \geq n_\alpha$,*

$$\mathbb{E}\left[\|Z_{n+1} - m\|^{2p}\right] \leq (1 - c_0\gamma_n)\,\mathbb{E}\left[\|Z_n - m\|^{2p}\right] + \frac{C_1}{n^{(p+1)\alpha}} + C_2\gamma_n\mathbb{E}\left[\|Z_n - m\|^{2p+2}\right]. \tag{5.12}$$

The proof is given in Section 5.6. The following lemma gives an upper bound of $\mathbb{E}\left[\|Z_{n+1} - m\|^{2p+2}\right]$ when inequality (5.10) is verified for all integer from 0 to $p-1$, i.e when the strong induction assumptions are verified.

Lemma 5.4.2. *Assume **(A1)** and **(A2)** hold, let $p \geq 2$, if inequality (5.10) is verified for all integer from 0 to $p-1$, there are a rank n_α and non-negative constants C_1', C_2' such that for all $n \geq n_\alpha$,*

$$\mathbb{E}\left[\|Z_{n+1} - m\|^{2p+2}\right] \leq \left(1 - \frac{2}{n}\right)^{p+1}\mathbb{E}\left[\|Z_n - m\|^{2p+2}\right] + \frac{C_1'}{n^{(p+2)\alpha}} + C_2'\gamma_n^2\mathbb{E}\left[\|Z_n - m\|^{2p}\right]. \tag{5.13}$$

The proof is given in 5.6. Note that for the sake of simplicity, we denote by the same way the ranks in Lemma 5.4.1 and Lemma 5.4.2.

5.4.2 Optimal rate of convergence in quadratic mean and L^p rates of converge of the averaged algorithm

As done in [CCZ13] and [Pel00], summing equalities (5.5) and applying Abel's transform, we get

$$n\Gamma_m\left(\overline{Z}_n - m\right) = \frac{T_1}{\gamma_1} - \frac{T_{n+1}}{\gamma_n} + \sum_{k=2}^{n} T_k\left(\frac{1}{\gamma_k} - \frac{1}{\gamma_{k-1}}\right) + \sum_{k=1}^{n}\delta_k + \sum_{k=1}^{n}\xi_{k+1}, \tag{5.14}$$

with $T_k := Z_k - m$. Using this decomposition and Theorem 5.4.1, we can derive the L^p rates of convergence of the averaged algorithm.

Theorem 5.4.2. *Assume* ***(A1)*** *and* ***(A2)*** *hold, for all integer $p \geq 1$, there is a positive constant A_p such that for all $n \geq 1$,*

$$\mathbb{E}\left[\left\|\overline{Z}_n - m\right\|^{2p}\right] \leq \frac{A_p}{n^p}. \tag{5.15}$$

The proof is given in Section 5.6. It heavily relies on Theorem 5.4.1 and on the following lemma which gives a bound of the p-th moments of the sum of (non necessarily independent) random variables. Note that this is probably not a new result but we were not able to find a proof in a published reference.

Lemma 5.4.3. *Let $Y_1, ..., Y_n$ be random variables taking values in a normed vector space such that for all positive constant q and for all $k \geq 1$, $\mathbb{E}\left[\|Y_k\|^q\right] < \infty$. Thus, for all constants $a_1, ..., a_n$ and for all integer p,*

$$\mathbb{E}\left[\left\|\sum_{k=1}^{n} a_k Y_k\right\|^p\right] \leq \left(\sum_{k=1}^{n} |a_k|\left(\mathbb{E}\left[\|Y_k\|^p\right]\right)^{\frac{1}{p}}\right)^p \tag{5.16}$$

The proof is given in a supplementary file. Finally, the following proposition ensures that the rate of convergence in quadratic mean given by Theorem 5.4.2 is the optimal one.

Proposition 5.4.1. *Assume* ***(A1)*** *and* ***(A2)*** *hold, there is a positive constant c such that for all $n \geq 1$,*

$$\mathbb{E}\left[\left\|\overline{Z}_n - m\right\|^2\right] \geq \frac{c}{n}.$$

Note that applying Hölder's inequality, previous proposition also ensures that the L^p rates of convergence given by Theorem 5.4.2 are the optimal ones.

5.5 Almost sure rates of convergence

It is proven in [CCZ13] that the Robbins-Monro algorithm converges almost surely to the geometric median. A direct application of Theorem 5.4.1 and Borel-Cantelli's lemma gives the following rates of convergence.

Theorem 5.5.1. *Assume **(A1)** and **(A2)** hold, for all $\beta < \alpha$,*

$$\|Z_n - m\| = o\left(\frac{1}{n^{\beta/2}}\right) \quad a.s. \tag{5.17}$$

The proof is given in a supplementary file. As a corollary, using decomposition (5.14) and Theorem 5.5.1, we get the following bound of the rate of convergence of the averaged algorithm :

Corollary 5.5.1. *Assume **(A1)** and **(A2)** hold, for all $\delta > 0$,*

$$\left\|\overline{Z}_n - m\right\| = o\left(\frac{(\ln n)^{\frac{1+\delta}{2}}}{\sqrt{n}}\right) \quad a.s. \tag{5.18}$$

The proof is given in a supplementary file.

Acknowledgements

The author thanks Hervé Cardot and Peggy Cénac for their patience, their trust, and their advice which were very helpful.

5.6 Proofs

5.6.1 Proofs of Section 5.4.1

First we recall some technical inequalities (see [Pet95] for example).

Lemma 5.6.1. *Let a, b, c be positive constants. Thus,*

$$ab \leq \frac{a^2}{2c} + \frac{b^2 c}{2}, \qquad\qquad a \leq \frac{c}{2} + \frac{a^2}{2c}.$$

Moreover let k, p be positive integers and $a_1, ..., a_p$ be positive constants. Thus,

$$\left(\sum_{j=1}^{p} a_j\right)^k \leq p^{k-1} \sum_{j=1}^{p} a_j^k.$$

Proof of Lemma 5.4.2. We suppose from now that for all $k \leq p-1$, there is a positive constant K_k such that for all $n \geq 1$,

$$\mathbb{E}\left[\|Z_n - m\|^{2k}\right] \leq \frac{K_k}{n^{k\alpha}}. \tag{5.19}$$

Using decomposition (5.4) and Cauchy-Schwarz's inequality, since by definition of ξ_{n+1} we have $\|\xi_{n+1}\| - 2\langle \Phi(Z_n), \xi_{n+1}\rangle \leq 1$,

$$\begin{aligned}\|Z_{n+1} - m\|^2 &= \|Z_n - m - \gamma_n \Phi(Z_n)\|^2 + \gamma_n^2 \|\xi_{n+1}\|^2 + 2\gamma_n \langle Z_n - m - \gamma_n \Phi(Z_n), \xi_{n+1}\rangle \\ &\leq \|Z_n - m - \gamma_n \Phi(Z_n)\|^2 + \gamma_n^2 + 2\gamma_n \langle Z_n - m, \xi_{n+1}\rangle .\end{aligned}$$

Let $V_n := \|Z_n - m - \gamma_n \Phi(Z_n)\|^2$. Using previous inequality,

$$\begin{aligned}\|Z_{n+1} - m\|^{2p+2} &\leq \left(V_n + \gamma_n^2 + 2\gamma_n \langle \xi_{n+1}, Z_n - m\rangle\right)^{p+1} \\ &= \left(V_n + \gamma_n^2\right)^{p+1} + 2(p+1)\gamma_n \langle \xi_{n+1}, Z_n - m\rangle \left(V_n + \gamma_n^2\right)^p \qquad (5.20) \\ &\quad + \sum_{k=2}^{p+1} \binom{p+1}{k} \left(2\gamma_n \langle \xi_{n+1}, Z_n - m\rangle\right)^k \left(V_n + \gamma_n^2\right)^{p+1-k} . \qquad (5.21)\end{aligned}$$

We shall upper bound the three terms in (5.20) and (5.21). Applying Cauchy-Schwarz's inequality and since almost surely $\|\Phi(Z_n)\| \leq C \|Z_n - m\|$,

$$\begin{aligned}V_n &= \|Z_n - m\|^2 - 2\gamma_n \langle Z_n - m, \Phi(Z_n)\rangle + \gamma_n^2 \|\Phi(Z_n)\|^2 \\ &\leq \|Z_n - m\|^2 + 2C\gamma_n \|Z_n - m\|^2 + \gamma_n^2 C^2 \|Z_n - m\|^2 \\ &\leq (1 + c_\gamma C)^2 \|Z_n - m\|^2 . \qquad (5.22)\end{aligned}$$

We now bound the expectation of the first term in (5.20). Indeed,

$$\begin{aligned}\mathbb{E}\left[\left(V_n + \gamma_n^2\right)^{p+1}\right] &= \mathbb{E}\left[V_n^{p+1}\right] + (p+1)\gamma_n^2 \mathbb{E}\left[V_n^p\right] + \sum_{k=0}^{p-1} \binom{p+1}{k} \gamma_n^{2(p+1-k)} \mathbb{E}\left[V_n^k\right] \\ &\leq \mathbb{E}\left[V_n^{p+1}\right] + (p+1)\left(1 + c_\gamma C\right)^{2p} \gamma_n^2 \mathbb{E}\left[\|Z_n - m\|^{2p}\right] \\ &\quad + \sum_{k=0}^{p-1} \binom{p+1}{k} \left(1 + c_\gamma C\right)^{2k} \gamma_n^{2(p+1-k)} \mathbb{E}\left[\|Z_n - m\|^{2k}\right] .\end{aligned}$$

Applying inequality (5.19),

$$\begin{aligned}\sum_{k=0}^{p-1} \binom{p+1}{k} \left(1 + c_\gamma C\right)^{2k} \gamma_n^{2(p+1-k)} \mathbb{E}\left[\|Z_n - m\|^{2k}\right] &\leq \sum_{k=0}^{p-1} \binom{p+1}{k} \left(1 + c_\gamma C\right)^{2k} \gamma_n^{2(p+1-k)} \frac{K_k}{n^{k\alpha}} \\ &\leq \sum_{k=0}^{p-1} \binom{p+1}{k} \frac{\left(1 + c_\gamma C\right)^{2k} c_\gamma^{2(p+1-k)} K_k}{n^{(2p+2-k)\alpha}} \\ &= O\left(\frac{1}{n^{(p+3)\alpha}}\right) .\end{aligned}$$

As a conclusion, there is a non-negative constant A_1 such that for all $n \geq 1$,

$$\mathbb{E}\left[\left(V_n + \gamma_n^2\right)^{p+1}\right] \leq \mathbb{E}\left[V_n^{p+1}\right] + (p+1)\left(1 + c_\gamma C\right)^{2p} \gamma_n^2 \mathbb{E}\left[\|Z_n - m\|^{2p}\right] + \frac{A_1}{n^{(p+3)\alpha}}. \quad (5.23)$$

We now bound the second term in (5.20). Using the facts that (ξ_{n+1}) is a sequence of martingale differences adapted to the filtration $(\mathcal{F}_n)$ and that Z_n is $\mathcal{F}_n$-measurable,

$$\mathbb{E}\left[2(p+1)\gamma_n \left\langle \xi_{n+1}, Z_n - m \right\rangle \left(V_n + \gamma_n^2\right)^p\right] = 0. \quad (5.24)$$

Finally, we bound the last term in (5.21), denoted by $(*)$. Since almost surely $\|\xi_n\| \leq 2$, applying Cauchy-Schwarz's inequality,

$$\begin{aligned}(*) &\leq \sum_{k=2}^{p+1} \sum_{j=0}^{p+1-k} \binom{p+1}{k} \binom{p+1-k}{j} 2^k \gamma_n^{2j+k} \|\xi_{n+1}\|^k \|Z_n - m\|^k V_n^{p+1-k-j} \\ &\leq \sum_{k=2}^{p+1} \sum_{j=0}^{p+1-k} \binom{p+1}{k} \binom{p+1-k}{j} 2^{2k} \gamma_n^{2j+k} \|Z_n - m\|^k V_n^{p+1-k-j} \quad (5.25)\end{aligned}$$

Since almost surely $V_n \leq \left(1 + c_\gamma C\right)^2 \|Z_n - m\|^2$ (see inequality (5.22)),

$$\begin{aligned}(*) &\leq \sum_{k=2}^{p+1} \sum_{j=0}^{p+1-k} \binom{p+1}{k} \binom{p+1-k}{j} 2^{2k} \gamma_n^{2j+k} \left(1 + c_\gamma C\right)^{2p+2-2k-2j} \|Z_n - m\|^{2p+2-k-2j} \\ &= \sum_{k=2}^{p+1} \sum_{j=1}^{p+1-k} \binom{p+1}{k} \binom{p+1-k}{j} 2^{2k} \gamma_n^{2j+k} \left(1 + c_\gamma C\right)^{2p+2-2k-2j} \|Z_n - m\|^{2p+2-k-2j} \quad (5.26) \\ &+ \sum_{k=3}^{p+1} \binom{p+1}{k} 2^{2k} \gamma_n^k \left(1 + c_\gamma C\right)^{2p+2-2k-2j} \|Z_n - m\|^{2p+2-k} \\ &+ 16 \binom{p+1}{2} \gamma_n^2 \left(1 + c_\gamma^2 C^2\right)^{2p} \|Z_n - m\|^{2p}. \quad (5.27)\end{aligned}$$

We bound the expectation of the two first terms on the right-hand side of (5.26). For the first

one, applying Cauchy-Schwarz's inequality,

$$\begin{aligned}
&\mathbb{E}\left[\sum_{k=2}^{p+1}\sum_{j=1}^{p+1-k}\binom{p+1}{k}\binom{p+1-k}{j}2^{2k}\gamma_n^{2j+k}\left(1+c_\gamma C\right)^{2p+2-2k-2j}\left\|Z_n-m\right\|^{2p+2-k-2j}\right]\\
&\leq\sum_{k=2}^{p+1}\sum_{j=1}^{p+1-k}\binom{p+1}{k}\binom{p+1-k}{j}\\
&2^{2k}\gamma_n^{2j+k}\left(1+c_\gamma C\right)^{2p+2-2k-2j}\sqrt{\mathbb{E}\left[\left\|Z_n-m\right\|^{2(p-j)}\right]\mathbb{E}\left[\left\|Z_n-m\right\|^{2(p-k-j+2)}\right]}.
\end{aligned}$$

Applying inequality (5.19),

$$\begin{aligned}
&\mathbb{E}\left[\sum_{k=2}^{p+1}\sum_{j=1}^{p+1-k}\binom{p+1}{k}\binom{p+1-k}{j}2^{2k}\gamma_n^{2j+k}\left(1+c_\gamma C\right)^{2p+2-2k-2j}\left\|Z_n-m\right\|^{2p+2-k-2j}\right]\\
&\leq\sum_{k=2}^{p+1}\sum_{j=1}^{p+1-k}\binom{p+1}{k}\binom{p+1-k}{j}2^{2k}\gamma_n^{2j+k}\left(1+c_\gamma C\right)^{2p+2-2k-2j}\frac{\sqrt{K_{p-j}}}{n^{\frac{p-j}{2}\alpha}}\frac{\sqrt{K_{p-k-j+2}}}{n^{\frac{p-k-j+2}{2}\alpha}}\\
&=o\left(\frac{1}{n^{(p+2)\alpha}}\right).
\end{aligned}$$

Similarly, for the second term on the right-hand side of (5.26), applying Cauchy-Schwarz's inequality, let

$$\begin{aligned}
(**)&:=\sum_{k=3}^{p+1}\binom{p+1}{k}2^{2k}\gamma_n^k\left(1+c_\gamma C\right)^{2p+2-2k}\mathbb{E}\left[\left\|Z_n-m\right\|^{2p+2-k}\right]\\
&\leq\sum_{k=4}^{p+1}\binom{p+1}{k}2^{2k}\gamma_n^k\left(1+c_\gamma C\right)^{2p+2-2k}\mathbb{E}\left[\left\|Z_n-m\right\|^{2p+2-k}\right]\\
&.+64\binom{p+1}{3}\left(1+c_\gamma C\right)^{2p-4}\mathbb{E}\left[\left\|Z_n-m\right\|^{2p-1}\right]\\
&\leq\sum_{k=4}^{p+1}\binom{p+1}{k}2^{2k}\gamma_n^k\left(1+c_\gamma C\right)^{2p+2-2k}\sqrt{\mathbb{E}\left[\left\|Z_n-m\right\|^{2(p+3-k)}\right]\mathbb{E}\left[\left\|Z_n-m\right\|^{2(p-1)}\right]}\\
&+64\binom{p+1}{3}\left(1+c_\gamma C\right)^{2p-4}\mathbb{E}\left[\left\|Z_n-m\right\|^{2p-1}\right].
\end{aligned}$$

Applying Lemma 5.6.1 and inequality (5.19)

$$\begin{aligned}(**) &\leq \sum_{k=4}^{p+1} \binom{p+1}{k} 2^{2k} \gamma_n^k \left(1 + c_\gamma C\right)^{2p+2-2k} \frac{\sqrt{K_{p+3-k} K_{p-1}}}{n^{(p+1-k/2)\alpha}} \\ &+ 32 \binom{p+1}{3} \left(1 + c_\gamma C\right)^{2p-4} \gamma_n^3 \left(\mathbb{E}\left[\|Z_n - m\|^{2p} \right] + \mathbb{E}\left[\|Z_n - m\|^{2p-2} \right] \right) \\ &= O\left(\frac{1}{n^{(p+2)\alpha}} \right) + 32 \binom{p+1}{3} \left(1 + c_\gamma C\right)^{2p-4} \gamma_n^3 \mathbb{E}\left[\|Z_n - m\|^{2p} \right].\end{aligned}$$

Finally, let us denote by $(***)$ the expectation of the term in (5.21), there is a positive constant A_2 such that for all $n \geq 1$,

$$\begin{aligned}(***) &\leq \frac{A_2}{n^{(p+2)\alpha}} + 16 \binom{p+1}{2} \left(1 + c_\gamma C\right) \gamma_n^2 \mathbb{E}\left[\|Z_n - m\|^{2p} \right] \\ &+ 32 \binom{p+1}{3} \left(1 + c_\gamma C\right)^{2p-4} \gamma_n^3 \mathbb{E}\left[\|Z_n - m\|^{2p} \right]. \end{aligned} \tag{5.28}$$

Applying inequalities (5.23),(5.24) and (5.28), there are positive constants C_1'', C_2' such that for all $n \geq 1$,

$$\mathbb{E}\left[\|Z_{n+1} - m\|^{2p+2} \right] \leq \mathbb{E}\left[V_n^{p+1} \right] + \frac{C_1''}{n^{(p+2)\alpha}} + C_2' \gamma_n^2 \mathbb{E}\left[\|Z_n - m\|^{2p} \right]. \tag{5.29}$$

In order to conclude, we need to bound $\mathbb{E}\left[V_n^{p+1} \right]$. Applying Lemma 5.2 in [CCGB15], there are a positive constant c and a rank n_α such that for all $n \geq n_\alpha$,

$$\mathbb{E}\left[V_n^{p+1} \mathbb{1}_{\{\|Z_n - m\| \leq c n^{1-\alpha}\}} \right] \leq \left(1 - \frac{2}{n}\right)^{p+1} \mathbb{E}\left[\|Z_n - m\|^{2p+2} \right]. \tag{5.30}$$

Finally, since there is a positive constant c_0 such that almost surely $\|Z_n - m\| \leq c_0 n^{1-\alpha}$ and since almost surely $V_n \leq \left(1 + c_\gamma C\right)^2 \|Z_n - m\|^2$,

$$\begin{aligned}\mathbb{E}\left[V_n^{p+1} \mathbb{1}_{\{\|Z_n - m\| \geq c n^{1-\alpha}\}} \right] &\leq \left(1 + c_\gamma C\right)^{2p+2} \mathbb{E}\left[\|Z_n - m\|^{2p+2} \mathbb{1}_{\{\|Z_n - m\| \geq c n^{1-\alpha}\}} \right] \\ &\leq \left(1 + c_\gamma C\right)^{2p+2} c_0^{2p+2} n^{(2p+2)(1-\alpha)} \mathbb{E}\left[\mathbb{1}_{\{\|Z_n - m\| \geq c n^{1-\alpha}\}} \right] \\ &= \left(1 + c_\gamma C\right)^{2p+2} c_0^{2p+2} n^{(2p+2)(1-\alpha)} \mathbb{P}\left(\|Z_n - m\| \geq c n^{1-\alpha} \right).\end{aligned}$$

Applying inequality (5.9) and Markov's inequality,

$$\begin{aligned}\mathbb{E}\left[V_n^{p+1}\mathbb{1}_{\{\|Z_n-m\|\geq cn^{1-\alpha}\}}\right] &\leq (1+c_\gamma C)^{2p+2}\, c_0^{2p+2} n^{(2p+2)(1-\alpha)} \frac{\mathbb{E}\left[\|Z_n-m\|^{2q}\right]}{(cn)^{2q(1-\alpha)}} \\ &\leq \frac{(1+c_\gamma C)^{2p+2}\, c_0^{2p+2} n^{(2p+2)(1-\alpha)}}{c^{2q(1-\alpha)}} \frac{M_q}{n^{2q(1-\alpha)}} \\ &= O\left(\frac{1}{n^{2q(1-\alpha)-(2p+2)(1-\alpha)}}\right).\end{aligned}$$

Taking $q \geq p+1+\frac{(p+2)\alpha}{2(1-\alpha)}$,

$$\mathbb{E}\left[V_n^{p+1}\mathbb{1}_{\{\|Z_n-m\|\geq cn^{1-\alpha}\}}\right] = O\left(\frac{1}{n^{(p+2)\alpha}}\right). \tag{5.31}$$

Finally, using inequalities (5.29) to (5.31), there is a positive constant C_1' such that for all $n \geq n_\alpha$,

$$\mathbb{E}\left[\|Z_{n+1}-m\|^{2p+2}\right] \leq \left(1-\frac{2}{n}\right)^{p+1} \mathbb{E}\left[\|Z_n-m\|^{2p+2}\right] + \frac{C_1'}{n^{(p+2)\alpha}} + C_2'\gamma_n^2 \mathbb{E}\left[\|Z_n-m\|^{2p}\right]. \tag{5.32}$$

□

Proof of Lemma 5.4.1. Since the eigenvalues of Γ_m belong to $[\lambda_{\min}, C]$, there are a rank n_α and a positive constant c' such that for all $n \geq n_\alpha$, we have $\|I_H - \gamma_n\Gamma_m\|_{op} \leq 1-\lambda_{\min}\gamma_n$ and $0 \leq (1-\lambda_{\min}\gamma_n)^2 + 4C^2\gamma_n^2 \leq 1-c'\gamma_n$. Using decomposition (5.5) and Cauchy-Schwarz's inequality, since $\|\delta_n\| \leq 2C\|Z_n-m\|$ and $\|\xi_{n+1}\|^2 - 2\langle\Phi(Z_n), \xi_{n+1}\rangle \leq 1$, we have for all $n \geq n_\alpha$,

$$\begin{aligned}\|Z_{n+1}-m\|^2 &\leq (1-\lambda_{\min}\gamma_n)^2 \|Z_n-m\|^2 + 2\gamma_n \langle \xi_{n+1}, Z_n - m - \gamma_n\Phi(Z_n)\rangle \\ &\quad - 2\gamma_n \langle (I_H-\gamma_n\Gamma_m)(Z_n-m), \delta_n\rangle + \gamma_n^2\|\delta_n\|^2 + \gamma_n^2\|\xi_{n+1}\|^2 \\ &\leq (1-c'\gamma_n)\|Z_n-m\|^2 + 2\gamma_n\|Z_n-m\|\,\|\delta_n\| + \gamma_n^2 + 2\gamma_n\langle Z_n-m, \xi_{n+1}\rangle.\end{aligned} \tag{5.33}$$

Thus, for all integers $p \geq 1$ and $n \geq n_\alpha$,

$$\begin{aligned}\mathbb{E}\left[\|Z_{n+1}-m\|^{2p}\right] \leq (1-c'\gamma_n)\,\mathbb{E}\left[\|Z_n-m\|^2\,\|Z_{n+1}-m\|^{2p-2}\right] \\ + 2\gamma_n\mathbb{E}\left[\|Z_n-m\|\,\|\delta_n\|\,\|Z_{n+1}-m\|^{2p-2}\right] + \gamma_n^2\mathbb{E}\left[\|Z_{n+1}-m\|^{2p-2}\right] \\ + 2\gamma_n\mathbb{E}\left[\langle Z_n-m,\xi_{n+1}\rangle\,\|Z_{n+1}-m\|^{2p-2}\right]. \qquad (5.34)\end{aligned}$$

In order to bound each term in previous inequality, we give a new upper bound of $\|Z_{n+1}-m\|^{2p-2}$. By convexity of G, we have almost surely $V_n \leq \|Z_n-m\|^2+\gamma_n^2$, and inequality (5.20) can be written as

$$\begin{aligned}\|Z_{n+1}-m\|^{2p-2} \leq \left(\|Z_n-m\|^2+\gamma_n^2\right)^{p-1} + 2(p-1)\gamma_n\langle\xi_{n+1},Z_n-m\rangle\left(\|Z_n-m\|^2+\gamma_n^2\right)^{p-2} \\ + \sum_{k=2}^{p-1}\binom{p-1}{k}|2\gamma_n\langle\xi_{n+1},Z_n-m\rangle|^k\left(\|Z_n-m\|^2+\gamma_n^2\right)^{p-1-k}.\end{aligned}$$

Applying Cauchy-Schwarz's inequality, since $\|\xi_{n+1}\| \leq 2$,

$$\begin{aligned}\|Z_{n+1}-m\|^{2p-2} \leq \left(\|Z_n-m\|^2+\gamma_n^2\right)^{p-1} + 2(p-1)\gamma_n\langle\xi_{n+1},Z_n-m\rangle\left(\|Z_n-m\|^2+\gamma_n^2\right)^{p-2} \qquad (5.35) \\ + \sum_{k=2}^{p-1}\binom{p-1}{k}2^{2k}\gamma_n^k\|Z_n-m\|^k\left(\|Z_n-m\|^2+\gamma_n^2\right)^{p-1-k}.\end{aligned}$$

Note that if $p \leq 2$, the last term on the right-hand side of previous inequality is equal to 0. Applying previous inequality, we can now bound each term in inequality (5.34).

Step 1 : Bounding $(1-c'\gamma_n)\,\mathbb{E}\left[\|Z_n-m\|^2\,\|Z_{n+1}-m\|^{2p-2}\right]$**.**

We will bound each term which appears when we multiply $(1-c'\gamma_n)\,\|Z_n-m\|^2$ by the bound given by inequality (5.35). First, applying inequalities (5.19),

$$\begin{aligned}&\mathbb{E}\left[(1-c'\gamma_n)\,\|Z_n-m\|^2\left(\|Z_n-m\|^2+\gamma_n^2\right)^{p-1}\right] \\ &= (1-c'\gamma_n)\,\mathbb{E}\left[\|Z_n-m\|^{2p}\right] + \sum_{k=0}^{p-2}\binom{p-1}{k}(1-c'\gamma_n)\,\gamma_n^{2(p-1-k)}\mathbb{E}\left[\|Z_n-m\|^{2k+2}\right] \\ &\leq (1-c'\gamma_n)\,\mathbb{E}\left[\|Z_n-m\|^{2p}\right] + \sum_{k=0}^{p-2}\binom{p-1}{k}(1-c'\gamma_n)\,c_\gamma^{2(p-1-k)}\frac{K_{k+1}}{n^{(2p-1-k)\alpha}}.\end{aligned}$$

Since for all $k \leq p-2$, we have $2p-1-k \geq p+1$, there is a positive constant B_1 such that for all $n \geq n_\alpha$,

$$\mathbb{E}\left[\left(1-c'\gamma_n\right)\|Z_n-m\|^2\left(\|Z_n-m\|^2+\gamma_n^2\right)^{p-1}\right] \leq \left(1-c'\gamma_n\right)\mathbb{E}\left[\|Z_n-m\|^{2p}\right]+\frac{B_1}{n^{(p+1)\alpha}}. \tag{5.36}$$

Moreover, using the facts that (ξ_n) is a martingale differences sequence adapted to the filtration $(\mathcal{F}_n)$, and that Z_n is $\mathcal{F}_n$-measurable,

$$\mathbb{E}\left[\left(1-c'\gamma_n\right)\|Z_n-m\|^2 2\left(p-1\right)\gamma_n\left\langle\xi_{n+1},Z_n-m\right\rangle\left(\|Z_n-m\|^2+\gamma_n^2\right)^{p-2}\right]=0. \tag{5.37}$$

We can now suppose that $p \geq 3$, since otherwise the last term in inequality (5.35) is equal to 0. Let

$$\begin{aligned}(\star) &:= \left(1-c'\gamma_n\right)\mathbb{E}\left[\|Z_n-m\|^2\sum_{k=2}^{p-1}\binom{p-1}{k}2^{2k}\gamma_n^k\|Z_n-m\|^k\left(\|Z_n-m\|^2+\gamma_n^2\right)^{p-1-k}\right]\\ &\leq \left(1-c'\gamma_n\right)\sum_{k=2}^{p-1}\binom{p-1}{k}2^{p-2+k}\gamma_n^k\left(\mathbb{E}\left[\|Z_n-m\|^{2p-k}\right]+\gamma_n^{2(p-1-k)}\mathbb{E}\left[\|Z_n-m\|^{k+2}\right]\right).\end{aligned}$$

Applying Cauchy-Schwarz's inequality,

$$\begin{aligned}(\star) \leq \left(1-c'\gamma_n\right)&\sum_{k=2}^{p-1}\binom{p-1}{k}2^{p-2+k}\gamma_n^k\\ &\left(\sqrt{\mathbb{E}\left[\|Z_n-m\|^{2(p-1)}\right]\mathbb{E}\left[\|Z_n-m\|^{2(p+1-k)}\right]}+\gamma_n^{2(p-1-k)}\sqrt{\mathbb{E}\left[\|Z_n-m\|^{2k}\right]\mathbb{E}\left[\|Z_n-m\|^4\right]}\right)\end{aligned}$$

Finally, applying inequality (5.19),

$$\begin{aligned}(\star) &\leq \left(1-c'\gamma_n\right)\sum_{k=2}^{p-1}\binom{p-1}{k}2^{p-2+k}\gamma_n^k\left(\frac{\sqrt{K_{p-1}K_{p+1-k}}}{n^{(p-k/2)\alpha}}+\gamma_n^{2(p-1-k)}\frac{\sqrt{K_kK_2}}{n^{\frac{(k+2)\alpha}{2}}}\right)\\ &= O\left(\frac{1}{n^{(p+1)\alpha}}\right),\end{aligned} \tag{5.38}$$

because for all $2 \leq k \leq p-1$ and $p \geq 3$, we have $p+k/2 \geq p+1$ and $2p-\frac{1}{2}k-1 \geq p+1$. Thus, there is a positive constant B_1' such that for all $n \geq n_\alpha$,

$$\mathbb{E}\left[\left(1-c'\gamma_n\right)\|Z_n-m\|^2\|Z_{n+1}-m\|^{2p-2}\right] \leq \left(1-c'\gamma_n\right)\mathbb{E}\left[\|Z_n-m\|^{2p}\right]+\frac{B_1'}{n^{(p+1)\alpha}}. \tag{5.39}$$

Step 2 : Bounding $2\gamma_n \mathbb{E}\left[\langle \xi_{n+1}, Z_n - m\rangle \left\|Z_{n+1} - m\right\|^{2p-2}\right]$**.**

Applying the fact that (ξ_n) is a martingale differences sequence adapted to the filtration $(\mathcal{F}_n)$ and applying inequality (5.35), let

$$\begin{aligned}(\star\star) &:= \mathbb{E}\left[2\gamma_n \langle \xi_{n+1}, Z_n - m\rangle \left\|Z_{n+1} - m\right\|^{2p-2}\right] \\ &\leq 4(p-1)\gamma_n^2 \mathbb{E}\left[\langle \xi_{n+1}, Z_n - m\rangle^2 \left(\left\|Z_n - m\right\|^2 + \gamma_n^2\right)^{p-2}\right]\end{aligned}$$

Since $\|\xi_{n+1}\| \leq 2$ and applying Cauchy-Schwarz's inequality,

$$\begin{aligned}(\star\star) &\leq 4(p-1)\gamma_n^2 \mathbb{E}\left[\left(\|\xi_{n+1}\| \left\|Z_n - m\right\|\right)^2 \left(\left\|Z_n - m\right\|^2 + \gamma_n^2\right)^{p-2}\right] \\ &\leq 16(p-1)\gamma_n^2 \mathbb{E}\left[\left\|Z_n - m\right\|^2 \left(\left\|Z_n - m\right\|^2 + \gamma_n^2\right)^{p-2}\right].\end{aligned}$$

With the help of Lemma 5.6.1,

$$(\star\star) \leq 2^{p+2}(p-1)\gamma_n^2 \left(\mathbb{E}\left[\left\|Z_n - m\right\|^{2(p-1)}\right] + \gamma_n^{2(p-2)} \mathbb{E}\left[\left\|Z_n - m\right\|^2\right]\right),$$

Applying previous inequality and inequality (5.19), there is a positive constant B_2' such that

$$\mathbb{E}\left[2\gamma_n \langle \xi_{n+1}, Z_n - m\rangle \left\|Z_{n+1} - m\right\|^{2p-2}\right] \leq \frac{B_2'}{n^{(p+1)\alpha}}. \tag{5.40}$$

Step 3 : Bounding $\gamma_n^2 \mathbb{E}\left[\left\|Z_{n+1} - m\right\|^{2p-2}\right]$**.**

Applying inequality (5.19),

$$\begin{aligned}\gamma_n^2 \mathbb{E}\left[\left\|Z_{n+1} - m\right\|^{2p-2}\right] &\leq \gamma_n^2 \frac{K_{p-1}}{(n+1)^{p-1}} \\ &= O\left(\frac{1}{n^{(p+1)\alpha}}\right).\end{aligned} \tag{5.41}$$

Step 4 : Bounding $2\gamma_n \mathbb{E}\left[\left\|Z_n - m\right\| \|\delta_n\| \left\|Z_{n+1} - m\right\|^{2p-2}\right]$**.**

As in step 1, we will bound each term which appears when we multiply $2\gamma_n \left\|Z_n - m\right\| \|\delta_n\|$ by the bound given by inequality (5.35). Since almost surely $\|\delta_n\| \leq 2C \left\|Z_n - m\right\|$, applying

inequality (5.38), one can check

$$\begin{aligned}
&2\gamma_n \mathbb{E}\left[\|Z_n - m\| \|\delta_n\| \sum_{k=2}^{p-1} \binom{p-1}{k} 2^{2k} \gamma_n^k \|Z_n - m\|^k \left(\|Z_n - m\|^2 + \gamma_n^2\right)^{p-1-k}\right] \\
&\leq 4C\gamma_n \mathbb{E}\left[\|Z_n - m\|^2 \sum_{k=2}^{p-1} \binom{p-1}{k} 2^{2k} \gamma_n^k \|Z_n - m\|^k \left(\|Z_n - m\|^2 + \gamma_n^2\right)^{p-1-k}\right] \\
&= o\left(\frac{1}{n^{(p+1)\alpha}}\right).
\end{aligned}$$

Moreover, since (ξ_n) is a martingale differences sequence adapted to the filtration $(\mathcal{F}_n)$,

$$\mathbb{E}\left[2\gamma_n \|Z_n - m\| \|\delta_n\| 2(p-1)\gamma_n \langle \xi_{n+1}, Z_n - m\rangle \left(\|Z_n - m\|^2 + \gamma_n^2\right)^{p-2}\right] = 0.$$

Finally, since almost surely $\|\delta_n\| \leq C_m \|Z_n - m\|^2$ and $\|\delta_n\| \leq 2C\|Z_n - m\|$, applying Lemma 5.6.1,

$$\begin{aligned}
(\star\star\star) &:= \mathbb{E}\left[2\gamma_n \|Z_n - m\| \|\delta_n\| \left(\|Z_n - m\|^2 + \gamma_n^2\right)^{p-1}\right] \\
&\leq 2^{p-1}\gamma_n \mathbb{E}\left[\|Z_n - m\|^{2p-1} \|\delta_n\|\right] + 2^{p-1}\gamma_n^{2p-1}\mathbb{E}\left[\|Z_n - m\| \|\delta_n\|\right] \\
&\leq 2^{p-1}C_m\gamma_n \mathbb{E}\left[\|Z_n - m\|^{2p+1}\right] + 2^{p-1}C\gamma_n^{2p-1}\mathbb{E}\left[\|Z_n - m\|^2\right].
\end{aligned}$$

Applying Lemma 5.6.1,

$$(\star\star\star) \leq \frac{1}{2}c'\gamma_n \mathbb{E}\left[\|Z_n - m\|^{2p}\right] + 2^{2p-2}\frac{C_m^2}{c'}\gamma_n \mathbb{E}\left[\|Z_n - m\|^{2p+2}\right] + O\left(\frac{1}{n^{(p+1)\alpha}}\right).$$

Thus, there are positive constants B_3', B_4' such that

$$2\mathbb{E}\left[\|Z_n - m\| \|\delta_n\| \|Z_{n+1} - m\|^{2p-2}\right] \leq \frac{1}{2}c'\gamma_n \mathbb{E}\left[\|Z_n - m\|^{2p}\right] + B_3'\gamma_n \mathbb{E}\left[\|Z_n - m\|^{2p+2}\right] + \frac{B_4'}{n^{(p+1)\alpha}}. \tag{5.42}$$

Step 5 : Conclusion. Taking $c_0 = \frac{1}{2}c'$, applying inequalities (5.39),(5.40),(5.41) and (5.42), there are positive constants C_1, C_2 such that for all $n \geq n_\alpha$,

$$\mathbb{E}\left[\|Z_{n+1} - m\|^{2p}\right] \leq (1 - c_0\gamma_n)\mathbb{E}\left[\|Z_n - m\|^{2p}\right] + \frac{C_1}{n^{(p+1)\alpha}} + C_2\gamma_n \mathbb{E}\left[\|Z_n - m\|^{2p+2}\right].$$

□

Proof of Theorem 5.4.1. We prove with the help of a complete induction that for all $p \geq 1$, and

for all $\beta \in (\alpha, \frac{p+2}{p}\alpha - \frac{1}{p})$, there are positive constants $K_p, C_{\beta,p}$ such that for all $n \geq 1$,

$$\mathbb{E}\left[\|Z_n - m\|^{2p}\right] \leq \frac{K_p}{n^{p\alpha}}, \qquad \mathbb{E}\left[\|Z_n - m\|^{2p+2}\right] \leq \frac{C_{\beta,p}}{n^{\beta p}}.$$

This result is proven in [CCGB15] for $p = 1$. Let $p \geq 2$ and let us suppose from now that for all integer $k \leq p-1$, there are positive constant K_k such that for all $n \geq 1$,

$$\mathbb{E}\left[\|Z_n - m\|^{2k}\right] \leq \frac{K_k}{n^{k\alpha}}. \tag{5.43}$$

We now split the end of the proof into two steps.

Step 1 : Calibration of the constants.

In order to simplify the demonstration thereafter, we introduce some constants and notations. Let β be a constant such that $\frac{p+2}{p}\alpha - \frac{1}{p} > \beta > \alpha$ and let $K'_p, K'_{p,\beta}$ be constants such that $K'_p \geq 2^{1+p\alpha} C_1 c_0^{-1} c_\gamma^{-1}$, ($C_1$ is defined in Lemma 5.4.1), and $2K'_p \geq K'_{p,\beta} \geq K'_p \geq 1$. By definition of β, there is a rank $n_{p,\beta} \geq n_\alpha$ (n_α is defined in Lemma 5.4.1 and in Lemma 5.4.2) such that for all $n \geq n_{p,\beta}$,

$$\begin{aligned} (1 - c_0\gamma_n)\left(\frac{n+1}{n}\right)^{p\alpha} + \frac{1}{2}c_0\gamma_n + \frac{2^{\alpha+\beta p+1} c_\gamma C_2}{(n+1)^{\alpha+(\beta-\alpha)p}} &\leq 1, \\ \left(1 - \frac{2}{n}\right)^{p+1}\left(\frac{n+1}{n}\right)^{p\beta} + \left(C'_1 + C'_2 c_\gamma^2\right) 2^{(p+2)\alpha} \frac{1}{(n+1)^{(p+2)\alpha - p\beta}} &\leq 1, \end{aligned} \tag{5.44}$$

with C_2 defined in Lemma 5.4.1 and C'_1, C'_2 are defined in Lemma 5.4.2. Because $\beta > \alpha$,

$$\begin{aligned} (1 - c_0\gamma_n)\left(\frac{n+1}{n}\right)^{p\alpha} + \frac{1}{2}c_0\gamma_n + \frac{2^{\alpha+\beta p+1} c_\gamma C_2}{(n+1)^{\alpha+(\beta-\alpha)p}} &= 1 - c_0\gamma_n + o\left(\frac{1}{n}\right) + \frac{1}{2}c_0\gamma_n + O\left(\frac{1}{n^{\alpha+(\beta-\alpha)p}}\right) \\ &= 1 - \frac{1}{2}c_0\gamma_n + o\left(\frac{1}{n^\alpha}\right). \end{aligned}$$

In the same way, since $\beta < \frac{p+2}{p}\alpha - \frac{1}{p}$, $p\beta < 2p+2$ and

$$\left(1 - \frac{2}{n}\right)^{p+1}\left(\frac{n+1}{n}\right)^{p\beta} + \left(C'_1 + C'_2 c_\gamma^2\right) 2^{(p+2)\alpha} \frac{1}{(n+1)^{(p+2)\alpha - p\beta}} = 1 - (2p + 2 - p\beta)\frac{1}{n} + o\left(\frac{1}{n}\right).$$

Step 2 : The induction.

Let us take $K'_p \geq n_{p,\beta}^{p\alpha}\mathbb{E}\left[\left\|Z_{n_{p,\beta}} - m\right\|^{2p}\right]$ and $K'_{p,\beta} \geq n_{p,\beta}^{p\alpha}\mathbb{E}\left[\left\|Z_{n_{p,\beta}} - m\right\|^{2p+2}\right]$, we will

prove by induction that for all $n \geq n_{p,\beta}$,

$$\mathbb{E}\left[\|Z_n - m\|^{2p}\right] \leq \frac{K'_p}{n^{p\alpha}}, \qquad \mathbb{E}\left[\|Z_n - m\|^{2p+2}\right] \leq \frac{K'_{p,\beta}}{n^{p\beta}}.$$

Applying Lemma 5.4.1 and by induction, since $2K'_p \geq K'_{p,\beta} \geq K'_p \geq 1$,

$$\begin{aligned}
\mathbb{E}\left[\|Z_{n+1} - m\|^{2p}\right] &\leq (1 - c_0\gamma_n)\,\mathbb{E}\left[\|Z_n - m\|^{2p}\right] + \frac{C_1}{n^{(p+1)\alpha}} + C_2\gamma_n\mathbb{E}\left[\|Z_n - m\|^{2p+2}\right] \\
&\leq (1 - c_0\gamma_n)\,\frac{K'_p}{n^{p\alpha}} + \frac{C_1}{n^{(p+1)\alpha}} + C_2\gamma_n\frac{K'_{p,\beta}}{n^{p\beta}} \\
&\leq (1 - c_0\gamma_n)\,\frac{K'_p}{n^{p\alpha}} + \frac{C_1}{n^{(p+1)\alpha}} + 2C_2\gamma_n\frac{K'_p}{n^{p\beta}}.
\end{aligned}$$

Factorizing by $\frac{K'_p}{(n+1)^{p\alpha}}$,

$$\begin{aligned}
\mathbb{E}\left[\|Z_{n+1} - m\|^{2p}\right] &\leq (1 - c_0\gamma_n)\left(\frac{n+1}{n}\right)^{p\alpha}\frac{K'_p}{(n+1)^{p\alpha}} + \left(\frac{n+1}{n}\right)^{p\alpha} C_1\frac{1}{(n+1)^{p\alpha}n^{\alpha}} \\
&\quad + 2c_\gamma C_2\left(\frac{n+1}{n}\right)^{\alpha+\beta p}\frac{K'_p}{(n+1)^{\beta p+\alpha}} \\
&\leq (1 - c_0\gamma_n)\left(\frac{n+1}{n}\right)^{p\alpha}\frac{K'_p}{(n+1)^{p\alpha}} + \frac{2^{p\alpha}C_1c_\gamma^{-1}\gamma_n}{(n+1)^{p\alpha}} + \frac{2^{\alpha+\beta p+1}c_\gamma C_2}{(n+1)^{\alpha+(\beta-\alpha)p}}\frac{K'_p}{(n+1)^{p\alpha}}.
\end{aligned}$$

Since $K'_p \geq 2^{1+p\alpha}C_1c_\gamma^{-1}c_0^{-1}$,

$$\begin{aligned}
\mathbb{E}\left[\|Z_{n+1} - m\|^{2p}\right] &\leq (1 - c_0\gamma_n)\left(\frac{n+1}{n}\right)^{p\alpha}\frac{K'_p}{(n+1)^{p\alpha}} + \frac{1}{2}\gamma_n c_0\frac{K'_p}{(n+1)^{p\alpha}} \\
&\quad + \frac{2^{\alpha+\beta p+1}c_\gamma C_2}{(n+1)^{\alpha+(\beta-\alpha)p}}\frac{K'_p}{(n+1)^{p\alpha}} \\
&\leq \left((1 - c_0\gamma_n)\left(\frac{n+1}{n}\right)^{p\alpha} + \frac{1}{2}c_0\gamma_n + \frac{2^{\alpha+\beta p+1}c_\gamma C_2}{(n+1)^{\alpha+(\beta-\alpha)p}}\right)\frac{K'_p}{(n+1)^{p\alpha}}.
\end{aligned}$$

By definition of $n_{p,\beta}$ (see (5.44)),

$$\mathbb{E}\left[\|Z_{n+1} - m\|^{2p}\right] \leq \frac{K'_p}{(n+1)^{p\alpha}}. \tag{5.45}$$

In the same way, applying Lemma 5.4.2 and by induction, since $K'_{p,\beta} \geq K'_p \geq 1$, for all

$n \geq n_{p,\beta}$,

$$\begin{aligned}\mathbb{E}\left[\|Z_{n+1}-m\|^{2p+2}\right] &\leq \left(1-\frac{2}{n}\right)^{p+1}\mathbb{E}\left[\|Z_n-m\|^{2p+2}\right]+\frac{C_1'}{n^{(p+2)\alpha}}+C_2'\gamma_n^2\mathbb{E}\left[\|Z_n-m\|^{2p}\right]\\ &\leq \left(1-\frac{2}{n}\right)^{p+1}\frac{K_{p,\beta}'}{n^{p\beta}}+\frac{C_1'}{n^{(p+2)\alpha}}+C_2'\gamma_n^2\frac{K_p'}{n^{p\alpha}}\\ &\leq \left(1-\frac{2}{n}\right)^{p+1}\frac{K_{p,\beta}'}{n^{p\beta}}+\frac{C_1'K_{p,\beta}'}{n^{(p+2)\alpha}}+C_2'\gamma_n^2\frac{K_{p,\beta}'}{n^{p\alpha}}.\end{aligned}$$

Factorizing by $\frac{K_{p,\beta}'}{(n+1)^{p\beta}}$,

$$\begin{aligned}\mathbb{E}\left[\|Z_{n+1}-m\|^{2p+2}\right] &\leq \left(1-\frac{2}{n}\right)^{p+1}\left(\frac{n+1}{n}\right)^{p\beta}\frac{K_{p,\beta}'}{(n+1)^{p\beta}}\\ &\quad+C_1'\left(\frac{n+1}{n}\right)^{(p+2)\alpha}\frac{1}{(n+1)^{(p+2)\alpha-p\beta}}\frac{K_{p,\beta}'}{(n+1)^{p\beta}}\\ &\quad+C_2'c_\gamma^2\left(\frac{n+1}{n}\right)^{(p+2)\alpha}\frac{1}{(n+1)^{(p+2)\alpha-p\beta}}\frac{K_{p,\beta}'}{(n+1)^{p\beta}}\\ &\leq \left(\left(1-\frac{2}{n}\right)^{p+1}\left(\frac{n+1}{n}\right)^{p\beta}+2^{(p+2)\alpha}\frac{C_1'+C_2'c_\gamma^2}{(n+1)^{(p+2)\alpha-p\beta}}\right)\frac{K_{p,\beta}'}{(n+1)^{p\beta}}.\end{aligned}$$

By definition of $n_{p,\beta}$,

$$\mathbb{E}\left[\|Z_{n+1}-m\|^{2p+2}\right] \leq \frac{K_{p,\beta}'}{(n+1)^{p\beta}}, \tag{5.46}$$

which concludes the induction. In order to conclude the proof, we just have to take $K_p \geq K_p', K_{p,\beta} \geq K_{p,\beta}'$, and

$$K_p \geq \max_{k<n_{p,\beta}} n^{k\alpha}\mathbb{E}\left[\|Z_k-m\|^{2p}\right], \qquad K_{p,\beta} \geq \max_{k<n_{p,\beta}} n^{k\beta}\mathbb{E}\left[\|Z_k-m\|^{2p+2}\right].$$

□

5.6.2 Proofs of Section 5.4.2

The following lemma give the L^p rates of convergence of the martingale term. Note that this is probably not a new result, but we were not able to find a proof in a published reference.

Lemma 5.6.2. *Let (ξ_n) be a sequence of martingale differences taking values in a Hilbert space H adapted to a filtration $(\mathcal{F}_n)$. Suppose that there is a non-negative constant M such that for all $n \geq 1$, $\|\xi_n\| \leq M$ almost surely. Then, for all integer $p \geq 1$, there is a positive constant C_p such that for all*

$n \geq 1$,

$$\mathbb{E}\left[\left\|\sum_{k=2}^{n} \xi_k\right\|^{2p}\right] \leq C_p n^p.$$

Démonstration. The proof consists in an induction on p and is given in a supplementary file. □

Proof of Theorem 5.4.2. We give there a succinct proof, and a more detailed one is given in a supplementary file. Using decomposition (5.14), let $\lambda_{\min} > 0$ be the smallest eigenvalue of Γ_m, we have with Lemma 5.6.1,

$$\begin{aligned}
\mathbb{E}\left[\left\|\overline{Z}_n - m\right\|^{2p}\right] \leq{}& \frac{5^{2p-1}}{\lambda_{\min}^{2p} n^{2p}} \mathbb{E}\left[\left\|\frac{T_1}{\gamma_1}\right\|^{2p}\right] + \frac{5^{2p-1}}{\lambda_{\min}^{2p} n^{2p}} \mathbb{E}\left[\left\|\frac{T_{n+1}}{\gamma_n}\right\|^{2p}\right] \\
&+ \frac{5^{2p-1}}{\lambda_{\min}^{2p} n^{2p}} \mathbb{E}\left[\left\|\sum_{k=2}^{n} T_k \left(\frac{1}{\gamma_k} - \frac{1}{\gamma_{k-1}}\right)\right\|^{2p}\right] + \frac{5^{2p-1}}{\lambda_{\min}^{2p} n^{2p}} \mathbb{E}\left[\left\|\sum_{k=1}^{n} \delta_n\right\|^{2p}\right] \\
&+ \frac{5^{2p-1}}{\lambda_{\min}^{2p} n^{2p}} \mathbb{E}\left[\left\|\sum_{k=1}^{n} \xi_{k+1}\right\|^{2p}\right].
\end{aligned}$$

We now bound each term at the right-hand side of previous inequality. Since Z_1 is almost surely bounded, we have $\frac{1}{n^{2p}} \mathbb{E}\left[\left\|\frac{T_1}{\gamma_1}\right\|^{2p}\right] = O\left(\frac{1}{n^{2p}}\right)$. Moreover, with Theorem 5.4.1,

$$\frac{1}{n^{2p}} \mathbb{E}\left[\frac{\|T_{n+1}\|^{2p}}{\gamma_n^{2p}}\right] = o\left(\frac{1}{n^p}\right), \tag{5.47}$$

since $\alpha < 1$. In the same way, since $\left|\frac{1}{\gamma_{k-1}} - \frac{1}{\gamma_k}\right| \leq 2\alpha c_\alpha^{-1} k^{\alpha-1}$, applying Lemma 5.4.3 and Theorem 5.4.1 ,

$$\frac{1}{n^{2p}} \mathbb{E}\left[\left\|\sum_{k=2}^{n} T_k \left(\frac{1}{\gamma_k} - \frac{1}{\gamma_{k-1}}\right)\right\|^{2p}\right] = O\left(\frac{1}{n^{(2-\alpha)p}}\right). \tag{5.48}$$

Finally, since $\|\delta_n\| \leq C_m \|Z_n - m\|^2$, applying Lemma 5.4.3 and Theorem 5.4.1,

$$\frac{1}{n^{2p}} \mathbb{E}\left[\left\|\sum_{k=1}^{n} \delta_n\right\|^{2p}\right] = O\left(\frac{1}{n^{2\alpha p}}\right). \tag{5.49}$$

Since $\alpha > 1/2$, we have $\frac{1}{n^{2p}} \mathbb{E}\left[\left\|\sum_{k=1}^{n} \delta_n\right\|^{2p}\right] = o\left(\frac{1}{n^p}\right)$. Finally, applying Lemma 5.6.2, there is

a positive constant C_p such that for all $n \geq 1$,

$$\frac{1}{n^{2p}}\mathbb{E}\left[\left\|\sum_{k=1}^{n}\xi_{k+1}\right\|^{2p}\right] = O\left(\frac{1}{n^p}\right). \tag{5.50}$$

We deduce from inequalities (5.47) to (5.50), that for all integer $p \geq 1$, there is a positive constant A_p such that for all $n \geq 1$,

$$\mathbb{E}\left[\left\|\overline{Z}_n - m\right\|^{2p}\right] \leq \frac{A_p}{n^p}. \tag{5.51}$$

□

Annexe B

Estimating the geometric median in Hilbert spaces with stochastic gradient algorithms : L^p and almost sure rates of convergence. Appendix

Résumé

Dans cette partie, on donne les preuves des lemmes techniques. De plus, la preuve du résultat sur les vitesses de convergence L^p de l'algorithme moyenné est donnée ainsi que celle des résultats sur les vitesses de convergences presque sûre des algorithmes. Plus précisément, on donne les preuves des Lemmes 5.4.3 et 5.6.2, qui permettent respectivement de majorer les moyennes L^p d'une somme de variables aléatoires et d'une somme de différences de martingales. On prouve alors le Théorème 5.4.2 et la Proposition 5.4.1, qui donnent les vitesses L^p de l'algorithme moyenné et assurent que ce sont les vitesses optimales. Finalement, les preuves du Théorème 5.5.1 et du Corollaire 5.5.1 (qui donnent les vitesses presque sûre des algorithmes) sont données.

B.1 Proofs of Section 5.4.2

Proof of Lemma 5.4.3. For all integers $p \geq 1$ and $n \geq 1$, there are positive constants c_b, $b \in \mathbb{N}^n$, such that for all non-negative real numbers y_k, $k = 1, ..., n$,

$$\left(\sum_{k=1}^{n} y_k\right)^p = \sum_{b=(b_1,...,b_n)\in\mathbb{N}^n, b_1+b_2+...+b_n=p} c_b y_1^{b_1} ... y_n^{b_n}. \tag{B.1}$$

As a particular case, applying a classical generalization of Hölder's inequality (see [Sma96], page 179, for example) ,

$$\begin{aligned}
\mathbb{E}\left[\left\|\sum_{k=1}^{n} a_k Y_k\right\|^p\right] &\leq \mathbb{E}\left[\left(\sum_{k=1}^{n} |a_k| \, \|Y_k\|\right)^p\right] \\
&= \sum_{b=(b_1,...,b_n)\in\mathbb{N}^n, b_1+b_2+...+b_n=p} c_b \, |a_1|^{b_1} ... |a_n|^{b_n} \, \mathbb{E}\left[\|Y_1\|^{b_1} ... \|Y_n\|^{b_n}\right] \\
&\leq \sum_{b=(b_1,...,b_n)\in\mathbb{N}^n, b_1+b_2+...+b_n=p} c_b \, |a_1|^{b_1} ... |a_n|^{b_n} \left(\mathbb{E}\left[\|Y_1\|^p\right]\right)^{\frac{b_1}{p}} ... \left(\mathbb{E}\left[\|Y_n\|^p\right]\right)^{\frac{b_n}{p}} \\
&= \sum_{b=(b_1,...,b_n)\in\mathbb{N}^n, b_1+b_2+...+b_n=p} c_b \left(|a_1| \left(\mathbb{E}\left[\|Y_1\|^p\right]\right)^{\frac{1}{p}}\right)^{b_1} ... \left(|a_n| \left(\mathbb{E}\left[\|Y_n\|^p\right]\right)^{\frac{1}{p}}\right)^{b_n} \\
&= \left(\sum_{k=1}^{n} |a_k| \left(\mathbb{E}\left[\|Y_k\|^p\right]\right)^{\frac{1}{p}}\right)^p.
\end{aligned}$$

□

Proof of Lemma 5.6.2. We prove Lemma 5.6.2 with the help of a strong induction on $p \geq 1$. First, if $p = 1$, since (ξ_n) is a sequence of martingale adapted to a filtration $(\mathcal{F}_n)$,

$$\begin{aligned}
\mathbb{E}\left[\left\|\sum_{k=2}^{n} \xi_k\right\|^2\right] &= \sum_{k=2}^{n} \mathbb{E}\left[\|\xi_k\|^2\right] + 2\sum_{k=2}^{n}\sum_{k'=k}^{n} \mathbb{E}\left[\langle \xi_k, \xi_{k'} \rangle\right] \\
&\leq (n-1)M^2 + 2\sum_{k=2}^{n}\sum_{k'=k}^{n} \mathbb{E}\left[\langle \xi_k, \mathbb{E}\left[\xi_{k'} | \mathcal{F}_{k'-1}\right] \rangle\right] \\
&= (n-1)M^2.
\end{aligned}$$

Let $p \geq 2$ and for all $n \geq 2$, $M_n := \sum_{k=2}^{n} \xi_k$. We suppose from now that for all $k \leq p-1$, there is a positive constant C_k such that for all $n \geq 2$,

$$\mathbb{E}\left[\|M_n\|^{2k}\right] \leq C_k (n-1)^k.$$

For all $n \geq 2$,

$$\begin{aligned}\|M_{n+1}\|^2 &= \|M_n\|^2 + 2\langle M_n, \xi_{n+1}\rangle + \|\xi_{n+1}\|^2 \\ &\leq \|M_n\|^2 + 2\langle M_n, \xi_{n+1}\rangle + M^2.\end{aligned}$$

Thus,

$$\begin{aligned}\|M_{n+1}\|^{2p} \leq \left(\|M_n\|^2 + M^2\right)^p + 2\langle M_n, \xi_{n+1}\rangle \left(\|M_n\|^2 + M^2\right)^{p-1} \\ + \sum_{k=2}^{p} \binom{p}{k} |2\langle M_n, \xi_{n+1}\rangle|^k \left(\|M_n\|^2 + M^2\right)^{p-k}. \end{aligned} \tag{B.2}$$

We now bound the expectation of the three terms on the right-hand side of previous inequality. First, by induction,

$$\begin{aligned}\mathbb{E}\left[\left(\|M_n\|^2 + M^2\right)^p\right] &= \mathbb{E}\left[\|M_n\|^{2p}\right] + \sum_{k=1}^{p} \binom{p}{k} M^{2k} \mathbb{E}\left[\|M_n\|^{2p-2k}\right] \\ &\leq \mathbb{E}\left[\|M_n\|^{2p}\right] + \sum_{k=1}^{p} \binom{p}{k} M^{2k} C_{p-k} n^{p-k} \\ &\leq \mathbb{E}\left[\|M_n\|^{2p}\right] + O\left(n^{p-1}\right).\end{aligned}$$

Moreover, since (ξ_n) is a sequence of martingale differences adapted to a filtration $(\mathcal{F}_n)$, and since M_n is $\mathcal{F}_n$-measurable,

$$\mathbb{E}\left[\langle M_n, \xi_{n+1}\rangle \left(\|M_n\|^2 + M^2\right)^{p-1}\right] = \mathbb{E}\left[\langle M_n, \mathbb{E}\left[\xi_{n+1}|\mathcal{F}_n\right]\rangle \left(\|M_n\|^2 + M^2\right)^{p-1}\right] = 0. \tag{B.3}$$

Finally, applying Cauchy-Schwarz's inequality and Lemma 5.6.1, since $\|\xi_n\| \leq M$, let

$$\begin{aligned}(*) &:= \sum_{k=2}^{p} \binom{p}{k} \mathbb{E}\left[|2\langle M_n, \xi_{n+1}\rangle|^k \left(\|M_n\|^2 + M^2\right)^{p-k}\right] \\ &\leq \sum_{k=2}^{p} \binom{p}{k} 2^k M^k \mathbb{E}\left[\|M_n\|^k \left(\|M_n\|^2 + M^2\right)^{p-k}\right] \\ &\leq \sum_{k=2}^{p} \binom{p}{k} 2^{p-1} M^k \left(\mathbb{E}\left[\|M_n\|^{2p-k}\right] + M^{2p-2k}\mathbb{E}\left[\|M_n\|^k\right]\right).\end{aligned}$$

Applying Cauchy-Schwarz's inequality and by induction,

$$\begin{aligned}(*) &\leq \sum_{k=2}^{p} \binom{p}{k} 2^{p-1} M^k \left(\sqrt{\mathbb{E}\left[\|M_n\|^{2p-2}\right] \mathbb{E}\left[\|M_n\|^{2(p+1-k)}\right]} + M^{2p-2k} \sqrt{\mathbb{E}\left[\|M_n\|^{2}\right] \mathbb{E}\left[\|M_n\|^{2k-2}\right]} \right) \\ &\leq \sum_{k=2}^{p} \binom{p}{k} 2^{p-1} M^k \left(\sqrt{C_{p-1} C_{p+1-k}} n^{p-k/2} + M^{2p-2k} \sqrt{C_1 C_{k-1}} n^{k/2} \right) \\ &= O\left(n^{p-1}\right), \end{aligned} \tag{B.4}$$

since $p \geq 2$. Thus, thanks to inequalities (B.2) to (B.4), there is a non-negative constant A_p such that for all $n \geq 1$,

$$\begin{aligned}\mathbb{E}\left[\|M_{n+1}\|^{2p}\right] &\leq \mathbb{E}\left[\|M_n\|^{2p}\right] + A_p n^{p-1} \\ &\leq \|\xi_2\|^{2p} + A_p \sum_{k=2}^{n} k^{p-1} \\ &\leq M^{2p} + A_p n^p, \end{aligned}$$

which concludes the induction and the proof. □

Proof of Theorem 5.4.2. Let us recall the following decomposition

$$n\Gamma_m \left(\overline{Z}_n - m\right) = \frac{T_1}{\gamma_1} - \frac{T_{n+1}}{\gamma_n} + \sum_{k=2}^{n} T_k \left(\frac{1}{\gamma_k} - \frac{1}{\gamma_{k-1}} \right) + \sum_{k=1}^{n} \delta_k + \sum_{k=1}^{n} \xi_{k+1}, \tag{B.5}$$

with $T_n = Z_n - m$. Let $\lambda_{\min} > 0$ be the smallest eigenvalue of Γ_m, we have with Lemma 5.6.1,

$$\begin{aligned}\mathbb{E}\left[\left\|\overline{Z}_n - m\right\|^{2p}\right] \leq{}& \frac{5^{2p-1}}{\lambda_{\min}^{2p} n^{2p}} \mathbb{E}\left[\left\|\frac{T_1}{\gamma_1}\right\|^{2p}\right] + \frac{5^{2p-1}}{\lambda_{\min}^{2p} n^{2p}} \mathbb{E}\left[\left\|\frac{T_{n+1}}{\gamma_n}\right\|^{2p}\right] \\ &+ \frac{5^{2p-1}}{\lambda_{\min}^{2p} n^{2p}} \mathbb{E}\left[\left\|\sum_{k=2}^{n} T_k \left(\frac{1}{\gamma_k} - \frac{1}{\gamma_{k-1}} \right)\right\|^{2p}\right] + \frac{5^{2p-1}}{\lambda_{\min}^{2p} n^{2p}} \mathbb{E}\left[\left\|\sum_{k=1}^{n} \delta_k\right\|^{2p}\right] \\ &+ \frac{5^{2p-1}}{\lambda_{\min}^{2p} n^{2p}} \mathbb{E}\left[\left\|\sum_{k=1}^{n} \xi_{k+1}\right\|^{2p}\right]. \end{aligned}$$

We now bound each term at the right-hand side of previous inequality. Since Z_1 is almost

surely bounded, we have $\frac{1}{n^{2p}}\mathbb{E}\left[\left\|\frac{T_1}{\gamma_1}\right\|^{2p}\right] = O\left(\frac{1}{n^{2p}}\right)$. Moreover, with Theorem 5.4.1,

$$\begin{aligned}\frac{1}{n^{2p}}\mathbb{E}\left[\frac{\|T_{n+1}\|^{2p}}{\gamma_n^{2p}}\right] &\leq \frac{1}{c_\gamma^{2p}}\frac{1}{n^{2p-2p\alpha}}\frac{K_1}{(n+1)^{p\alpha}} \\ &= o\left(\frac{1}{n^p}\right), \end{aligned} \tag{B.6}$$

since $\alpha < 1$. In the same way, since $\left|\frac{1}{\gamma_{k-1}} - \frac{1}{\gamma_k}\right| \leq 2\alpha c_\alpha^{-1} k^{\alpha-1}$, applying Lemma 5.4.3 and Theorem 5.4.1,

$$\begin{aligned}\frac{1}{n^{2p}}\mathbb{E}\left[\left\|\sum_{k=2}^n T_k\left(\frac{1}{\gamma_k} - \frac{1}{\gamma_{k-1}}\right)\right\|^{2p}\right] &\leq \frac{1}{n^{2p}}\left(\sum_{k=2}^n \left(\frac{1}{\gamma_k} - \frac{1}{\gamma_{k-1}}\right)\left(\mathbb{E}\left[\|T_k\|^{2p}\right]\right)^{\frac{1}{2p}}\right)^{2p} \\ &\leq \frac{1}{n^{2p}}\left(\sum_{k=2}^n 2\alpha c_\gamma^{-1} k^{\alpha-1}\left(\frac{K_p}{k^{\alpha p}}\right)^{\frac{1}{2p}}\right)^{2p} \\ &\leq \frac{2^{2p}\alpha^{2p} c_\gamma^{-2p} K_p}{n^{2p}}\left(\sum_{k=2}^n \frac{1}{k^{1-\alpha/2}}\right)^{2p} \\ &= O\left(\frac{1}{n^{(2-\alpha)p}}\right). \end{aligned} \tag{B.7}$$

Finally, since $\|\delta_n\| \leq C_m \|Z_n - m\|^2$, applying Lemma 5.4.3 and Theorem 5.4.1,

$$\begin{aligned}\frac{1}{n^{2p}}\mathbb{E}\left[\left\|\sum_{k=1}^n \delta_k\right\|^{2p}\right] &\leq \frac{1}{n^{2p}}\left(\sum_{k=1}^n \left(\mathbb{E}\left[\|\delta_k\|^{2p}\right]\right)^{\frac{1}{2p}}\right)^{2p} \\ &\leq \frac{C_m^{2p}}{n^{2p}}\left(\sum_{k=1}^n \left(\mathbb{E}\left[\|Z_k - m\|^{4p}\right]\right)^{\frac{1}{2p}}\right)^{2p} \\ &\leq \frac{C_m^{2p} K_{2p}}{n^{2p}}\left(\sum_{k=1}^n \frac{1}{k^\alpha}\right)^{2p} \\ &= O\left(\frac{1}{n^{2\alpha p}}\right). \end{aligned} \tag{B.8}$$

Since $\alpha > 1/2$, we have $\frac{1}{n^{2p}}\mathbb{E}\left[\|\sum_{k=1}^n \delta_k\|^{2p}\right] = o\left(\frac{1}{n^p}\right)$. Finally, applying Lemma 5.6.2, there is

a positive constant C_p such that for all $n \geq 1$,

$$\frac{1}{n^{2p}}\mathbb{E}\left[\left\|\sum_{k=1}^{n}\xi_{k+1}\right\|^{2p}\right] \leq \frac{1}{n^{2p}}C_p(n+1)^p$$
$$= O\left(\frac{1}{n^p}\right). \tag{B.9}$$

We deduce from inequalities (B.6) to (B.9), that for all integer $p \geq 1$, there is a positive constant A_p such that for all $n \geq 1$,

$$\mathbb{E}\left[\left\|\overline{Z}_n - m\right\|^{2p}\right] \leq \frac{A_p}{n^p}. \tag{B.10}$$

□

Proof of Proposition 5.4.1. We now give a lower bound of $\mathbb{E}\left[\left\|\overline{Z}_n - m\right\|^2\right]$. One can check that $\frac{1}{n}\sum_{k=1}^{n}\xi_{k+1}$ is the dominant term in decomposition (B.5). Indeed, decomposition (B.5) can be written as

$$\Gamma_m\left(\overline{Z}_n - m\right) = \frac{1}{n}\sum_{k=1}^{n}\xi_{k+1} + \frac{1}{n}R_n, \tag{B.11}$$

with

$$R_n := \frac{T_1}{\gamma_1} - \frac{T_{n+1}}{\gamma_n} + \sum_{k=2}^{n}T_k\left(\frac{1}{\gamma_k} - \frac{1}{\gamma_{k-1}}\right) + \sum_{k=1}^{n}\delta_k.$$

Applying inequalities (B.6), (B.7) and (B.8), one can check that

$$\frac{1}{n^2}\mathbb{E}\left[\left\|R_n\right\|^2\right] = o\left(\frac{1}{n}\right). \tag{B.12}$$

Moreover,

$$\mathbb{E}\left[\left\|\Gamma_m\left(\overline{Z}_n - m\right)\right\|^2\right] = \frac{1}{n^2}\mathbb{E}\left[\left\|\sum_{k=1}^{n}\xi_{k+1}\right\|^2\right] + \frac{1}{n^2}\mathbb{E}\left[\left\|R_n\right\|^2\right] + \frac{2}{n^2}\mathbb{E}\left[\left\langle\sum_{k=1}^{n}\xi_{k+1}, R_n\right\rangle\right] \tag{B.13}$$

Applying Cauchy-Schwarz's inequality and Lemma 5.4.3, there is a positive constant C_1 such

that for all $n \geq 1$,

$$\begin{aligned}
\frac{2}{n^2}\mathbb{E}\left[\left|\left\langle \sum_{k=1}^{n} \xi_{k+1}, R_n \right\rangle\right|\right] &\leq 2\mathbb{E}\left[\frac{1}{n^2}\left\|\sum_{k=1}^{n} \xi_{k+1}\right\| \|R_n\|\right] \\
&\leq 2\sqrt{\frac{1}{n^2}\mathbb{E}\left[\left\|\sum_{k=1}^{n} \xi_{k+1}\right\|^2\right]}\sqrt{\frac{1}{n^2}\mathbb{E}\left[\|R_n\|^2\right]} \\
&\leq \frac{2\sqrt{C_1}}{\sqrt{n}}\sqrt{\frac{1}{n^2}\mathbb{E}\left[\|R_n\|^2\right]} \\
&= o\left(\frac{1}{n}\right).
\end{aligned}$$

Moreover, since $\mathbb{E}\left[\|\xi_{n+1}\|^2\right] = 1 - \mathbb{E}\left[\|\Phi(Z_n)\|^2\right]$ (see [CCZ13] for details), using the fact that (ξ_n) is a sequence of martingale differences adapted to the filtration $(\mathcal{F}_n)$, we get

$$\begin{aligned}
\frac{1}{n^2}\mathbb{E}\left[\left\|\sum_{k=1}^{n} \xi_{k+1}\right\|^2\right] &= \frac{1}{n^2}\sum_{k=1}^{n}\mathbb{E}\left[\|\xi_{k+1}\|^2\right] + 2\sum_{k=1}^{n}\sum_{k'=k+1}^{n}\mathbb{E}\left[\langle \xi_{k+1}, \xi_{k'+1}\rangle\right] \\
&= \frac{1}{n} - \frac{1}{n^2}\sum_{k=1}^{n}\mathbb{E}\left[\|\Phi(Z_k)\|^2\right] + 2\sum_{k=1}^{n}\sum_{k'=k+1}^{n}\mathbb{E}\left[\langle \xi_{k+1}, \mathbb{E}\left[\xi_{k'+1}|\mathcal{F}_{k'}\right]\rangle\right] \\
&= \frac{1}{n} - \frac{1}{n^2}\sum_{k=1}^{n}\mathbb{E}\left[\|\Phi(Z_k)\|^2\right].
\end{aligned}$$

Moreover, since $\|\Phi(Z_n)\| \leq C\|Z_n - m\|$, applying Theorem 5.4.1, we have,

$$\begin{aligned}
\frac{1}{n^2}\sum_{k=1}^{n}\mathbb{E}\left[\|\Phi(Z_k)\|^2\right] &\leq \frac{C^2}{n^2}\sum_{k=1}^{n}\mathbb{E}\left[\|Z_k - m\|^2\right] \\
&\leq \frac{C^2 K_1}{n^2}\sum_{k=1}^{n}\frac{1}{k^\alpha} \\
&= o\left(\frac{1}{n}\right).
\end{aligned}$$

Finally,

$$\mathbb{E}\left[\left\|\Gamma_m\left(\overline{Z}_n - m\right)\right\|^2\right] = \frac{1}{n} + o\left(\frac{1}{n}\right). \tag{B.14}$$

Thus, since the largest eigenvalue of Γ_m satisfies $\lambda_{\max} \leq C$, there is a rank n_α such that for all $n \geq n_\alpha$,

$$\mathbb{E}\left[\left\|\overline{Z}_n - m\right\|^2\right] \geq \frac{1}{2C^2 n}.$$

Let $c'' := \min\left\{\min_{1\leq k\leq n_\alpha}\left\{k\mathbb{E}\left[\left\|\overline{Z}_k - m\right\|^2\right]\right\}, \frac{1}{2C^2}\right\}$, for all $n \geq 1$,

$$\mathbb{E}\left[\left\|\overline{Z}_n - m\right\|^2\right] \geq \frac{c''}{n}. \tag{B.15}$$

□

B.2 Proofs of Section 5.5

Proof of Theorem 5.5.1. Let $\beta' \in (1/2, 1)$ such that $\beta' < \alpha$. In order to apply Borel-Cantelli's Lemma, we will prove that

$$\sum_{n\geq 1}\mathbb{P}\left(\|Z_n - m\| \geq \frac{1}{n^{\beta'/2}}\right) < \infty. \tag{B.16}$$

Applying Theorem 5.4.1, for all $p \geq 1$, for all $n \geq 1$,

$$\begin{aligned}\mathbb{P}\left(\|Z_n - m\| \geq \frac{1}{n^{\beta'/2}}\right) &\leq \mathbb{E}\left[\|Z_n - m\|^{2p}\right] n^{p\beta'} \\ &\leq \frac{K_p}{n^{p(\alpha-\beta')}}.\end{aligned}$$

Since $\beta' < \alpha$, we can take $p > \frac{1}{\alpha-\beta'}$ and we get

$$\sum_{n\geq 1}\mathbb{P}\left(\|Z_n - m\| \geq \frac{1}{n^{\beta'/2}}\right) \leq \sum_{n\geq 1}\frac{K_p}{n^{p(\alpha-\beta')}} < \infty.$$

Applying Borel-Cantelli's Lemma,

$$\|Z_n - m\| = O\left(n^{-\frac{\beta'}{2}}\right) \quad a.s, \tag{B.17}$$

for all $\beta' < \alpha$. In a particular case, for all $\beta < \alpha$,

$$\|Z_n - m\| = o\left(n^{-\frac{\beta}{2}}\right) \quad a.s. \tag{B.18}$$

□

Proof of Corollary 5.5.1. Let us recall decomposition (B.5) of the averaged algorithm :

$$\Gamma_m\left(\overline{Z}_n - m\right) = \frac{1}{n}\left(\frac{T_1}{\gamma_1} - \frac{T_{n+1}}{\gamma_n} + \sum_{k=2}^{n} T_k\left(\frac{1}{\gamma_k} - \frac{1}{\gamma_{k-1}}\right) + \sum_{k=1}^{n}\delta_k + \sum_{k=1}^{n}\xi_{k+1}\right).$$

We will give the almost sure rate of convergence of each term. First, since Z_1 is bounded, we have $\left\|\frac{T_1}{n\gamma_1}\right\| = O\left(\frac{1}{n}\right)$ almost surely. Applying Theorem 5.5.1, let $\beta' < \alpha$,

$$\begin{aligned}\left\|\frac{T_{n+1}}{n\gamma_n}\right\| &= o\left(\frac{n^{-\frac{\beta'}{2}}}{n^{1-\alpha}}\right) \quad a.s \\ &= o\left(\frac{1}{\sqrt{n}}\right) \quad a.s.\end{aligned}$$

Indeed, we obtain the last equality by taking $\alpha > \beta' > 2\alpha - 1$, which is possible since $\alpha < 1$. Moreover, since $\left|\gamma_k^{-1} - \gamma_{k-1}^{-1}\right| \leq 2\alpha c_\gamma^{-1} k^{\alpha-1}$, let $\beta' < \alpha$, applying Theorem 5.5.1,

$$\begin{aligned}\left\|\frac{1}{n}\sum_{k=2}^{n} T_k\left(\frac{1}{\gamma_k} - \frac{1}{\gamma_{k-1}}\right)\right\| &\leq \frac{1}{n}\sum_{k=1}^{n}\|T_k\|\left|\frac{1}{\gamma_{k-1}} - \frac{1}{\gamma_k}\right| \\ &= o\left(\frac{1}{n}\sum_{k=2}^{n} k^{\alpha-\beta'/2-1}\right) \quad a.s \\ &= o\left(\frac{n^{\alpha-\beta'/2}}{n}\right) \quad a.s \\ &= o\left(\frac{1}{\sqrt{n}}\right) \quad a.s.\end{aligned}$$

Indeed, we get the last equality taking $\beta' > 2\alpha - 1$. Moreover, since $\|\delta_n\| \leq C_m \|Z_n - m\|^2$,

for all $\beta' < \alpha$,

$$\begin{aligned}\left\|\frac{1}{n}\sum_{k=1}^{n}\delta_k\right\| &\leq \frac{1}{n}\sum_{k=1}^{n}\|\delta_k\| \\ &\leq \frac{C_m}{n}\sum_{k=1}^{n}\|Z_k - m\|^2 \\ &= o\left(\frac{1}{n}\sum_{k=1}^{n}\frac{1}{k^{\beta'}}\right) \quad a.s \\ &= o\left(\frac{1}{n^{\beta'}}\right) \quad a.s \\ &= o\left(\frac{1}{\sqrt{n}}\right) \quad a.s.\end{aligned}$$

Indeed, we obtain the last equality by taking $\alpha > \beta' > 1/2$. Finally, since $\mathbb{E}\left[\|\sum_{k=1}^{n}\xi_{k+1}\|^2\right] = n + o(n)$ (see [CCGB15] and proof of Theorem 5.4.2), applying the law of large numbers for martingales (see Theorem 1.3.15 in [Duf97]), for all $\delta > 0$,

$$\frac{1}{n}\sum_{k=1}^{n}\xi_{n+1} = o\left(\frac{(\ln n)^{\frac{1+\delta}{2}}}{\sqrt{n}}\right) \quad a.s, \tag{B.19}$$

which concludes the proof. □

Remark B.2.1. *Note that the law of large numbers for martingales in [Duf97] is not given for general Hilbert spaces. Nevertheless, in our context, this law of large numbers can be extended. We just have to prove that for all positive constant δ, $U_n := \frac{1}{\sqrt{n(\ln(n))^{1+\delta}}}\|\sum_{k=1}^{n}\xi_{k+1}\|$ converges almost surely to a finite random variable. Since (ξ_n) is a sequence of martingale differences adapted to the filtration $(\mathcal{F}_n)$, and since $\mathbb{E}\left[\|\xi_{n+1}\|^2\,|\mathcal{F}_n\right] \leq 1$,*

$$\begin{aligned}\mathbb{E}\left[U_{n+1}^2|\mathcal{F}_n\right] &= \frac{n(\ln(n))^{1+\delta}}{(n+1)(\ln(n+1))^{1+\delta}}U_n^2 + \frac{1}{(n+1)(\ln(n+1))^{1+\delta}}\mathbb{E}\left[\|\xi_{n+1}\|^2\,|\mathcal{F}_n\right] \\ &\leq U_n^2 + \frac{1}{(n+1)(\ln(n+1))^{1+\delta}}.\end{aligned}$$

Thus, applying Robbins-Siegmund Theorem (see [Duf97]), (U_n) converges almost surely to a finite random variable, which concludes the proof.

Deuxième partie

Estimation récursive de la Median Covariation Matrix dans les espaces de Hilbert et application à l'Analyse des Composantes Principales en ligne

Chapitre 6

Fast Estimation of the Median Covariation Matrix with Application to Online Robust Principal Components Analysis

Résumé

Nous avons vu au Chapitre 2 que la médiane géométrique est un indicateur de position robuste et on peut lui associer la "Median Covariation Matrix", qui est un indicateur de dispersion multivarié robuste pouvant être généralisé sans aucune difficulté aux données fonctionnelles. A l'aide des estimateurs de la médiane définis par (4.3) et (4.4), on introduit un algorithme de gradient stochastique et sa version moyennée (voir le Chapitre 1) pour estimer la "Median Covariation Matrix". On donne les propriétés de convergence asymptotiques de ces estimateurs récursifs sous des conditions faibles. Plus précisément on donne la forte consistance des algorithmes (Théorème 6.3.2) avant de donner les vitesses de convergences en moyenne quadratiques de l'algorithme de gradient (Théorème 6.3.3) et de son moyenné (Théorème 6.3.4). L'estimation des composantes principales peut elle aussi se faire en ligne, et cette approche peut être très utiles pour la détections de données atypiques en ligne. L'étude de simulations montre clairement que cet indicateur robuste est une alternative compétitive au "minimum covariance determinant" ([RvD99]) quand les données sont à valeurs dans des espaces de petite dimension, et à l'analyse en composantes principales robustes basée sur la

This Chapter is based on a work with Hervé Cardot ([CGB15]).

"projection pursuit" (voir [CRG05]) et sur les projection sphériques ([LMS+99], [Ger08]) dans le cas où les données sont à valeurs dans des espaces de grande dimension. Une illustration sur un gros échantillon à valeurs dans un espace de grande dimension, consistant en les audiences TV individuelles mesurées à chaque minute sur une période de 24 heures, confirme l'intérêt de considérer l'analyse robuste des composantes principales basée sur la Median Covariation Matrix.

Abstract

The geometric median covariation matrix is a robust multivariate indicator of dispersion which can be extended without any difficulty to functional data. We define estimators, based on recursive algorithms, that can be simply updated at each new observation and are able deal rapidly with large samples of high dimensional data without being obliged to store all the data in memory. Asymptotic convergence properties of the recursive algorithms are studied under weak conditions. The computation of the principal components can also be performed online and this approach can also be useful for online outlier detection. A simulation study clearly shows that this robust indicator is a competitive alternative to minimum covariance determinant when the dimension of the data is small and robust principal components analysis based on projection pursuit and spherical projections for high dimension data. An illustration on a large sample and high dimensional dataset consisting of individual TV audiences measured at a minute scale over a period of 24 hours confirms the interest of considering the robust principal components analysis based on the median covariation matrix.

6.1 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] or [RS05]). When both the dimension and the sample size are large, outlying observations may be difficult to detect automatically. 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] and [MMY06]).

The most popular approaches are probably the minimum covariance determinant estimator (see [RvD99]) and the robust projection pursuit (see [CRG05] and [CFO07]). Robust PCA based on projection pursuit has been extended to deal with functional data in [HU07] and [BBT$^+$11]. Adopting another point of view, robust modifications of the covariance matrix, based on projection of the data onto the unit sphere, have been proposed in [LMS$^+$99] (see also [Ger08] and [TKO12]).

We consider in this work another robust way of measuring association between variables, that can be extended directly to functional data. It is based on the notion of median covariation matrix (MCM) which is defined as the minimizer of an expected loss criterion based on the Hilbert-Schmidt norm (see [KP12] for a first definition in a more general M-estimation setting). It can be seen as a geometric median (see [Kem87] or [MNO10]) in the particular Hilbert spaces of square matrices (or operators for functional data) equipped with the Frobenius (or Hilbert-Schmidt) norm. The MCM is non negative and unique under weak conditions. As shown in [KP12] it also has the same eigenspace as the usual covariance matrix when the distribution of the data is symmetric and the second order moment is finite. Being a spatial median in a particular Hilbert space of matrices, the MCM is also a robust indicator of central location, among the covariance matrices, which has a 50 % breakdown point (see [Kem87] or [MMY06]) as well as a bounded gross sensitivity error (see [CCZ13]).

The aim of this work is twofold. It provides efficient recursive estimation algorithms of the MCM that are able to deal with large samples of high dimensional data. By this recursive property, these algorithms can naturally deal with data that are observed sequentially and provide a natural update of the estimators at each new observation. Another advantage compared to classical approaches is that such recursive algorithms will not require to store all the data. Secondly, this work also aims at highlighting the interest of considering the median covariation matrix to perform principal components analysis of high dimensional contaminated data.

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 [Wei37a] and [VZ00]) can be directly used to estimate effectively both the geometric median and the median covariation matrix. When both the dimension and the sample size are large this static algorithm which requires to store all the data may be inappropriate and ineffective. We show how the algorithm developed by [CCZ13] for the geometric median in Hilbert spaces can be adapted to estimate recursively and simultaneously the median as well as the median covariation matrix. Then an averaging step ([PJ92]) of the two initial recursive estimators of the median and the MCM permits to improve the accuracy of the initial stochastic gradient algorithms. We also explain how the eigenelements of the estimator of the MCM can be updated online without being obliged to perform a new spectral decomposition at each new observation.

The paper is organized as follows. The median covariation matrix as well as the recursive estimators are defined in Section 2. In Section 3, almost sure and quadratic mean consistency results are given for variables taking values in general separable Hilbert spaces. The proofs, which are based on new induction steps compared to [CCZ13], allow to get better convergence rates in quadratic mean even if this new framework is much more complicated because two averaged non linear algorithms are running simultaneously. One can also note that the techniques generally employed to deal with two time scale Robbins Monro algorithms (see [MP06] for the multivariate case) require assumptions on the rest of the Taylor expansion and the finite dimension of the data that are too restrictive in our framework. In Section 4, a comparison with some classic robust PCA techniques is made on simulated data. The interest of considering the MCM is also highlighted on the analysis of individual TV audiences, a large sample of high dimensional data which, because of its dimension, can not be analyzed in a reasonable time with classical robust PCA approaches. The main parts of the proofs are described in Section 5. Perspectives for future research are discussed in Section 6. Some technical parts of the proofs as well as a description of Weiszfeld's algorithm in our context are gathered in an Appendix.

6.2 Population point of view and recursive estimators

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

We consider a random variable X that takes values in H and define its center $m \in H$ as follows :

$$m := \arg\min_{u \in H} \mathbb{E}\left[\|X - u\| - \|X\|\right]. \tag{6.1}$$

The solution $m \in H$ is often called the geometric median of X. It is uniquely defined under broad assumptions on the distribution of X (see [Kem87]) which can be expressed as follows.

Assumption 1. *There exist two linearly independent unit vectors* $(u_1, u_2) \in H^2$, *such that*

$$\mathbb{V}(\langle u, X\rangle) > 0, \quad \textit{for } u \in \{u_1, u_2\}.$$

If the distribution of $X - m$ is symmetric around zero and if X admits a first moment that is finite then the geometric median is equal to the expectation of X, $m = \mathbb{E}[X]$. Note however that the general definition (6.1) does not require to assume that the first order moment of $\|X\|$ is finite since $|\mathbb{E}[\|X - u\| - \|X\|]| \leq \|u\|$.

6.2.1 The (geometric) median covariation matrix (MCM)

We now consider the special vector space, denoted by $\mathcal{S}(H)$, of $d \times d$ matrices when $H = \mathbb{R}^d$, or for general separable Hilbert spaces H, the vector space of linear operators mapping $H \to H$. Denoting by $\{e_j, j \in J\}$ an orthonormal basis in H, the vector space $\mathcal{S}(H)$ equipped with the following inner product :

$$\langle A, B\rangle_F = \sum_{j\in J} \langle Ae_j, Be_j\rangle \tag{6.2}$$

is also a separable Hilbert space. In $\mathcal{S}(\mathbb{R}^d)$, we have equivalently

$$\langle A, B\rangle_F = \mathrm{tr}\left(A^T B\right), \tag{6.3}$$

where A^T is the transpose matrix of A. The induced norm is the well known Frobenius norm (also called Hilbert-Schmidt norm) and is denoted by $\|.\|_F$.

When X has finite second order moments, with expectation $\mathbb{E}[X] = \mu$, the covariance matrix of X, $\mathbb{E}\left[(X-\mu)(X-\mu)^T\right]$ can be defined as the minimum argument, over all the elements belonging to $\mathcal{S}(H)$, of the functional $G_{\mu,2} : \mathcal{S}(H) \to \mathbb{R}$,

$$G_{\mu,2}(\Gamma) = \mathbb{E}\left[\left\|(X-\mu)(X-\mu)^T - \Gamma\right\|_F^2 - \left\|(X-\mu)(X-\mu)^T\right\|_F^2\right].$$

Note that in general Hilbert spaces with inner product $\langle ., .\rangle$, operator $(X-\mu)(X-\mu)^T$ should be understood as the operator $u \in H \mapsto \langle u, X-\mu\rangle(X-\mu)$. The MCM is obtained by removing the squares in previous function in order to get a more robust indicator of "covariation".

For $\alpha \in H$, define $G_\alpha : \mathcal{S}(H) \to \mathbb{R}$ by

$$G_\alpha(V) := \mathbb{E}\left[\left\|(X-\alpha)(X-\alpha)^T - V\right\|_F - \left\|(X-\alpha)(X-\alpha)^T\right\|_F\right]. \tag{6.4}$$

The median covariation matrix, denoted by Γ_m, is defined as the minimizer of $G_m(V)$ over all elements $V \in \mathcal{S}(H)$. The second term at the right-hand side of (6.4) prevents from having to introduce hypotheses on the existence of the moments of X. Introducing the random variable $Y := (X-m)(X-m)^T$ that takes values in $\mathcal{S}(H)$, the MCM is unique provided that the support of Y is not concentrated on a line and Assumption 1 can be rephrased as follows in $\mathcal{S}(H)$,

Assumption 2. *There exist two linearly independent unit vectors $(V_1, V_2) \in \mathcal{S}(H)^2$, such that*

$$\mathbb{V}(\langle V, Y\rangle_F) > 0, \quad \textit{for } V \in \{V_1, V_2\}.$$

We can remark that Assumption 1 and Assumption 2 are strongly connected. Indeed, if Assumption 1 holds, then $\mathbb{V}(\langle u, X\rangle) > 0$ for $u \in \{u_1, u_2\}$. Consider the rank one matrices $V_1 = u_1 u_1^T$ and $V_2 = u_2 u_2^T$, we have $\langle V_1, Y\rangle_F = \langle u_1, X-m\rangle^2$ which has a strictly positive variance when the distribution of X has no atom. More generally $\mathbb{V}(\langle V_1, Y\rangle_F) > 0$ unless there is a scalar $a > 0$ such that $\mathbb{P}[\langle u_1, X-m\rangle = a] = \mathbb{P}[\langle u_1, X-m\rangle = -a] = \frac{1}{2}$ (assuming also that $\mathbb{P}[X-m=0] = 0$).

Furthermore it can be deduced easily that the MCM, which is a geometric median in the particular Hilbert spaces of Hilbert-Schmidt operators, is a robust indicator with a 50% breakdown point (see [Kem87]) and a bounded sensitive gross error (see [CCZ13]).

We also assume that

Assumption 3. *There is a constant C such that for all $h \in H$ and all $V \in \mathcal{S}(H)$*

$$(a): \quad \mathbb{E}\left[\left\|(X-h)(X-h)^T - V\right\|_F^{-1}\right] \le C.$$
$$(b): \quad \mathbb{E}\left[\left\|(X-h)(X-h)^T - V\right\|_F^{-2}\right] \le C.$$

This assumption implicitly forces the distribution of $(X-h)(X-h)^T$ to have no atoms. In the case where $H = \mathbb{R}^d$, it is more "likely" to be satisfied when the dimension d of the data is large (see [**?**] and [CCZ13] for a discussion). Note that it could be weakened as in [CCZ13] by allowing points, necessarily different from the MCM Γ_m, to have strictly positive masses.

Considering the particular case $V = 0$, Assumption 3(a) implies that for all $h \in H$,

$$\mathbb{E}\left[\frac{1}{\|X-h\|^2}\right] \leq C, \tag{6.5}$$

and this is not restrictive when the dimension d of H is equal or larger than 3.

Under Assumption 3(a), the functional G_h is twice Fréchet differentiable, with gradient

$$\nabla G_h(V) = -\mathbb{E}\left[\frac{(X-h)(X-h)^T - V}{\|(X-h)(X-h)^T - V\|_F}\right]. \tag{6.6}$$

and Hessian operator, $\nabla_h^2 G(V) : \mathcal{S}(H) \to \mathcal{S}(H)$,

$$\nabla_h^2 G(V) = \mathbb{E}\left[\frac{1}{\|Y(h)-V\|_F}\left(I_{S(H)} - \frac{(Y(h)-V)\otimes_F (Y(h)-V)}{\|Y(h)-V\|_F^{\,2}}\right)\right]. \tag{6.7}$$

where $Y(h) = (X-h)(X-h)^T$, $I_{S(H)}$ is the identity operator on $\mathcal{S}(H)$ and $A \otimes_F B(V) = \langle A, V\rangle_F B$ for any elements A, B and V belonging to $\mathcal{S}(H)$.

Furthermore, Γ_m is also defined as the unique zero of the non linear equation :

$$\nabla G_m(\Gamma_m) = 0. \tag{6.8}$$

Remarking that previous equality can be rewritten as follows,

$$\Gamma_m = \frac{1}{\mathbb{E}\left[\frac{1}{\|(X-m)(X-m)^T - \Gamma_m\|_F}\right]}\mathbb{E}\left[\frac{(X-m)(X-m)^T}{\|(X-m)(X-m)^T - \Gamma_m\|_F}\right], \tag{6.9}$$

it is clear that Γ_m is a bounded, symmetric and non negative operator in $\mathcal{S}(H)$.

As stated in Proposition 2 of [KP12], operator Γ_m has an important stability property when the distribution of X is symmetric, with finite second moment, *i.e* $\mathbb{E}\left[\|X\|^2\right] < \infty$. Indeed, the covariance operator of X, $\Sigma = \mathbb{E}\left[(X-m)(X-m)^T\right]$, which is well defined in this case, and Γ_m share the same eigenvectors : if e_j is an eigenvector of Σ with corresponding eigenvalue λ_j, then $\Gamma_m e_j = \tilde{\lambda}_j e_j$, for some non negative value $\tilde{\lambda}_j$. This important result means that for Gaussian and more generally symmetric distribution (with finite second order moments), the covariance operator and the median covariation operator have the same eigenspaces. Note that it is also conjectured in [KP12] that the order of the eigenfunctions is also the same.

6.2.2 Efficient recursive algorithms

We suppose now that we have i.i.d. copies $X_1, \ldots, X_n, \ldots$ of random variables with the same law as X.

For simplicity, we temporarily suppose that the median m of X is known. We consider a sequence of (learning) weights $\gamma_n = c_\gamma / n^\alpha$, with $c_\gamma > 0$ and $1/2 < \alpha < 1$ and we define the recursive estimation procedure as follows

$$W_{n+1} = W_n + \gamma_n \frac{(X_{n+1} - m)(X_{n+1} - m)^T - W_n}{\|(X_{n+1} - m)(X_{n+1} - m)^T - W_n\|_F} \tag{6.10}$$

$$\overline{W}_{n+1} = \overline{W}_n - \frac{1}{n+1}\left(\overline{W}_n - W_{n+1}\right). \tag{6.11}$$

This algorithm can be seen as a particular case of the averaged stochastic gradient algorithm studied in [CCZ13]. Indeed, the first recursive algorithm (6.10) is a stochastic gradient algorithm,

$$\mathbb{E}\left[\frac{(X_{n+1} - m)(X_{n+1} - m)^T - W_n}{\|(X_{n+1} - m)(X_{n+1} - m)^T - W_n\|_F} | \mathcal{F}_n\right] = \nabla G_m(W_n)$$

where $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$ is the σ-algebra generated by $X_1, \ldots, X_n$ whereas the final estimator $\overline{W}_n$ is obtained by averaging the past values of the first algorithm. The averaging step (see [PJ92]), *i*.e. the computation of the arithmetical mean of the past values of a slowly convergent estimator (see Proposition 6.3.1 below), permits to obtain a new and efficient estimator converging at a parametric rate, with the same asymptotic variance as the empirical risk minimizer (see Theorem 6.3.1 below).

In most of the cases the value of m is unknown so that it also required to estimate the median. To build an estimator of Γ_m, it is possible to estimate simultaneously m and Γ_m by considering two averaged stochastic gradient algorithms that are running simultaneously. For $n \geq 1$,

$$m_{n+1} = m_n + \gamma_n^{(m)} \frac{X_{n+1} - m_n}{\|X_{n+1} - m_n\|}$$

$$\overline{m}_{n+1} = \overline{m}_n - \frac{1}{n+1}\left(\overline{m}_n - m_{n+1}\right) \tag{6.12}$$

$$V_{n+1} = V_n + \gamma_n \frac{(X_{n+1} - \overline{m}_n)(X_{n+1} - \overline{m}_n)^T - V_n}{\|(X_{n+1} - \overline{m}_n)(X_{n+1} - \overline{m}_n)^T - V_n\|_F} \tag{6.13}$$

$$\overline{V}_{n+1} = \overline{V}_n - \frac{1}{n+1}\left(\overline{V}_n - V_{n+1}\right), \tag{6.14}$$

where the averaged recursive estimator $\overline{m}_{n+1}$ of the median m is controlled by a sequence of descent steps $\gamma_n^{(m)}$. The learning rates are generally chosen as follows, $\gamma_n^{(m)} = c_m n^{-\alpha}$, where

the tuning constants satisfy $c_m \in [2, 20]$ and $1/2 < \alpha < 1$.

6.2.3 Online estimation of the principal components

It is also possible to approximate recursively the q eigenvectors (unique up to sign) of Γ_m associated to the q largest eigenvalues without being obliged to perform a spectral decomposition of $\overline{V}_{n+1}$ at each new observation. Many recursive strategies can be employed (see [CD15] for a review on various recursive estimation procedures of the eigenelements of a covariance matrix). Because of its simplicity and its accuracy, we consider the following one :

$$u_{j,n+1} = u_{j,n} + \frac{1}{n+1}\left(\overline{V}_{n+1}\frac{u_{j,n}}{\|u_{j,n}\|} - u_{j,n}\right), \quad j = 1, \ldots, q \tag{6.15}$$

combined with an orthogonalization by deflation of $u_{1,n+1}, \ldots u_{q,n+1}$. This recursive algorithm is based on ideas developed by [WZH03] that are related to the power method for extracting eigenvectors. If we assume that the q first eigenvalues $\lambda_1 > \cdots > \lambda_q$ are distinct, the estimated eigenvectors $u_{1,n+1}, \ldots u_{q,n+1}$, which are uniquely determined up to sign change, tend to $\lambda_1 u_1, \ldots, \lambda_q u_q$.

Once the eigenvectors are computed, it is possible to compute the principal components as well as indices of outlyingness for each new observation (see [HRVA08] for a review of outliers detection with multivariate approaches).

6.2.4 Practical issues, complexity and memory

The recursive algorithms (6.13) and (6.14) require each $O(d^2)$ elementary operations at each update. With the additional online estimation given in (6.15) of the q eigenvectors associated to the q largest eigenvalues, $O(qd^2)$ additional operations are required. The orthogonalization procedure only requires $O(q^2 d)$ elementary operations.

Note that the use of classical Newton-Raphson algorithms for estimating the MCM (see [FFC12]) can not be envisaged for high dimensional data since the computation or the approximation of the Hessian matrix would require $O(d^4)$ elementary operations. The well known and fast Weiszfeld's algorithm requires $O(nd^2)$ elementary operations for each sample with size n. However, the estimation cannot be updated automatically if the data arrive sequentially. Another drawback compared to the recursive algorithms studied in this paper is that all the data must be stored in memory, which is of order $O(nd^2)$ elements whereas the recursive technique require an amount of memory of order $O(d^2)$.

The performances of the recursive algorithms depend on the values of tuning parameters

c_γ, c_m and α. The value of parameter α is often chosen to be $\alpha = 2/3$ or $\alpha = 3/4$. Previous empirical studies (see [CCZ13] and [CCC10]) have shown that, thanks to the averaging step, estimator $\overline{m}_n$ performs well and is not too sensitive to the choice of c_m, provided that the value of c_m is not too small. An intuitive explanation could be that here the recursive process is in some sense "self-normalized" since the deviations at each iteration in (6.10) have unit norm and finding some universal values for c_m is possible. Usual values for c_m and c_γ are in the interval $[2, 20]$. When n is fixed, this averaged recursive algorithm is about 30 times faster than the Weiszfeld's approach (see [CCZ13]).

6.3 Asymptotic properties

When m is known, $\overline{W}_n$ can be seen as an averaged stochastic gradient estimator of the geometric median in a particular Hilbert space and the asymptotic weak convergence of such estimator has been studied in [CCZ13]. They have shown that :

Theorem 6.3.1. *([CCZ13], Theorem 3.4).*
If assumptions 1-3(a) hold, then as n tends to infinity,

$$\sqrt{n}\left(\overline{W}_n - \Gamma_m\right) \rightsquigarrow \mathcal{N}(0, \Delta)$$

where $\rightsquigarrow$ stands for convergence in distribution and $\Delta = \left(\nabla_m^2(\Gamma_m)\right)^{-1} \Psi \left(\nabla_m^2(\Gamma_m)\right)^{-1}$ is the limiting covariance operator, with $\Psi = \mathbb{E}\left[\frac{(Y(m)-\Gamma_m)\otimes_F(Y(m)-\Gamma_m)}{\|Y(m)-\Gamma_m\|_F^2}\right]$.

As explained in [CCZ13], the estimator $\overline{W}_n$ is efficient in the sense that it has the same asymptotic distribution as the empirical risk minimizer related to $G_m(V)$ (see for the derivation of its asymptotic normality in [MNO10] in the multivariate case and [CC14] in a more general functional framework).

Using the delta method for weak convergence in Hilbert spaces (see [DPR82] or [CGER07]), one can deduce, from Theorem 6.3.1, the asymptotic normality of the estimated eigenvectors of $\overline{W}_n$. It can also be proven (see [GB15]), under Assumptions 1-3, that there is a positive constant K such that for all $n \geq 1$,

$$\mathbb{E}\left[\left\|\overline{W}_n - \Gamma_m\right\|_F^2\right] \leq \frac{K}{n}.$$

Note finally that non asymptotic bounds for the deviation of $\overline{W}_n$ around Γ_m can be derived readily with the general results given in [CCGB15].

The more realistic case in which m must also be estimated is more complicated because

$\overline{V}_n$ depends on $\overline{m}_n$ which is also estimated recursively with the same data. We first state the strong consistency of the estimators V_n and $\overline{V}_n$.

Theorem 6.3.2. *If assumptions 1-3(b) hold, we have*

$$\lim_{n\to\infty} \|V_n - \Gamma_m\|_F = 0 \quad a.s.$$

and

$$\lim_{n\to\infty} \left\|\overline{V}_n - \Gamma_m\right\|_F = 0 \quad a.s.$$

The obtention of the rate convergence of the averaged recursive algorithm relies on a fine control of the asymptotic behavior of the Robbins-Monro algorithms, as stated in the following proposition.

Theorem 6.3.3. *If assumptions 1-3(b) hold, there is a positive constant C' and for all $\beta \in (\alpha, 2\alpha)$, there is a positive constant C_β such that for all $n \geq 1$,*

$$\mathbb{E}\left[\|V_n - \Gamma_m\|_F^2\right] \leq \frac{C'}{n^\alpha},$$
$$\mathbb{E}\left[\|V_n - \Gamma_m\|_F^4\right] \leq \frac{C_\beta}{n^\beta}.$$

The obtention of an upper bound for the rate of convergence at the order four of the Robbins-Monro algorithm is crucial in the proofs. Furthermore, the following proposition ensures that the exhibited rate in quadratic mean is the optimal one.

Proposition 6.3.1. *Under assumptions 1-3(b), there is a positive constant c' such that for all $n \geq 1$,*

$$\mathbb{E}\left[\|V_n - \Gamma_m\|_F^2\right] \geq \frac{c'}{n^\alpha}.$$

Finally, the following theorem is the most important theoretical result of this work. It shows that, in spite of the fact that it only considers the observed data one by one, the averaged recursive estimation procedure gives an estimator which has a classical parametric $\sqrt{n}$ rate of convergence in the Hilbert-Schmidt norm.

Theorem 6.3.4. *Under Assumptions 1-3(b), there is a positive constant K' such that for all $n \geq 1$,*

$$\mathbb{E}\left[\left\|\overline{V}_n - \Gamma_m\right\|_F^2\right] \leq \frac{K'}{n}.$$

Assuming the eigenvalues of Γ_m are of multiplicity one, it can be deduced from Theorem 6.3.4 and Lemma 4.3 in [Bos00], the convergence in quadratic mean of the eigenvectors of $\overline{V}_n$ towards the corresponding (up to sign) eigenvector of Γ_m .

6.4 An illustration on simulated and real data

A small comparison with other classical robust PCA techniques is performed in this section considering data in relatively high dimension but samples with moderate sizes. This permits to compare our approach with classical robust PCA techniques, which are generally not designed to deal with large samples of high dimensional data. In our comparison, we have employed the following well known robust techniques : robust projection pursuit (see [CRG05] and [CFO07]), minimum covariance determinant (MCD, see [RvD99]) and spherical PCA (see [LMS$^+$99]). The computations were made in the R language ([R D10]), with the help of packages `pcaPP` and `rrcov`. Our codes are available on request.

If the size of the data $n \times d$ is not too large, an effective way for estimating Γ_m is to employ Weiszfeld's algorithm (see [Wei37a] and [VZ00] as well the Supplementary file for a description of the algorithms in our particular situation). Note that other optimization algorithms which may be preferred in small dimension (see [FFC12]) have not been considered here since they would require the computation of an Hessian matrix whose size is d^4 and this would lead to much slower algorithms. Note finally that all these alternative algorithms do not admit a natural updating scheme when the data arrive sequentially so that they should be completely ran again at each new observation.

6.4.1 Simulation protocol

Independent realizations of a random variable $Y \in \mathbb{R}^d$ are drawn, where

$$Y \quad = \quad (1 - O(\delta))X + O(\delta)\epsilon, \tag{6.16}$$

is a mixture of two distributions and X, O and ϵ are independent random variables. The random vector X has a centered Gaussian distribution in $\mathbb{R}^d$ with covariance matrix $[\Sigma]_{\ell,j} = \min(\ell, j)/d$ and can be thought as a discretized version of a Brownian sample path in $[0, 1]$. The multivariate contamination comes from ϵ, with different rates of contamination controlled by the Bernoulli variable $O(\delta)$, independent from X and ϵ, with $\mathbb{P}(O(\delta) = 1) = \delta$ and $\mathbb{P}(O(\delta) = 0) = 1 - \delta$. Three different scenarios (see Figure 6.1) are considered for the distribution of ϵ :

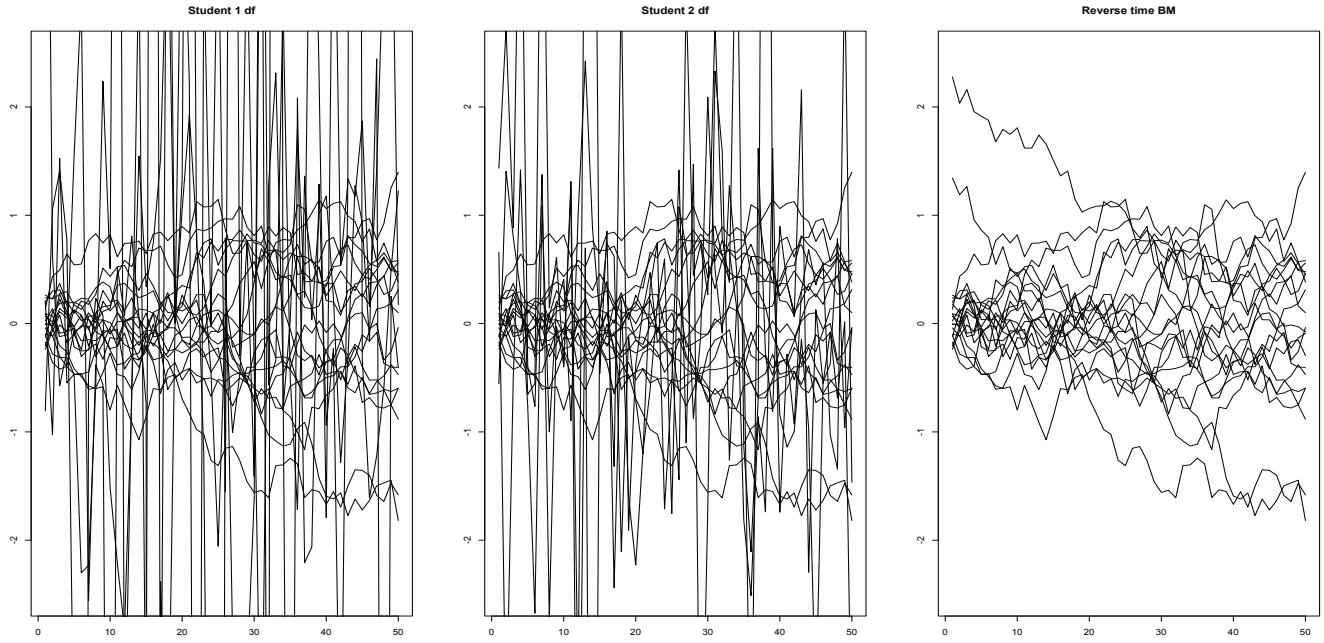

FIGURE 6.1 – A sample of $n = 20$ trajectories when $d = 50$ and $\delta = 0.10$ for the three different contamination scenarios : Student t with 1 degree of freedom, Student t with 2 degrees of freedom and reverse time Brownian motion (from left to right).

— 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 tuning coefficients $c_m = c_\gamma = 2$ and a speed rate of $\alpha = 3/4$. Note that the values of these tuning parameters have not been particularly optimised. We have noted that the simulation results were very stable, and did not depend much on the value of c_m and c_γ for $c_m, c_\gamma \in [1, 20]$.

The estimation error of the eigenspaces associated to the largest eigenvalues is evaluated by considering the squared Frobenius norm between the associated orthogonal projectors.

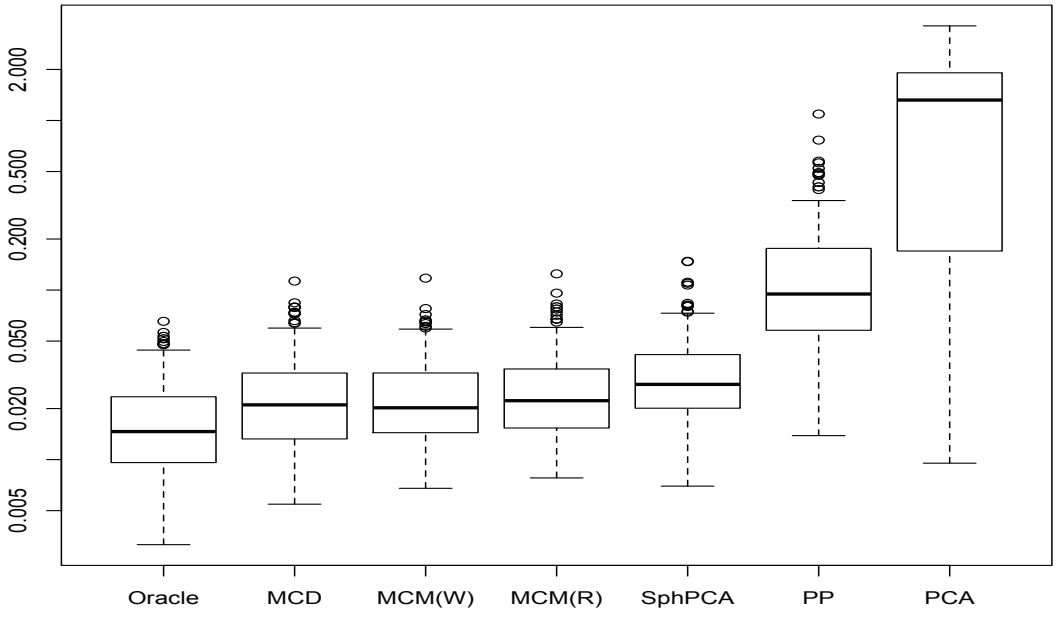

FIGURE 6.2 – Estimation errors (at a logarithmic scale) over 200 Monte Carlo replications, for $n = 200$, $d = 50$ and a contamination by a t distribution with 2 degrees of freedom with $\delta = 0.02$. MCM(W) stands for the estimation performed by the Weiszfeld's algorithm whereas MCM(R) denotes the averaged recursive approach.

Denoting by $\mathbf{P}_q$ the orthogonal projector onto the space generated by the q eigenvectors of the covariance matrix Σ associated to the q largest eigenvalues and by $\widehat{\mathbf{P}}_q$ an estimation, we consider the following loss criterion,

$$\begin{aligned} R(\widehat{\mathbf{P}}_q, \mathbf{P}_q) &= \mathrm{tr}\left[\left(\widehat{\mathbf{P}}_q - \mathbf{P}_q\right)^T \left(\widehat{\mathbf{P}}_q - \mathbf{P}_q\right)\right] \\ &= 2q - 2\mathrm{tr}\left[\widehat{\mathbf{P}}_q \mathbf{P}_q\right]. \end{aligned} \tag{6.17}$$

Note that we always have $R(\widehat{\mathbf{P}}_q, \mathbf{P}_q) \leq 2q$ and $R(\widehat{\mathbf{P}}_q, \mathbf{P}_q) = 2q$ means that the eigenspaces generated by the true and the estimated eigenvectors are orthogonal.

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 6.1 – 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.

6.4.2 Comparison with classical robust PCA techniques

We first compare the performances of the two estimators of the MCM based on the Weiszfeld's algorithm and the recursive algorithms (see (6.14)) with more classical robust PCA techniques.

We generated samples of Y with size $n = 200$ 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 ($q = 2$), according to criterion (6.17), are given in Table 6.1. In Figure 6.2, the distribution of the estimation error $R(\widehat{\mathbf{P}}_q, \mathbf{P}_q)$ is drawn for the different approaches.

We can make the following remarks. At first note that even when the level of contami-

nation 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 designed to deal with very large samples, performs well even for such moderate sample sizes (see also Figure 6.2).

6.4.3 Online estimation of the principal components

We now consider an experiment in high dimension, $d = 1000$, and evaluate the ability of the recursive algorithms defined in (6.15) to estimate recursively the eigenvectors of Γ_m associated to the largest eigenvalues. Note that due to the high dimension of the data and limited computation time, we only make comparison of the recursive robust techniques with the classical PCA. For this we generate growing samples and compute, for each sample size the approximation error of the different (fast) strategies to the true eigenspace generated by the q eigenvectors associated to the q largest eigenvalues of Γ_m.

We have drawn in Figure 6.3, the evolution of the mean (over 100 replications) approximation error $R(\mathbf{P}_q, \hat{\mathbf{P}}_q)$, for a dimension $q = 3$, as a function of the sample size for samples contaminated by a 2 degrees of freedom Student t distribution with a rate $\delta = 0.1$. An important fact is that the recursive algorithm which approximates recursively the eigenelements behaves very well and we can see nearly no difference between the spectral decomposition of $\overline{V}_n$ (denoted by MCM in Figure 6.3) and the estimates produced with the sequential algorithm (6.15) for sample sizes larger than a few hundreds. We can also note that the error made by the classical PCA is always very high and does not decrease with the sample size.

6.4.4 Robust PCA of TV audience

The last example is a high dimension and large sample case. Individual TV audiences are measured, by the French company Médiamétrie, every minutes for a panel of $n = 5422$ people over a period of 24 hours, $d = 1440$ (see [CCM12] for a more detailed presentation of the data). With a classical PCA, the first eigenspace represents 24.4% of the total variability, whereas the second one reproduces 13.5% of the total variance, the third one 9.64% and the fourth one 6.79%. Thus, more than 54% of the variability of the data can be captured in a four dimensional space. Taking account of the large dimension of the data, these values indicate

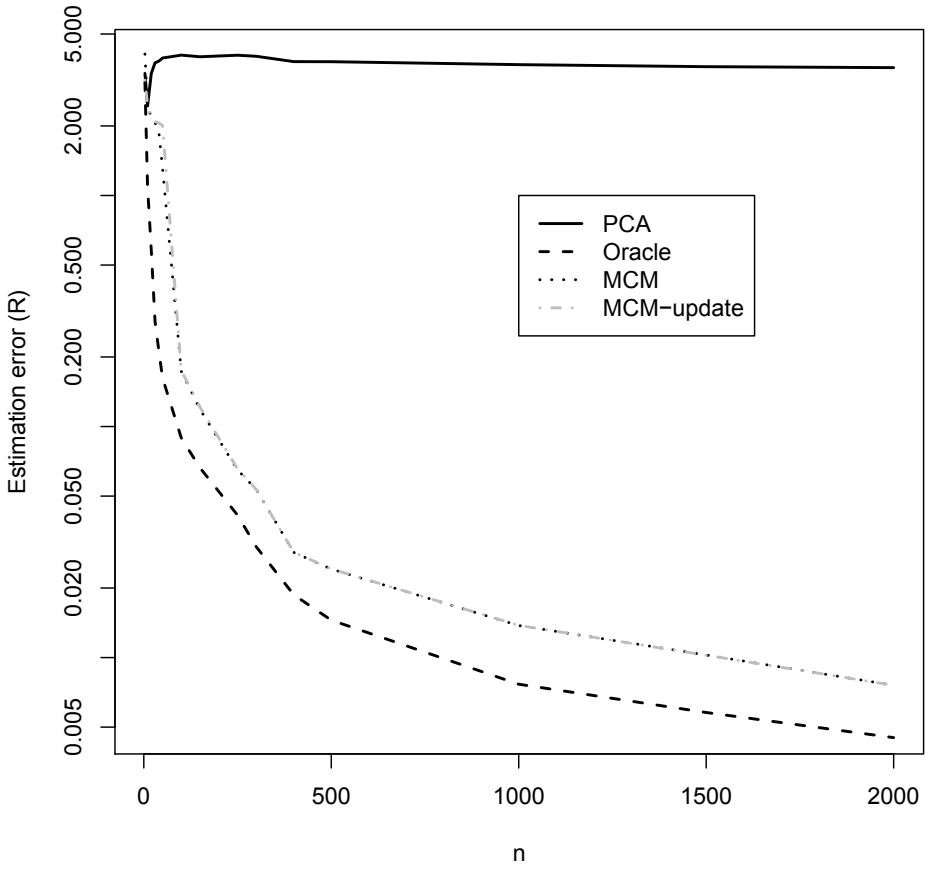

FIGURE 6.3 – Estimation errors of the eigenspaces (criterion $R(\widehat{\mathbf{P}}_q)$) with $d = 1000$ and $q = 3$ for classical PCA, the oracle PCA and the recursive MCM estimator with recursive estimation of the eigenelements (MCM-update) and with static estimation (based on the spectral decomposition of $\overline{V}_n$) of the eigenelements (MCM).

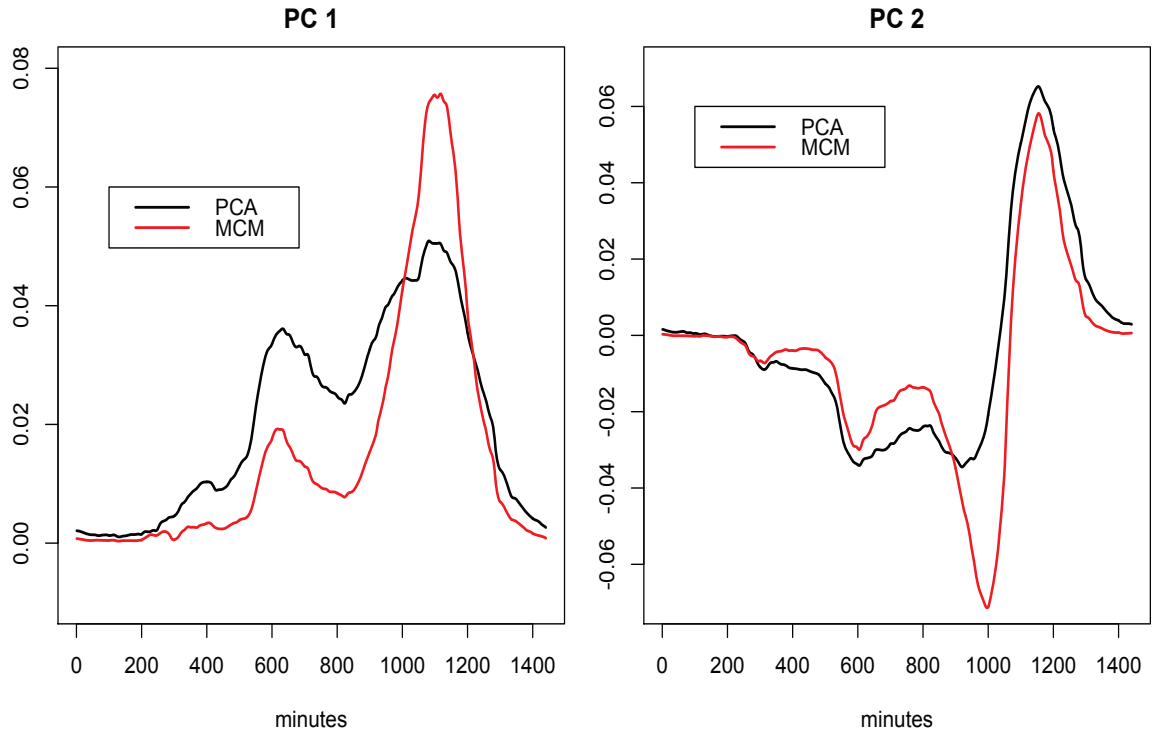

FIGURE 6.4 – TV audience data measured the 6th September 2010, at the minute scale. Comparison of the principal components of the classical PCA (black) and robust PCA based on the Median Covariation Matrix (red). First eigenvectors on the left, second eigenvectors on the right.

a high temporal correlation.

Because of the large dimension of the data, the Weiszfeld's algorithm as well as the other robust PCA techniques can not be used anymore in reasonable time with a personal computer. The MCM has been computed thanks to the recursive algorithm given in (6.14) in approximately 3 minutes on a laptop in the R language (without any specific C routine).

As seen in Figure 6.4, the first two eigenvectors obtained by a classical PCA and the robust PCA based on the MCM are rather different. This is confirmed by the relatively large distance between the two corresponding eigenspaces, $R(\widehat{P}_2^{PCA}, \widehat{P}_2^{MCM}) = 0.56$. The first robust eigenvector puts the stress on the time period comprised between 1000 minutes and 1200 minutes whereas the first non robust eigenvector focuses, with a smaller intensity, on a larger period of time comprised between 600 and 1200 minutes. The second robust eigenvector differentiates between people watching TV during the period between 890 and 1050 minutes (negative value of the second principal component) and people watching TV between

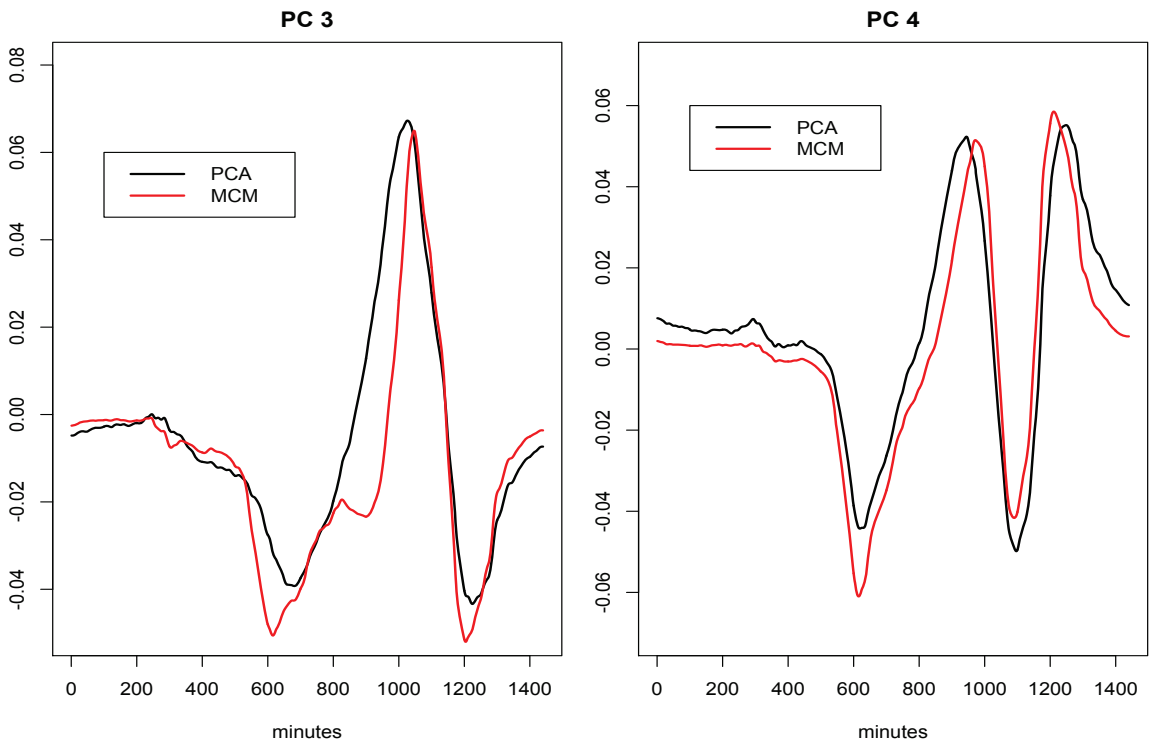

FIGURE 6.5 – TV audience data measured the 6th September 2010, at the minute scale. Comparison of the principal components of the classical PCA (black) and robust PCA based on the MCM (red). Third eigenvectors on the left, fourth eigenvectors on the right.

minutes 1090 and 1220 (positive value of the second principal component). Rather surprisingly, the third and fourth eigenvectors of the non robust and robust covariance matrices look quite similar (see Figure 6.5).

6.5 Proofs

We give in this Section the proofs of Theorems 6.3.2, 6.3.3 and 6.3.4. These proofs rely on several technical Lemmas whose proofs are given in the Supplementary file.

6.5.1 Proof of Theorem 6.3.2

Let us recall the Robbins-Monro algorithm, defined recursively by

$$\begin{aligned} V_{n+1} &= V_n + \gamma_n \frac{\left(X_{n+1} - \overline{m}_n\right)\left(X_{n+1} - \overline{m}_n\right)^T - V_n}{\left\|\left(X_{n+1} - \overline{m}_n\right)\left(X_{n+1} - \overline{m}_n\right)^T - V_n\right\|_F} \\ &= V_n - \gamma_n U_{n+1}, \end{aligned}$$

with $U_{n+1} := -\frac{\left(X_{n+1}-\overline{m}_n\right)\left(X_{n+1}-\overline{m}_n\right)^T - V_n}{\left\|\left(X_{n+1}-\overline{m}_n\right)\left(X_{n+1}-\overline{m}_n\right)^T - V_n\right\|_F}$. Since $\mathcal{F}_n := \sigma\left(X_1, ..., X_n\right)$, we have $\mathbb{E}\left[U_{n+1}|\mathcal{F}_n\right] = \nabla G_{\overline{m}_n}(V_n)$. Thus $\xi_{n+1} := \nabla_{\overline{m}_n} G(V_n) - U_{n+1}$, (ξ_n) is a sequence of martingale differences adapted to the filtration $(\mathcal{F}_n)$. Indeed, $\mathbb{E}\left[\xi_{n+1}|\mathcal{F}_n\right] = \nabla G_{\overline{m}_n}(V_n) - \mathbb{E}\left[U_{n+1}|\mathcal{F}_n\right] = 0$. The algorithm can be written as follows

$$V_{n+1} = V_n - \gamma_n \nabla G_{\overline{m}_n}(V_n) + \gamma_n \xi_{n+1}.$$

Moreover, it can be considered as a stochastic gradient algorithm because it can be decomposed as follows :

$$V_{n+1} = V_n - \gamma_n \left(\nabla G_{\overline{m}_n}(V_n) - \nabla G_{\overline{m}_n}(\Gamma_m)\right) + \gamma_n \xi_{n+1} - \gamma_n r_n, \tag{6.18}$$

with $r_n := \nabla G_{\overline{m}_n}(\Gamma_m) - \nabla G_m(\Gamma_m)$. Finally, linearizing the gradient,

$$V_{n+1} - \Gamma_m = \left(I_{\mathcal{S}(H)} - \gamma_n \nabla_m^2 G(\Gamma_m)\right)\left(V_n - \Gamma_m\right) + \gamma_n \xi_{n+1} - \gamma_n r_n - \gamma_n r'_n - \gamma_n \delta_n, \tag{6.19}$$

with

$$\begin{aligned} r'_n &:= \left(\nabla^2_{\overline{m}_n} G\left(\Gamma_m\right) - \nabla^2_m G\left(\Gamma_m\right)\right)\left(V_n - \Gamma_m\right), \\ \delta_n &:= \nabla G_{\overline{m}_n}\left(V_n\right) - \nabla G_{\overline{m}_n}\left(\Gamma_m\right) - \nabla^2_{\overline{m}_n} G\left(\Gamma_m\right)\left(V_n - \Gamma_m\right). \end{aligned}$$

The following lemma gives upper bounds of these remainder terms. Its proof is given in the Supplementary file.

Lemma 6.5.1. *Under assumptions 1-3(b), we can bound the three remainder terms. First,*

$$\left\|\delta_n\right\|_F \leq 6C \left\|V_n - \Gamma_m\right\|_F^2. \tag{6.20}$$

In the same way, for all $n \geq 1$,

$$\|r_n\|_F \leq 4\left(\sqrt{C} + C\sqrt{\|\Gamma_m\|_F}\right)\|\overline{m}_n - m\|. \tag{6.21}$$

Finally, for all $n \geq 1$,

$$\left\|r_n'\right\|_F \leq 12\left(C\sqrt{\|\Gamma_m\|_F} + C^{3/4}\right)\|\overline{m}_n - m\|\,\|V_n - \Gamma_m\|_F. \tag{6.22}$$

We deduce from decomposition (6.18) that for all $n \geq 1$,

$$\begin{aligned}\|V_{n+1} - \Gamma_m\|_F^2 = \|V_n - \Gamma_m\|_F^2 &- 2\gamma_n\left\langle V_n - \Gamma_m, \nabla G_{\overline{m}_n}(V_n) - \nabla G_{\overline{m}_n}(\Gamma_m)\right\rangle_F \\ &+ \gamma_n^2\|\nabla G_{\overline{m}_n}(V_n) - \nabla G_{\overline{m}_n}(\Gamma_m)\|_F^2 \\ &+ \gamma_n^2\|\xi_{n+1}\|_F^2 + 2\gamma_n\left\langle V_n - \Gamma_m - \gamma_n\left(\nabla G_{\overline{m}_n}(V_n) - \nabla G_{\overline{m}_n}(\Gamma_m)\right), \xi_{n+1}\right\rangle_F \\ &+ \gamma_n^2\|r_n\|_F^2 - 2\gamma_n\left\langle r_n, V_n - \Gamma_m\right\rangle_F - 2\gamma_n^2\left\langle r_n, \xi_{n+1} - \nabla G_{\overline{m}_n}(V_n) + \nabla G_{\overline{m}_n}(\Gamma_m)\right\rangle_F.\end{aligned}$$

Note that for all $h \in H$ and $V \in \mathcal{S}(H)$ we have $\|\nabla G_h(V)\|_F \leq 1$. Furthermore, $\|r_n\|_F \leq 2$ and $\|\xi_{n+1}\|_F \leq 2$. Using the fact that (ξ_n) is a sequence of martingale differences adapted to the filtration $(\mathcal{F}_n)$,

$$\begin{aligned}\mathbb{E}\left[\|V_{n+1} - \Gamma_m\|_F^2\,|\mathcal{F}_n\right] \leq \|V_n - \Gamma_m\|_F^2 &- 2\gamma_n\left\langle V_n - \Gamma_m, \nabla_{\overline{m}_n} G\left(V_n\right) - \nabla_{\overline{m}_n} G\left(\Gamma_m\right)\right\rangle_F \\ &+ 28\gamma_n^2 - 2\gamma_n\left\langle r_n, V_n - \Gamma_m\right\rangle_F.\end{aligned}$$

Let $\alpha_n = n^{-\beta}$, with $\beta \in (1-\alpha, \alpha)$, we have

$$\begin{aligned}\mathbb{E}\left[\|V_{n+1} - \Gamma_m\|_F^2\,|\mathcal{F}_n\right] \leq (1 + \gamma_n\alpha_n)\|V_n - \Gamma_m\|_F^2 &- 2\gamma_n\left\langle V_n - \Gamma_m, \nabla_{\overline{m}_n} G\left(V_n\right) - \nabla_{\overline{m}_n} G\left(\Gamma_m\right)\right\rangle_F \\ &+ 28\gamma_n^2 + \frac{\gamma_n}{\alpha_n}\|r_n\|_F^2.\end{aligned} \tag{6.23}$$

Moreover, applying Lemma 6.5.1 and Theorem 5.1 in [GB15], we get for all positive constant δ,

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

Thus, since $2\gamma_n\left\langle V_n - \Gamma_m, \nabla_{\overline{m}_n} G\left(V_n\right) - \nabla_{\overline{m}_n} G\left(\Gamma_m\right)\right\rangle_F \geq 0$, the Robbins-Siegmund Theorem (see [Duf97] for instance) ensures that $\|V_n - \Gamma_m\|_F$ converges almost surely to a finite random

variable and

$$\sum_{n\geq 1} \gamma_n \left\langle V_n - \Gamma_m, \nabla_{\overline{m}_n} G\left(V_n\right) - \nabla_{\overline{m}_n} G\left(\Gamma_m\right)\right\rangle_F < +\infty \quad a.s.$$

Furthermore, by induction, inequality (6.23) becomes

$$\mathbb{E}\left[\left\|V_{n+1} - \Gamma_m\right\|_F^2\right] \leq \left(\prod_{k=1}^{\infty}\left(1+\gamma_k\alpha_k\right)\right)\mathbb{E}\left[\left\|V_1 - \Gamma_m\right\|_F^2\right] + 28\left(\prod_{k=1}^{\infty}\left(1+\gamma_k\alpha_k\right)\right)\sum_{k=1}^{\infty}\gamma_k^2 + \left(\prod_{k=1}^{\infty}\left(1+\gamma_k\alpha_k\right)\right)\sum_{k=1}^{\infty}\frac{\gamma_k}{\alpha_k}\mathbb{E}\left[\left\|r_k\right\|_F^2\right].$$

Since $\beta < \alpha$, applying Theorem 4.2 in [GB15] and Lemma 6.1, there is a positive constant C_0 such that

$$\sum_{k=1}^{\infty}\frac{\gamma_k}{\alpha_k}\mathbb{E}\left[\left\|r_k\right\|_F^2\right] = C_0\sum_{k=1}^{\infty}k^{-\alpha-1-\beta} < +\infty.$$

Thus, there is a positive constant M such that for all $n \geq 1$, $\mathbb{E}\left[\left\|V_n - \Gamma_m\right\|_F^2\right] \leq M$. Since $\overline{m}_n$ converges almost surely to m, one can conclude the proof of the almost sure consistency of V_n with the same arguments as in the proof of Theorem 3.1 in [CCZ13] and the convexity properties given in the Section B of the supplementary file.

Finally, the almost sure consistency of $\overline{V}_n$ is obtained by a direct application of Topelitz's lemma (see *e.g.* Lemma 2.2.13 in [Duf97]).

6.5.2 Proof of Theorem 6.3.3

The proof of Theorem 6.3.3 relies on properties of the p-th moments of V_n for all $p \geq 1$ given in the following three Lemmas. These properties enable us, with the application of Markov's inequality, to control the probability of the deviations of the Robbins Monro algorithm from Γ_m.

Lemma 6.5.2. *Under assumptions 1-3(b), for all integer p, there is a positive constant M_p such that for all $n \geq 1$,*

$$\mathbb{E}\left[\left\|V_n - \Gamma_m\right\|_F^{2p}\right] \leq M_p.$$

Lemma 6.5.3. *Under assumptions 1-3(b), there are positive constants C_1, C_1', C_2, C_3 such that for all $n \geq 1$,*

$$\mathbb{E}\left[\left\|V_n - \Gamma_m\right\|^2\right] \leq C_1 e^{-C_1' n^{1-\alpha}} + \frac{C_2}{n^{\alpha}} + C_3 \sup_{E(n/2)+1\leq k\leq n-1}\mathbb{E}\left[\left\|V_k - \Gamma_m\right\|^4\right],$$

where $E(x)$ is the integer part of the real number x.

Lemma 6.5.4. *Under assumptions 1-3(b), for all integer $p' \geq 1$, there are a rank $n_{p'}$ and positive constants $C_{1,p'}, C_{2,p'}, C_{3,p'}, c_{p'}$ such that for all $n \geq n_{p'}$,*

$$\mathbb{E}\left[\|V_{n+1} - \Gamma_m\|_F^4\right] \leq \left(1 - c_{p'}\gamma_n n^{-\frac{1-\alpha}{p'}}\right)\mathbb{E}\left[\|V_n - \Gamma_m\|_F^4\right] + \frac{C_{1,p'}}{n^{3\alpha}} + \frac{C_{2,p'}}{n^{2\alpha}}\mathbb{E}\left[\|V_n - \Gamma_m\|_F^2\right] + \frac{C_{3,p'}}{n^{3\alpha - 3\frac{1-\alpha}{p'}}}.$$

We can now prove Theorem 6.3.3.

Let us choose an integer p' such that $p' > 3/2$. Thus, $2 + \alpha - 3\frac{1-\alpha}{p'} \geq 3\alpha$, and applying Lemma 6.5.4, there are positive constants $C_{1,p'}, C_{2,p'}, c_{p'}$ and a rank $n_{p'}$ such that for all $n \geq n_{p'}$,

$$\mathbb{E}\left[\|V_{n+1} - \Gamma_m\|_F^4\right] \leq \left(1 - c_{p'}\gamma_n n^{-\frac{1-\alpha}{p'}}\right)\mathbb{E}\left[\|V_n - \Gamma_m\|_F^4\right] + \frac{C_{1,p'}}{n^{3\alpha}} + \frac{C_{2,p'}}{n^{2\alpha}}\mathbb{E}\left[\|V_n - \Gamma_m\|_F^2\right]. \tag{6.24}$$

Let us now choose $\beta \in (\alpha, 2\alpha)$ and p' such that $p' > \frac{1-\alpha}{2\alpha-\beta}$. Note that $3\alpha - \beta > \alpha + \frac{1-\alpha}{p'}$. One can check that there is a rank $n'_{p'} \geq n_{p'}$ such that for all $n \geq n'_{p'}$,

$$(n+1)^\alpha C_1 e^{-C_1' n^{1-\alpha}} + \frac{1}{2} + C_3 2^{\beta+1}\frac{1}{(n+1)^{\beta-\alpha}} \leq 1,$$
$$\left(1 - c_{p'}\gamma_n n^{-\frac{1-\alpha}{p'}}\right)\left(\frac{n+1}{n}\right)^\beta + 2^{3\alpha}\frac{C_{1,p'} + C_{2,p'}}{(n+1)^{3\alpha-\beta}} \leq 1.$$

With the help of a strong induction, we are going to prove the announced results, that is to say that there are positive constants $C_{p'}, C_\beta$ such that $2C_{p'} \geq C_\beta \geq C_{p'} \geq 1$ and $C_{p'} \geq 2^{\alpha+1}C_2$ (with C_2 defined in Lemma 6.5.3), such that for all $n \geq 1$,

$$\mathbb{E}\left[\|V_n - \Gamma_m\|_F^2\right] \leq \frac{C_{p'}}{n^\alpha},$$
$$\mathbb{E}\left[\|V_n - \Gamma_m\|_F^4\right] \leq \frac{C_\beta}{n^\beta}.$$

First, let us choose $C_{p'}$ and C_β such that

$$C_{p'} \geq \max_{k \leq n'_{p'}}\left\{k^\alpha \mathbb{E}\left[\|V_k - \Gamma_m\|_F^2\right]\right\},$$
$$C_\beta \geq \max_{k \leq n'_{p'}}\left\{k^\beta \mathbb{E}\left[\left\|V_{n'_{p'}} - \Gamma_m\right\|_F^4\right]\right\}.$$

Thus, for all $k \leq n'_{p'}$,

$$\mathbb{E}\left[\|V_k - \Gamma_m\|_F^2\right] \leq \frac{C_{p'}}{k^\alpha},$$
$$\mathbb{E}\left[\|V_k - \Gamma_m\|_F^4\right] \leq \frac{C_\beta}{k^\beta}.$$

We suppose from now that $n \geq n'_{p'}$ and that previous inequalities are verified for all $k \leq n-1$. Applying Lemma 6.5.2 and by induction,

$$\begin{aligned}
\mathbb{E}\left[\|V_{n+1} - \Gamma_m\|_F^2\right] &\leq C_1 e^{-C'_1 n^{1-\alpha}} + \frac{C_2}{n^\alpha} + C_3 \sup_{E((n+1)/2)+1 \leq k \leq n} \left\{\mathbb{E}\left[\|V_k - \Gamma_m\|_F^4\right]\right\} \\
&\leq C_1 e^{-C'_1 n^{1-\alpha}} + \frac{C_2}{n^\alpha} + C_3 \sup_{E((n+1)/2)+1 \leq k \leq n} \left\{\frac{C_\beta}{k^\beta}\right\} \\
&\leq C_1 e^{-C'_1 n^{1-\alpha}} + \frac{C_2}{n^\alpha} + C_3 2^\beta \frac{C_\beta}{n^\beta}.
\end{aligned}$$

Since $2C_{p'} \geq C_\beta \geq C_{p'} \geq 1$ and since $C_{p'} \geq 2^{\alpha+1} C_2$, factorizing by $\frac{C_{p'}}{(n+1)^\alpha}$,

$$\begin{aligned}
\mathbb{E}\left[\|V_{n+1} - \Gamma_m\|_F^2\right] &\leq C_{p'} C_1 e^{-C'_1 n^{1-\alpha}} + C_{p'} 2^{-\alpha-1} \frac{1}{n^\alpha} + C_3 2^\beta \frac{2C_{p'}}{n^\beta} \\
&\leq \frac{C'_p}{(n+1)^\alpha} (n+1)^\alpha C_1 e^{-C'_1 n^{1-\alpha}} + 2^{-\alpha} \left(\frac{n}{n+1}\right)^\alpha \frac{C_{p'}}{2(n+1)^\alpha} + \frac{C_3 2^{\beta+1}}{(n+1)^{\beta-\alpha}} \frac{C_{p'}}{(n+1)^\alpha} \\
&\leq \frac{C'_p}{(n+1)^\alpha} C_1 (n+1)^\alpha e^{-C'_1 n^{1-\alpha}} + \frac{1}{2} \frac{C_{p'}}{(n+1)^\alpha} + C_3 2^{\beta+1} \frac{1}{(n+1)^{\beta-\alpha}} \frac{C_{p'}}{(n+1)^\alpha} \\
&\leq \left((n+1)^\alpha C_1 e^{-C'_1 n^{1-\alpha}} + \frac{1}{2} + C_3 2^{\beta+1} \frac{1}{(n+1)^{\beta-\alpha}}\right) \frac{C_{p'}}{(n+1)^\alpha}.
\end{aligned}$$

By definition of $n'_{p'}$,

$$\mathbb{E}\left[\|V_{n+1} - \Gamma_m\|_F^2\right] \leq \frac{C_{p'}}{(n+1)^\alpha}. \tag{6.25}$$

In the same way, applying Lemma 6.5.4 and by induction,

$$\begin{aligned}
\mathbb{E}\left[\|V_{n+1} - \Gamma_m\|_F^4\right] &\leq \left(1 - c_{p'} \gamma_n n^{-\frac{1-\alpha}{p'}}\right) \mathbb{E}\left[\|V_n - \Gamma_m\|_F^4\right] + \frac{C_{1,p'}}{n^{3\alpha}} + \frac{C_{2,p'}}{n^{2\alpha}} \mathbb{E}\left[\|V_n - \Gamma_m\|_F^2\right] \\
&\leq \left(1 - c_{p'} \gamma_n n^{-\frac{1-\alpha}{p'}}\right) \frac{C_\beta}{n^\beta} + \frac{C_{1,p'}}{n^{3\alpha}} + \frac{C_{2,p'}}{n^{2\alpha}} \frac{C_{p'}}{n^\alpha}.
\end{aligned}$$

Since $C_\beta \geq C_{p'} \geq 1$, factorizing by $\frac{C_\beta}{(n+1)^\beta}$,

$$\begin{aligned}
\mathbb{E}\left[\|V_{n+1} - \Gamma_m\|_F^4\right] &\leq \left(1 - c_{p'}\gamma_n n^{-\frac{1-\alpha}{p'}}\right)\frac{C_\beta}{n^\beta} + \left(C_{1,p'} + C_{2,p'}\right)\frac{C_\beta}{n^{3\alpha}} \\
&\leq \left(1 - c_{p'}\gamma_n n^{-\frac{1-\alpha}{p'}}\right)\left(\frac{n+1}{n}\right)^\beta \frac{C_\beta}{n^\beta} + 2^{3\alpha}\frac{C_{1,p'} + C_{2,p'}}{(n+1)^{3\alpha-\beta}}\frac{C_\beta}{(n+1)^\beta} \\
&\leq \left(\left(1 - c_{p'}\gamma_n n^{-\frac{1-\alpha}{p'}}\right)\left(\frac{n+1}{n}\right)^\beta + 2^{3\alpha}\frac{C_{1,p'} + C_{2,p'}}{(n+1)^{3\alpha-\beta}}\right)\frac{C_\beta}{(n+1)^\beta}.
\end{aligned}$$

By definition of $n'_{p'}$,

$$\mathbb{E}\left[\|V_{n+1} - \Gamma_m\|_F^4\right] \leq \frac{C_\beta}{(n+1)^\beta}, \tag{6.26}$$

which concludes the induction and the proof.

6.5.3 Proof of Theorem 6.3.4

In order to prove Theorem 6.3.4, we first recall the following Lemma.

Lemma 6.5.5 ([GB15]). *Let $Y_1, ..., Y_n$ be random variables taking values in a normed vector space such that for all positive constant q and for all $k \geq 1$, $\mathbb{E}\left[\|Y_k\|^q\right] < \infty$. Then, for all real numbers $a_1, ..., a_n$ and for all integer p, we have*

$$\mathbb{E}\left[\left\|\sum_{k=1}^n a_k Y_k\right\|^p\right] \leq \left(\sum_{k=1}^n |a_k| \left(\mathbb{E}\left[\|Y_k\|^p\right]\right)^{\frac{1}{p}}\right)^p \tag{6.27}$$

We can now prove Theorem 6.3.4. Let us rewrite decomposition (6.19) as follows

$$\nabla_m^2 G\left(\Gamma_m\right)\left(V_n - \Gamma_m\right) = \frac{T_n}{\gamma_n} - \frac{T_{n+1}}{\gamma_n} + \xi_{n+1} - r_n - r'_n - \delta_n, \tag{6.28}$$

with $T_n := V_n - \Gamma_m$. As in [Pel00], we sum these equalities, apply Abel's transform and divide by n to get

$$\nabla_m^2 G\left(\Gamma_m\right)\left(\overline{V}_n - \Gamma_m\right) = \frac{1}{n}\left(\frac{T_1}{\gamma_1} - \frac{T_{n+1}}{\gamma_{n+1}} + \sum_{k=2}^n T_k\left(\frac{1}{\gamma_k} - \frac{1}{\gamma_{k-1}}\right) - \sum_{k=1}^n \delta_k - \sum_{k=1}^n r_k - \sum_{k=1}^n r'_k + \sum_{k=1}^n \xi_{k+1}\right).$$

We now bound the quadratic mean of each term at the right-hand side of previous equality.

First, we have $\frac{1}{n^2}\mathbb{E}\left[\left\|\frac{T_1}{\gamma_1}\right\|_F^2\right] = o\left(\frac{1}{n}\right)$. Applying Theorem 6.3.3,

$$\frac{1}{n^2}\mathbb{E}\left[\left\|\frac{T_{n+1}}{\gamma_n}\right\|_F^2\right] \leq \frac{1}{n^2}\frac{C' c_\gamma^{-2}}{n^{-\alpha}} = o\left(\frac{1}{n}\right).$$

Moreover, since $\left|\gamma_k^{-1} - \gamma_{k-1}^{-1}\right| \leq 2\alpha c_\gamma^{-1} k^{\alpha-1}$, the application of Lemma 6.5.5 and Theorem 6.3.3 gives

$$\begin{aligned}
\frac{1}{n^2}\mathbb{E}\left[\left\|\sum_{k=2}^{n}\left(\gamma_k^{-1} - \gamma_{k-1}^{-1}\right) T_k\right\|_F^2\right] &\leq \frac{1}{n^2}\left(\sum_{k=2}^{n}\left|\gamma_k^{-1} - \gamma_{k-1}^{-1}\right|\sqrt{\mathbb{E}\left[\|T_k\|_F^2\right]}\right)^2 \\
&\leq \frac{1}{n^2} 4\alpha^2 c_\gamma^{-2} C' \left(\sum_{k=2}^{n}\frac{1}{k^{1-\alpha/2}}\right)^2 \\
&= O\left(\frac{1}{n^{2-\alpha}}\right) \\
&= o\left(\frac{1}{n}\right),
\end{aligned}$$

since $\alpha < 1$. In the same way, since $\|\delta_n\|_F \leq 6C\|T_n\|_F^2$, applying Lemma 6.5.5 and Theorem 6.3.3 with $\beta > 1$,

$$\begin{aligned}
\frac{1}{n^2}\mathbb{E}\left[\left\|\sum_{k=1}^{n}\delta_k\right\|_F^2\right] &\leq \frac{1}{n^2}\left(\sum_{k=1}^{n}\sqrt{\mathbb{E}\left[\|\delta_k\|_F^2\right]}\right)^2 \\
&\leq \frac{36C^2}{n^2}\left(\sum_{k=1}^{n}\sqrt{\mathbb{E}\left[\|T_k\|_F^4\right]}\right)^2 \\
&\leq \frac{36C^2 C_\beta}{n^2}\left(\sum_{k=1}^{n}\frac{1}{k^{\beta/2}}\right)^2 \\
&= O\left(\frac{1}{n^\beta}\right) \\
&= o\left(\frac{1}{n}\right),
\end{aligned}$$

Moreover, let $D := 12\left(\sqrt{C} + C\sqrt{\|\Gamma_m\|_F}\right)$. Since $\|r_n\|_F \leq D\|\overline{m}_n - m\|$, and since there is a

positive constant C'' such that for all $n \geq 1$, $\mathbb{E}\left[\|\overline{m}_n - m\|^2\right] \leq C'' n^{-1}$,

$$\begin{aligned}\frac{1}{n^2}\mathbb{E}\left[\left\|\sum_{k=1}^{n} r_k\right\|_F^2\right] &\leq \frac{1}{n^2}\left(\sum_{k=1}^{n}\sqrt{\mathbb{E}\left[\|r_k\|_F^2\right]}\right)^2 \\ &\leq \frac{D^2}{n^2}\left(\sum_{k=1}^{n}\sqrt{\mathbb{E}\left[\|\overline{m}_n - m\|^2\right]}\right) \\ &\leq \frac{D^2 C''}{n^2}\left(\sum_{k=1}^{n}\frac{1}{k^{1/2}}\right)^2 \\ &= O\left(\frac{1}{n}\right).\end{aligned}$$

Since $\|r'_n\|_F \leq C_0 \|\overline{m}_n - m\| \|V_n - \Gamma_m\|_F^2$ with $C_0 := 12\left(C\sqrt{\|\Gamma_m\|_F} + C^{3/4}\right)$, Cauchy-Schwarz's inequality and Lemma 6.5.5 give

$$\begin{aligned}\frac{1}{n^2}\mathbb{E}\left[\left\|\sum_{k=1}^{n} r'_n\right\|_F^2\right] &\leq \frac{1}{n^2}\left(\sum_{k=1}^{n}\sqrt{\mathbb{E}\left[\|r'_n\|_F^2\right]}\right)^2 \\ &\leq \frac{C_0^2}{n^2}\left(\sum_{k=1}^{n}\sqrt{\mathbb{E}\left[\|\overline{m}_n - m\|^2 \|V_n - \Gamma_m\|_F^2\right]}\right)^2 \\ &\leq \frac{C_0^2}{n^2}\left(\sum_{k=1}^{n}\left(\mathbb{E}\left[\|\overline{m}_n - m\|^4\right]\right)^{\frac{1}{4}}\left(\mathbb{E}\left[\|V_n - \Gamma_m\|_F^4\right]\right)^{\frac{1}{4}}\right)^2.\end{aligned}$$

Applying Theorem 4.2 in [GB15] and Theorem 3.3,

$$\begin{aligned}\frac{1}{n^2}\mathbb{E}\left[\left\|\sum_{k=1}^{n} r'_n\right\|_F^2\right] &\leq \frac{C_0^2\sqrt{C_\beta}\sqrt{K_2}}{n^2}\left(\sum_{k=1}^{n}\frac{1}{k^{\beta/4+1/2}}\right)^2 \\ &= O\left(\frac{1}{n^{1+\beta/2}}\right) \\ &= o\left(\frac{1}{n}\right),\end{aligned}$$

since $\beta > 0$. Finally, one can easily check that $\mathbb{E}\left[\|\xi_{n+1}\|_F^2\right] \leq 1$, and since (ξ_n) is a sequence

of martingale differences adapted to the filtration $(\mathcal{F}_n)$,

$$\begin{aligned}\frac{1}{n^2}\mathbb{E}\left[\left\|\sum_{k=1}^{n}\xi_{k+1}\right\|_F^2\right] &= \frac{1}{n^2}\left(\sum_{k=1}^{n}\mathbb{E}\left[\left\|\xi_{k+1}\right\|_F^2\right]+2\sum_{k=1}^{n}\sum_{k'=k+1}^{n}\mathbb{E}\left[\left\langle\xi_{k+1},\xi_{k'+1}\right\rangle_F\right]\right)\\ &= \frac{1}{n^2}\left(\sum_{k=1}^{n}\mathbb{E}\left[\left\|\xi_{k+1}\right\|_F^2\right]+2\sum_{k=1}^{n}\sum_{k'=k+1}^{n}\mathbb{E}\left[\left\langle\xi_{k+1},\mathbb{E}\left[\xi_{k'+1}\middle|\mathcal{F}_{k'}\right]\right\rangle_F\right]\right)\\ &= \frac{1}{n^2}\sum_{k=1}^{n}\mathbb{E}\left[\left\|\xi_{k+1}\right\|_F^2\right]\\ &\leq \frac{1}{n}.\end{aligned}$$

Thus, there is a positive constant K such that for all $n \geq 1$,

$$\mathbb{E}\left[\left\|\nabla_m^2 G\left(\Gamma_m\right)\left(\overline{V}_n-\Gamma_m\right)\right\|_F^2\right] \leq \frac{K}{n}.$$

Let $\lambda_{\min}$ be the smallest eigenvalue of $\nabla_m^2 G\left(\Gamma_m\right)$. We have, with Proposition B.1 in the supplementary file, that $\lambda_{\min} > 0$ and the announced result is proven,

$$\mathbb{E}\left[\left\|\overline{V}_n-\Gamma_m\right\|_F^2\right] \leq \frac{K}{\lambda_{\min}^2 n}.$$

6.6 Concluding remarks

The simulation study and the illustration on real data indicate that performing robust principal components analysis via the median covariation matrix, which can bring new information compared to classical PCA, is an interesting alternative to more classical robust principal components analysis techniques. The use of recursive algorithms permits to perform robust PCA on very large datasets, in which outlying observations may be hard to detect. Another interest of the use of such sequential algorithms is that estimation of the median covariation matrix as well as the principal components can be performed online with automatic update at each new observation and without being obliged to store all the data in memory.

A deeper study of the asymptotic behaviour of the recursive algorithms would certainly deserve further investigations. Proving the asymptotic normality and obtaining the limiting variance of the sequence of estimators $\overline{V}_n$ when m is unknown would be of great interest. It a challenging issue that is beyond the scope of the paper and would require to study the joint weak convergence of the two simultaneous recursive averaged estimators of m and Γ_m.

The use of the MCM could be interesting to robustify the estimation in many different statistical models, particularly with functional data. For example, it could be employed as an alternative to robust functional projection pursuit in robust functional time series prediction or for robust estimation in functional linear regression, with the introduction of the median cross-covariation matrix.

Acknowledgements. We thank the company Médiamétrie for allowing us to illustrate our methodologies with their data. We also thank Dr. Peggy Cénac for a careful reading of the proofs.

Annexe C

Fast Estimation of the Median Covariation Matrix with Application to Online Robust Principal Components Analysis. Appendix

Résumé

Dans cette partie, nous commençons par rappeler comment estimer la médiane géométrique et la Median Covariation Matrix à l'aide de l'algorithme de Weiszfeld. Ensuite, nous donnons des propriétés de convexité de la fonction que l'on veut minimiser, avant de rappeler quelques décompositions des algorithmes. Enfin, les preuves des lemmes et propositions techniques sont données. Plus précisément, on donne la preuve du Lemme 6.5.1, qui permet de majorer les termes de reste. De plus, on prouve les Lemmes 6.5.2, 6.5.3 et 6.5.4 qui permettent d'obtenir la vitesse de convergence en moyenne quadratique de l'algorithme de gradient stochastique.

C.1 Estimating the median covariation matrix with Weiszfeld's algorithm

Suppose we have a fixed size sample $X_1, \ldots, X_n$ and we want to estimate the geometric median.

The iterative Weiszfeld's algorithm relies on the fact that the solution m_n^* of the following optimization problem

$$\min_{\mu \in H} \sum_{i=1}^{n} \|X_i - \mu\|$$

satisfies, when $m_n^* \neq X_i$, for all $i = 1, \ldots, n$

$$m_n^* = \sum_{i=1}^{n} w_i \left(m_n^*\right) \; X_i$$

where the weights $w_i(x)$ are defined by

$$w_i(x) = \frac{\|X_i - x\|^{-1}}{\sum_{j=1}^{n} \|X_j - x\|^{-1}}.$$

Weiszfeld's algorithm is based on the following iterative scheme. Consider first a pilot estimator $\widehat{m}^{(0)}$ of m. At step (e), a new approximation $\widehat{m}_n^{(e+1)}$ to m is given by

$$\widehat{m}_n^{(e+1)} = \sum_{i=1}^{n} w_i \left(\widehat{m}_n^{(e)}\right) \; X_i. \tag{C.1}$$

The iterative procedure is stopped when $\left\|\widehat{m}_n^{(e+1)} - \widehat{m}_n^{(e)}\right\| \leq \epsilon$, for some precision ϵ known in advance. The final value of the algorithm is denoted by $\widehat{m}_n$.

The estimator of the MCM is computed similarly. Suppose $\widehat{\Gamma}^{(e)}$ has been calculated at step (e), then at step $(e+1)$, the new approximation $\widehat{\Gamma}^{(e+1)}$ to Γ_m is defined by

$$\widehat{\Gamma}_n^{(e+1)} = \sum_{i=1}^{n} W_i \left(\widehat{\Gamma}^{(e)}\right) (X_i - \widehat{m}_n)(X_i - \widehat{m}_n)^T. \tag{C.2}$$

The procedure is stopped when $\left\|\widehat{\Gamma}^{(e+1)} - \widehat{\Gamma}^{(e)}\right\|_F \leq \epsilon$, for some precision ϵ fixed in advance.

Note that by construction, this algorithm leads to an estimated median covariation matrix that is always non negative.

C.2 Convexity results

In this section, we first give and recall some convexity properties of functional G_h. The following one gives some information on the spectrum of the Hessian of G.

Proposition C.2.1. *Under assumptions 1-3(b), for all $h \in H$ and $V \in \mathcal{S}(H)$, $\mathcal{S}(H)$ admits an orthonormal basis composed of eigenvectors of $\nabla_h^2 G(V)$. Let us denote by $\{\lambda_{h,V,i}, i \in \mathbb{N}\}$ the set of eigenvalues of $\nabla_h^2 G(V)$. For all $i \in \mathbb{N}$,*

$$0 \leq \lambda_{h,V,i} \leq C.$$

Moreover, there is a positive constant c_m such that for all $i \in \mathbb{N}$,

$$0 < c_m \leq \lambda_{m,\Gamma_m,i} \leq C.$$

Finally, by continuity, there are positive constants ϵ, ϵ' such that for all $h \in \mathcal{B}(m, \epsilon)$ and $V \in \mathcal{B}(\Gamma_m, \epsilon')$, and for all $i \in \mathbb{N}$,

$$\frac{1}{2} c_m \leq \lambda_{h,V,i} \leq C.$$

The proof is very similar to the one in [CCZ13] and consequently it is not given here. Furthermore, as in [CCGB15], it ensures the local strong convexity as shown in the following corollary.

Corollary C.2.1. *Under assumptions 1-3(b), for all positive constant A, there is a positive constant c_A such that for all $V \in \mathcal{B}(\Gamma_m, A)$ and $h \in \mathcal{B}(m, \epsilon)$,*

$$\langle \nabla_h G(V) - \nabla_h G(\Gamma_m), V - \Gamma_m \rangle_H \geq c_A \left\| V - \Gamma_m \right\|_F^2.$$

Finally, the following lemma gives an upper bound on the remainder term in the Taylor's expansion of the gradient.

Lemma C.2.1. *Under assumptions 1-3(b), for all $h \in H$ and $V \in \mathcal{S}(H)$,*

$$\left\| \nabla G_h(V) - \nabla G_h(\Gamma_m) - \nabla_h^2 G(\Gamma_m)(V - \Gamma_m) \right\|_F \leq 6C \left\| V - \Gamma_m \right\|_F^2. \tag{C.3}$$

Proof of Lemma C.2.1. Let $\delta_{V,h} := \nabla G_h(V) - \nabla G_h(\Gamma_m) - \nabla_h^2 G(\Gamma_m)(V - \Gamma_m)$, since

$\nabla G_h(V) - \nabla G_h(\Gamma_m) = \int_0^1 \nabla_h^2 G(\Gamma_m + t(V-\Gamma_m))(V-\Gamma_m)\,dt$, we have

$$\begin{aligned}\|\delta_{V,h}\|_F &= \left\| \int_0^1 \nabla_h^2 G(\Gamma_m + t(V-\Gamma_m))((V-\Gamma_m)\,dt - \nabla_h^2 G(\Gamma_m)(V-\Gamma_m) \right\|_F \\ &\leq \int_0^1 \left\| \nabla_h^2 G(\Gamma_m + t(V-\Gamma_m))((V-\Gamma_m) - \nabla_h^2 G(\Gamma_m)(V-\Gamma_m) \right\|_F dt.\end{aligned}$$

As in the proof of Lemma 5.1 in [CCGB15], under assumptions 1-3(b), one can check that for all $h \in H$, and $t \in [0,1]$,

$$\left\| \nabla_h^2 G(\Gamma_m + t(V-\Gamma_m))((V-\Gamma_m) - \nabla_h^2 G(\Gamma_m)(V-\Gamma_m) \right\|_F \leq 6C \|V-\Gamma_m\|_F^2,$$

which concludes the proof. □

C.3 Decompositions of the Robbins-Monro algorithm and proof of Lemma 6.5.1

Let us recall that the Robbins-Monro algorithm is defined recursively by

$$\begin{aligned}V_{n+1} &= V_n + \gamma_n \frac{(X_{n+1}-\overline{m}_n)(X_{n+1}-\overline{m}_n)^T - V_n}{\left\|(X_{n+1}-\overline{m}_n)(X_{n+1}-\overline{m}_n)^T - V_n\right\|_F} \\ &= V_n - \gamma_n U_{n+1},\end{aligned}$$

with $U_{n+1} := -\frac{(X_{n+1}-\overline{m}_n)(X_{n+1}-\overline{m}_n)^T - V_n}{\left\|(X_{n+1}-\overline{m}_n)(X_{n+1}-\overline{m}_n)^T - V_n\right\|_F}$. Let us remark that $\xi_{n+1} := \nabla_{\overline{m}_n} G(V_n) - U_{n+1}$, (ξ_n) is a sequence of martingale differences adapted to the filtration $(\mathcal{F}_n)$ and the algorithm can be written as follows

$$V_{n+1} = V_n - \gamma_n (\nabla G_{\overline{m}_n}(V_n) - \nabla G_{\overline{m}_n}(\Gamma_m)) + \gamma_n \xi_{n+1} - \gamma_n r_n, \tag{C.4}$$

with $r_n := \nabla G_{\overline{m}_n}(\Gamma_m) - \nabla G_m(\Gamma_m)$. Finally, let is consider the following linearization of the gradient,

$$V_{n+1} - \Gamma_m = \left(I_{\mathcal{S}(H)} - \gamma_n \nabla_m^2 G(\Gamma_m)\right)(V_n - \Gamma_m) + \gamma_n \xi_{n+1} - \gamma_n r_n - \gamma_n r_n' - \gamma_n \delta_n, \tag{C.5}$$

with

$$r'_n := \left(\nabla^2_{\overline{m}_n} G\left(\Gamma_m\right) - \nabla^2_m G\left(\Gamma_m\right)\right)\left(V_n - \Gamma_m\right),$$
$$\delta_n := \nabla G_{\overline{m}_n}\left(V_n\right) - \nabla G_{\overline{m}_n}\left(\Gamma_m\right) - \nabla^2_{\overline{m}_n} G\left(\Gamma_m\right)\left(V_n - \Gamma_m\right).$$

Proof of Lemma 6.5.1. The bound of $\|\delta_n\|$ is a corollary of Lemma C.2.1.

Bounding $\|r_n\|$

Let us recall that for all $h \in H$, $Y(h) := (X-h)(X-h)^T$. We now define for all $h \in H$ the random function $\varphi_h : [0,1] \longrightarrow \mathcal{S}(H)$ defined for all $t \in [0,1]$ by

$$\varphi_h(t) := \frac{Y(m+th) - \Gamma_m}{\|Y(m+th) - \Gamma_m\|_F}.$$

Note that $r_n = \mathbb{E}\left[\varphi_{\overline{m}_n - m}(0) - \varphi_{\overline{m}_n - m}(1) \middle| \mathcal{F}_n\right]$. Thus, by dominated convergence,

$$\|r_n\|_F \leq \sup_{t\in[0,1]} \mathbb{E}\left[\left\|\varphi'_{\overline{m}_n - m}(t)\right\|_F \middle| \mathcal{F}_n\right].$$

Moreover, one can check that for all $h \in H$,

$$\begin{aligned}
\varphi'_h(t) = &- \frac{h\left(X - m - th\right)^T}{\|Y(m+th) - \Gamma_m\|_F} - \frac{\left(X - m - th\right)h^T}{\|Y(m+th) - \Gamma_m\|_F} \\
&+ \left\langle Y(m+th) - \Gamma_m, h\left(X - m - th\right)^T \right\rangle_F \frac{Y(m+th) - \Gamma_m}{\|Y(m+th) - \Gamma_m\|_F^3} \\
&+ \left\langle Y(m+th) - \Gamma_m, \left(X - m - th\right)h^T \right\rangle_F \frac{Y(m+th) - \Gamma_m}{\|Y(m+th) - \Gamma_m\|_F^3}.
\end{aligned}$$

We now bound each term on the right-hand side of previous equality. First, applying Cauchy-Schwarz's inequality and using the fact that for all $h, h' \in H$, $\left\|hh'^T\right\|_F = \|h\|\,\|h'\|$,

$$\begin{aligned}
\mathbb{E}\left[\frac{\left\|h\left(X - m - th\right)^T\right\|_F}{\|Y(m+th) - \Gamma_m\|_F}\right] &\leq \|h\|\,\mathbb{E}\left[\frac{\|X - m - th\|}{\|Y(m+th) - \Gamma_m\|_F}\right] \\
&\leq \|h\|\,\mathbb{E}\left[\frac{\sqrt{\|Y(m+th)\|_F}}{\|Y(m+th) - \Gamma_m\|_F}\right] \\
&\leq \|h\|\left(\mathbb{E}\left[\frac{\sqrt{\|\Gamma_m\|_F}}{\|Y(m+th) - \Gamma_m\|_F}\right] + \mathbb{E}\left[\frac{1}{\sqrt{\|Y(m+th) - \Gamma_m\|_F}}\right]\right).
\end{aligned}$$

Thus, since $\mathbb{E}\left[\frac{1}{\|Y(m+th)-\Gamma_m\|_F}\right] \leq C$,

$$\mathbb{E}\left[\frac{\left\|h\left(X-m-th\right)^T\right\|_F}{\|Y(m+th)-\Gamma_m\|_F}\right] \leq \|h\|\left(C\sqrt{\|\Gamma_m\|_F}+\sqrt{C}\right). \tag{C.6}$$

In the same way,

$$\mathbb{E}\left[\frac{\left\|\left(X-m-th\right)h^T\right\|_F}{\|Y(m+th)-\Gamma_m\|_F}\right] \leq \|h\|\left(C\sqrt{\|\Gamma_m\|_F}+\sqrt{C}\right). \tag{C.7}$$

Applying Cauchy-Schwarz's inequality,

$$\begin{aligned}\mathbb{E}\left[\left|\left\langle Y(m+th)-\Gamma_m, h\left(X-m-th\right)^T\right\rangle_F\right|\frac{\|Y(m+th)-\Gamma_m\|_F}{\|Y(m+th)-\Gamma_m\|_F^3}\right] &\leq \mathbb{E}\left[\frac{\left\|h\left(X-m-th\right)^T\right\|_F}{\|Y(m+th)-\Gamma_m\|_F}\right]\\ &\leq \|h\|\,\mathbb{E}\left[\frac{\|X-m-th\|}{\|Y(m+th)-\Gamma_m\|_F}\right]\\ &\leq \|h\|\,\mathbb{E}\left[\frac{\sqrt{\|Y(m+th)\|_F}}{\|Y(m+th)-\Gamma_m\|_F}\right].\end{aligned}$$

Thus, since $\mathbb{E}\left[\frac{1}{\|Y(m+th)-\Gamma_m\|_F}\right] \leq C$, and since for all positive constants a, b, $\sqrt{a+b} \leq \sqrt{a}+\sqrt{b}$,

$$\begin{aligned}\|h\|\,\mathbb{E}\left[\frac{\sqrt{\|Y(m+th)\|_F}}{\|Y(m+th)-\Gamma_m\|_F}\right] &\leq \|h\|\left(\mathbb{E}\left[\frac{\sqrt{\|\Gamma_m\|_F}}{\|Y(m+th)-\Gamma_m\|_F}\right]+\mathbb{E}\left[\frac{1}{\sqrt{\|Y(m+th)-\Gamma_m\|_F}}\right]\right)\\ &\leq \|h\|\left(C\sqrt{\|\Gamma_m\|_F}+\sqrt{C}\right).\end{aligned}$$

Finally,

$$\mathbb{E}\left[\left|\left\langle Y(m+th)-\Gamma_m, h\left(X-m-th\right)^T\right\rangle_F\right|\frac{\|Y(m+th)-\Gamma_m\|_F}{\|Y(m+th)-\Gamma_m\|_F^3}\right] \leq \|h\|\left(C\sqrt{\|\Gamma_m\|_F}+\sqrt{C}\right), \tag{C.8}$$

$$\mathbb{E}\left[\left|\left\langle Y(m+th)-\Gamma_m, \left(X-m-th\right)h^T\right\rangle_F\right|\frac{\|Y(m+th)-\Gamma_m\|_F}{\|Y(m+th)-\Gamma_m\|_F^3}\right] \leq \|h\|\left(C\sqrt{\|\Gamma_m\|_F}+\sqrt{C}\right). \tag{C.9}$$

Applying inequalities (C.6) to (C.9) with $h = \overline{m}_n - m$, the announced result is proven,

$$\|r_n\|_F \leq 4\left(\sqrt{C} + C\sqrt{\|\Gamma_m\|_F}\right)\|\overline{m}_n - m\|\,.$$

Bounding $\|r'_n\|$

For all $h \in H$ and $V \in \mathcal{S}(H)$, we define the random function $\varphi_{h,V} : [0,1] \longrightarrow \mathcal{S}(H)$ such that for all $t \in [0,1]$,

$$\varphi_{h,V}(t) := \frac{1}{\|Y(m+th) - \Gamma_m\|_F}\left(I_{\mathcal{S}(H)} - \frac{(Y(m+th) - \Gamma_m) \otimes_F (Y(m+th) - \Gamma_m)}{\|Y(m+th) - \Gamma_m\|_F^2}\right)(V)\,.$$

Note that $r'_n = \mathbb{E}\left[\varphi_{\overline{m}_n - m, V_n - \Gamma_m}(1) - \varphi_{\overline{m}_n - m, V_n - \Gamma_m}(0)\middle|\mathcal{F}_n\right]$. By dominated convergence,

$$\|r'_n\|_F \leq \sup_{t\in[0,1]} \mathbb{E}\left[\|\varphi'_{\overline{m}_n - m, V_n - \Gamma_m}(t)\|_F \middle|\mathcal{F}_n\right].$$

Moreover, as for the bound of $\|r_n\|$, one can check, with an application of Cauchy-Schwarz's inequality, that for all $h \in H$, $V \in \mathcal{S}(H)$, and $t \in [0,1]$,

$$\begin{aligned}
\varphi'_{h,V}(t) &\leq 6\frac{\|Y(m+th) - \Gamma_m\|_F \left\|h^T(X - m - th)\right\|_F}{\|Y(m+th) - \Gamma_m\|_F^3}\|V\|_F \\
&\quad + 6\frac{\|Y(m+th) - \Gamma_m\|_F \left\|h(X - m - th)^T\right\|_F}{\|Y(m+th) - \Gamma_m\|_F^5}\|(Y(m+th) - \Gamma_m) \otimes_F (Y(m+th) - \Gamma_m)(V)\|_F \\
&\leq 12\frac{\left\|h(X - m - th)^T\right\|_F}{\|Y(m+th) - \Gamma_m\|_F^2}\|V\|_F\,.
\end{aligned}$$

Finally,

$$\begin{aligned}
\mathbb{E}\left[\frac{\left\|h(X - m - th)^T\right\|_F}{\|Y(m+th) - \Gamma_m\|_F^2}\|V\|_F\right] &\leq \mathbb{E}\left[\frac{\|h\|\,\|X - m - th\|}{\|Y(m+th) - \Gamma_m\|_F^2}\|V\|_F\right] \\
&\leq \|h\|\,\|V\|_F\,\mathbb{E}\left[\frac{\sqrt{\|\Gamma_m\|_F}}{\|Y(m+th) - \Gamma_m\|_F^2}\right] \\
&\quad + \|h\|\,\|V\|_F\,\mathbb{E}\left[\frac{1}{\|Y(m+th) - \Gamma_m\|_F^{3/2}}\right] \\
&\leq \left(C\sqrt{\|\Gamma_m\|_F} + C^{3/4}\right)\|h\|\,\|V\|_F\,. \qquad \text{(C.10)}
\end{aligned}$$

Then the announced result follows from an application of inequality (C.10) with $h = \overline{m}_n - m$

and $V = V_n - \Gamma_m$,

$$\left\|r_n'\right\| \leq 12\left(C\sqrt{\left\|\Gamma_m\right\|_F} + C^{3/4}\right)\left\|\overline{m}_n - m\right\|\left\|V_n - \Gamma_m\right\|_F.$$

□

C.4 Proofs of Lemma 6.5.2, 6.5.3 and 6.5.4

Proof of Lemma 6.5.2. Using decomposition (C.4),

$$\begin{aligned}\left\|V_{n+1} - \Gamma_m\right\|_F^2 = \left\|V_n - \Gamma_m\right\|_F^2 &- 2\gamma_n\left\langle V_n - \Gamma_m, \nabla G_{\overline{m}_n}(V_n) - \nabla G_{\overline{m}_n}(\Gamma_m)\right\rangle_F \\ &+ \gamma_n^2\left\|\nabla G_{\overline{m}_n}(V_n) - \nabla G_{\overline{m}_n}(\Gamma_m)\right\|_F^2 \\ &+ \gamma_n^2\left\|\xi_{n+1}\right\|_F^2 + 2\gamma_n\left\langle V_n - \Gamma_m - \gamma_n\left(\nabla G_{\overline{m}_n}(V_n) - \nabla G_{\overline{m}_n}(\Gamma_m)\right), \xi_{n+1}\right\rangle_F \\ &+ \gamma_n^2\left\|r_n\right\|_F^2 - 2\gamma_n\left\langle r_n, V_n - \Gamma_m\right\rangle_F - 2\gamma_n^2\left\langle r_n, \xi_{n+1} - \nabla G_{\overline{m}_n}(V_n) + \nabla G_{\overline{m}_n}(\Gamma_m)\right\rangle_F.\end{aligned}$$

Note that for all $h \in H$ and $V \in \mathcal{S}(H)$ we have $\left\|\nabla G_h(V)\right\|_F \leq 1$. Moreover, $\left\|r_n\right\|_F \leq 2$ and $\left\|\xi_{n+1}\right\|_F \leq 2$. Since for all $h \in H$, G_h is a convex function, we get with Cauchy-Schwarz's inequality,

$$\left\|V_{n+1} - \Gamma_m\right\|_F^2 a \leq \left\|V_n - \Gamma_m\right\|_F^2 + 36\gamma_n^2 + 2\gamma_n\left\langle\xi_{n+1}, V_n - \Gamma_m\right\rangle_F - 2\gamma_n\left\langle r_n, V_n - \Gamma_m\right\rangle_F. \quad \text{(C.11)}$$

Let $C' := 4\left(\sqrt{C} + C\sqrt{\left\|\Gamma_m\right\|_F}\right)$, let us recall that $\left\|r_n\right\|_F \leq C'\left\|\overline{m}_n - m\right\|$. We now prove by induction that for all integer $p \geq 1$, there is a positive constant M_p such that for all $n \geq 1$, $\mathbb{E}\left[\left\|V_n - \Gamma_m\right\|_F^{2p}\right] \leq M_p$.

The case $p = 1$ has been studied in the proof of Theorem 3.2. Let $p \geq 2$ and suppose from now that for all $k \leq p-1$, there is a positive constant M_k such that for all $n \geq 1$,

$$\mathbb{E}\left[\left\|V_n - \Gamma_m\right\|_F^{2k}\right] \leq M_k.$$

Bounding $\mathbb{E}\left[\left\|V_n - \Gamma_m\right\|_F^{2p}\right]$.
Let us apply inequality (C.11), for all $p \geq 2$ and use the fact that (ξ_n) is a sequence of mar-

tingales differences adapted to the filtration $(\mathcal{F}_n)$,

$$\mathbb{E}\left[\|V_{n+1}-\Gamma_m\|_F^{2p}\right] \leq \mathbb{E}\left[\left(\|V_n-\Gamma_m\|_F^2+36\gamma_n^2+2\gamma_n\|r_n\|_F\|V_n-\Gamma_m\|_F\right)^p\right]$$
$$+\sum_{k=2}^{p}\binom{p}{k}\mathbb{E}\left[\left(2\gamma_n\langle V_n-\Gamma_m,\xi_{n+1}\rangle_F\right)^k\left(\|V_n-\Gamma_m\|_F^2+36\gamma_n^2+2\gamma_n\|r_n\|_F\|V_n-\Gamma_m\|_F\right)^{p-k}\right]. \tag{C.12}$$

Let us denote by $(*)$ the second term on the right-hand side of inequality (C.12). Applying Cauchy-Schwarz's inequality and since $\|\xi_{n+1}\|_F \leq 2$,

$$(*) = \sum_{k=2}^{n}\binom{p}{k}\mathbb{E}\left[\left(2\gamma_n\langle V_n-\Gamma_m,\xi_{n+1}\rangle\right)^k\left(\|V_n-\Gamma_m\|_F^2+36\gamma_n^2+2\gamma_n\|r_n\|_F\|V_n-\Gamma_m\|_F\right)^{p-k}\right]$$
$$\leq \sum_{k=2}^{p}\binom{p}{k}2^{2k}\gamma_n^k\mathbb{E}\left[\|V_n-\Gamma_m\|_F^k\left(\|V_n-\Gamma_m\|_F^2+36\gamma_n^2+2\gamma_n\|r_n\|_F\|V_n-\Gamma_m\|_F\right)^{p-k}\right].$$

With the help of Lemma C.5.1,

$$(*) \leq \sum_{k=2}^{p}2^{2k}3^{p-k-1}\gamma_n^k\mathbb{E}\left[\|V_n-\Gamma_m\|_F^{2p-k}\right]+\sum_{k=2}^{p}2^{2k}3^{p-k-1}36^{p-k}\gamma_n^{2p-k}\mathbb{E}\left[\|V_n-\Gamma_m\|_F^k\right]$$
$$+\sum_{k=2}^{p}2^{p+k}3^{p-k-1}\gamma_n^p\mathbb{E}\left[\|r_n\|_F^{p-k}\|V_n-\Gamma_m\|_F^p\right].$$

Applying Cauchy-Schwarz's inequality,

$$\sum_{k=2}^{p}2^{2k}3^{p-k-1}\gamma_n^k\mathbb{E}\left[\|V_n-\Gamma_m\|_F^{2p-k}\right] = \sum_{k=2}^{p}2^{2k}3^{p-k-1}\gamma_n^k\mathbb{E}\left[\|V_n-\Gamma_m\|_F^{p-1}\|V_n-\Gamma_m\|_F^{p+1-k}\right]$$
$$\leq \sum_{k=2}^{p}2^{2k}3^{p-k-1}\gamma_n^k \sqrt{\mathbb{E}\left[\|V_n-\Gamma_m\|_F^{2(p-1)}\right]}\sqrt{\mathbb{E}\left[\|V_n-\Gamma_m\|_F^{2(p+1-k)}\right]}.$$

By induction,

$$\sum_{k=2}^{p}2^{2k}3^{p-k-1}\gamma_n^k\mathbb{E}\left[\|V_n-\Gamma_m\|_F^{2p-k}\right] \leq \sum_{k=2}^{p}2^{2k}3^{p-k-1}\gamma_n^k\sqrt{M_{p-1}}\sqrt{M_{p+1-k}}$$
$$= O\left(\gamma_n^2\right). \tag{C.13}$$

In the same way, applying Cauchy-Schwarz's inequality and by induction,

$$\begin{aligned}\sum_{k=2}^{p} 2^{2k}3^{p-k-1}36^{p-k}\gamma_n^{2p-k}\mathbb{E}\left[\|V_n-\Gamma_m\|_F^k\right] &= \sum_{k=2}^{p} 2^{2k}3^{p-k-1}36^{p-k}\gamma_n^{2p-k}\mathbb{E}\left[\|V_n-\Gamma_m\|_F\|V_n-\Gamma_m\|_F^{k-1}\right]\\ &\leq \sum_{k=2}^{p} 2^{2k}3^{p-k-1}36^{p-k}\gamma_n^{2p-k}\sqrt{M_1}\sqrt{M_{k-1}}\\ &= O\left(\gamma_n^2\right), \end{aligned} \tag{C.14}$$

since $p \geq 2$. Similarly, since $\|r_n\|_F \leq 2$ and since $p \geq 2$, applying Cauchy-Schwarz's inequality and by induction,

$$\begin{aligned}\sum_{k=2}^{p} 2^{p+k}3^{p-k-1}\gamma_n^p\mathbb{E}\left[\|r_n\|_F^{p-k}\|V_n-\Gamma_m\|_F^p\right] &\leq \sum_{k=2}^{p} 2^{2p}3^{p-k-1}\gamma_n^p\mathbb{E}\left[\|V_n-\Gamma_m\|_F^p\right]\\ &\leq \sum_{k=2}^{p} 2^{2p}3^{p-k-1}\gamma_n^p\sqrt{M_1}\sqrt{M_{p-1}}\\ &= O\left(\gamma_n^2\right). \end{aligned} \tag{C.15}$$

Finally, applying inequalities (C.13) to (C.15), there is a positive constant A_1' such that for all $n \geq 1$,

$$\mathbb{E}\left[\sum_{k=2}^{p}\binom{p}{k}\left(2\gamma_n\left\langle V_n-\Gamma_m,\xi_{n+1}\right\rangle_F\right)^k\left(\|V_n-\Gamma_m\|_F^2+36\gamma_n^2+2\gamma_n\|r_n\|_F\|V_n-\Gamma_m\|_F\right)^{p-k}\right] \leq A_1'\gamma_n^2. \tag{C.16}$$

We now denote by $(**)$ the first term at the right-hand side of inequality (C.12). With the help of Lemma C.5.1 and applying Cauchy-Schwarz's inequality,

$$\begin{aligned}(**) &\leq \mathbb{E}\left[\|V_n-\Gamma_m\|_F^{2p}\right]+\sum_{k=1}^{p}\binom{p}{k}\mathbb{E}\left[\left(36\gamma_n^2+2\gamma_n\left\langle r_n,V_n-\Gamma_m\right\rangle_F\right)^k\|V_n-\Gamma_m\|_F^{2p-2k}\right]\\ &\leq \mathbb{E}\left[\|V_n-\Gamma_m\|_F^{2p}\right]+\sum_{k=1}^{p}\binom{p}{k}2^{k-1}\mathbb{E}\left[\left(36^k\gamma_n^{2k}+2^k\gamma_n^k\|r_n\|_F^k\|V_n-\Gamma_m\|_F^k\right)\|V_n-\Gamma_m\|_F^{2p-2k}\right]. \end{aligned}$$

Moreover, let

$$\begin{aligned}(***) &:= \sum_{k=1}^{p}\binom{p}{k}2^{k-1}\mathbb{E}\left[\left(36^k\gamma_n^{2k}+2^k\gamma_n^k\|r_n\|_F^k\|V_n-\Gamma_m\|_F^k\right)\|V_n-\Gamma_m\|_F^{2p-2k}\right]\\ &= \sum_{k=1}^{p}\binom{p}{k}2^{k-1}36^k\gamma_n^{2k}\mathbb{E}\left[\|V_n-\Gamma_m\|_F^{2p-2k}\right]+\sum_{k=1}^{p}\binom{p}{k}2^{2k-1}\gamma_n^k\mathbb{E}\left[\|r_n\|_F^k\|V_n-\Gamma_m\|_F^{2p-k}\right]. \end{aligned}$$

By induction,

$$\begin{aligned}\sum_{k=1}^{p}\binom{p}{k}2^{k-1}36^k\gamma_n^{2k}\mathbb{E}\left[\|V_n-\Gamma_m\|_F^{2p-2k}\right] &= \sum_{k=1}^{p}\binom{p}{k}2^{k-1}36^k\gamma_n^{2k}M_{p-k}\\ &= O\left(\gamma_n^2\right).\end{aligned}$$

Moreover,

$$\begin{aligned}\sum_{k=1}^{p}\binom{p}{k}2^{2k-1}\gamma_n^k\mathbb{E}\left[\|r_n\|_F^k\|V_n-\Gamma_m\|_F^{2p-k}\right] &= \sum_{k=2}^{p}\binom{p}{k}2^{2k-1}\gamma_n^k\mathbb{E}\left[\|r_n\|_F^k\|V_n-\Gamma_m\|_F^{2p-k}\right]\\ &\quad+2p\gamma_n\mathbb{E}\left[\|r_n\|_F\|V_n-\Gamma_m\|_F^{2p-1}\right].\end{aligned}$$

Applying Cauchy-Schwarz's inequality and by induction, since $\|r_n\|_F\leq 2$,

$$\begin{aligned}\sum_{k=2}^{p}\binom{p}{k}2^{2k-1}\gamma_n^k\mathbb{E}\left[\|r_n\|_F^k\|V_n-\Gamma_m\|_F^{2p-k}\right] &\leq \sum_{k=2}^{p}\binom{p}{k}2^{3k-1}\gamma_n^k\mathbb{E}\left[\|V_n-\Gamma_m\|_F^{2p-k}\right]\\ &\leq \sum_{k=2}^{p}\binom{p}{k}2^{3k-1}\gamma_n^k\sqrt{M_{p+1-k}}\sqrt{M_{p-1}}\\ &= O\left(\gamma_n^2\right).\end{aligned}$$

Moreover, applying Theorem 4.2 in [GB15] and Hölder's inequality, since $\|r_n\|_F \leq C'\ \|\overline{m}_n-m\|$,

$$\begin{aligned}2p\gamma_n\mathbb{E}\left[\|r_n\|_F\|V_n-\Gamma_m\|_F^{2p-1}\right] &\leq 2C'p\gamma_n\mathbb{E}\left[\|\overline{m}_n-m\|\,\|V_n-\Gamma_m\|_F^{2p-1}\right]\\ &\leq 2C'p\gamma_n\left(\mathbb{E}\left[\|\overline{m}_n-m\|^{2p}\right]\right)^{\frac{1}{2p}}\left(\mathbb{E}\left[\|V_n-\Gamma_m\|_F^{2p}\right]\right)^{\frac{2p-1}{2p}}\\ &\leq 2C'p\gamma_n\frac{K_p^{\frac{1}{2p}}}{n^{1/2}}\left(\mathbb{E}\left[\|V_n-\Gamma_m\|_F^{2p}\right]\right)^{\frac{2p-1}{2p}}.\end{aligned}$$

Finally,

$$\begin{aligned}2C'p\gamma_n\frac{K_p^{\frac{1}{2p}}}{n^{1/2}}\left(\mathbb{E}\left[\|V_n-\Gamma_m\|_F^{2p}\right]\right)^{\frac{2p-1}{2p}} &\leq 2C'p\gamma_n\frac{K_p^{\frac{1}{2p}}}{n^{1/2}}\max\left\{1,\mathbb{E}\left[\|V_n-\Gamma_m\|_F^{2p}\right]\right\}\\ &\leq 2C'p\gamma_n\frac{K_p^{\frac{1}{2p}}}{n^{1/2}}\left(1+\mathbb{E}\left[\|V_n-\Gamma_m\|_F^{2p}\right]\right).\end{aligned}$$

Thus, there are positive constants A_0'', A_1'' such that

$$(**) \leq \left(1 + A_0'' \frac{1}{n^{\alpha+1/2}}\right) \mathbb{E}\left[\|V_n - \Gamma_m\|_F^{2p}\right] + A_1'' \frac{1}{n^{\alpha+1/2}}. \tag{C.17}$$

Finally, thanks to inequalities (C.16) and (C.17), there are positive constants A_0', A_1' such that

$$\begin{aligned}
\mathbb{E}\left[\|V_{n+1} - \Gamma_m\|_F^{2p}\right] &\leq \left(1 + A_0' \frac{1}{n^{\alpha+1/2}}\right) \mathbb{E}\left[\|V_n - \Gamma_m\|_F^{2p}\right] + A_1' \frac{1}{n^{\alpha+1/2}} \\
&\leq \prod_{k=1}^{n} \left(1 + A_0' \frac{1}{k^{\alpha+1/2}}\right) \mathbb{E}\left[\|V_1 - \Gamma_m\|_F^{2p}\right] \\
&+ \sum_{k=1}^{n} \prod_{j=k+1}^{n} \left(1 + A_0' \frac{1}{j^{\alpha+1/2}}\right) A_1' \frac{1}{k^{\alpha+1/2}} \\
&\leq \prod_{k=1}^{\infty} \left(1 + A_0' \frac{1}{k^{\alpha+1/2}}\right) \mathbb{E}\left[\|V_1 - \Gamma_m\|_F^{2p}\right] \\
&+ \prod_{j=1}^{\infty} \left(1 + A_0' \frac{1}{j^{\alpha+1/2}}\right) \sum_{k=1}^{\infty} A_1' \frac{1}{k^{\alpha+1/2}} \\
&\leq M_p,
\end{aligned}$$

which concludes the induction and the proof.

□

Proof of Lemma 6.5.3. Let us define the following linear operators :

$$\begin{aligned}
\alpha_n &:= I_{\mathcal{S}(H)} - \gamma_n \nabla_m^2 G(\Gamma_m), \\
\beta_n &:= \prod_{k=1}^{n} \alpha_k = \prod_{k=1}^{n} \left(I_{\mathcal{S}(H)} - \gamma_k \nabla_m^2 G(\Gamma_m)\right), \\
\beta_0 &:= I_{\mathcal{S}(H)}.
\end{aligned}$$

Using decomposition (C.5) and by induction, for all $n \geq 1$,

$$V_n - \Gamma_m = \beta_{n-1}\left(V_1 - \Gamma_m\right) + \beta_{n-1} M_n - \beta_{n-1} R_n - \beta_{n-1} R_n' - \beta_{n-1} \Delta_n, \tag{C.18}$$

with

$$M_n := \sum_{k=1}^{n-1} \gamma_k \beta_k^{-1} \xi_{k+1}, \qquad R_n := \sum_{k=1}^{n-1} \gamma_k \beta_k^{-1} r_k,$$
$$R'_n := \sum_{k=1}^{n-1} \gamma_k \beta_k^{-1} r'_k, \qquad \Delta_n := \sum_{k=1}^{n-1} \gamma_k \beta_k^{-1} \delta_k.$$

We now study the asymptotic behavior of the linear operators β_n and $\beta_{n-1}\beta_k^{-1}$. As in [CCZ13], one can check that there are positive constants c_0, c_1 such that for all integers $k, n \geq 1$ with $k \leq n-1$,

$$\|\beta_{n-1}\|_{op} \leq c_0 e^{-\lambda_{\min} \sum_{k=1}^{n} \gamma_n}, \qquad \left\|\beta_{n-1}\beta_k^{-1}\right\|_{op} \leq c_1 e^{-\lambda_{\min} \sum_{j=k}^{n} \gamma_j}, \tag{C.19}$$

where $\|.\|_{op}$ is the usual spectral norm for linear operators. We now bound the quadratic mean of each term in decomposition (C.18).

Step 1 : the quasi deterministic term $\beta_{n-1}(V_1 - \Gamma_m)$.
Applying inequality (C.19), there is a positive constant c'_0 such that

$$\begin{aligned}
\mathbb{E}\left[\|\beta_{n-1}(V_1 - \Gamma_m)\|_F^2\right] &\leq \|\beta_{n-1}\|_{op}^2 \, \mathbb{E}\left[\|V_1 - \Gamma_m\|_F^2\right] \\
&\leq c_0 e^{-2\lambda_{\min} \sum_{k=1}^{n} \gamma_n} \mathbb{E}\left[\|V_1 - \Gamma_m\|_F^2\right] \\
&\leq c_0 e^{-c'_0 n^{1-\alpha}} \mathbb{E}\left[\|V_1 - \Gamma_m\|_F^2\right].
\end{aligned} \tag{C.20}$$

This term converges exponentially fast to 0.

Step 2 : the martingale term $\beta_{n-1}M_n$.

Since (ξ_n) is a sequence of martingale differences adapted to the filtration $(\mathcal{F}_n)$,

$$\begin{aligned}
\mathbb{E}\left[\|\beta_{n-1}M_n\|_F^2\right] &= \sum_{k=1}^{n-1}\mathbb{E}\left[\left\|\beta_{n-1}\beta_k^{-1}\gamma_k\xi_{k+1}\right\|_F^2\right] \\
&\quad + 2\sum_{k=1}^{n-1}\sum_{k'=k+1}^{n-1}\gamma_k\gamma_{k'}\mathbb{E}\left[\left\langle\beta_{n-1}\beta_k^{-1}\xi_{k+1},\beta_{n-1}\beta_{k'}^{-1}\xi_{k'+1}\right\rangle_F\right] \\
&= \sum_{k=1}^{n-1}\mathbb{E}\left[\left\|\beta_{n-1}\beta_k^{-1}\gamma_k\xi_{k+1}\right\|_F^2\right] \\
&\quad + 2\sum_{k=1}^{n-1}\sum_{k'=k+1}^{n-1}\gamma_k\gamma_{k'}\mathbb{E}\left[\left\langle\beta_{n-1}\beta_k^{-1}\xi_{k+1},\beta_{n-1}\beta_{k'}^{-1}\mathbb{E}\left[\xi_{k'+1}|\mathcal{F}_{k'}\right]\right\rangle_F\right] \\
&= \sum_{k=1}^{n-1}\mathbb{E}\left[\left\|\beta_{n-1}\beta_k^{-1}\gamma_k\xi_{k+1}\right\|_F^2\right].
\end{aligned}$$

Moreover, as in [CCGB15], Lemma C.5.2 ensures that there is a positive constant C_1' such that for all $n \geq 1$,

$$\mathbb{E}\left[\|\beta_{n-1}M_n\|_F^2\right] \leq \frac{C_1'}{n^\alpha}. \tag{C.21}$$

Step 3 : the first remainder term $\beta_{n-1}R_n$.
Remarking that $\|r_n\|_F \leq 4\left(\sqrt{C}+C\sqrt{\|\Gamma_m\|_F}\right)\|\overline{m}_n - m\|$,

$$\begin{aligned}
\mathbb{E}\left[\|\beta_{n-1}R_n\|_F^2\right] &\leq \mathbb{E}\left[\left(\sum_{k=1}^{n-1}\gamma_k\left\|\beta_{n-1}\beta_k^{-1}\right\|_{op}\|r_k\|_F\right)^2\right] \\
&\leq 16\left(\sqrt{C}+\sqrt{\|\Gamma_m\|_F}\right)^2\mathbb{E}\left[\left(\sum_{k=1}^{n-1}\gamma_k\left\|\beta_{n-1}\beta_k^{-1}\right\|_{op}\|\overline{m}_k - m\|\right)^2\right].
\end{aligned}$$

Applying Lemma 4.3 and Theorem 4.2 in [GB15],

$$\begin{aligned}
\mathbb{E}\left[\|\beta_{n-1}R_n\|_F^2\right] &\leq 16\left(\sqrt{C}+C\sqrt{\|\Gamma_m\|_F}\right)^2\left(\sum_{k=1}^{n-1}\gamma_k\left\|\beta_{n-1}\beta_k^{-1}\right\|_{op}\sqrt{\mathbb{E}\left[\|\overline{m}_k - m\|^2\right]}\right)^2 \\
&\leq 16\left(\sqrt{C}+C\sqrt{\|\Gamma_m\|_F}\right)^2 K_1\left(\sum_{k=1}^{n-1}\gamma_k\left\|\beta_{n-1}\beta_k^{-1}\right\|_{op}\frac{1}{k^{1/2}}\right)^2.
\end{aligned}$$

Applying inequality (C.19),

$$\mathbb{E}\left[\left\|\beta_{n-1}R_n\right\|_F^2\right] \leq 16\left(\sqrt{C}+C\sqrt{\Gamma_m}\right)^2 K_1\left(\sum_{k=1}^{n-1}\gamma_k e^{-\sum_{j=k}^n\gamma_j}\frac{1}{k^{1/2}}\right)^2$$
$$\leq 16\left(\sqrt{C}+C\sqrt{\Gamma_m}\right)^2 K_1\left(\sum_{k=1}^{n}\gamma_k e^{-\sum_{j=k}^n\gamma_j}\frac{1}{k^{1/2}}\right)^2.$$

Splitting the sum into two parts and applying Lemma C.5.2, we have

$$\mathbb{E}\left[\left\|\beta_{n-1}R_n\right\|_F^2\right] \leq 32\left(\sqrt{C}+C\sqrt{\left\|\Gamma_m\right\|_F}\right)^2 K_1\left(\sum_{k=1}^{E(n/2)}\gamma_k e^{-\sum_{j=k}^n\gamma_j}\frac{1}{k^{1/2}}\right)^2$$
$$+32\left(\sqrt{C}+C\sqrt{\left\|\Gamma_m\right\|_F}\right)^2 K_1\left(\sum_{k=E(n/2)+1}^{n}\gamma_k e^{-\sum_{j=k}^n\gamma_j}\frac{1}{k^{1/2}}\right)^2$$
$$=O\left(\frac{1}{n}\right).$$

Thus, there is a positive constant C_2' such that for all $n \geq 1$,

$$\mathbb{E}\left[\left\|\beta_{n-1}R_n\right\|_F^2\right] \leq \frac{C_2'}{n}. \tag{C.22}$$

Step 4 : the second remainder term $\beta_{n-1}R_n'$.
Let us recall that for all $n \geq 1$, $\left\|r_n'\right\|_F \leq 12D\left\|\overline{m}_n - m\right\|\left\|V_n - \Gamma_m\right\|_F$ with $D := C\sqrt{\left\|\Gamma_m\right\|_F} + C^{3/4}$. Thus,

$$\mathbb{E}\left[\left\|\beta_{n-1}R_n'\right\|_F^2\right] \leq \mathbb{E}\left[\left(\sum_{k=1}^{n-1}\gamma_k\left\|\beta_{n-1}\beta_k^{-1}\right\|_{op}\left\|r_k'\right\|_F\right)^2\right]$$
$$\leq 144D^2\mathbb{E}\left[\left(\sum_{k=1}^{n-1}\gamma_k\left\|\beta_{n-1}\beta_k^{-1}\right\|_{op}\left\|\overline{m}_k - m\right\|\left\|V_k - \Gamma_m\right\|_F\right)^2\right].$$

Applying Lemma 4.3 in [GB15],

$$\mathbb{E}\left[\left\|\beta_{n-1}R_n'\right\|_F^2\right] \leq 144D^2\left(\sum_{k=1}^{n-1}\gamma_k\left\|\beta_{n-1}\beta_k^{-1}\right\|_{op}\sqrt{\mathbb{E}\left[\left\|\overline{m}_k - m\right\|^2\left\|V_k - \Gamma_m\right\|_F^2\right]}\right)^2.$$

Thanks to Lemma 6.5.2, there is a positive constant M_2 such that for all $n \geq 1$, $\mathbb{E}\left[\left\|V_n - \Gamma_m\right\|_F^4\right] \leq M_2$.

Thus, applying Cauchy-Schwarz's inequality and Theorem 4.2 in [GB15],

$$\begin{aligned}\mathbb{E}\left[\left\|\beta_{n-1}R_n'\right\|_F^2\right] &\leq 144D^2\left(\sum_{k=1}^{n-1}\gamma_k\left\|\beta_{n-1}\beta_k^{-1}\right\|_{op}\left(\mathbb{E}\left[\|\overline{m}_k-m\|^4\right]\right)^{\frac{1}{4}}\left(\mathbb{E}\left[\|V_k-\Gamma_m\|_F^4\right]\right)^{\frac{1}{4}}\right)^2\\ &\leq 144D^2\sqrt{M_2K_2}\left(\sum_{k=1}^{n-1}\gamma_k\left\|\beta_{n-1}\beta_k^{-1}\right\|_{op}\frac{1}{k^{1/2}}\right)^2.\end{aligned}$$

As in step 3, splitting the sum into two parts, one can check that there is a positive constant C_1'' such that for all $n \geq 1$,

$$\mathbb{E}\left[\left\|\beta_{n-1}R_n'\right\|_F^2\right] \leq \frac{C_1''}{n}. \tag{C.23}$$

Step 5 : the third remainder term : $\beta_{n-1}\Delta_n$
Since $\|\delta_n\|_F \leq 6C\|V_n-\Gamma_m\|_F^2$, applying Lemma 4.3 in [GB15],

$$\begin{aligned}\mathbb{E}\left[\|\beta_{n-1}\Delta_n\|_F^2\right] &\leq \mathbb{E}\left[\left(\sum_{k=1}^{n-1}\gamma_k\left\|\beta_{n-1}\beta_k^{-1}\right\|_{op}\|\delta_k\|_F\right)^2\right]\\ &\leq 36C^2\mathbb{E}\left[\left(\sum_{k=1}^{n-1}\gamma_k\left\|\beta_{n-1}\beta_k^{-1}\right\|_{op}\|V_k-\Gamma_m\|_F^2\right)^2\right]\\ &\leq 36C^2\left(\sum_{k=1}^{n-1}\gamma_k\left\|\beta_{n-1}\beta_k^{-1}\right\|_{op}\sqrt{\mathbb{E}\left[\|V_k-\Gamma_m\|_F^4\right]}\right)^2.\end{aligned}$$

Thanks to Lemma 6.5.2, there is a positive constant M_2 such that for all $n \geq 1$, $\mathbb{E}\left[\|V_n-\Gamma_m\|_F^4\right] \leq M_2$. Thus, splitting the sum into two parts and applying inequalities (C.19) and Lemma C.5.2, there are positive constant c_0', C_2' such that for all $n \geq 1$,

$$\begin{aligned}\mathbb{E}\left[\|\beta_{n-1}\Delta_n\|_F^2\right] &\leq 72C^2M_2^2\left(\sum_{k=1}^{E(n/2)}\gamma_k e^{-\sum_{j=k}^n\gamma_j}\right)^2\\ &\quad+72C^2\sup_{E(n/2)+1\leq k\leq n-1}\left\{\mathbb{E}\left[\|V_k-\Gamma_m\|_F^4\right]\right\}\left(\sum_{k=E(n/2)+1}^{n}\gamma_k e^{-\sum_{j=k}^n\gamma_j}\right)^2\\ &\leq C_2'\sup_{E(n/2)+1\leq k\leq n-1}\left\{\mathbb{E}\left[\|V_k-\Gamma_m\|_F^4\right]\right\}+O\left(e^{-2c_0'n^{1-\alpha}}\right).\end{aligned}$$

Thus, there is a positive constant C_0' such that for all $n \geq 1$,

$$\mathbb{E}\left[\|\beta_{n-1}\Delta_n\|_F^2\right] \leq C_0'e^{-2c_0'n^{1-\alpha}}+C_2'\sup_{E(n/2)+1\leq k\leq n-1}\left\{\mathbb{E}\left[\|V_k-\Gamma_m\|_F^4\right]\right\}. \tag{C.24}$$

Conclusion :
Applying Lemma C.5.1 and decomposition (C.18), for all $n \geq 1$,

$$\begin{aligned}\mathbb{E}\left[\|V_n - \Gamma_m\|_F^2\right] \leq 5\mathbb{E}\left[\|\beta_{n-1}(V_1 - \Gamma_m)\|_F^2\right] + 5\mathbb{E}\left[\|\beta_{n-1}M_n\|_F^2\right] + 5\mathbb{E}\left[\|\beta_{n-1}R_n\|_F^2\right] \\ + 5\mathbb{E}\left[\left\|\beta_{n-1}R'_n\right\|_F^2\right] + 5\mathbb{E}\left[\|\beta_{n-1}\Delta_n\|_F^2\right].\end{aligned}$$

Applying inequalities (C.20) to (C.24), there are positive constants C_1, C_1', C_2, C_3 such that for all $n \geq 1$,

$$\mathbb{E}\left[\|V_n - \Gamma_m\|^2\right] \leq C_1 e^{-C_1' n^{1-\alpha}} + \frac{C_2}{n^\alpha} + C_3 \sup_{E(n/2)+1 \leq k \leq n-1} \mathbb{E}\left[\|V_k - \Gamma_m\|_F^4\right].$$

□

Proof of Lemma 6.5.4. Let us define $W_n := V_n - \Gamma_m - \gamma_n\left(\nabla G_{\overline{m}_n}(V_n) - \nabla G_{\overline{m}_n}(\Gamma_m)\right)$ and use decomposition (C.4),

$$\begin{aligned}\|V_{n+1} - \Gamma_m\|_F^2 = \|W_n\|_F^2 + \gamma_n^2\|\xi_{n+1}\|_F^2 + \gamma_n^2\|r_n\|_F^2 + 2\gamma_n\langle\xi_{n+1}, V_n - \Gamma_m\rangle_F + 2\gamma_n^2\langle\xi_{n+1}, \nabla G_{\overline{m}_n}(V_n)\rangle_F \\ - 2\gamma_n^2\langle r_n, \nabla G_{\overline{m}_n}(V_n) - \nabla G_{\overline{m}_n}(\Gamma_m)\rangle_F - 2\gamma_n\langle r_n, V_n - \Gamma_m\rangle_F.\end{aligned}$$

Since $\|\xi_{n+1}\|_F \leq 2$, $\|r_n\|_F \leq 2$ and the fact that for all $h \in H$, $V \in \mathcal{S}(H)$, $\nabla_h G(V) \leq 1$, we get with an application of Cauchy-Schwarz's inequality

$$\|V_{n+1} - \Gamma_m\|_F^2 \leq \|W_n\|_F^2 + 2\gamma_n\langle\xi_{n+1}, V_n - \Gamma_m\rangle_F + 2\gamma_n\|r_n\|_F\|V_n - \Gamma_m\|_F + 20\gamma_n^2.$$

Thus, since (ξ_n) is a sequence of martingale differences adapted to the filtration $(\mathcal{F}_n)$, and since $\|W_n\|_F^2 \leq \left(1 + C^2c_\gamma^2\right)\|V_n - \Gamma_m\|_F^2$ (this inequality follows from Proposition C.2.1 and from the fact that for all $h \in H$, G_h is a convex application),

$$\begin{aligned}\mathbb{E}\left[\|V_{n+1} - \Gamma_m\|_F^4\right] \leq \mathbb{E}\left[\|W_n\|_F^4\right] + 2\gamma_n\mathbb{E}\left[\|r_n\|_F\|W_n\|_F^2\|V_n - \Gamma_m\|_F\right] \\ + 40\left(1 + C^2c_\gamma^2\right)\gamma_n^2\mathbb{E}\left[\|V_n - \Gamma_m\|_F^2\right] \\ + 4\gamma_n^2\mathbb{E}\left[\langle\xi_{n+1}, V_n - \Gamma_m\rangle_F^2\right] + 400\gamma_n^4 + 40\gamma_n^3\mathbb{E}\left[\|r_n\|_F\|V_n - \Gamma_m\|_F^2\right] \\ + 4\gamma_n^2\mathbb{E}\left[\|r_n\|_F^2\|V_n - \Gamma_m\|_F^2\right].\end{aligned}$$

Since $\|\xi_{n+1}\|_F \leq 2$ and $\|r_n\|_F \leq 2$, applying Cauchy-Schwarz's inequality, there are positive

constants C_1', C_2' such that for all $n \geq 1$,

$$\mathbb{E}\left[\|V_{n+1} - \Gamma_m\|_F^4\right] \leq \mathbb{E}\left[\|W_n\|_F^4\right] + 2\gamma_n \mathbb{E}\left[\|r_n\|_F \|W_n\|_F^2 \|V_n - \Gamma_m\|_F\right] + \frac{C_1'}{n^{3\alpha}} + \frac{C_2'}{n^{2\alpha}} \mathbb{E}\left[\|V_n - \Gamma_m\|_F^2\right]. \tag{C.25}$$

We now bound the two first terms at the right-hand side of inequality (C.25).

Step 1 : bounding $\mathbb{E}\left[\|W_n\|_F^4\right]$.
Since $\nabla G_{\overline{m}_n}(V_n) - \nabla G_{\overline{m}_n}(\Gamma_m) = \int_0^1 \nabla^2_{\overline{m}_n} G\left(\Gamma_m + t\left(V_n - \Gamma_m\right)\right)\left(V_n - \Gamma_m\right) dt$, applying Proposition C.2.1, one can check that

$$\begin{aligned}\|W_n\|^2 &= \|V_n - \Gamma_m\|_F^2 - 2\gamma_n \left\langle V_n - \Gamma_m, \nabla G_{\overline{m}_n}(V_n) - \nabla G_{\overline{m}_n}(\Gamma_m)\right\rangle_H + \gamma_n^2 \|\nabla G_{\overline{m}_n}(V_n) - \nabla G_{\overline{m}_n}(\Gamma_m)\|_F^2 \\ &\leq \left(1 + C^2\gamma_n^2\right)\|V_n - \Gamma_m\|_F^2 - 2\gamma_n \left\langle V_n - \Gamma_m, \nabla G_{\overline{m}_n}(V_n) - \nabla G_{\overline{m}_n}(\Gamma_m)\right\rangle_H.\end{aligned}$$

Since for all $h \in H$, G_h is a convex application, $\|W_n\|_F^2 \leq \left(1 + c_\gamma^2 C^2\right)\|V_n - \Gamma_m\|_F^2$. Let p' be a positive integer. We now introduce the sequence of events $\left(A_{n,p'}\right)_{n\in\mathbb{N}}$ defined for all $n \geq 1$ by

$$A_{n,p'} := \left\{\omega \in \Omega, \quad \|V_n(\omega) - \Gamma_m\|_F \leq n^{\frac{1-\alpha}{p'}}, \quad \text{and} \quad \|\overline{m}_n(\omega) - m\| \leq \epsilon\right\}, \tag{C.26}$$

with ϵ defined in Proposition C.2.1. For the sake of simplicity, we consider that ϵ' defined in Proposition C.2.1 verifies $\epsilon' \leq 1$. Applying Proposition C.2.1, let

$$\begin{aligned}B_n :&= \left\langle \nabla G_{\overline{m}_n}(V_n) - \nabla G_{\overline{m}_n}(\Gamma_m), V_n - \Gamma_m\right\rangle_F \mathbf{1}_{A_{n,p'}} \mathbf{1}_{\{\|V_n - \Gamma_m\|_F \leq \epsilon'\}} \\ &= \int_0^1 \left\langle \nabla^2_{\overline{m}_n} G\left(\Gamma_m + t\left(V_n - \Gamma_m\right)\right)\left(V_n - \Gamma_m\right), V_n - \Gamma_m\right\rangle_F \mathbf{1}_{\{\|V_n - \Gamma_m\|_F \leq \epsilon'\}} \mathbf{1}_{A_{n,p'}} dt \\ &\geq \frac{1}{2} c_m \|V_n - \Gamma_m\|_F^2 \mathbf{1}_{\{\|V_n - \Gamma_m\|_F \leq \epsilon'\}} \mathbf{1}_{A_{n,p'}}.\end{aligned} \tag{C.27}$$

In the same way, since $G_{\overline{m}_n}$ is convex, let

$$\begin{aligned}B_n' :&= \left\langle \nabla G_{\overline{m}_n}(V_n) - \nabla G_{\overline{m}_n}(\Gamma_m), V_n - \Gamma_m\right\rangle_F \mathbf{1}_{A_{n,p'}} \mathbf{1}_{\{\|V_n - \Gamma_m\|_F > \epsilon'\}} \\ &= \int_0^1 \left\langle \nabla^2_{\overline{m}_n}\left(\Gamma_m + t\left(V_n - \Gamma_m\right)\right)\left(V_n - \Gamma_m\right), V_n - \Gamma_m\right\rangle \mathbf{1}_{\{\|V_n - \Gamma_m\|_F > \epsilon'\}} \mathbf{1}_{A_{n,p'}} dt \\ &\geq \int_0^{\frac{\epsilon'}{\|V_n - \Gamma_m\|_F}} \left\langle \nabla^2_{\overline{m}_n}\left(\Gamma_m + t\left(V_n - \Gamma_m\right)\right)\left(V_n - \Gamma_m\right), V_n - \Gamma_m\right\rangle \mathbf{1}_{\{\|V_n - \Gamma_m\|_F > \epsilon'\}} \mathbf{1}_{A_{n,p'}} dt\end{aligned}$$

Applying Proposition C.2.1,

$$\begin{aligned} B'_n &\geq \int_0^{\frac{\epsilon'}{\|V_n-\Gamma_m\|_F}} \frac{1}{2} c_m \|V_n - \Gamma_m\|_F^2 \mathbf{1}_{\{\|V_n-\Gamma_m\|_F>\epsilon'\}} \mathbf{1}_{A_{n,p'}} dt \\ &\geq \frac{\epsilon' c_m}{2\|V_n-\Gamma_m\|_F} \|V_n - \Gamma_m\|_F^2 \mathbf{1}_{\{\|V_n-\Gamma_m\|_F>\epsilon'\}} \mathbf{1}_{A_{n,p'}} \\ &\geq \frac{\epsilon' c_m}{2} n^{-\frac{1-\alpha}{p'}} \|V_n - \Gamma_m\|_F^2 \mathbf{1}_{\{\|V_n-\Gamma_m\|_F>\epsilon'\}} \mathbf{1}_{A_{n,p'}}. \end{aligned} \tag{C.28}$$

There is a rank $n'_{p'}$ such that for all $n \geq n'_{p'}$, we have $\frac{\epsilon' c_m}{2} n^{-\frac{1-\alpha}{p}} \leq \frac{1}{2} c_m$. Thus, applying inequalities (C.27) and (C.28), for all $n \geq n'_{p'}$,

$$\|W_n\|_F^2 \mathbf{1}_{A_{n,p'}} \leq \left(1 - \frac{\epsilon' c_m}{2} \gamma_n n^{-\frac{1-\alpha}{p'}}\right) \|V_n - \Gamma_m\|_F^2 \mathbf{1}_{A_{n,p'}}.$$

Thus, there are a positive constant $c_{p'}$ and a rank $n_{p'}$ such that for all $n \geq n_{p'}$,

$$\begin{aligned} \mathbb{E}\left[\|W_n\|_F^4 \mathbf{1}_{A_{n,p'}}\right] &\leq \left(1 - \frac{\epsilon' c_m}{2} \gamma_n n^{-\frac{1-\alpha}{p'}}\right)^2 \mathbb{E}\left[\|V_n - \Gamma_m\|_F^4 \mathbf{1}_{A_{n,p'}}\right] \\ &\leq \left(1 - 2c_{p'} \gamma_n n^{-\frac{1-\alpha}{p'}}\right) \mathbb{E}\left[\|V_n - \Gamma_m\|_F^4\right]. \end{aligned} \tag{C.29}$$

Now, we must get an upper bound for $\mathbb{E}\left[\|W_n\|_F^4 \mathbf{1}_{A^c_{n,p'}}\right]$. Since $\|W_n\|_F^2 \leq \left(1 + c_\gamma^2 C^2\right) \|V_n - \Gamma_m\|_F^2$ and since there is a positive constant c_0 such that for all $n \geq 1$,

$$\|V_n - \Gamma_m\|_F \leq \|V_1 - \Gamma_m\|_F + \sum_{k=1}^{n} \gamma_k \leq c_0 n^{1-\alpha}$$

we have

$$\begin{aligned} \mathbb{E}\left[\|W_n\|_F^4 \mathbf{1}_{A^c_{n,p'}}\right] &\leq \left(1 + c_\gamma^2 C^2\right)^2 \mathbb{E}\left[\|V_n - \Gamma_m\|_F^4 \mathbf{1}_{A^c_{n,p'}}\right] \\ &\leq \left(1 + c_\gamma^2 C^2\right)^2 c_0^4 n^{4-4\alpha} \mathbb{P}\left[A^c_{n,p'}\right] \\ &\leq \left(1 + c_\gamma^2 C^2\right)^2 c_0^4 n^{4-4\alpha} \left(\mathbb{P}\left[\|\overline{m}_n - m\| \geq \epsilon\right] + \mathbb{P}\left[\|V_n - \Gamma_m\|_F \geq n^{\frac{1-\alpha}{p'}}\right]\right). \end{aligned}$$

Applying Markov's inequality, Theorem 4.2 in [GB15] and Lemma 6.5.2,

$$\begin{aligned}\mathbb{E}\left[\|W_n\|_F^4 \mathbf{1}_{A^c_{n,p'}}\right] &\leq \left(1+c_\gamma^2 C^2\right)^2 c_0^4 n^{4-4\alpha}\left(\frac{\mathbb{E}\left[\|\overline{m}_n - m\|^{2p''}\right]}{\epsilon^{2p''}} + \frac{\mathbb{E}\left[\|V_n-\Gamma_m\|_F^{2q}\right]}{n^{2q\frac{1-\alpha}{p'}}}\right)\\ &\leq \frac{K_{p''}}{\epsilon^{2p''}}\left(1+c_\gamma^2 C^2\right)^2 c_0^4 n^{4-4\alpha-p''} + \left(1+c_\gamma^2 C^2\right)^2 c_0^4 M_q n^{4-4\alpha-2q\frac{1-\alpha}{p'}}.\end{aligned}$$

Taking $p'' \geq 4-\alpha$ and $q \geq p'\frac{4-\alpha}{2(1-\alpha)}$,

$$\mathbb{E}\left[\|W_n\|_F^4 \mathbf{1}_{A^c_{n,p'}}\right] = O\left(\frac{1}{n^{3\alpha}}\right). \tag{C.30}$$

Thus, applying inequalities (C.29) and (C.30), there are positive constants $c_{p'}$, $C_{1,p'}$ and a rank $n_{p'}$ such that for all $n \geq n_{p'}$,

$$\mathbb{E}\left[\|W_n\|_F^4\right] \leq \left(1-2c_{p'}\gamma_n n^{-\frac{1-\alpha}{p'}}\right)\mathbb{E}\left[\|V_n-\Gamma_m\|_F^4\right] + \frac{C_{1,p'}}{n^{3\alpha}}. \tag{C.31}$$

Step 2 : bounding $2\gamma_n \mathbb{E}\left[\|r_n\|_F \|W_n\|_F^2 \|V_n-\Gamma_m\|_F\right]$.
Since $\|W_n\|_F^2 \leq \left(1+c_\gamma^2 C^2\right)\|V_n-\Gamma_m\|_F^2$, applying Lemma C.5.1, let

$$\begin{aligned}D_n :&= 2\gamma_n \mathbb{E}\left[\|r_n\|_F \|W_n\|_F^2 \|V_n-\Gamma_m\|_F\right]\\ &\leq 2\left(1+c_\gamma^2 C^2\right)\gamma_n \mathbb{E}\left[\|r_n\|_F \|V_n-\Gamma_m\|_F^3\right]\\ &\leq \frac{2}{c_{p'}}\left(1+c_\gamma^2 C^2\right)^2 \gamma_n n^{\frac{1-\alpha}{p'}}\mathbb{E}\left[\|r_n\|_F^2\|V_n-\Gamma_m\|_F^2\right] + \frac{1}{2}c_{p'}\gamma_n n^{-\frac{1-\alpha}{p'}}\mathbb{E}\left[\|V_n-\Gamma_m\|_F^4\right]\\ &\leq \frac{2}{c_{p'}^2}\left(1+c_\gamma^2 C^2\right)^4 \gamma_n n^{3\frac{1-\alpha}{p'}}\mathbb{E}\left[\|r_n\|_F^4\right] + c_{p'}\gamma_n n^{-\frac{1-\alpha}{p'}}\mathbb{E}\left[\|V_n-\Gamma_m\|_F^4\right].\end{aligned}$$

Since $\|r_n\|_F \leq \left(\sqrt{C}+C\sqrt{\|\Gamma_m\|_F}\right)\|\overline{m}_n - m\|_F$ and applying Theorem 4.2 in [GB15],

$$\begin{aligned}D_n &\leq \frac{2}{c_{p'}^2}\left(1+c_\gamma^2 C^2\right)^4\left(\sqrt{C}+C\sqrt{\|\Gamma_m\|_F}\right)^4 \gamma_n n^{3\frac{1-\alpha}{p'}}\mathbb{E}\left[\|\overline{m}_n-m\|^4\right] + c_{p'}\gamma_n n^{-\frac{1-\alpha}{p'}}\mathbb{E}\left[\|V_n-\Gamma_m\|_F^4\right]\\ &\leq \frac{2}{c_{p'}^2}K_2\left(1+c_\gamma^2 C^2\right)^4\left(\sqrt{C}+C\sqrt{\|\Gamma_m\|_F}\right)^4 \gamma_n n^{3\frac{1-\alpha}{p'}}\frac{1}{n^2} + c_{p'}\gamma_n n^{-\frac{1-\alpha}{p'}}\mathbb{E}\left[\|V_n-\Gamma_m\|_F^4\right]\\ &= c_{p'}\gamma_n n^{-\frac{1-\alpha}{p'}}\mathbb{E}\left[\|V_n-\Gamma_m\|_F^4\right] + O\left(\frac{1}{n^{2+\alpha-3(1-\alpha)/p'}}\right).\end{aligned} \tag{C.32}$$

Step 3 : Conclusion.
Applying inequalities (C.25), (C.31) and (C.32), there are a rank $n_{p'}$ and positive constants $c_{p'}, C_{1,p'}, C_{2,p'}, C_{3,p'}$ such that for all $n \geq n_{p'}$,

$$\mathbb{E}\left[\|V_{n+1} - \Gamma_m\|_F^4\right] \leq \left(1 - c_{p'}\gamma_n n^{-\frac{1-\alpha}{p'}}\right)\mathbb{E}\left[\|V_n - \Gamma_m\|_F^4\right] + \frac{C_{1,p'}}{n^{3\alpha}} + \frac{C_{2,p'}}{n^{2\alpha}}\mathbb{E}\left[\|V_n - \Gamma_m\|_F^2\right] + \frac{C_{3,p'}}{n^{2+\alpha-3\frac{1-\alpha}{p'}}}.$$

□

C.5 Some technical inequalities

First, the following lemma recalls some well-known inequalities.

Lemma C.5.1. *Let a, b, c be positive constants. Then,*

$$ab \leq \frac{a^2}{2c} + \frac{b^2 c}{2},$$
$$a \leq \frac{c}{2} + \frac{a^2}{2c}.$$

Moreover, let k, p be positive integers and $a_1, ..., a_p$ be positive constants. Then,

$$\left(\sum_{j=1}^{p} a_j\right)^k \leq p^{k-1}\sum_{j=1}^{p} a_j^k.$$

The following lemma gives the asymptotic behavior for some specific sequences of descent steps.

Lemma C.5.2. *Let α, β be non-negative constants such that $0 < \alpha < 1$, and (u_n), (v_n) be two sequences defined for all $n \geq 1$ by*

$$u_n := \frac{c_u}{n^\alpha}, \qquad v_n := \frac{c_v}{n^\beta},$$

with $c_u, c_v > 0$. Thus, there is a positive constant c_0 such that for all $n \geq 1$,

$$\sum_{k=1}^{E(n/2)} e^{-\sum_{j=k}^{n} u_j} u_k v_k = O\left(e^{-c_0 n^{1-\alpha}}\right), \tag{C.33}$$

$$\sum_{k=E(n/2)+1}^{n} e^{-\sum_{j=k}^{n} u_j} u_k v_k = O\left(v_n\right), \tag{C.34}$$

where $E(.)$ is the integer part function.

Proof of Lemma C.5.2. We first prove inequality (C.33). For all $n \geq 1$,

$$\begin{aligned}\sum_{k=1}^{E(n/2)} e^{-\sum_{j=k}^{n} u_j} u_k v_k &= c_u c_v \sum_{k=1}^{E(n/2)} e^{-\sum_{j=k}^{n} u_j} \frac{1}{k^{\alpha+\beta}} \\ &\leq c_u c_v \sum_{k=1}^{E(n/2)} e^{-c_u \sum_{j=k}^{n} \frac{1}{j^{\alpha}}}.\end{aligned}$$

Moreover, for all $k \leq E(n/2)$,

$$\begin{aligned}c_u \sum_{j=k}^{n} \frac{1}{j^{\alpha}} &\geq c_u \frac{n}{2} \frac{1}{n^{\alpha}} \\ &\geq \frac{c_u}{2} n^{1-\alpha}.\end{aligned}$$

Thus,

$$\sum_{k=1}^{E(n/2)} e^{-\sum_{j=k}^{n} u_j} u_k v_k \leq c_u c_v n e^{-\frac{c_u}{2} n^{1-\alpha}}.$$

We now prove inequality (C.34). With the help of an integral test for convergence,

$$\begin{aligned}\sum_{j=k}^{n} u_j &= c_u \sum_{j=k}^{n} \frac{1}{j^{\alpha}} \\ &\geq c_u \int_{k}^{n+1} \frac{1}{t^{\alpha}} dt \\ &\geq \frac{c_u}{1-\alpha} \left((n+1)^{1-\alpha} - k^{-\alpha}\right).\end{aligned}$$

Thus,

$$\sum_{k=E(n/2)+1}^{n} e^{-\sum_{j=k}^{n} u_j} u_k v_k \leq c_u c_v e^{-(n+1)^{1-\alpha}} \sum_{k=E(n/2)+1}^{n} e^{k^{1-\alpha}} k^{-\alpha-\beta}$$

With the help of an integral test for convergence, there is a rank $n_{u,v}$ (for sake of simplicity,

we consider that $n_{u,v} = 1$) such that for all $n \geq n_{u,v}$,

$$\begin{aligned}\sum_{k=E(n/2)+1}^{n} e^{k^{1-\alpha}} k^{-\alpha-\beta} &\leq \int_{E(n/2)+1}^{n+1} e^{t^{1-\alpha}} t^{-\alpha-\beta} dt \\ &\leq \frac{1}{1-\alpha} \left[e^{t^{1-\alpha}} t^{-\beta} \right]_{E(n/2)+1}^{n} + \beta \int_{E(n/2)+1}^{n} e^{t^{1-\alpha}} t^{-1-\beta} dt \\ &= e^{(n+1)^{1-\alpha}(n+1)^{-\beta}} + o\left(\int_{E(n/2)+1}^{n+1} e^{t^{1-\alpha}} t^{-\alpha-\beta} dt \right),\end{aligned}$$

since $\alpha < 1$. Thus,

$$\sum_{k=E(n/2)+1}^{n} e^{k^{1-\alpha}} k^{-\alpha-\beta} = O\left(e^{n^{1-\alpha} n^{-\beta}} \right).$$

As a conclusion, we have

$$\begin{aligned}\sum_{k=E(n/2)+1}^{n} e^{-\sum_{j=k}^{n} u_j} u_k v_k &= O\left(e^{-(n+1)^{1-\alpha} + n^{1-\alpha}} v_n \right) \\ &= O\left(v_n \right).\end{aligned}$$

□

Troisième partie

Vitesse de convergence des algorithmes de Robbins-Monro et de leur moyenné

Chapitre 7

L^p and almost sure rates of convergence of averaged stochastic gradient algorithms and applications to robust statistics

Résumé

On a rappelé au Chapitre 1 différents cadres de travail pour obtenir les vitesses de convergence des algorithmes de gradient stochastiques moyennés. On a notamment présenté le cadre introduit par [Pel98] et [Pel00], pour lequel les vitesses de convergence presque sûre de ces algorithmes sont établies. Cependant, ces résultats n'avaient été démontré que dans le cas d'espaces de dimension finie. De plus, on a présenté au Chapitre 1 le cadre introduit par [BM13], qui permet d'obtenir la vitesse de convergence en moyenne quadratique des algorithmes dans des espaces de Hilbert, mais ce, avec des hypothèses très restrictives sur la fonction que l'on veut minimiser.

Dans ce contexte, en s'appuyant sur les méthodes de démonstration mises en place au Chapitre 4 et améliorées aux Chapitres 5 et 6, on donne un cadre de travail dans les espaces de Hilbert, moins restrictif que ceux introduits par [BM13] et [Bac14], qui nous permet d'établir les vitesses de convergence presque sûre des algorithmes (Théorème 7.3.2) ainsi que leurs vitesses L^p (Théorèmes 7.3.3 et 7.3.4)

Abstract

It is more and more usual to deal with large samples taking values in high dimensional spaces such that functional spaces. Moreover, one usual stochastic optimization problem is to minimize a convex function depending on a random variable. In this context, Robbins-Monro algorithms and their averaged version are good candidate to approximate the solution of these kind of problems. Indeed, they usually do not need too much computational efforts, do not need to store all the data, which is crucial when we deal with big data, and can be simply updated, which is interesting when the data arrive sequentially. The aim of this work is to give a general framework which is sufficient to get asymptotic and non asymptotic rates of convergence of stochastic gradient algorithms as well as of their averaged version.

7.1 Introduction

With the development of automatic sensors, it is more and more usual to deal with large samples of observations taking values in high dimensional spaces such as functional spaces. In this context, it may be possible that usual methods in stochastic approximation encounter some computational issues due to the large dimension and size of the data. One usual stochastic convex optimization problem is to minimize the following kind of function

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

where H is a Hilbert space and X is a random variable taking value in a space $\mathcal{X}$ and $g : \mathcal{X} \times H \longrightarrow \mathbb{R}$. Moreover, the functional $G : H \longrightarrow \mathbb{R}$ is convex. Many examples exist in the literature such as the L^1-median (see [Kem87] or [CCZ13] among others), geometric quantiles (see [Cha96]) or several regressions (see [BM13]). One usual method to approximate the minimizer of this kind of function, given a sample $X_1, ..., X_n$, is to consider the empirical problem generated by the sample, i.e to consider the M-estimate (see the books of [HR09] and [MMY06] among others)

$$\widehat{m}_n := \arg\min_h \frac{1}{n}\sum_{k=1}^{n} g\left(X_k, h\right),$$

and to estimate $\widehat{m}_n$ using usual deterministic optimization methods (see [VZ00] and [BS15] for the special case of the median). Nevertheless, one of the most important problem of this method is that it requires to store all the data, which can be exhaustive if we deal with large samples taking values in high dimensional spaces. Thus, in order to overcome this, stochastic gradient algorithms introduced by [RM51] are efficient candidates. Indeed, commonly, they do not need too much computational efforts, do not require to store all the data and can be simply updated, which represents a real interest when the data arrive sequentially.

The literature is very large on this domain (see the books of [Duf97], [KY03] among others) and a method to improve their convergence, which consists in averaging the Robbins-Monro estimators, was introduced by [PJ92]. Many asymptotic results exist (see [Duf97], [Pel98], or [Pel00] for instance) but they often depend on the dimension of the space and the proofs can not be directly adapted for infinite dimensional spaces such as functional spaces.

Moreover, an asymptotic result such as a Central Limit Theorem does not give any clue of how far the distribution of the estimator is from its asymptotic law for a fixed sample size n. Through non asymptotic results, the aim is to obtain finite sample guarantees with high probability, which is always desirable for statisticians who deal with real data. Nice argu-

ments for considering non asymptotic properties are given in [Rud14], for example. Note that the obtainment of such results often requires much more efforts and assumptions compare to classical weak convergence results. Moreover, as far as we know, there are only few results in the literature on non asymptotic rates of convergence. However, in recent works, [CCGB15] and [GB15] give a deep non asymptotic study of the estimators of the median, giving non asymptotic confidence balls as well as general L^p rates of convergence. In the same way, [BM13] and [Bac14] give some general conditions to get the rate of convergence in quadratic mean of averaged stochastic gradient algorithms.

The aim of this work is, in a first time, to give assumptions which will enable us to get asymptotic results such as almost sure rates of convergence of the Robbins-Monro algorithm as well as of its averaged version in general Hilbert spaces. Nevertheless, as mentioned above, asymptotic results are often non sufficient, and we propose to generalize the method introduced by [CCGB15] and improved by [GB15] and [CGB15] to get the L^p rates of convergence of the algorithms. These assumptions consist in assuming the local strong convexity of the function we would like to minimize, and in having a control on the loss of this strong convexity. This allows, in a first time, to obtain the L^p rates of convergence of the Robbins-Monro algorithm. In a second time, this enables us to get the rates of convergence of the averaged algorithm.

The paper is organized as follows. Section 7.2 introduces the framework, assumptions, the algorithms and some convexity properties of the function we would like to minimize are given. In Section 7.3, the strong efficiency of the algorithms will be given as well as their almost sure and L^p rates of convergence. Two examples of application are given in Section 7.4. First, we will be interested in estimating geometric quantiles in Hilbert spaces, which are a generalization of the real quantiles introduced by [Cha96]. An important fact is that they are robust indicators which can be useful in statistical depth and outliers detection (see [Ser06], [CDPB09] or [HP06]). In a second time, these algorithms and results can be applied in several regressions an we will focus on an example of robust logistic regression. Finally, the proofs are postponed in Section 7.5 and in a supplementary file.

7.2 The algorithms and assumptions

7.2.1 Assumptions and general framework

Let H be a separable Hilbert space such as $\mathbb{R}^d$ or $L^2(I)$ for some closed interval $I \subset \mathbb{R}$. We denote by $\langle .,. \rangle$ its inner product and by $\|.\|$ the associated norm. Let X be a random variable taking values in a space $\mathcal{X}$, and let $G : H \longrightarrow \mathbb{R}$ be the function we would like to minimize,

defined for all $h \in H$ by

$$G(h) := \mathbb{E}\left[g(X,h)\right], \tag{7.1}$$

where $g : \mathcal{X} \times H \longrightarrow \mathbb{R}$. Moreover, let us suppose that the functional G is convex. We consider from now that the following assumptions are fulfilled :

(A1) The functional g is Frechet-differentiable for the second variable almost everywhere. Moreover, G is differentiable and denoting by Φ its gradient, there exists $m \in H$ such that

$$\Phi(m) := \nabla G(m) = 0.$$

(A2) The functional G is twice continuously differentiable almost everywhere and for all positive constant A, there is a positive constant C_A such that for all $h \in \mathcal{B}(m, A)$,

$$\|\Gamma_h\|_{op} \leq C_A,$$

where Γ_h is the Hessian of the functional G at h and $\|.\|_{op}$ is the usual spectral norm for linear operators.

(A3) There exists a positive constant ϵ such that for all $h \in \mathcal{B}(m, \epsilon)$, there is a basis of H composed of eigenvectors of Γ_h. Moreover, let us denote by $\lambda_{\min}$ the limit inf of the eigenvalues of Γ_m, then $\lambda_{\min}$ is positive. Finally, for all $h \in \mathcal{B}(m, \epsilon)$, and for all eigenvalue λ_h of Γ_h, we have $\lambda_h \geq \frac{\lambda_{\min}}{2} > 0$.

(A4) There are positive constants ϵ, C_ϵ such that for all $h \in \mathcal{B}(m, \epsilon)$,

$$\|\Phi(h) - \Gamma_m(h-m)\| \leq C_\epsilon \|h-m\|^2.$$

(A5) Let $f : \mathcal{X} \times H \longrightarrow \mathbb{R}_+$ and let C be a positive constant such that for almost every $x \in \mathcal{X}$ and for all $h \in H$, $\|\nabla_h g(x,h)\| \leq f(x,h) + C\|h-m\|$ almost surely, and

(a) There is a positive constant L_1 such that for all $h \in H$,

$$\mathbb{E}\left[f(X,h)^2\right] \leq L_1.$$

(b) For all integer q, there is a positive constant L_q such that for all $h \in H$,

$$\mathbb{E}\left[f(X,h)^{2q}\right] \leq L_q.$$

Note that for the sake of simplicity, we often denote by the same way the different constants. We now make some comments on the assumptions. First, note that no convexity assumption on the function g is required.

Assumption **(A1)** is crucial to introduce a stochastic gradient algorithm. Moreover, note that one can relate to deterministic optimization's literature to ensure the existence of a minimizer of the functional G or equivalently to ensure the existence of a zero of the gradient.

Assumptions **(A2)** and **(A3)** give some properties on the spectrum of the Hessian and ensure that the functional G is locally strongly convex. Note that assumption **(A3)** can be resumed as $\lambda_{\min}(\Gamma_m) > 0$, where $\lambda_{\min}(.)$ is the function which gives the smallest eigenvalue (or the lim inf of the eigenvalues in infinite dimensional spaces) of a linear operator, if the functional $h \mapsto \lambda_{\min}(\Gamma_h)$ is continuous on a neighborhood of m. Note that in a space of finite dimension, assumption **(A3)** reduces to an assumption on the existence of the Hessian almost everywhere if the functional $h \longmapsto \Gamma_h$ is continuous.

Moreover, assumption **(A4)** allows to bound the remainder term in the Taylor's expansion of the gradient. Note that since the functional G is twice continuously differentiable and since $\Phi(m) = 0$, it comes $\Phi(h) = \int_0^1 \Gamma_{m+t(h-m)}(h-m)dt$, and in a particular case, $\Phi(h) - \Gamma_m(h-m) = \int_0^1 \left(\Gamma_{m+t(h-m)}(h-m) - \Gamma_m(h-m)\right) dt$. Thus, assumption **(A4)** can be verified by giving a neighborhood of m for each there is a positive constant C_ϵ such for all h in this neighborhood, if we consider the function $\varphi_h : [0,1] \longrightarrow H$ defined for all $t \in [0,1]$ by $\varphi_h(t) := \Gamma_{m+t(h-m)}(h-m)$, then for all $t \in [0,1]$,

$$\left\|\varphi_h'(t)\right\| \leq C_\epsilon \left\|h-m\right\|^2.$$

Assumption **(A5)** enables us to bound the gradient under conditions on the functional f. More precisely, assumption **(A5a)** allows to get the almost sure rates of convergence while assumptions **(A5b)** enables us to obtain the L^p rates of convergence, which represents a significant relaxation of the usual conditions needed to get non asymptotic results. For example, a main difference with [Bac14] is that, instead of having a bounded gradient, we split this bound into two parts : one which has to admits q-th moments for all q, and one which depends on the estimation error. Moreover, note that is is possible to replace assumption **(A5)** by

(A5a') There is a positive constant L^1 such that for all $h \in H$,

$$\mathbb{E}\left[\left\|\nabla_h g(X,h)\right\|^2\right] \leq L_1\left(1 + \left\|h-m\right\|^2\right).$$

(A5b') For all integer q, there is a positive constant L_q such that for all $h \in H$,

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

Remark 7.2.1. *Note that we have analogous assumptions to the usual ones in finite dimension (see [Pel00] among others) but our proofs remain true in infinite dimension, which was not the case in previous works. Moreover, note that these assumptions represent a real improvement compare to the ones introduced in [BM13] since apart from assumption* ***(B5)****, we just introduce local hypothesis and not uniform ones.*

7.2.2 The algorithms

Let $X_1, ..., X_n, ...$ be independent random variables with the same law as X. The Robbins-Monro (or stochastic gradient) algorithm is defined iteratively by

$$\begin{aligned} Z_{n+1} &= Z_n - \gamma_n \nabla_h g\left(X_{n+1}, Z_n\right) \\ &=: Z_n - \gamma_n U_{n+1}, \end{aligned} \tag{7.2}$$

where Z_1 is chosen bounded and $U_{n+1} := \nabla_h g\left(X_{n+1}, Z_n\right)$. Moreover, (γ_n) is a decreasing sequence of positive real numbers which verifies the following usual assumptions for almost sure convergence of Robbins-Monro algorithms (see [Duf97])

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

The term U_{n+1} can be considered as a random perturbation of the gradient Φ at Z_n. Indeed, let $(\mathcal{F}_n)$ be the sequence of σ-algebra defined for all $n \geq 1$ by $\mathcal{F}_n := \sigma\left(X_1, ..., X_n\right) = \sigma\left(Z_1, ..., Z_n\right)$, then

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

In order to improve the convergence, we now introduce the averaged algorithm (see [PJ92]) defined recursively by

$$\overline{Z}_{n+1} = \overline{Z}_n + \frac{1}{n+1}\left(Z_{n+1} - \overline{Z}_n\right), \tag{7.3}$$

with $\overline{Z}_1 = Z_1$. This also can be written as follows

$$\overline{Z}_n = \frac{1}{n}\sum_{k=1}^{n} Z_k.$$

7.2.3 Some convexity properties

We now give some convexity properties of the functional G. First, since $\Phi(m) = 0$ and since G is twice continuously differentiable almost everywhere, note that

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

The first proposition gives the local strong convexity of the functional G.

Proposition 7.2.1. *Assume **(A1)** to **(A3)** and **(A5a)** hold. For all positive constant A, there is a positive constant c_A such that for all $h \in \mathcal{B}(m, A)$,*

$$\langle \Phi(h), h-m \rangle \geq c_A \|h-m\|^2 .$$

Moreover, there is a positive constant C such that for all $h \in H$,

$$|\langle \Phi(h), h-m \rangle| \leq C \|h-m\|^2 .$$

*This result remain true replacing assumption **(A5a)** by **(A5a')**.*

The following corollary ensures that our problem have a unique solution.

Corollary 7.2.1. *Assume **(A1)** to **(A3)** and **(A5a)** hold. Then, m is the unique solution of the equation*

$$\Phi(h) = 0,$$

and in a particular case, m is the unique minimizer of the functional G.

Remark 7.2.2. *Note that Assumption **(A3)** and Proposition 7.2.1 enables us to invert the Hessian at m and to have a control on the "loss" of the strong convexity. Then, assumption **(A3)** could be replaced by*

(A3') *There is a basis composed of eigenvectors of Γ_m and its smallest eigenvalue $\lambda_{\min}$ (or the $\liminf$ of the eigenvalues in the case of infinite dimensional spaces) is positive. Moreover there is a positive constant c such that for all $A > 0$ and for all $h \in \mathcal{B}(m, A)$,*

$$\langle \Phi(h), h-m \rangle \geq \frac{c}{A} \|h-m\|^2 .$$

Finally, the last propositions gives an uniform bound of the remainder term in the Taylor's expansion of the gradient.

Proposition 7.2.2. *Assume **(A1)**, **(A2)** and **(A5a)** hold. Then, for all $h \in H$, there is a positive constant C_m such that for all $h \in H$,*

$$\|\Phi(h) - \Gamma_m(h-m)\| \leq C_m \|h-m\|^2 .$$

*This result remain true replacing assumptions **(A3)** and/or **(A5a)** by **(A3')** and/or **(A5a')**.*

7.3 Rates of convergence

In this section, we consider a step sequence $(\gamma_n)_{n\geq 1}$ of the form $\gamma_n := c_\gamma n^{-\alpha}$ with $c_\gamma > 0$ and $\alpha \in (1/2, 1)$. Note that taking $\alpha = 1$ could be possible with a good choice of the constant c_γ (taking $c_\gamma > \frac{1}{\lambda_{\min}}$ for instance). Nevertheless, the averaging step enables us to get the optimal rate of convergence with a smaller variance than the Robbins-Monro algorithm with a fastly decreasing step sequence $\gamma_n = c_\gamma n^{-1}$ (see [PJ92], [Pel98] and [Pel00] for more details).

7.3.1 Almost sure rates of convergence

In this section, we focus on the almost sure rates of convergence of the algorithms defined in (7.2) and (7.3). First, the following theorem gives the strong consistency of the algorithms.

Theorem 7.3.1. *Suppose **(A1)** to **(A3)** and **(A5a)** hold. Then,*

$$\lim_{n\to\infty} \|Z_n - m\| = 0 \quad a.s,$$
$$\lim_{n\to\infty} \left\|\overline{Z}_n - m\right\| = 0 \quad a.s.$$

*This result remain true replacing assumptions **(A3)** and/or **(A5a)** by **(A3')** and/or **(A5a')**.*

The following theorem gives the almost sure rate of convergence of the Robbins-Monro algorithm as well as the averaged algorithm's one under the additional assumption **(A4)**.

Theorem 7.3.2. *Suppose **(A1)** to **(A5a)** hold. For all $\delta, \delta' > 0$,*

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

*This result remain true replacing assumptions **(A3)** and/or **(A5a)** by **(A3')** and/or **(A5a')**.*

Note that similar results are given in [Pel98], but only in finite dimension. More precisely, the given proofs are not available if the dimension of H is infinite. For example, these methods rely on the fact that the Hessian of the functional G admits finite dimensional eigenspaces, which is not necessarily true for general Hilbert spaces. Another problem is that norms are not equivalent in infinite dimensional spaces, and consequently, the Hilbert-Schmidt norm for linear operators does not necessarily exist while the spectral norm does.

7.3.2 L^p rates of convergence

In this section, we focus on the L^p rates of convergence of the algorithms. The proofs are postponed in Section 7.5. The idea is to give non asymptotic results without focusing only on the rate of convergence in quadratic mean. Indeed, recent works (see [CGB15] and [GB15] for instance), exhibits the fact that having L^p rates of convergence can be very useful. More precisely, these kind of results can be useful to get non asymptotic results when we need to inject an estimator in an algorithm. A simple example is to consider the usual iterative estimator of the variance. First, we give the L^p rates of convergence of the Robbins-Monro algorithm.

Theorem 7.3.3. *Assume* ***(A1)*** *to* ***(A5b)*** *hold. Then, for all integer p, there is a positive constant K_p such that for all $n \geq 1$,*

$$\mathbb{E}\left[\|Z_n - m\|^{2p}\right] \leq \frac{K_p}{n^{p\alpha}}. \tag{7.4}$$

This result remain true replacing assumptions ***(A3)*** *and/or* ***(A5b)*** *by* ***(A3')*** *and/or* ***(A5b')****.*

Remark that it is also possible to get the rate of convergence in quadratic mean taking only a finite numbers of q such that $\mathbb{E}\left[f\left(X,h\right)^q\right] \leq L_q$ (with f defined in assumption **(A5)**), but it would require to introduce additional conditions on α (see the remark below).

Finally, the last theorem gives the L^p rates of convergence of the averaged algorithm.

Theorem 7.3.4. *Assume* ***(A1)*** *to* ***(A5b)*** *hold. Then, for all integer p, there is a positive constant K'_p such that for all $n \geq 1$,*

$$\mathbb{E}\left[\left\|\overline{Z}_n - m\right\|^{2p}\right] \leq \frac{K'_p}{n^p}.$$

This result remain true replacing assumptions ***(A3)*** *and/or* ***(A5b)*** *by* ***(A3')*** *and/or* ***(A5b')****.*

As proved in [CCGB15] and [GB15], these rates of convergence are the optimal ones for the Robbins-Monro algorithm and its averaged version. An extension of this work could be, under the same assumptions, to verify the asymptotic normality of the averaged algorithm as well as a to give a recursive and fast estimator of the covariance.

Remark 7.3.1. *Let $\beta \in (1,2]$, one can obtain the same rate of convergence in quadratic mean and almost sure rate of convergence of the Robbins-Monro algorithm replacing assumption **(A4)** by*

$$\|\Phi(h) - \Gamma_m(h-m)\| \leq C_\beta \|h-m\|^\beta$$

for all $h \in \mathcal{B}(m,\epsilon)$, and get the same rate of convergence in quadratic mean and almost sure rate of convergence for the averaged algorithm taking a step sequence of the form $\gamma_n := c_\gamma n^{-\alpha}$ with $\alpha \in (\beta^{-1},1)$.

Remark 7.3.2. *Let p be a positive integer, note that it is possible to get the L^{2p} rates of convergence of the Robbins-Monro algorithm just supposing that there is a positive integer q such that $q > 2p+2$ and a positive constant L_q such that $\mathbb{E}\left[f(X,h)^{2q}\right] \leq L_q$ (or such that $\mathbb{E}\left[\nabla_h g(X,h)\right] \leq L_q\left(1+\|h-m\|^{2q}\right)$) and taking a step sequence of the form $\gamma_n := c_\gamma n^{-\alpha}$ with $\alpha \in \left(\frac{1}{2}, \frac{q}{p+2+q}\right)$.*

7.4 Application

7.4.1 An application in general separable Hilbert spaces : the geometric quantile

Let H be a separable Hilbert space and let X be a random variable taking values in H. The geometric quantile m^v of X corresponding to a direction v, where $v \in H$ and $\|v\| < 1$, is defined by

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

Note that if $v = 0$, the geometric quantile m^0 corresponds to the geometric median (see [Hal48] or [Kem87] for instance). Let G_v be the function we would like to minimize, it is defined for all $h \in H$ by $G_v := \mathbb{E}\left[\|X-h\| + \langle X-h, v\rangle\right]$. Moreover, since $\|v\| < 1$,

$$\lim_{\|h\|\to\infty} G_v(h) = +\infty,$$

and G_v admits so a minimizer m^v which is also a solution of the following equation

$$\Phi_v(h) := \nabla G_v(h) = -\mathbb{E}\left[\frac{X-h}{\|X-h\|}\right] - v = 0,$$

and assumption **(A1)** is so verified. Then, the Robbins-Monro algorithm and its averaged version are defined recusively by

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

with $m_1^v = \overline{m}_1^v$ chosen bounded. In order to ensure the uniqueness of the geometric quantiles and the convergence of these estimators, we consider from now that the following assumptions are fulfilled :

(B1) The random variable X is not concentrated on a straight line : for all $h \in H$, there is $h' \in H$ such that $\langle h, h' \rangle = 0$ and

$$\mathrm{Var}\left(\langle X, h' \rangle \right) > 0.$$

(B2) The random variable X is not concentrated around single points : for all positive constant A, there is a positive constant C_A such that for all $h \in H$,

$$\mathbb{E}\left[\frac{1}{\|X - h\|} \right] \leq C_A, \qquad \mathbb{E}\left[\frac{1}{\|X - h\|^2} \right] \leq C_A.$$

Note that, as for the median, Assumption **(B2)** is not restrictive since we deal with a high dimensional space. For example, if $H = \mathbb{R}^d$ with $d \geq 3$, as discussed in [Cha92] and [CCZ13], this condition is satisfied since X admits a density which is bounded on every compact subset of $\mathbb{R}^d$. Finally, this assumption ensures the existence of the Hessian of G_v, which is defined for all $h \in H$ by

$$\nabla^2 G_v(h) = \mathbb{E}\left[\frac{1}{\|X - h\|} \left(I_H - \frac{X - h}{\|X - h\|} \otimes \frac{X - h}{\|X - h\|} \right) \right].$$

Then, Corollary 4.2.1 ensures that if assumptions **(B1)** and **(B2)** are fulfilled, assumptions **(A2)** and **(A3)** are verified. Moreover, Lemma 4.5.1 ensures that assumption **(A4)** is fulfilled. Finally, for all positive integer $p \geq 1$ and $h \in H$,

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

Then, assumptions **(A5a)** and **(A5b)** are also verified. Thus, the estimators of the geometric quantiles verify Theorem 7.3.1 to 7.3.4 .

7.4.2 An application in $\mathbb{R}^d$: a robust logistic regression

Let $d \geq 1$ and $H = \mathbb{R}^d$. Let (X, Y) be a couple of random variables taking values in $H \times \{-1, 1\}$. We want to minimize the functional G_r defined for all $h \in \mathbb{R}^d$ by (see [Bac14])

$$G_r(h) := \mathbb{E}\left[\log\left(\cosh\left(Y - \langle X, h\rangle\right)\right)\right].$$

In order to ensure the existence and uniqueness of the solution, we consider from now that the following assumptions are fulfilled :

(B1') There exists m^r such that $\nabla G_r(m^r) = 0$.

(B2') The Hessian of the functional G_r at m^r is positive.

(B3'a) The random variable X admits a 2-th moment.

(B3'b) For all integer p, the random variable X admits a p-th moment.

Assumption **(B1')** ensures the existence of a solution while **(B2')** gives its uniqueness. Assumption **(B3a)** ensures that the function is twice Fréchet-differentiable and its gradient and hessian are defined for all $h \in \mathbb{R}^d$ by

$$\nabla G_r(h) = \mathbb{E}\left[\frac{-\sinh\left(Y - \langle X, h\rangle\right)}{\cosh\left(Y - \langle X, h\rangle\right)} X\right],$$
$$\nabla^2 G_r(h) = \mathbb{E}\left[\frac{1}{\left(\cosh\left(Y - \langle X, h\rangle\right)\right)^2} X \otimes X\right].$$

Thus, assumption **(B2')** is verified, for example, since there are positive constants M, M' such that the matrix $\mathbb{E}\left[X \otimes X \mathbb{1}_{\{\|X\| \leq M\}} \mathbb{1}_{\{\|Y\| \leq M'\}}\right]$ is positive. Then, the solution m^r can be estimated recursively as follows :

$$m^r_{n+1} = m^r_n + \gamma_n \frac{\sinh\left(Y_{n+1} - \langle X_{n+1}, m^r_n\rangle\right)}{\cosh\left(Y_{n+1} - \langle X_{n+1}, m^r_n\rangle\right)} X_{n+1},$$
$$\overline{m}^r_{n+1} = \overline{m}^r_n + \frac{1}{n+1}\left(m^r_{n+1} - \overline{m}^r_n\right),$$

with $\overline{m}^r_1 = m^r_1$ bounded. Under assumptions **(B1')** to **(B3'a)**, assumptions **(A1)** to **(A5a)** are satisfied and Theorems 7.3.1 and 7.3.2 are verified. Under additional assumption **(B3'b)**, assumption **(A5b)** is satisfied and Theorems 7.3.3 and 7.3.4 are verified.

Remark 7.4.1. *Remark that these results remain true for several cases of regression. For example, one*

can consider the logistic regression

$$m^l := \arg\min_{h\in\mathbb{R}^d} \mathbb{E}\left[\log\left(1+\exp\left(-Y\left\langle X,h\right\rangle\right)\right)\right],$$

with (X,Y) *taking values in* $\mathbb{R}^d\times\{-1,1\}$. *Then, we can introduce algorithms of the form*

$$\begin{aligned} m^l_{n+1} &= m^l_n + \gamma_n \frac{\exp\left(-Y_{n+1}\left\langle X_{n+1}, m^l_n\right\rangle\right)}{1+\exp\left(-Y_{n+1}\left\langle X_{n+1}, m^l_n\right\rangle\right)} Y_{n+1}X_{n+1}, \\ \overline{m}^l_{n+1} &= \overline{m}^l_n + \frac{1}{n+1}\left(m^l_{n+1}-\overline{m}^l_n\right). \end{aligned}$$

7.5 Proofs

7.5.1 Some decompositions of the algorithms

Let us recall that the Robbins-Monro algorithm is defined by

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

with $U_{n+1} := \nabla_h g\left(X_{n+1}, Z_n\right)$. Then, let $\xi_{n+1} := \Phi(Z_n) - U_{n+1}$, the algorithm can be decomposed as follows :

$$Z_{n+1} - m = Z_n - m - \gamma_n \Phi(Z_n) + \gamma_n \xi_{n+1}. \tag{7.5}$$

Note that (ξ_n) is a martingale differences sequence adapted to the filtration $(\mathcal{F}_n)$. Furthermore, linearizing the gradient, the algorithm can be written as

$$Z_{n+1} - m = \left(I_H - \gamma_n \Gamma_m\right)\left(Z_n - m\right) + \gamma_n \xi_{n+1} - \gamma_n \delta_n, \tag{7.6}$$

where $\delta_n := \Phi(Z_n) - \Gamma_m\left(Z_n - m\right)$ is the remainder term in the Taylor's expansion of the gradient. Note that thanks to Proposition 7.2.2, there is a positive constant C_m such that for all $n \geq 1$, $\|\delta_n\| \leq C_m \|Z_n - m\|^2$. Finally, by induction, we have the following usual decomposition

$$Z_n - m = \beta_{n-1}\left(Z_1 - m\right) + \beta_{n-1} M_n - \beta_{n-1} R_n, \tag{7.7}$$

with

$$\beta_{n-1} := \prod_{k=1}^{n-1} \left(I_H - \gamma_k \Gamma_m\right), \qquad M_n := \sum_{k=1}^{n-1} \gamma_k \beta_k^{-1} \xi_{k+1},$$
$$\beta_0 := I_H, \qquad R_n := \sum_{k=1}^{n-1} \gamma_k \beta_k^{-1} \delta_k.$$

In the same way, in order to get the rates of convergence, we need to exhibit a new decomposition of the averaged algorithm. In this aim, equality (7.6) can be written as

$$\Gamma_m \left(Z_n - m\right) = \frac{Z_n - m}{\gamma_n} - \frac{Z_{n+1} - m}{\gamma_n} + \xi_{n+1} - \delta_n.$$

As in [Pel00], summing these equalities, applying Abel's transform and dividing by n, we have

$$\Gamma_m \left(\overline{Z}_n - m\right) = \frac{1}{n}\left(\frac{Z_1 - m}{\gamma_1} - \frac{Z_{n+1} - m}{\gamma_n} + \sum_{k=2}^{n}\left(\frac{1}{\gamma_k} - \frac{1}{\gamma_{k-1}}\right)\left(Z_k - m\right) - \sum_{k=1}^{n} \delta_k\right) + \frac{1}{n}\sum_{k=1}^{n} \xi_{k+1}. \tag{7.8}$$

7.5.2 Proof of Section 7.3.1

Proof of Theorem 7.3.1. Using decomposition (7.5) and since (ξ_n) is a sequence of martingale differences adapted to the filtration $(\mathcal{F}_n)$,

$$\mathbb{E}\left[\left\|Z_{n+1} - m\right\|^2 | \mathcal{F}_n\right] = \left\|Z_n - m\right\|^2 - 2\gamma_n \left\langle Z_n - m, \Phi(Z_n)\right\rangle + \gamma_n^2 \left\|\Phi(Z_n)\right\|^2 + \gamma_n^2 \mathbb{E}\left[\left\|\xi_{n+1}\right\|^2 | \mathcal{F}_n\right].$$

Moreover, with Assumption **(A5a)**,

$$\begin{aligned}
\mathbb{E}\left[\left\|\xi_{n+1}\right\|^2 | \mathcal{F}_n\right] &= \mathbb{E}\left[\left\|U_{n+1}\right\|^2 | \mathcal{F}_n\right] - 2\left\langle \mathbb{E}\left[U_{n+1} | \mathcal{F}_n\right], \Phi(Z_n)\right\rangle + \left\|\Phi(Z_n)\right\|^2 \\
&\leq \mathbb{E}\left[\left(f(X_{n+1}, Z_n) + C\left\|Z_n - m\right\|\right)^2 | \mathcal{F}_n\right] - \left\|\Phi(Z_n)\right\|^2 \\
&\leq 2\mathbb{E}\left[f(X_{n+1}, Z_n)^2 | \mathcal{F}_n\right] + 2C^2\left\|Z_n - m\right\|^2 - \left\|\Phi(Z_n)\right\|^2 \\
&\leq 2L_1 + 2C^2\left\|Z_n - m\right\|^2 - \left\|\Phi(Z_n)\right\|^2.
\end{aligned}$$

Thus,

$$\mathbb{E}\left[\left\|Z_{n+1} - m\right\|^2 | \mathcal{F}_n\right] \leq \left(1 + 2C^2\gamma_n^2\right)\left\|Z_n - m\right\|^2 - 2\gamma_n \left\langle \Phi(Z_n), Z_n - m\right\rangle + 2\gamma_n^2 L_1.$$

Since $\langle \Phi(Z_n), Z_n - m \rangle \geq 0$ and $\sum_{n\geq 1} \gamma_n^2 < +\infty$, Robbins-Siegmund theorem (see [Duf97] for example) ensures that $\|Z_n - m\|$ converges almost surely to a finite random variable and that

$$\sum_{n\geq 1} \gamma_n \langle \Phi(Z_n), Z_n - m \rangle < +\infty \quad a.s.$$

Moreover, since $\langle \Phi(Z_n), Z_n - m \rangle \geq 0$, by induction, there is a positive constant M such that for all $n \geq 1$,

$$\begin{aligned}
\mathbb{E}\left[\|Z_{n+1} - m\|^2\right] &\leq \left(1 + 2C^2\gamma_n^2\right) \mathbb{E}\left[\|Z_n - m\|^2\right] + 2\gamma_n^2 L_1 \\
&\leq \left(\prod_{k\geq 1}\left(1 + 2C^2\gamma_k^2\right)\right) \mathbb{E}\left[\|Z_1 - m\|^2\right] + 2L_1 \left(\prod_{k\geq 1}\left(1 + 2C^2\gamma_k^2\right)\right) \sum_{k\geq 1} \gamma_k^2 \\
&\leq M.
\end{aligned}$$

Thus, one can conclude the proof in the same way as in the proof of Theorem 3.1 in [CCZ13] for instance. Finally, one can apply Toeplitz's lemma (see [Duf97], Lemma 2.2.13) to get the strong consistency of the averaged algorithm.

□

In order to get the almost sure rates of convergence of the Robbins-Monro algorithm, we now introduce a technical lemma which gives the rate of convergence of the martingale term $\beta_{n-1} M_n$ in decomposition (7.7).

Lemma 7.5.1. *For all $\delta > 0$,*

$$\|\beta_{n-1} M_n\|^2 = o\left(\frac{(\ln n)^\delta}{n^\alpha}\right) \quad a.s.$$

Proof of Lemma 7.5.1. Since (ξ_n) is a sequence of martingale differences adapted to the filtration $(\mathcal{F}_n)$, and since $M_{n+1} = M_n + \gamma_n \beta_n^{-1} \xi_{n+1}$,

$$\begin{aligned}
\mathbb{E}\left[\|\beta_n M_{n+1}\|^2 |\mathcal{F}_n\right] &= \|\beta_n M_n\|^2 + 2\gamma_n \langle \beta_n M_n, \mathbb{E}\left[\xi_{n+1} | \mathcal{F}_n\right] \rangle + \gamma_n^2 \mathbb{E}\left[\|\xi_{n+1}\|^2 |\mathcal{F}_n\right] \\
&= \|\beta_n M_n\|^2 + \gamma_n^2 \mathbb{E}\left[\|\xi_{n+1}\|^2 |\mathcal{F}_n\right] \\
&\leq \|I_H - \gamma_n \Gamma_m\|_{op}^2 \|\beta_{n-1} M_n\|^2 + \gamma_n^2 \mathbb{E}\left[\|\xi_{n+1}\|^2 |\mathcal{F}_n\right].
\end{aligned}$$

Since each eigenvalue λ of Γ_m verifies $0 < \lambda_{\min} \leq \lambda \leq C$ and since (γ_n) converges to 0, there

is a rank n_0 such that for all $n \geq n_0$, $\|I_H - \gamma_n \Gamma_m\|_{op} \leq 1 - \lambda_{\min}\gamma_n$. Thus, for all $n \geq n_0$,

$$\mathbb{E}\left[\|\beta_n M_{n+1}\|^2 |\mathcal{F}_n\right] \leq (1 - \lambda_{\min}\gamma_n)^2 \|\beta_{n-1} M_n\|^2 + \gamma_n^2 \mathbb{E}\left[\|\xi_{n+1}\|^2 |\mathcal{F}_n\right].$$

Let $\delta > 0$, for all $n \geq 1$, let $V_n := \frac{n^{2\alpha-1}}{(\ln n)^{1+\delta}} \|\beta_{n-1} M_n\|^2$, then for all $n \geq n_0$,

$$\begin{aligned}
\mathbb{E}\left[V_{n+1}|\mathcal{F}_n\right] &\leq (1 - \lambda_{\min}\gamma_n)^2 \frac{(n+1)^{2\alpha-1}}{(\ln(n+1))^{1+\delta}} \|\beta_{n-1} M_n\|^2 + \frac{(n+1)^{2\alpha-1}}{(\ln(n+1))^{1+\delta}} \gamma_n^2 \mathbb{E}\left[\|\xi_{n+1}\|^2 |\mathcal{F}_n\right] \\
&= (1 - \lambda_{\min}\gamma_n)^2 \left(\frac{n+1}{n}\right)^{2\alpha-1} \left(\frac{\ln n}{\ln(n+1)}\right)^{1+\delta} V_n \\
&+ \frac{(n+1)^{2\alpha-1}}{(\ln(n+1))^{1+\delta}} \gamma_n^2 \mathbb{E}\left[\|\xi_{n+1}\|^2 |\mathcal{F}_n\right].
\end{aligned}$$

Moreover, there are a positive constant c and a rank n_0' (let us take $n_0' \geq n_0$) such that for all $n \geq n_0'$,

$$\left(1 - \lambda_{\min} c_\gamma n^{-\alpha}\right) \left(\frac{n+1}{n}\right)^{2\alpha-1} \left(\frac{\ln n}{\ln(n+1)}\right)^{1+\delta} \leq 1 - cn^{-\alpha}.$$

Moreover, $cn^{-\alpha} V_n = c\frac{n^{\alpha-1}}{(\ln n)^{1+\delta}} \|\beta_{n-1} M_n\|^2$. Thus, for all $n \geq n_0'$,

$$\mathbb{E}\left[V_{n+1}|\mathcal{F}_n\right] \leq V_n + \frac{(n+1)^{2\alpha-1}}{(\ln(n+1))^{1+\delta}} \gamma_n^2 \mathbb{E}\left[\|\xi_{n+1}\|^2 |\mathcal{F}_n\right] - c\frac{n^{\alpha-1}}{(\ln n)^{1+\delta}} \|\beta_{n-1} M_n\|^2. \tag{7.9}$$

Moreover, since $\mathbb{E}\left[\|\xi_{n+1}\|^2 |\mathcal{F}_n\right] \leq 2L_1 + 2C \|Z_n - m\|^2$ and since $\|Z_n - m\|$ converges almost surely to 0, the application of the Robbins-Siegmund theorem ensures that (V_n) converges almost surely to a finite random variable and ensures that

$$\sum_{n \geq n_0'} \frac{n^{\alpha-1}}{(\ln n)^{1+\delta}} \|\beta_{n-1} M_n\|^2 < \infty \quad a.s.$$

Previous inequality can also be written as

$$\sum_{n \geq n_0'} \frac{1}{n \ln n} \left(\frac{n^\alpha}{(\ln n)^\delta} \|\beta_{n-1} M_n\|^2\right) < \infty \quad a.s,$$

so that we necessarily have, applying Toeplitz's lemma for example,

$$\frac{n^\alpha}{(\ln n)^\delta} \|\beta_{n-1} M_n\|^2 \xrightarrow[n\to\infty]{a.s} 0. \tag{7.10}$$

□

Remark 7.5.1. *Note that this proof is the main difference with [Pel00]. Indeed, in order to prove the same result, many methods were used but they cannot be applied directly if H was a infinite dimensional space. Theorem 7.3.2 is quite straightforward.*

Proof of Theorem 7.3.2. **Rate of convergence of the Robbins-Monro algorithm**

Applying decomposition (7.7), as in [Pel98], let

$$\Delta_n = \beta_{n-1}(Z_1 - m) - \beta_{n-1}R_n = (Z_n - m) - \beta_{n-1}M_n.$$

We have

$$\begin{aligned}\Delta_{n+1} &= Z_{n+1} - m - \beta_n M_{n+1} \\ &= (I_H - \gamma_n\Gamma_m)(Z_n - m) + \gamma_n\xi_{n+1} - \gamma_n\delta_n - \gamma_n\xi_{n+1} - (I_H - \gamma_n\Gamma_m)\beta_{n-1}M_n \\ &= (I_H - \gamma_n\Gamma_m)\Delta_n - \gamma_n\delta_n.\end{aligned}$$

Thus, applying a lemma of stabilization (see [Duf96] Lemma 4.1.1 for instance), and since $\|\delta_n\| \leq C_m \|Z_n - m\|^2$ almost surely,

$$\|\Delta_n\| = O(\|\delta_n\|) = O\left(\|Z_n - m\|^2\right) \quad a.s.$$

Finally, since (Z_n) converges almost surely to m, $\|\Delta_n\| = o(\|Z_n - m\|)$ almost surely and

$$\begin{aligned}\|Z_n - m\| &\leq \|\beta_{n-1}M_n\| + \|\Delta_n\| \\ &= o\left(\frac{(\ln n)^{\delta/2}}{n^{\alpha/2}}\right) + o(\|Z_n - m\|) \quad a.s,\end{aligned}$$

which concludes the proof.

Rate of convergence of the averaged algorithm

With the help of decomposition (7.8),

$$\begin{aligned}\left\|\overline{Z}_n - m\right\|^2 \leq \frac{5}{\lambda_{\min}^2 n^2}\frac{\|Z_1 - m\|^2}{\gamma_1^2} + \frac{5}{\lambda_{\min}^2 n^2}\frac{\|Z_{n+1} - m\|^2}{\gamma_n^2} + \frac{5}{\lambda_{\min}^2 n^2}\left\|\sum_{k=1}^{n}\delta_k\right\|^2 \\ + \frac{5}{\lambda_{\min}^2 n^2}\left\|\sum_{k=2}^{n}(Z_k - m)\left(\frac{1}{\gamma_k} - \frac{1}{\gamma_{k-1}}\right)\right\|^2 + \frac{5}{\lambda_{\min}^2 n^2}\left\|\sum_{k=1}^{n}\xi_{k+1}\right\|^2.\end{aligned}$$

As in [GB15], applying the almost sure rate of convergence of the Robbins-Monro algorithm, one can check that

$$\frac{1}{n^2}\frac{\|Z_1 - m\|}{\gamma_1} = o\left(\frac{1}{n}\right) \quad a.s,$$
$$\frac{1}{n^2}\frac{\|Z_{n+1} - m\|^2}{\gamma_n^2} = o\left(\frac{1}{n}\right) \quad a.s,$$
$$\frac{1}{n^2}\left\|\sum_{k=2}^{n}(Z_k - m)\left(\frac{1}{\gamma_k} - \frac{1}{\gamma_{k-1}}\right)\right\|^2 = o\left(\frac{1}{n}\right) \quad a.s,$$
$$\left\|\sum_{k=1}^{n}\delta_k\right\|^2 = o\left(\frac{1}{n}\right) \quad a.s.$$

Let $\delta > 0$ and $M'_n := \frac{\sqrt{n}}{\sqrt{(\ln n)^{1+\delta}}}\left\|\frac{1}{n}\sum_{k=1}^{n}\xi_{k+1}\right\| = \frac{1}{\sqrt{n(\ln n)^{1+\delta}}}\left\|\sum_{k=1}^{n}\xi_{k+1}\right\|$. Since (ξ_n) is a martingale differences sequence adapted to the filtration $(\mathcal{F}_n)$, and since

$$\begin{aligned}\mathbb{E}\left[\|\xi_{n+2}\|^2 |\mathcal{F}_{n+1}\right] &\leq 2\mathbb{E}\left[f(X_{n+2}, Z_{n+1})^2|\mathcal{F}_{n+1}\right] + 2C^2\|Z_{n+1} - m\|^2 \\ &\leq 2L_1 + 2C^2\|Z_{n+1} - m\|^2,\end{aligned}$$

we have

$$\begin{aligned}\mathbb{E}\left[M'^2_{n+1}|\mathcal{F}_{n+1}\right] &= \frac{n(\ln n)^{1+\delta}}{(n+1)(\ln(n+1))^{1+\delta}}M'^2_n + \frac{1}{(n+1)(\ln(n+1))^{1+\delta}}\mathbb{E}\left[\|\xi_{n+2}\|^2 |\mathcal{F}_{n+1}\right] \\ &\leq M'^2_n + \frac{1}{(n+1)(\ln(n+1))^{1+\delta}}\left(2L_1 + 2C^2\|Z_{n+1} - m\|^2\right).\end{aligned}$$

Since $\|Z_{n+1} - m\|$ converges almost surely to 0, applying Robbins-Siegmund theorem, M'^2_n converges almost surely to a finite random variable, which concludes the proof. □

7.5.3 Proof of Theorem 7.3.3

In order to prove Theorem 7.3.3 with the help of a strong induction on p, we have to introduce some technical lemmas. Note that these lemmas remain true replacing assumptions **(A3)** and/or **(A5b)** by **(A3')** and/or **(A5b')** but the proofs are only given for the first assumptions.

The first lemma gives a bound of the $2p$-th moment when inequality (7.4) is verified for all integer from 0 to $p-1$.

Lemma 7.5.2. *Assume* ***(A1)*** *to* ***(A5b)*** *hold. Let p be a positive integer, and suppose that for all*

$k \leq p-1$, there is a positive constant K_k such that for all $n \geq 1$,

$$\mathbb{E}\left[\|Z_n - m\|^{2k}\right] \leq \frac{K_k}{n^{k\alpha}}. \tag{7.11}$$

Thus, there are positive constants c_0, C_1, C_2 and a rank n_α such that for all $n \geq n_\alpha$,

$$\mathbb{E}\left[\|Z_{n+1} - m\|^{2p}\right] \leq (1 - c_0\gamma_n)\,\mathbb{E}\left[\|Z_n - m\|^{2p}\right] + \frac{C_1}{n^{(p+1)\alpha}} + C_2\gamma_n\mathbb{E}\left[\|Z_n - m\|^{2p+2}\right].$$

Then, the second lemma gives an upper bound of the $(2p+2)$-th moment when inequality (7.4) is verified for all integer from 0 to $p-1$.

Lemma 7.5.3. *Assume **(A1)** to **(A3)** and **(A5b)** hold. Let p be a positive integer, and suppose that for all $k \leq p-1$, there is a positive constant K_k such that for all $n \geq 1$,*

$$\mathbb{E}\left[\|Z_n - m\|^{2k}\right] \leq \frac{K_k}{n^{k\alpha}}.$$

Thus, there are positive constants C_1', C_2' and a rank n_α such that for all $n \geq n_\alpha$,

$$\mathbb{E}\left[\|Z_{n+1} - m\|^{2p+2}\right] \leq \left(1 - \frac{2}{n}\right)^{p+1}\mathbb{E}\left[\|Z_n - m\|^{2p+2}\right] + \frac{C_1'}{n^{(p+2)\alpha}} + C_2'\gamma_n^2\mathbb{E}\left[\|Z_n - m\|^{2p}\right].$$

Finally, the last lemma enables us to give a bound of the probability of the Robbins-Monro algorithm to go far away from m, which is crucial in order to prove Lemma 7.5.3.

Lemma 7.5.4. *Assume **(A1)** to **(A3)** and **(A5b)** hold. Then, for all integer $p \geq 1$, there is a positive constant M_p such that for all $n \geq 1$,*

$$\mathbb{E}\left[\|Z_n - m\|^{2p}\right] \leq M_p.$$

Proof of Theorem 7.3.3. As in [GB15], we will prove with the help of a strong induction that for all integer $p \geq 1$, and for all $\beta \in \left(\alpha, \frac{p+2}{p}\alpha - \frac{1}{p}\right)$, there are positive constants $K_p, C_{\beta,p}$ such that for all $n \geq 1$,

$$\mathbb{E}\left[\|Z_n - m\|^{2p}\right] \leq \frac{K_p}{n^{p\alpha}},$$
$$\mathbb{E}\left[\|Z_n - m\|^{2p+2}\right] \leq \frac{C_{\beta,p}}{n^{\beta p}}.$$

Applying Lemma 7.5.4, Lemma 7.5.2 and Lemma 7.5.3, as soon as the initialization is satisfied, the proof is strictly analogous to the proof of Theorem 4.1 in [GB15]. Thus, we will just prove that for $p = 1$ and for all $\beta \in (\alpha, 3\alpha - 1)$, there are positive constants $K_1, C_{\beta,1}$ such that

for all $n \geq 1$,

$$\mathbb{E}\left[\|Z_n - m\|^2\right] \leq \frac{K_1}{n^\alpha},$$
$$\mathbb{E}\left[\|Z_n - m\|^4\right] \leq \frac{C_{\beta,1}}{n^\beta}.$$

We now split the end of the proof into two steps.

Step 1 : Calibration of the constants. In order to simplify the demonstration thereafter, we now introduce some notations. Let $K_1', C_{\beta,1}'$ be positive constants such that $K_1' \geq 2^{1+\alpha} C_1 c_0^{-1} c_\gamma^{-1}$, ($c_0, C_1$ are defined in Lemma 7.5.2), and $2K_1' \geq C_{\beta,1}' \geq K_1' \geq 1$. By definition of β, there is a rank $n_\beta \geq n_\alpha$ (n_α is defined in Lemma 7.5.2 and in Lemma 7.5.3) such that for all $n \geq n_\beta$,

$$(1 - c_0\gamma_n)\left(\frac{n+1}{n}\right)^\alpha + \frac{1}{2}c_0\gamma_n + \frac{2^{\alpha+\beta+1}c_\gamma C_2}{(n+1)^\beta} \leq 1,$$
$$\left(1 - \frac{2}{n}\right)^2 \left(\frac{n+1}{n}\right)^\beta + \left(C_1' + C_2' c_\gamma^2\right) 2^{3\alpha} \frac{1}{(n+1)^{3\alpha-\beta}} \leq 1,$$

with C_2 defined in Lemma 7.5.2 and C_1', C_2' defined in Lemma 7.5.3. The rank n_β exists because since $\beta > \alpha$,

$$\begin{aligned}(1 - c_0\gamma_n)\left(\frac{n+1}{n}\right)^\alpha + \frac{1}{2}c_0\gamma_n + \frac{2^{\alpha+\beta+1}c_\gamma C_2}{(n+1)^\beta} &= 1 - c_0\gamma_n + \frac{\alpha}{n} + \frac{1}{2}c_0\gamma_n + O\left(\frac{1}{n^\beta}\right) \\ &= 1 - \frac{1}{2}c_0\gamma_n + o\left(\frac{1}{n^\alpha}\right).\end{aligned}$$

Moreover, since $\beta < 3\alpha - 1$, we have $\beta < 2$, and

$$\begin{aligned}\left(1 - \frac{2}{n}\right)^2 \left(\frac{n+1}{n}\right)^\beta + \left(C_1' + C_2' c_\gamma^2\right) 2^{3\alpha} \frac{1}{(n+1)^{3\alpha-\beta}} &= 1 - (4-2\beta)\frac{1}{n} + o\left(\frac{1}{n}\right) + O\left(\frac{1}{n^{3\alpha-\beta}}\right) \\ &= 1 - (4-2\beta)\frac{1}{n} + o\left(\frac{1}{n}\right).\end{aligned}$$

Step 2 : The induction on n. Let us take $K_1' \geq \max_{1 \leq k \leq n_\beta} \left\{k^\alpha \mathbb{E}\left[\|Z_k - m\|^2\right]\right\}$ and

$C'_{\beta,1} \geq \max_{1\leq k\leq n_\beta}\left\{k^\beta \mathbb{E}\left[\|Z_k - m\|^4\right]\right\}$. We now prove by induction that for all $n \geq n_\beta$,

$$\mathbb{E}\left[\|Z_n - m\|^2\right] \leq \frac{K'_1}{n^\alpha},$$
$$\mathbb{E}\left[\|Z_n - m\|^4\right] \leq \frac{C_{\beta,1}}{n^\beta}.$$

Applying Lemma 7.5.2 and by induction, since $2K'_1 \geq C'_{\beta,1} \geq K'_1 \geq 1$,

$$\begin{aligned}\mathbb{E}\left[\|Z_{n+1} - m\|^2\right] &\leq (1 - c_0\gamma_n)\,\mathbb{E}\left[\|Z_n - m\|^{2p}\right] + \frac{C_1}{n^{2\alpha}} + C_2\gamma_n\mathbb{E}\left[\|Z_n - m\|^{2p+2}\right]\\ &\leq (1 - c_0\gamma_n)\,\frac{K'_1}{n^\alpha} + \frac{C_1}{n^{2\alpha}} + C_2\gamma_n\frac{C'_{\beta,1}}{n^\beta}\\ &\leq (1 - c_0\gamma_n)\,\frac{K'_1}{n^\alpha} + \frac{C_1}{n^{2\alpha}} + 2C_2\gamma_n\frac{K'_1}{n^\beta}.\end{aligned}$$

Factorizing by $\frac{K'_1}{(n+1)^\alpha}$,

$$\begin{aligned}\mathbb{E}\left[\|Z_{n+1} - m\|^2\right] &\leq (1 - c_0\gamma_n)\left(\frac{n+1}{n}\right)^\alpha\frac{K'_1}{(n+1)^\alpha} + \left(\frac{n+1}{n}\right)^\alpha C_1\frac{1}{(n+1)^\alpha n^\alpha}\\ &\quad + 2c_\gamma C_2\left(\frac{n+1}{n}\right)^{\alpha+\beta}\frac{K'_1}{(n+1)^{\alpha+\beta}}\\ &\leq (1 - c_0\gamma_n)\left(\frac{n+1}{n}\right)^\alpha\frac{K'_1}{(n+1)^\alpha} + \frac{2^\alpha C_1 c_\gamma^{-1}\gamma_n}{(n+1)^\alpha} + \frac{2^{\alpha+\beta+1}c_\gamma C_2}{(n+1)^\beta}\frac{K'_1}{(n+1)^\alpha}.\end{aligned}$$

Taking $K'_1 \geq 2^{1+\alpha}C_1 c_\gamma^{-1}c_0^{-1}$,

$$\begin{aligned}\mathbb{E}\left[\|Z_{n+1} - m\|^2\right] &\leq (1 - c_0\gamma_n)\left(\frac{n+1}{n}\right)^\alpha\frac{K'_1}{(n+1)^\alpha} + \frac{1}{2}\gamma_n c_0\frac{K'_1}{(n+1)^\alpha} + \frac{2^{\alpha+\beta+1}c_\gamma C_2}{(n+1)^\beta}\frac{K'_1}{(n+1)^\alpha}\\ &\leq \left((1 - c_0\gamma_n)\left(\frac{n+1}{n}\right)^\alpha + \frac{1}{2}c_0\gamma_n + \frac{2^{\alpha+\beta+1}c_\gamma C_2}{(n+1)^\beta}\right)\frac{K'_1}{(n+1)^\alpha}.\end{aligned}$$

By definition of n_β,

$$\mathbb{E}\left[\|Z_{n+1} - m\|^2\right] \leq \frac{K'_1}{(n+1)^\alpha}. \tag{7.12}$$

In the same way, one can check by induction and applying Lemma 7.5.3 that

$$\mathbb{E}\left[\|Z_{n+1} - m\|^4\right] \leq \left(\left(1 - \frac{2}{n}\right)^2\left(\frac{n+1}{n}\right)^\beta + 2^{3\alpha}\frac{C'_1 + C'_2c_\gamma^2}{(n+1)^{3\alpha-\beta}}\right)\frac{C'_{\beta,1}}{(n+1)^\beta}.$$

By definition of n_β,

$$\mathbb{E}\left[\|Z_{n+1}-m\|^4\right] \leq \frac{C'_{\beta,1}}{n^\beta}, \tag{7.13}$$

which concludes the induction on n, and one can conclude the induction on p and the proof in a similar way as in [GB15]. □

7.5.4 Proof of Theorem 7.3.4

Proof of Theorem 7.3.4. Let $\lambda_{\min}$ be the smallest eigenvalue of Γ_m, with the help of decomposition (7.8), for all integer $p \geq 1$,

$$\begin{aligned}\mathbb{E}\left[\left\|\overline{Z}_n-m\right\|^{2p}\right] \leq{}& \frac{5^{2p-1}}{\lambda_{\min}^{2p}n^{2p}}\frac{\mathbb{E}\left[\|Z_1-m\|^{2p}\right]}{\gamma_1^{2p}} + \frac{5^{2p-1}}{\lambda_{\min}^{2p}n^{2p}}\frac{\mathbb{E}\left[\|Z_{n+1}-m\|^{2p}\right]}{\gamma_n^{2p}}\\ &+ \frac{5^{2p-1}}{\lambda_{\min}^{2p}n^{2p}}\mathbb{E}\left[\left\|\sum_{k=1}^n \delta_k\right\|^{2p}\right] + \frac{5^{2p-1}}{\lambda_{\min}^{2p}n^{2p}}\mathbb{E}\left[\left\|\sum_{k=2}^n (Z_k-m)\left(\frac{1}{\gamma_k}-\frac{1}{\gamma_{k-1}}\right)\right\|^{2p}\right]\\ &+ \frac{5^{2p-1}}{\lambda_{\min}^{2p}n^{2p}}\mathbb{E}\left[\left\|\sum_{k=1}^n \xi_{k+1}\right\|^{2p}\right].\end{aligned}$$

As in [GB15], applying Theorem 7.3.3 and Lemma 4.1 in [GB15], one can check that

$$\begin{aligned}&\frac{1}{n^{2p}}\frac{\mathbb{E}\left[\|Z_1-m\|^{2p}\right]}{\gamma_1^{2p}} = O\left(\frac{1}{n^{2p}}\right),\\ &\frac{1}{n^{2p}}\frac{\mathbb{E}\left[\|Z_{n+1}-m\|^{2p}\right]}{\gamma_n^{2p}} = O\left(\frac{1}{n^{(2-\alpha)p}}\right),\\ &\frac{1}{n^{2p}}\mathbb{E}\left[\left\|\sum_{k=2}^n (Z_k-m)\left(\frac{1}{\gamma_k}-\frac{1}{\gamma_{k-1}}\right)\right\|^{2p}\right] = O\left(\frac{1}{n^{(2-\alpha)p}}\right),\\ &\frac{1}{n^{2p}}\mathbb{E}\left[\left\|\sum_{k=1}^n \delta_k\right\|^{2p}\right] = O\left(\frac{1}{n^{2\alpha p}}\right).\end{aligned}$$

We now prove with the help of a strong induction that for all integer $p \geq 1$, there is a positive constant C_p such that

$$\mathbb{E}\left[\left\|\sum_{k=1}^n \xi_{k+1}\right\|^{2p}\right] \leq C_p n^p.$$

Step 1 : Initialization of the induction. Since (ξ_n) is martingale differences sequence adapted to the filtration $(\mathcal{F}_n)$,

$$\begin{aligned}\mathbb{E}\left[\left\|\sum_{k=1}^{n}\xi_{k+1}\right\|^2\right] &= \sum_{k=1}^{n}\mathbb{E}\left[\|\xi_{k+1}\|^2\right] + 2\sum_{k=1}^{n}\sum_{k'=k+1}^{n}\mathbb{E}\left[\langle\xi_{k+1},\xi_{k'+1}\rangle\right] \\ &= \sum_{k=1}^{n}\mathbb{E}\left[\|\xi_{k+1}\|^2\right] + 2\sum_{k=1}^{n}\sum_{k'=k+1}^{n}\mathbb{E}\left[\langle\xi_{k+1},\mathbb{E}\left[\xi_{k'+1}|\mathcal{F}_{k'}\right]\rangle\right] \\ &= \sum_{k=1}^{n}\mathbb{E}\left[\|\xi_{k+1}\|^2\right].\end{aligned}$$

Moreover, since $\mathbb{E}\left[\|\xi_{n+1}\|^2|\mathcal{F}_n\right] \leq \mathbb{E}\left[\|U_{n+1}\|^2|\mathcal{F}_n\right] \leq 2\mathbb{E}\left[f(X_{n+1},Z_n)^2|\mathcal{F}_n\right] + 2C^2\|Z_n-m\|^2$, applying Theorem 7.3.3,

$$\begin{aligned}\mathbb{E}\left[\left\|\sum_{k=1}^{n}\xi_{k+1}\right\|^2\right] &\leq 2\sum_{k=1}^{n}\mathbb{E}\left[f(X_{k+1},Z_k)^2|\mathcal{F}_k\right] + 2C^2\sum_{k=1}^{n}\mathbb{E}\left[\|Z_k-m\|^2\right] \\ &\leq 2\sum_{k=1}^{n}L_1 + O\left(n^{1-\alpha}\right) \\ &\leq C_1 n.\end{aligned}$$

Step 2 : the induction. Let $p \geq 2$, we suppose from now that for all $p' \leq p-1$, there is a positive constant $C_{p'}$ such that for all $n \geq 1$,

$$\mathbb{E}\left[\left\|\sum_{k=1}^{n}\xi_{k+1}\right\|^{2p'}\right] \leq C_{p'}n^{p'}.$$

Moreover,

$$\left\|\sum_{k=1}^{n+1}\xi_{k+1}\right\|^2 = \left\|\sum_{k=1}^{n}\xi_{k+1}\right\|^2 + 2\left\langle\sum_{k=1}^{n}\xi_{k+1},\xi_{n+2}\right\rangle + \|\xi_{n+2}\|^2.$$

Thus, let $M_n := \sum_{k=1}^{n}\xi_{k+1}$, with the help of previous equality and applying Cauchy-Schwarz's inequality,

$$\begin{aligned}\|M_{n+1}\|^{2p} \leq \left(\|M_n\|^2 + \|\xi_{n+2}\|^2\right)^p + 2\left\langle M_n,\xi_{n+2}\right\rangle\left(\|M_n\|^2 + \|\xi_{n+2}\|^2\right)^{p-1} \\ + \sum_{k=2}^{p}\binom{p}{k}2^k\|M_n\|^k\|\xi_{n+2}\|^k\left(\|M_n\|^2 + \|\xi_{n+2}\|^2\right)^{p-k}.\end{aligned}$$

We now bound the expectation of the three terms on the right-hand side of previous inequality. First, since

$$\begin{aligned}\|U_{n+1}\| &\leq f\left(X_{n+1}, Z_n\right) + C\left\|Z_n - m\right\|, \\ \|\Phi(Z_n)\| &\leq \sqrt{L_1} + C\left\|Z_n - m\right\|,\end{aligned}$$

we have

$$\begin{aligned}\mathbb{E}\left[\left\|\xi_{n+2}\right\|^{2k}|\mathcal{F}_{n+1}\right] &\leq 3^{2k-1}\left(\mathbb{E}\left[f(X_{n+2}, Z_n)^{2k}|\mathcal{F}_{n+1}\right] + 2^{2k}C^{2k}\left\|Z_{n+1} - m\right\|^{2k} + L_1^k\right) \\ &\leq 3^{2k-1}\left(L_k + L_1^k + 2^{2k}C^{2k}\left\|Z_{n+1} - m\right\|^{2k}\right).\end{aligned}$$

Then, since M_n is F_{n+1}-measurable,

$$\begin{aligned}\mathbb{E}\left[\left(\left\|M_n\right\|^2 + \left\|\xi_{n+2}\right\|^2\right)^p\right] &\leq \mathbb{E}\left[\left\|M_n\right\|^{2p}\right] + \sum_{k=1}^{p}\binom{p}{k}\mathbb{E}\left[\mathbb{E}\left[\left\|\xi_{n+2}\right\|^{2k}|\mathcal{F}_n\right]\left\|M_n\right\|^{2p-2k}\right] \\ &\leq \mathbb{E}\left[\left\|M_n\right\|^{2p}\right] + \sum_{k=1}^{p}\binom{p}{k}3^{2k-1}\left(L_k + L_1^k\right)\mathbb{E}\left[\left\|M_n\right\|^{2p-2k}\right] \\ &\quad + \sum_{k=1}^{p}\binom{p}{k}3^{2k-1}2^{2k}C^{2k}\mathbb{E}\left[\left\|Z_{n+1} - m\right\|^{2k}\left\|M_n\right\|^{2p-2k}\right]\end{aligned}$$

By induction,

$$\sum_{k=1}^{p}\binom{p}{k}3^{2k-1}\left(L_k + L_1^k\right)\mathbb{E}\left[\left\|M_n\right\|^{2p-2k}\right] \leq \sum_{k=1}^{p}\binom{p}{k}3^{2k-1}\left(L_k + L_1^k\right)C_{p-k}n^{p-k} = O\left(n^{p-1}\right).$$

Moreover, since for all positive real number a and for all positive integer q, $a \leq 1 + a^q$, applying Hölder's inequality and by induction, let

$$\begin{aligned}(\star) :&= \sum_{k=1}^{p}\binom{p}{k}3^{2k-1}2^{2k}C^{2k}\mathbb{E}\left[\left\|Z_{n+1} - m\right\|^{2k}\left\|M_n\right\|^{2p-2k}\right] \\ &\leq \sum_{k=1}^{p}\binom{p}{k}3^{2k-1}2^{2k}C^{2k}\mathbb{E}\left[\left\|M_n\right\|^{2p-2k}\right] + \sum_{k=1}^{p}\binom{p}{k}3^{2k-1}2^{2k}C^{2k}\mathbb{E}\left[\left\|Z_{n+1} - m\right\|^{2qk}\left\|M_n\right\|^{2p-2k}\right] \\ &\leq \sum_{k=1}^{p}\binom{p}{k}3^{2k-1}2^{2k}C^{2k}\left(\mathbb{E}\left[\left\|Z_{n+1} - m\right\|^{2qp}\right]\right)^{\frac{k}{p}}\left(\mathbb{E}\left[\left\|M_n\right\|^{2p}\right]\right)^{\frac{2p-2k}{2p}} + O\left(n^{p-1}\right).\end{aligned}$$

Note that $\left(\mathbb{E}\left[\left\|M_n\right\|^{2p}\right]\right)^{\frac{2p-2k}{2p}} \leq 1 + \mathbb{E}\left[\left\|M_n\right\|^{2p}\right]$. Thus, taking $q \geq 2$ and applying Theo-

rem 7.3.3, there are positive constants C_0, C_1' such that

$$\begin{aligned}(\star) &\leq \sum_{k=1}^{p} \binom{p}{k} 3^{2k-1} 2^{2k} C^{2k} \left(K_{qp}\right)^{\frac{k}{p}} \frac{1}{n^{qk\alpha}} \left(1 + \mathbb{E}\left[\|M_n\|^{2p}\right]\right) + O\left(n^{p-1}\right) \\ &\leq C_0 \gamma_n^2 \mathbb{E}\left[\|M_n\|^{2p}\right] + C_1' n^{p-1}.\end{aligned}$$

Finally, there are positive constants C_0, C_1 such that

$$\mathbb{E}\left[\left(\|M_n\|^2 + \|\xi_{n+2}\|^2\right)^p\right] \leq \left(1 + C_0\gamma_n^2\right) \mathbb{E}\left[\|M_n\|^{2p}\right] + C_1 n^{p-1}. \tag{7.14}$$

Moreover, since (ξ_n) is a martingale differences sequence adapted to the filtration $(\mathcal{F}_n)$ and applying Lemma A.1 in [GB15],

$$\begin{aligned}2\mathbb{E}\left[\langle M_n, \xi_{n+2}\rangle \left(\|M_n\|^2 + \|\xi_{n+2}\|^2\right)^{p-1}\right] &= 2\sum_{k=1}^{p-1} \binom{p-1}{k} \mathbb{E}\left[\langle M_n, \xi_{n+2}\rangle \|\xi_{n+2}\|^{2k} \|M_n\|^{2p-2-2k}\right] \\ &\leq \sum_{k=1}^{p-1} \binom{p-1}{k} \mathbb{E}\left[\|\xi_{n+2}\|^{2k+2} \|M_n\|^{2p-2-2k}\right] \\ &\quad + \sum_{k=1}^{p-1} \binom{p-1}{k} \mathbb{E}\left[\|\xi_{n+2}\|^{2k} \|M_n\|^{2p-2k}\right]\end{aligned}$$

Since $p \geq 2$ and by induction, as for $(\star)$, one can check that there are positive constants C_0', C_1' such that for all $n \geq 1$,

$$2\mathbb{E}\left[\langle M_n, \xi_{n+2}\rangle \left(\|M_n\|^2 + \|\xi_{n+2}\|^2\right)^{p-1}\right] \leq C_0' \gamma_n^2 \mathbb{E}\left[\|M_n\|^{2p}\right] + C_1' n^{p-1}. \tag{7.15}$$

Moreover, let

$$\begin{aligned}(\star\star) :&= \sum_{k=2}^{p} \binom{p}{k} 2^k \mathbb{E}\left[\|M_n\|^k \|\xi_{n+2}\|^k \left(\|M_n\|^2 + \|\xi_{n+2}\|^2\right)^{p-k}\right] \\ &\leq \sum_{k=2}^{p} \binom{p}{k} 2^{p-1} \mathbb{E}\left[\|M_n\|^k \|\xi_{n+2}\|^k \left(\|M_n\|^{2p-2k} + \|\xi_{n+2}\|^{2p-2k}\right)\right] \\ &\leq \sum_{k=2}^{p} \binom{p}{k} 2^{p-1} \mathbb{E}\left[\|\xi_{n+2}\|^k \|M_n\|^{2p-k}\right] + \sum_{k=2}^{p} \binom{p}{k} 2^{p-1} \mathbb{E}\left[\|M_n\|^k \|\xi_{n+2}\|^{2p-k}\right].\end{aligned}$$

We now bound the two terms on the right-hand side of previous inequality. First, let

$$\begin{aligned}(\star\star') &:= \sum_{k=2}^{p} \binom{p}{k} 2^{p-1} \mathbb{E}\left[\|M_n\|^k \|\xi_{n+2}\|^{2p-k}\right] \\ &\leq \sum_{k=2}^{p} \binom{p}{k} 2^{p-3} \mathbb{E}\left[\left(\|M_n\|^2 + \|M_n\|^{2k-2}\right)\left(\|\xi_{n+2}\|^{2p-2k+2} + \|\xi_{n+2}\|^{2p-2}\right)\right]\end{aligned}$$

As for $(\star)$, one can check that there are positive constants C_0'', C_1'' such that for all $n \geq 1$,

$$(\star\star') \leq C_0'' \gamma_n^2 \mathbb{E}\left[\|M_n\|^{2p}\right] + C_1'' n^{p-1}.$$

In the same way, let

$$\begin{aligned}(\star\star'') &:= \sum_{k=2}^{p} \binom{p}{k} 2^{p-1} \mathbb{E}\left[\|\xi_{n+2}\|^k \|M_n\|^{2p-k}\right] \\ &\leq \sum_{k=2}^{p} \binom{p}{k} 2^{p-3} \mathbb{E}\left[\left(\|\xi_{n+2}\|^2 + \|\xi_{n+2}\|^{2k-2}\right)\left(\|M_n\|^{2p-2k+2} + \|M_n\|^2\right)\right]\end{aligned}$$

As for $(\star)$, there are positive constants C_0''', C_1''' such that

$$(\star\star'') \leq C_0''' \gamma_n^2 \mathbb{E}\left[\|M_n\|^{2p}\right] + C_1''' n^{p-1},$$

and in a particular case

$$(\star\star) \leq \left(C_0'' + C_0'''\right) \gamma_n^2 \mathbb{E}\left[\|M_n\|^{2p}\right] + \left(C_1'' + C_1'''\right) n^{p-1}. \tag{7.16}$$

Thus, thanks to inequalities (7.14) to (7.16), there are positive constants B_0, B_1 such that for all $n \geq 1$,

$$\begin{aligned}\mathbb{E}\left[\|M_{n+1}\|^{2p}\right] &\leq \left(1 + B_0 \gamma_n^2\right) \mathbb{E}\left[\|M_n\|^{2p}\right] + B_1 n^{p-1} \\ &\leq \left(\prod_{k=1}^{\infty} \left(1 + B_0 \gamma_k^2\right)\right) \mathbb{E}\left[\|M_1\|^{2p}\right] + \left(\prod_{k=1}^{\infty} \left(1 + B_0 \gamma_k^2\right)\right) \sum_{k=1}^{n} B_1 k^{p-1} \\ &\leq \left(\prod_{k=1}^{\infty} \left(1 + B_0 \gamma_k^2\right)\right) \mathbb{E}\left[\|M_1\|^{2p}\right] + \left(\prod_{k=1}^{\infty} \left(1 + B_0 \gamma_k^2\right)\right) B_1 n^p,\end{aligned}$$

which concludes the induction and the proof. □

Annexe D

L^p and almost sure rates of convergence of averaged stochastic gradient algorithms and applications to robust statistics. Supplementary proofs

Résumé

Dans cette partie, on commence par prouver les Propositions 7.2.1 et 7.2.2, qui permettent d'assurer la forte convexité locale de la fonction que l'on veut minimiser ainsi que de majorer le gradient ainsi que le terme de reste dans la décomposition de Taylor du gradient. Enfin, on donne les preuves des Lemmes techniques 7.5.2, 7.5.3 et 7.5.4, qui permettent de majorer les moments d'ordre $2p$ et $2p+2$, ce qui est crucial pour obtenir les vitesses de convergence de l'algorithme de Robbins-Monro.

D.1 Proofs of Propositions 7.2.1 and 7.2.2 and recall on the decompositions of the algorithms

D.1.1 Proofs of Propositions 7.2.1 and 7.2.2

Proof of Proposition 7.2.1. If $h \in \mathcal{B}(m, \epsilon)$, under assumptions **(A2)** and **(A3)** and by dominated convergence,

$$\begin{aligned}\langle \Phi(h), h-m \rangle &= \left\langle \int_0^1 \Gamma_{m+t(h-m)}(h-m)dt, h-m \right\rangle \\ &= \int_0^1 \left\langle \Gamma_{m+t(h-m)}(h-m), h-m \right\rangle dt \\ &\geq \frac{\lambda_{\min}}{2} \|h-m\|^2 .\end{aligned}$$

In the same way, if $\|h-m\| > \epsilon$, since G is convex, under assumptions **(A2)** and **(A3)** and by dominated convergence,

$$\begin{aligned}\langle \Phi(h), h-m \rangle &= \left\langle \int_0^1 \Gamma_{m+t(h-m)}(h-m)dt, h-m \right\rangle \\ &= \int_0^1 \left\langle \Gamma_{m+t(h-m)}(h-m), h-m \right\rangle dt \\ &\geq \int_0^{\frac{\epsilon}{\|h-m\|}} \left\langle \Gamma_{m+t(h-m)}(h-m), h-m \right\rangle dt \\ &\geq \int_0^{\frac{\epsilon}{\|h-m\|}} \frac{\lambda_{\min}}{2} \|h-m\|^2 dt \\ &= \frac{\lambda_{\min}\epsilon}{2} \|h-m\| .\end{aligned}$$

Thus, let A be a positive constant and $h \in \mathcal{B}(m, A)$,

$$\langle \Phi(h), h-m \rangle \geq c_A \|h-m\|^2 ,$$

with $c_A := \min\left\{\frac{\lambda_{\min}}{2}, \frac{\lambda_{\min}\epsilon}{2A}\right\}$. We now give an upper bound of this term. First, thanks to assumption **(A2)**, let A be a positive constant, for all $h \in \mathcal{B}(m, A)$,

$$\begin{aligned}
\langle \Phi(h), h-m\rangle &= \int_0^1 \left\langle \Gamma_{m+t(h-m)}(h-m), h-m \right\rangle dt \\
&\leq \int_0^1 \left\| \Gamma_{m+t(h-m)}(h-m) \right\| \|h-m\| \, dt \\
&\leq \int_0^1 \left\| \Gamma_{m+t(h-m)} \right\|_{op} \|h-m\|^2 \, dt \\
&\leq C_A \|h-m\|^2 .
\end{aligned}$$

Moreover, applying Cauchy-Schwarz's inequality and thanks to assumption **(A5a)**, for all $h \in H$ such that $\|h-m\| \geq A$,

$$\begin{aligned}
|\langle \Phi(h), h-m\rangle| &\leq \|\Phi(h)\| \, \|h-m\| \\
&\leq (\mathbb{E}[f(X,h)] + C\|h-m\|)\|h-m\| \\
&\leq \sqrt{L_1}\|h-m\| + C\|h-m\|^2 \\
&\leq \left(\frac{\sqrt{L_1}}{A} + C\right)\|h-m\|^2 ,
\end{aligned}$$

which concludes the proof. □

Proof of Proposition 7.2.2. Let us recall that there are positive constants ϵ, C_ϵ such that for all $h \in \mathcal{B}(m, \epsilon)$,

$$\|\Phi(h) - \Gamma_m(h-m)\| \leq C_\epsilon \|h-m\|^2 .$$

Let $h \in H$ such that $\|h-m\| \geq \epsilon$. Then, thanks to assumptions **(A2)** and **(A3)**,

$$\begin{aligned}
\|\Phi(h) - \Gamma_m(h-m)\| &\leq \|\Phi(h)\| + \|\Gamma_m\|_{op} \|h-m\| \\
&\leq (\mathbb{E}[f(X,h)] + C\|h-m\|) + C_0\|h-m\| \\
&\leq \left(\frac{\sqrt{L_1}}{\epsilon^2} + \frac{C}{\epsilon} + \frac{C_0}{\epsilon}\right)\|h-m\|^2 ,
\end{aligned}$$

which concludes the proof. □

D.1.2 Decomposition of the algorithm

Let us recall that the Robbins-Monro algorithm can be written as

$$Z_{n+1} - m = Z_n - m - \gamma_n \Phi(Z_n) + \gamma_n \xi_{n+1}, \tag{D.1}$$

where $\Phi(Z_n)$ is the gradient of the function G at Z_n, and $\xi_{n+1} := \Phi(Z_n) - \nabla_h g\left(X_{n+1}, Z_n\right)$. Moreover, let us recall that denoting by $\left(\mathcal{F}_n\right)_{n\geq 1}$ the sequence of σ-algebra defined for all $n \geq 1$ by $\sigma_n := \sigma\left(X_1, ..., X_n\right)$, then (ξ_n) is a sequence of martingale differences. Finally, linearizing the gradient, the Robbins-Monro algorithm can be written as

$$Z_{n+1} - m = \left(I_H - \gamma_n \Gamma_m\right)\left(Z_n - m\right) + \gamma_n \xi_{n+1} - \gamma_n \delta_n, \tag{D.2}$$

where $\delta_n := \Phi(Z_n) - \Gamma_m\left(Z_n - m\right)$ is the remainder term in the Taylor's expansion of the gradient.

D.2 Proof of Lemma 7.5.4

Proof of Lemma 7.5.4. We prove Lemma 7.5.4 with the help of a strong induction on p. The case $p = 1$ is already done in the proof of Theorem 3.1. We suppose from now that $p \geq 2$ and that for all $k \leq p-1$, there is a positive constant M_k such that for all $n \geq 1$,

$$\mathbb{E}\left[\left\|Z_n - m\right\|^{2k}\right] \leq M_k.$$

Let $V_n := Z_n - m - \gamma_n \Phi(Z_n)$, and with the help of decomposition (D.1)

$$\begin{aligned}\left\|Z_{n+1} - m\right\|^2 &= \left\|V_n\right\|^2 + \gamma_n^2 \left\|\xi_{n+1}\right\|^2 + 2\gamma_n \left\langle V_n, \xi_{n+1}\right\rangle \\ &\leq \left\|V_n\right\|^2 + \gamma_n^2 \left\|U_{n+1}\right\|^2 + 2\gamma_n \left\langle Z_n - m, \xi_{n+1}\right\rangle.\end{aligned}$$

Thus, applying Cauchy-Schwarz's inequaltiy

$$\begin{aligned}\left\|Z_{n+1} - m\right\|^{2p} \leq \left(\left\|V_n\right\|^2 + \gamma_n^2 \left\|U_{n+1}\right\|^2\right)^p + 2p\gamma_n \left\langle Z_n - m, \xi_{n+1}\right\rangle \left(\left\|V_n\right\|^2 + \gamma_n^2 \left\|U_{n+1}\right\|^2\right)^{p-1} \\ + \sum_{k=2}^{p} \binom{p}{k} 2^k \gamma_n^k \left\|Z_n - m\right\|^k \left\|\xi_{n+1}\right\|^k \left(\left\|V_n\right\|^2 + \gamma_n^2 \left\|U_{n+1}\right\|^2\right)^{p-k}.\end{aligned} \tag{D.3}$$

Moreover, since $\|U_{n+1}\| \leq f(X_{n+1}, Z_n) + C\|Z_n - m\|$, it comes $\|\Phi(Z_n)\| \leq C\|Z_n - m\| + \sqrt{L_1}$, and

$$\begin{aligned}\|\xi_{n+1}\| &\leq \|U_{n+1}\| + \|\Phi(Z_n)\| \\ &\leq f(X_{n+1}, Z_n) + 2C\|Z_n - m\| + \sqrt{L_1}.\end{aligned}$$

Applying Lemma A.1 in [GB15], for all positive integer k,

$$\|U_{n+1}\|^k \leq 2^{k-1} f(X_{n+1}, Z_n) + 2^{k-1} C^k \|Z_n - m\|^k \quad a.s, \tag{D.4}$$

$$\|\xi_{n+1}\|^k \leq 3^{k-1} f(X_{n+1}, Z_n)^k + 3^{k-1} 2^k C^k \|Z_n - m\|^k + 3^{k-1} L_1^{\frac{k}{2}} \quad a.s. \tag{D.5}$$

Moreover, since $\langle \Phi(Z_n), Z_n - m \rangle \geq 0$ and since $\|\Phi(Z_n)\| \leq C\|Z_n - m\| + \sqrt{L_1}$,

$$\begin{aligned}\|V_n\|^2 &\leq \left(1 + 2C^2\gamma_n^2\right)\|Z_n - m\|^2 + 2\gamma_n^2 L_1, \\ \|V_n\|^2 &\leq \left(1 + 2C^2 c_\gamma^2\right)\|Z_n - m\|^2 + 2\gamma_n^2 L_1.\end{aligned}$$

We now bound each term on the right-hand side of inequality (D.3).

Bounding $\mathbb{E}\left[\left(\|V_n\|^2 + \gamma_n^2\|U_{n+1}\|^2\right)^p\right]$. Applying previous inequality and inequality (D.4), let

$$\begin{aligned}(*) :&= \mathbb{E}\left[\left(\|V_n\|^2 + \gamma_n^2\|U_{n+1}\|^2\right)^p\right] \\ &= \mathbb{E}\left[\|V_n\|^{2p}\right] + \sum_{k=1}^{p}\binom{p}{k}\gamma_n^{2k}\mathbb{E}\left[\|V_n\|^{2p-2k}\,\mathbb{E}\left[\|U_{n+1}\|^{2k}|\mathcal{F}_n\right]\right] \\ &\leq \mathbb{E}\left[\|V_n\|^{2p}\right] + \sum_{k=1}^{p}\binom{p}{k}\gamma_n^{2p} 2^{2p-2} L_1^{p-k}\mathbb{E}\left[\mathbb{E}\left[f(X_{n+1}, Z_n)^{2k}|\mathcal{F}_n\right] + C^{2k}\|Z_n - m\|^{2k}\right] \\ &+ \sum_{k=1}^{p}\binom{p}{k}\gamma_n^{2k} 2^{p+k-2}\left(1 + 2C^2c_\gamma^2\right)^{p-k}\mathbb{E}\left[\|Z_n - m\|^{2p-2k}\,\mathbb{E}\left[f(X_{n+1}, Z_n)^{2k}|\mathcal{F}_n\right]\right] \\ &+ \sum_{k=1}^{p}\binom{p}{k}\gamma_n^{2k} 2^{p+k-2}\left(1 + 2C^2c_\gamma^2\right)^{p-k}\mathbb{E}\left[\|Z_n - m\|^{2p-2k} C^{2k}\|Z_n - m\|^{2k}\right].\end{aligned}$$

Moreover, since $\mathbb{E}\left[f(X_{n+1}, Z_n)^{2k}|\mathcal{F}_n\right] \leq L_k$ and by induction, there are positive constants

A_0, A_1 such that

$$
\begin{aligned}
(*) &\leq \mathbb{E}\left[\|V_n\|^{2p}\right] + \gamma_n^{2p} 2^{2p-2}\left(L_p + C^{2p}\mathbb{E}\left[\|Z_n - m\|^{2p}\right]\right) + \sum_{k=1}^{p-1}\binom{p}{k}\gamma_n^{2p}2^{2p-2}L_1^{p-k}\left(L_k + C^{2k}M_k\right) \\
&+ \gamma_n^{2p}2^{2p-2}L_p + \sum_{k=1}^{p-1}\binom{p}{k}2^{p+k-2}\gamma_n^{2k}\left(1+2C^2c_\gamma^2\right)^{p-k}L_k M_{p-k} \\
&+ \sum_{k=1}^{p}\binom{p}{k}2^{p+k-2}\gamma_n^{2k}\left(1+2C^2c_\gamma^2\right)^{p-k}\mathbb{E}\left[\|Z_n - m\|^{2p}\right] \\
&\leq \mathbb{E}\left[\|V_n\|^{2p}\right] + A_0\gamma_n^2\mathbb{E}\left[\|Z_n - m\|^{2p}\right] + A_1\gamma_n^2. \qquad \text{(D.6)}
\end{aligned}
$$

Since $\|V_n\|^2 \leq \left(1+2C^2\gamma_n^2\right)\|Z_n - m\|^2 + 2L_1\gamma_n^2$ and by induction,

$$
\begin{aligned}
\mathbb{E}\left[\|V_n\|^{2p}\right] &\leq \left(1+2C^2\gamma_n^2\right)^p\mathbb{E}\left[\|Z_n - m\|^{2p}\right] + \sum_{k=1}^{p}\binom{p}{k}\left(1+2C^2\gamma_n^2\right)^{p-k}2^k L_1^k\gamma_n^{2k}\mathbb{E}\left[\|Z_n - m\|^{2p-2k}\right] \\
&= \left(1+2C^2\gamma_n^2\right)^p\mathbb{E}\left[\|Z_n - m\|^{2p}\right] + \sum_{k=1}^{p}\binom{p}{k}\left(1+2C^2\gamma_n^2\right)^{p-k}2^k L_1^k\gamma_n^{2k}M_{p-k} \\
&\leq \left(1+2C^2\gamma_n^2\right)^p\mathbb{E}\left[\|Z_n - m\|^{2p}\right] + O\left(\gamma_n^2\right).
\end{aligned}
$$

Then, there are positive constants A_2, A_3 such that

$$
(*) \leq \left(1 + A_2\gamma_n^2\right)\mathbb{E}\left[\|Z_n - m\|^{2p}\right] + A_3\gamma_n^2. \qquad \text{(D.7)}
$$

Bounding $2p\gamma_n\mathbb{E}\left[\langle\xi_{n+1}, Z_n - m\rangle\left(\|V_n\|^2 + \gamma_n^2\|U_{n+1}\|^2\right)^{p-1}\right]$**.** Since (ξ_n) is a martingale differences sequence adapted to the filtration $(\mathcal{F}_n)$, and since V_n is $\mathcal{F}_n$-measurable, let

$$
\begin{aligned}
(**) :&= \gamma_n\mathbb{E}\left[\langle\xi_{n+1}, Z_n - m\rangle\left(\|V_n\|^2 + \gamma_n^2\|U_{n+1}\|^2\right)^{p-1}\right] \\
&= \gamma_n\mathbb{E}\left[\langle\mathbb{E}\left[\xi_{n+1}|\mathcal{F}_n\right], Z_n - m\rangle\|V_n\|^2\right] \\
&+ \gamma_n\mathbb{E}\left[\langle\xi_{n+1}, Z_n - m\rangle\sum_{k=1}^{p-1}\binom{p-1}{k}\gamma_n^{2k}\|V_n\|^{2p-2-2k}\|U_{n+1}\|^{2k}\right] \\
&= \gamma_n\mathbb{E}\left[\langle\xi_{n+1}, Z_n - m\rangle\sum_{k=1}^{p-1}\binom{p-1}{k}\gamma_n^{2k}\|V_n\|^{2p-2-2k}\|U_{n+1}\|^{2k}\right].
\end{aligned}
$$

Since $|\langle \xi_{n+1}, Z_n - m\rangle| \leq \frac{1}{2}\left(\|\xi_{n+1}\|^2 + \|Z_n - m\|^2\right)$ and $\|V_n\|^2 \leq \left(1 + 2c_\gamma^2 C^2\right)\|Z_n - m\|^2 + 2L_1\gamma_n^2$,

$$
\begin{aligned}
(**) \leq{}& \frac{1}{2}\sum_{k=1}^{p-1}\binom{p-1}{k}\gamma_n^{2k+1}2^{p-2-k}\left(1+2C^2c_\gamma^2\right)^{p-1-k}\mathbb{E}\left[\|Z_n-m\|^{2p-2-2k}\|U_{n+1}\|^{2k}\|\xi_{n+1}\|^2\right]\\
&+\frac{1}{2}\sum_{k=1}^{p-1}\binom{p-1}{k}\gamma_n^{2k+1}2^{p-2-k}\left(1+C^2c_\gamma^2\right)^{p-1-k}\mathbb{E}\left[\|Z_n-m\|^{2p-2k}\|U_{n+1}\|^{2k}\right]\\
&+\frac{1}{2}\sum_{k=1}^{p-1}\binom{p-1}{k}\gamma_n^{2p-1}2^{2p-3-2k}L_1^{p-1-k}\left(\mathbb{E}\left[\|U_{n+1}\|^{2k}\|\xi_{n+1}\|^2\right]+\mathbb{E}\left[\|U_{n+1}\|^{2k}\|Z_n-m\|^2\right]\right).
\end{aligned}
$$

Moreover, since $p \geq 2$, applying inequalities (D.4) and (D.5) and by induction, as for $(*)$, one can check that there are positive constants A_1', A_2' such that

$$
(**) \leq A_1'\gamma_n^3\mathbb{E}\left[\|Z_n-m\|^{2p}\right] + A_2'\gamma_n^3. \tag{D.8}
$$

Bounding $\sum_{k=2}^{p}\binom{p}{k}2^k\gamma_n^k\mathbb{E}\left[\|Z_n-m\|^k\|\xi_{n+1}\|^k\left(\|V_n\|^2+\gamma_n^2\|U_{n+1}\|^2\right)^{p-k}\right]$. Applying Lemma A.1 in [GB15], and since $\|Z_n-m\|^k\|\xi_{n+1}\|^k \leq \frac{1}{2}\left(\|Z_n-m\|^{2k}+\|\xi_{n+1}\|^{2k}\right)$, let

$$
\begin{aligned}
(***) :={}& \sum_{k=2}^{p}\binom{p}{k}2^k\gamma_n^k\mathbb{E}\left[\|Z_n-m\|^k\|\xi_{n+1}\|^k\left(\|V_n\|^2+\gamma_n^2\|U_{n+1}\|^2\right)^{p-k}\right]\\
\leq{}& \sum_{k=2}^{p}\binom{p}{k}2^{p-2}\gamma_n^k\mathbb{E}\left[\|V_n\|^{2p-2k}\|Z_n-m\|^{2k}+\gamma_n^{2p-2k}\|Z_n-m\|^{2k}\|U_{n+1}\|^{2p-2k}\right]\\
&+\sum_{k=2}^{p}\binom{p}{k}2^{p-2}\gamma_n^k\mathbb{E}\left[\|V_n\|^{2p-2k}\|\xi_{n+1}\|^{2k}+\gamma_n^{2p-2k}\|U_{n+1}\|^{2p-2k}\|\xi_{n+1}\|^{2k}\right]
\end{aligned}
$$

Applying inequalities (D.4) and (D.5) and by induction, as for $(*)$, one can check that there are positive constants A_1'', A_2'' such that

$$
(***) \leq A_1''\gamma_n^2\mathbb{E}\left[\|Z_n-m\|^{2p}\right] + A_2''\gamma_n^2. \tag{D.9}
$$

Conclusion. Applying inequalities (D.7) to (D.9) and by induction, there are positive constants

B_1, B_2 such that

$$\begin{aligned}\mathbb{E}\left[\|Z_{n+1}-m\|^{2p}\right] &\leq (1+B_1\gamma_n^2)\,\mathbb{E}\left[\|Z_n-m\|^{2p}\right]+B_2\gamma_n^2\\ &\leq \left(\prod_{k=1}^{\infty}(1+B_1\gamma_k^2)\right)\mathbb{E}\left[\|Z_1-m\|^{2p}\right]+B_2\left(\prod_{k=1}^{\infty}(1+B_1\gamma_k^2)\right)\sum_{k=1}^{\infty}\gamma_k^2\\ &\leq M_p,\end{aligned}$$

which concludes the induction and the proof.

□

D.3 Proof of Lemma 7.5.3

Proof of Lemma 7.5.3. Let $p \geq 1$, we suppose from now that for all integer $k < p$, there is a positive constant K_k such that for all $n \geq 1$,

$$\mathbb{E}\left[\|Z_n-m\|^{2k}\right] \leq \frac{K_k}{n^{k\alpha}}. \tag{D.10}$$

As in previous proof, let us recall that

$$\begin{aligned}\|Z_{n+1}-m\|^{2p+2} \leq \left(\|V_n\|^2+\gamma_n^2\|U_{n+1}\|^2\right)^{p+1} &+2(p+1)\gamma_n\langle Z_n-m,\xi_{n+1}\rangle\left(\|V_n\|^2+\gamma_n^2\|U_{n+1}\|^2\right)^{p}\\ &+\sum_{k=2}^{p+1}\binom{p+1}{k}2^k\gamma_n^k\|Z_n-m\|^k\|\xi_{n+1}\|^k\left(\|V_n\|^2+\gamma_n^2\|U_{n+1}\|^2\right)^{p+1-k},\end{aligned} \tag{D.11}$$

with $V_n := Z_n - m - \gamma_n\Phi(Z_n)$. We now bound the expectation of each term on the right-hand side of previous inequality.

Bounding $(*) := \mathbb{E}\left[\left(\|V_n\|^2+\gamma_n^2\|U_{n+1}\|^2\right)^{p+1}\right]$**.** As in the proof of Lemma 7.5.4, we have

$$\begin{aligned}(*) \leq\ &\mathbb{E}\left[\|V_n\|^{2p+2}\right]+\sum_{k=1}^{p+1}\binom{p+1}{k}2^{p+k-1}\gamma_n^{2k}\left(1+2C^2c_\gamma^2\right)^{p+1-k}C^{2k}\mathbb{E}\left[\|Z_n-m\|^{2p+2}\right]\\ &+\sum_{k=1}^{p+1}\binom{p+1}{k}2^{p+k-1}\gamma_n^{2k}\left(1+2C^2c_\gamma^2\right)^{p+1-k}L_k\mathbb{E}\left[\|Z_n-m\|^{2p+2-2k}\right]\\ &+\sum_{k=1}^{p+1}\binom{p+1}{k}2^{2p}L_1^{p+1-k}\gamma_n^{2p+2}C^{2k}\mathbb{E}\left[\|Z_n-m\|^{2k}\right]+\sum_{k=1}^{p+1}\binom{p+1}{k}2^{2p}L_1^{p+1-k}\gamma_n^{2p+2}L_k\end{aligned}$$

Since for all integer $k \leq p-1$, there is a positive constant K_k such that for all $n \geq 1$, $\mathbb{E}\left[\|Z_n - m\|^{2k}\right] \leq K_k n^{-k\alpha}$, there are positive constants A_0, A_0', A_0'' such that

$$(*) \leq \mathbb{E}\left[\|V_n\|^{2p+2}\right] + A_0\gamma_n^2\mathbb{E}\left[\|Z_n - m\|^{2p+2}\right] + A_0'\gamma_n^2\mathbb{E}\left[\|Z_n - m\|^{2p}\right] + \frac{A_0''}{n^{(p+2)\alpha}}. \quad \text{(D.12)}$$

Bounding $(**) := \mathbb{E}\left[2(p+1)\gamma_n \langle Z_n - m, \xi_{n+1}\rangle \left(\|V_n\|^2 + \gamma_n^2 \|U_{n+1}\|^2\right)^p\right]$**.** As in the proof of Lemma 7.5.4, since (ξ_n) is a sequence of martingale differences adapted to the filtration $(\mathcal{F}_n)$,

$$\begin{aligned}
(**) &= 2(p+1)\gamma_n \sum_{k=1}^{p} \binom{p}{k} \gamma_n^{2k} \langle Z_n - m, \xi_{n+1}\rangle \|V_n\|^{2p-2k} \|U_{n+1}\|^{2k} \\
&\leq (p+1)\gamma_n \sum_{k=1}^{p} \binom{p}{k} \gamma_n^{2k} 2^{p-k-1} \left(1 + 2C^2c_\gamma^2\right)^{p-k} \mathbb{E}\left[\|Z_n - m\|^{2p-2k} \|U_{n+1}\|^{2k} \|\xi_{n+1}\|^2\right] \\
&+ (p+1)\gamma_n \sum_{k=1}^{p} \binom{p}{k} \gamma_n^{2k} 2^{p-k-1} \left(1 + 2C^2c_\gamma^2\right)^{p-k} \mathbb{E}\left[\|Z_n - m\|^{2p+2-2k} \|U_{n+1}\|^{2k}\right] \\
&+ (p+1)\gamma_n \sum_{k=1}^{p} \binom{p}{k} \gamma_n^{2p} 2^{2p-2k-1} L_1^{p-k} \left(\mathbb{E}\left[\|U_{n+1}\|^{2k} \|\xi_{n+1}\|^2\right] + \mathbb{E}\left[\|U_{n+1}\|^{2k} \|Z_n - m\|^2\right]\right).
\end{aligned}$$

Moreover, applying inequalities (D.4) and (D.5), let

$$\begin{aligned}
(\star) :&= \gamma_n \sum_{k=1}^{p} \binom{p}{k} 2^{p-k-1} \gamma_n^{2k} \left(1 + 2C^2c_\gamma^2\right)^{p-k} \mathbb{E}\left[\|Z_n - m\|^{2p-2k} \|U_{n+1}\|^{2k} \|\xi_{n+1}\|^2\right] \\
&\leq \sum_{k=1}^{p} \binom{p}{k} 2^{p+k-2} 3\gamma_n^{2k+1} \left(1 + 2C^2c_\gamma^2\right)^{p-k} \\
&\quad \mathbb{E}\left[\|Z_n - m\|^{2p-2k} f\left(X_{n+1}, Z_n\right)^2 \left(f\left(X_{n+1}, Z_n\right)^{2k} + C^{2k} \|Z_n - m\|^{2k}\right)\right] \\
&\quad + \sum_{k=1}^{p} \binom{p}{k} 2^{p+k-1} 3\gamma_n^{2k+1} \left(1 + 2C^2c_\gamma^2\right)^{p-k} C^2\mathbb{E}\left[\|Z_n - m\|^{2p-2k+2} f(X_{n+1}, Z_n)^{2k}\right] \\
&\quad + \sum_{k=1}^{p} \binom{p}{k} 2^{p+k-1} 3\gamma_n^{2k+1} \left(1 + 2C^2c_\gamma^2\right)^{p-k} C^2\mathbb{E}\left[\|Z_n - m\|^{2p-2k+2} C^{2k} \|Z_n - m\|^{2k}\right] \\
&\quad + \sum_{k=1}^{p} \binom{p}{k} 2^{p+k-2} 3\gamma_n^{2k+1} \left(1 + 2C^2c_\gamma^2\right)^{p-k} L_1\mathbb{E}\left[\|Z_n - m\|^{2p-2k} f\left(X_{n+1}, Z_n\right)^{2k}\right] \\
&\quad + \sum_{k=1}^{p} \binom{p}{k} 2^{p+k-2} 3\gamma_n^{2k+1} \left(1 + 2C^2c_\gamma^2\right)^{p-k} L_1\mathbb{E}\left[\|Z_n - m\|^{2p-2k} C^{2k} \|Z_n - m\|^{2k}\right].
\end{aligned}$$

Since $\mathbb{E}\left[f(X_{n+1},Z_n)^{2k+2}\middle|\mathcal{F}_n\right] \leq L_{k+1}$, and since for all $k \leq p-1$, $\mathbb{E}\left[\|Z_n-m\|^{2k}\right] \leq K_k n^{-k\alpha}$, there are positive constants A_1, A_2, A_3 such that

$$(\star) \leq A_1\gamma_n^3\mathbb{E}\left[\|Z_n-m\|^{2p+2}\right] + A_2\gamma_n^3\mathbb{E}\left[\|Z_n-m\|^{2p}\right] + \frac{A_3}{n^{(p+2)\alpha}}.$$

With analogous calculus, one can check that there are positive constants A_1', A_2', A_3' such that

$$(\star\star) \leq A_1'\gamma_n^3\mathbb{E}\left[\|Z_n-m\|^{2p+2}\right] + A_2'\gamma_n^3\mathbb{E}\left[\|Z_n-m\|^{2p}\right] + \frac{A_3'}{n^{(p+2)\alpha}}. \tag{D.13}$$

Bounding $(\star\star\star) := \sum_{k=2}^{p+1}\binom{p+1}{k}2^k\gamma_n^k\mathbb{E}\left[\|Z_n-m\|^k\|\xi_{n+1}\|^k\left(\|V_n\|^2+\gamma_n^2\|U_{n+1}\|^2\right)^{p+1-k}\right]$. First, thanks to inequality (D.5),

$$\begin{aligned}\|Z_n-m\|^k\|\xi_{n+1}\|^k &\leq 3^{k-1}\|Z_n-m\|^k\left(f(X_{n+1},Z_n)^k + 2^kC^k\|Z_n-m\|^k + L_1^{k/2}\right)\\ &\leq 6^{k-1}2C^k\|Z_n-m\|^{2k} + 3^{k-1}\|Z_n-m\|^2\left(f(X_{n+1},Z_n)^{2k}+L_1^k\right)\\ &\quad + \frac{3^{k-1}}{2}\|Z_n-m\|^{2k-2}.\end{aligned}$$

Then,

$$\begin{aligned}(\star\star\star) &\leq \sum_{k=2}^{p+1}\binom{p+1}{k}2^{p+k}3^{k-1}\gamma_n^kC^k\mathbb{E}\left[\|Z_n-m\|^{2k}\left(\|V_n\|^{2p+2-2k}+\gamma_n^{2p+2-2k}\|U_{n+1}\|^{2p+2k-2k}\right)\right]\\ &+\sum_{k=2}^{p+1}\binom{p+1}{k}2^p3^{k-1}\gamma_n^k\mathbb{E}\left[\|Z_n-m\|^2\left(f(X_{n+1},Z_n)^{2k}+L_1^k\right)\|V_n\|^{2p+2-2k}\right]\\ &+\sum_{k=2}^{p+1}\binom{p+1}{k}2^p3^{k-1}\gamma_n^k\mathbb{E}\left[\|Z_n-m\|^2\left(f(X_{n+1},Z_n)^{2k}+L_1^k\right)\gamma_n^{2p+2-2k}\|U_{n+1}\|^{2p+2k-2k}\right]\\ &+\sum_{k=2}^{p+1}\binom{p+1}{k}2^{p-1}3^{k-1}\gamma_n^k\mathbb{E}\left[\|Z_n-m\|^{2k-2}\left(\|V_n\|^{2p+2-2k}+\gamma_n^{2p+2-2k}\|U_{n+1}\|^{2p+2k-2k}\right)\right]\end{aligned}$$

With analogous calculus to the previous ones, one can check that there are positive constants A_1'', A_2'', A_3'' such that

$$(\star\star\star) \leq A_1''\gamma_n^2\mathbb{E}\left[\|Z_n-m\|^{2p+2}\right] + A_2''\gamma_n^2\mathbb{E}\left[\|Z_n-m\|^{2p}\right] + \frac{A_3''}{n^{(p+2)\alpha}}. \tag{D.14}$$

Thus, applying inequalities (D.12) to (D.14), there are positive constants B_0, B_1, B_2 such that

$$\mathbb{E}\left[\|Z_{n+1}-m\|^{2p+2}\right] \leq \mathbb{E}\left[\|V_n\|^{2p+2}\right] + B_0\gamma_n^2\mathbb{E}\left[\|Z_n-m\|^{2p+2}\right] + B_1\gamma_n^2\mathbb{E}\left[\|Z_n-m\|^{2p}\right] + \frac{B_2}{n^{(p+2)\alpha}}. \tag{D.15}$$

Then, in order to conclude the proof, we just have to bound $\mathbb{E}\left[\|V_n\|^{2p}\right]$.

Bounding $\mathbb{E}\left[\|V_n\|^{2p+2}\right]$. As in [CCGB15] (see Lemma 7.5.2), applying Proposition 7.2.1, one can check that there is a positive constant c and a rank n'_α such that for all $n \geq n'_\alpha$,

$$\begin{aligned} C\|Z_n-m\|^2 \mathbb{1}_{\{\|Z_n-m\|\leq cn^{1-\alpha}\}} &\geq \langle \Phi(Z_n), Z_n-m\rangle \mathbb{1}_{\{\|Z_n-m\|\leq cn^{1-\alpha}\}} \\ &\geq \frac{4}{c_\gamma n^{1-\alpha}}\|Z_n-m\|^2 \mathbb{1}_{\{\|Z_n-m\|\leq cn^{1-\alpha}\}}. \end{aligned}$$

Then, since $\|\Phi(Z_n)\|^2 \leq 2C^2\|Z_n-m\|^2 + 2L_1\gamma_n^2$, there is a rank n''_α such that for all $n \geq n''_\alpha$,

$$\|Z_n-m-\gamma_n\Phi(Z_n)\|^2 \mathbb{1}_{\{\|Z_n-m\|\leq cn^{1-\alpha}\}} \leq \left(1-\frac{3}{n}\right)\|Z_n-m\|^2 \mathbb{1}_{\{\|Z_n-m\|\leq cn^{1-\alpha}\}} + 2L_1\gamma_n^2.$$

Then, one can check that there are positive constants A'''_1, A'''_2 such that

$$\begin{aligned} &\mathbb{E}\left[\|Z_n-m-\gamma_n\Phi(Z_n)\|^{2p+2}\mathbb{1}_{\{\|Z_n-m\|\leq cn^{1-\alpha}\}}\right] \\ &\leq \sum_{k=0}^{p+1}\binom{p+1}{k}2^{p+1-k}L_1^{p+1-k}\gamma_n^{2(p+1-k)}\left(1-\frac{3}{n}\right)^k\mathbb{E}\left[\|Z_n-m\|^{2k}\right] \\ &\leq \left(1-\frac{3}{n}\right)^{p+1}\mathbb{E}\left[\|Z_n-m\|^{2p+2}\right] + A'''_1\gamma_n^2\mathbb{E}\left[\|Z_n-m\|^{2p}\right] \\ &+ \frac{A'''_2}{n^{(p+2)\alpha}}. \end{aligned}$$

Moreover, applying Cauchy-Schwarz's inequality, Markov's inequality and Lemma 7.5.4,

$$\begin{aligned} \mathbb{E}\left[\|Z_n-m\|^{2p+2}\mathbb{1}_{\{\|Z_n-m\|\geq cn^{1-\alpha}\}}\right] &\leq \sqrt{\mathbb{E}\left[\|Z_n-m\|^{4p+4}\right]}\sqrt{\mathbb{P}\left[\|Z_n-m\|\geq cn^{1-\alpha}\right]} \\ &\leq \sqrt{M_{2p+2}}\frac{\sqrt{\mathbb{E}\left[\|Z_n-m\|^{2q}\right]}}{c^q n^{q(1-\alpha)}} \\ &\leq \sqrt{M_{2p+2}}\frac{\sqrt{M_q}}{c^q n^{q(1-\alpha)}}, \end{aligned}$$

and one can conclude the proof applying inequality (D.15), taking $q \geq \frac{(p+2)\alpha}{1-\alpha}$ and taking a rank n_α such that for all $n \geq n_\alpha$, $\left(1-\frac{3}{n}\right)^{p+1} + \left(B_0 + A_1'''\right)\gamma_n^2 \leq \left(1-\frac{2}{n}\right)^{p+1}$. □

Remark D.3.1. *Note that in order to get the rate of convergence in quadratic mean of the Robbins-Monro algorithm, i.e in the case where $p = 1$, we just have to suppose that there are a positive integer $q \geq \frac{3\alpha}{1-\alpha}$ and a positive constant L_q such that for all $h \in H$, $\mathbb{E}\left[f\left(X,h\right)^{2q}\right] \leq L_q$.*

D.4 Proof of Lemma 7.5.2

Proof of Lemma 7.5.2. Using decomposition (D.2) and Cauchy-Schwarz's inequality, there are a positive constant c' and a rank n'_α such that for all $n \geq n'_\alpha$,

$$\begin{aligned}\left\|Z_{n+1}-m\right\|^2 &\leq \left\|I_H - \gamma_n \Gamma_m\right\|_{op}^2 \left\|Z_n - m\right\|^2 + \gamma_n^2 \left\|\xi_{n+1}\right\|^2 + 2\gamma_n \left\langle Z_n - m - \gamma_n \Phi(Z_n), \xi_{n+1}\right\rangle \\ &\quad + 2\gamma_n \left\langle Z_n - m, \delta_n\right\rangle \\ &\leq \left(1 - c'\gamma_n\right)^2 \left\|Z_n - m\right\|^2 + \gamma_n^2 \left\|U_{n+1}\right\|^2 + 2\gamma_n \left\langle Z_n - m, \xi_{n+1}\right\rangle + 2\gamma_n \left\|Z_n - m\right\| \left\|\delta_n\right\|.\end{aligned}$$

If $p = 1$, since there is a positive constant C_m such that for all $n \geq 1$, $\left\|\delta_n\right\| \leq C_m \left\|Z_n - m\right\|^2$, we have

$$2\left\|\delta_n\right\| \left\|Z_n - m\right\| \leq \frac{c'}{2}\gamma_n \left\|Z_n - m\right\|^2 + 2\frac{C_m^2}{c'} \left\|Z_n - m\right\|^4,$$

and since (ξ_n) is a martingale differences sequence adapted to the filtration $(\mathcal{F}_n)$, applying inequality (D.4), for all $n \geq n'_\alpha$,

$$\begin{aligned}\mathbb{E}\left[\left\|Z_{n+1}-m\right\|^2\right] &\leq \left(1 - c'\gamma_n\right)\mathbb{E}\left[\left\|Z_n - m\right\|^2\right] + 2\gamma_n^2 \mathbb{E}\left[\mathbb{E}\left[f\left(X_{n+1}, Z_n\right)\right)^2 |\mathcal{F}_n\right] + C^2 \left\|Z_n - m\right\|^2\right] \\ &\quad + \frac{c'}{2}\gamma_n \mathbb{E}\left[\left\|Z_n - m\right\|^2\right] + 2\gamma_n \frac{C_m^2}{c'}\mathbb{E}\left[\left\|Z_n - m\right\|^4\right] \\ &\leq \left(1 - \frac{c'}{2}\gamma_n + 2C^2\gamma_n^2\right)\mathbb{E}\left[\left\|Z_n - m\right\|^2\right] + 2\gamma_n^2 L_1 + 2\gamma_n \frac{C_m^2}{c'}\mathbb{E}\left[\left\|Z_n - m\right\|^4\right],\end{aligned}$$

and one can conclude the proof for $p = 1$ taking a rank n_α and a positive constant c such that for all $n \geq n_\alpha$, $1 - \frac{c'}{2}\gamma_n + 2C^2\gamma_n^2 \leq 1 - c\gamma_n$.

We suppose from now that $p \geq 2$. For all $n \geq n'_\alpha$,

$$\begin{aligned}\mathbb{E}\left[\|Z_{n+1}-m\|^{2p}\right] \leq (1-c'\gamma_n)\,\mathbb{E}\left[\|Z_n-m\|^2\,\|Z_{n+1}-m\|^{2p-2}\right] \\ + 2\gamma_n\mathbb{E}\left[\|Z_n-m\|\,\|\delta_n\|\,\|Z_{n+1}-m\|^{2p-2}\right] \\ + \gamma_n^2\mathbb{E}\left[\|U_{n+1}\|^2\,\|Z_{n+1}-m\|^{2p-2}\right] + 2\gamma_n\mathbb{E}\left[\langle Z_n-m,\xi_{n+1}\rangle\,\|Z_{n+1}-m\|^{2p-2}\right].\end{aligned} \tag{D.16}$$

Moreover, let us recall

$$\begin{aligned}\|Z_{n+1}-m\|^{2p-2} \leq \left(\|V_n\|^2+\gamma_n^2\,\|U_{n+1}\|^2\right)^{p-1} \\ + 2(p-1)\gamma_n\,\langle Z_n-m,\xi_{n+1}\rangle\left(\|V_n\|^2+\gamma_n^2\,\|U_{n+1}\|^2\right)^{p-2} \\ + \sum_{k=2}^{p-1}\binom{p-1}{k}2^k\gamma_n^k\,\|Z_n-m\|^k\,\|\xi_{n+1}\|^k\left(\|V_n\|^2+\gamma_n^2\,\|U_{n+1}\|^2\right)^{p-1-k},\end{aligned}$$

with $V_n := Z_n - m - \gamma_n\Phi(Z_n)$.

We now bound each term on the right-hand side of inequality (D.16).

Bounding $(1-c'\gamma_n)\,\mathbb{E}\left[\|Z_n-m\|^2\,\|Z_{n+1}-m\|^{2p+2}\right]$
First, since $\|V_n\|^2 \leq \left(1+2C^2\gamma_n^2\right)\|Z_n-m\|^2 + 2L_1\gamma_n^2$, let

$$\begin{aligned}(*) :&= \left(1-c'\gamma_n\right)\mathbb{E}\left[\|Z_n-m\|^2\left(\|V_n\|^2+\gamma_n^2\,\|U_{n+1}\|^2\right)^{p-1}\right] \\ &\leq \left(1-c'\gamma_n\right)\mathbb{E}\left[\|Z_n-m\|^2\left(\left(1+2C^2\gamma_n^2\right)\|Z_n-m\|^2+2L_1\gamma_n^2+\gamma_n^2\,\|U_{n+1}\|^2\right)^{p-1}\right] \\ &\leq \left(1-c'\gamma_n\right)\left(1+2C^2\gamma_n^2\right)^{p-1}\mathbb{E}\left[\|Z_n-m\|^{2p}\right] \\ &+ \sum_{k=0}^{p-2}\binom{p-1}{k}\left(1-c'\gamma_n\right)\gamma_n^{2(p-1-k)}\left(1+2C^2c_\gamma^2\right)^k\mathbb{E}\left[\|Z_n-m\|^{2k+2}\left(2L_1+\|U_{n+1}\|^2\right)^{p-1-k}\right].\end{aligned}$$

Applying inequality (D.4), $2L_1 + \|U_{n+1}\|^2 \leq 2\left(L_1 + C^2\,\|Z_n-m\|^2 + f\left(X_{n+1},Z_n\right)^2\right)$, and

since for all $n \geq n_\alpha$ we have $1 - c'\gamma_n \leq 1$,

$$
\begin{aligned}
(*) \leq & \left(1 - c'\gamma_n\right)\left(1 + 2C^2\gamma_n^2\right)^{p-1} \mathbb{E}\left[\|Z_n - m\|^{2p}\right] \\
& + \sum_{k=0}^{p-2} \binom{p-1}{k} \gamma_n^{2(p-1-k)} 6^{p-k-1} 3^{-1} \left(1 + 2C^2 c_\gamma^2\right)^k \mathbb{E}\left[\mathbb{E}\left[f(X_{n+1}, Z_n)^{2(p-1-k)} | \mathcal{F}_n\right] \|Z_n - m\|^{2k+2}\right] \\
& + \sum_{k=0}^{p-2} \binom{p-1}{k} \gamma_n^{2(p-1-k)} 6^{p-k-1} 3^{-1} \left(1 + 2C^2 c_\gamma^2\right)^k C^{2(p-1-k)} \mathbb{E}\left[\|Z_n - m\|^{2p}\right] \\
& + \sum_{k=0}^{p-2} \binom{p-1}{k} \gamma_n^{2(p-1-k)} 6^{p-k-1} 3^{-1} \left(1 + 2C^2 c_\gamma^2\right)^k L_1^{p-1-k} \mathbb{E}\left[\|Z_n - m\|^{2k+2}\right].
\end{aligned}
$$

Applying inequality (D.10), since $p \geq 2$ and since for all $k \leq p-2$, we have $2p - 1 - k \geq p + 1$, one can check that there is a positive constant A_1 such that

$$
(*) \leq \left(1 - c'\gamma_n + A_1\gamma_n^2\right) \mathbb{E}\left[\|Z_n - m\|^{2p}\right] + O\left(\frac{1}{n^{(p+1)\alpha}}\right). \tag{D.17}
$$

In the same way, applying Cauchy-Schwarz's inequality, let

$$
\begin{aligned}
(*)' :&= 2(p-1)\left(1 - c'\gamma_n\right)\gamma_n \mathbb{E}\left[\|Z_n - m\|^2 \langle Z_n - m, \xi_{n+1}\rangle \left(\|V_n\|^2 + \gamma_n^2 \|U_{n+1}\|^2\right)^{p-2}\right] \\
&\leq 2(p-1)\left(1 - c'\gamma_n\right)\gamma_n \mathbb{E}\left[\|Z_n - m\|^2 \langle Z_n - m, \xi_{n+1}\rangle \|V_n\|^{2(p-2)}\right] \\
&+ 2(p-1)\left(1 - c'\gamma_n\right) \sum_{k=1}^{p-2} \binom{p-2}{k} \gamma_n^{2k+1} \mathbb{E}\left[\|Z_n - m\|^3 \|\xi_{n+1}\| \|V_n\|^{2(p-2-k)} \|U_{n+1}\|^{2k}\right]
\end{aligned}
$$

Note that the last term on the right-hand side of previous inequality is equal to 0 if $p = 2$. Since (ξ_n) is a martingale differences sequence adapted to the filtration $(\mathcal{F}_n)$,

$$
2(p-1)\left(1 - c'\gamma_n\right)\gamma_n \mathbb{E}\left[\|Z_n - m\|^2 \langle V_n, \xi_{n+1}\rangle \|V_n\|^{2(p-2)}\right] = 0.
$$

Moreover, since $\|V_n\|^2 \leq \left(1 + 2C^2 c_\gamma^2\right)\|Z_n - m\|^2 + 2L_1\gamma_n^2$, and since

$$
\|Z_n - m\|^3 \|\xi_{n+1}\| \leq \frac{1}{2}\|Z_n - m\|^4 + \frac{1}{2}\|Z_n - m\|^2 \|\xi_{n+1}\|^2,
$$

we have

$$
\begin{aligned}
(*)' \leq (p-1)\sum_{k=1}^{p-2}\binom{p-2}{k}2^{p-3-k}\left(1+2C^2c_\gamma^2\right)^{p-2-k}\gamma_n^{2k+1}\mathbb{E}\left[\|Z_n-m\|^{2(p-k)}\|U_{n+1}\|^{2k}\right] \\
+(p-1)\sum_{k=1}^{p-2}\binom{p-2}{k}2^{p-3-k}\left(1+2C^2c_\gamma^2\right)^{p-2-k}\gamma_n^{2k+1}\mathbb{E}\left[\|Z_n-m\|^{2(p-1-k)}\|U_{n+1}\|^{2k}\|\xi_{n+1}\|^2\right] \\
+(p-1)\sum_{k=1}^{p-2}\binom{p-2}{k}2^{2p-5-2k}\gamma_n^{2p-1}L_1^{p-2-k}\mathbb{E}\left[\|Z_n-m\|^4\|U_{n+1}\|^{2k}\right] \\
+(p-1)\sum_{k=1}^{p-2}\binom{p-2}{k}2^{2p-5-2k}\gamma_n^{2p-1}L_1^{p-2-k}\mathbb{E}\left[\|Z_n-m\|^2\|U_{n+1}\|^{2k}\|\xi_{n+1}\|^2\right].
\end{aligned}
$$

As for $(*)$, applying inequalities (D.10), (D.4) and (D.5), one can check that there is a positive constant A_2 such that

$$
(*)' \leq A_2\gamma_n^2\mathbb{E}\left[\|Z_n-m\|^{2p}\right]+O\left(\frac{1}{n^{(p+1)\alpha}}\right). \tag{D.18}
$$

In the same way, since $\|V_n\| \leq \left(1+2c_\gamma^2C^2\right)\|Z_n-m\|^2+2L_1\gamma_n^2$ and since

$$
\|Z_n-m\|^k\|\xi_{n+1}\|^k \leq \frac{1}{2}\|Z_n-m\|^{2k-2}+\frac{1}{2}\|Z_n-m\|^2\|\xi_{n+1}\|^{2k},
$$

we have

$$
\begin{aligned}
(*)'' :&= \left(1-c'\gamma_n\right)\sum_{k=2}^{p-1}\binom{p-1}{k}2^k\gamma_n^k\mathbb{E}\left[\|Z_n-m\|^{k+2}\|\xi_{n+1}\|^k\left(\|V_n\|^2+\gamma_n^2\|U_{n+1}\|^2\right)^{p-1-k}\right] \\
&\leq \frac{1}{2}\sum_{k=2}^{p-1}\binom{p-1}{k}2^k\gamma_n^k \\
&\mathbb{E}\left[\|Z_n-m\|^{2k}\left(\left(1+2C^2c_\gamma^2\right)\|Z_n-m\|^2+2L_1\gamma_n^2+\gamma_n^2\|U_{n+1}\|^2\right)^{p-1-k}\right] \\
&+\frac{1}{2}\sum_{k=2}^{p-1}\binom{p-1}{k}2^k\gamma_n^k \\
&\mathbb{E}\left[\|Z_n-m\|^4\|\xi_{n+1}\|^{2k}\left(\left(1+2C^2c_\gamma^2\right)\|Z_n-m\|^2+2L_1\gamma_n^2+\gamma_n^2\|U_{n+1}\|^2\right)^{p-1-k}\right].
\end{aligned}
$$

With analogous calculus to the previous ones, applying inequality (D.10), one can check that

there are positive constants A_3, A_4 such that

$$(*)'' \leq A_3\gamma_n^2\mathbb{E}\left[\|Z_n - m\|^{2p}\right] + A_4\gamma_n^2\mathbb{E}\left[\|Z_n - m\|^{2p+2}\right] + O\left(\frac{1}{n^{(p+1)\alpha}}\right). \tag{D.19}$$

Finally, applying inequalities (D.17) to (D.19), there are positive constants B_0, B_1, B_2 such that

$$\begin{aligned}\mathbb{E}\left[\left(1 - c'\gamma_n\right)\|Z_n - m\|^2\|Z_{n+1} - m\|^{2p-2}\right] \leq \left(1 - c'\gamma_n + B_0\gamma_n^2\right)\mathbb{E}\left[\|Z_n - m\|^{2p}\right] \\ + B_2\gamma_n^2\mathbb{E}\left[\|Z_n - m\|^{2p+2}\right] + \frac{B_1}{n^{(p+1)\alpha}}.\end{aligned}$$

Bounding $2\gamma_n\mathbb{E}\left[\|Z_n - m\|\,\|\delta_n\|\,\|Z_{n+1} - m\|^{2p-2}\right]$**.** First, let

$$\begin{aligned}(*) :&= 2\gamma_n\mathbb{E}\left[\|Z_n - m\|\,\|\delta_n\|\left(\|V_n\|^2 + \gamma_n^2\|U_{n+1}\|^2\right)^{p-1}\right] \\ &\leq 2^{p-1}\gamma_n\mathbb{E}\left[\|Z_n - m\|\,\|\delta_n\|\,\|V_n\|^{2p-2}\right] + 2^{p-1}\gamma_n^{2p-1}\mathbb{E}\left[\|Z_n - m\|\,\|\delta_n\|\,\|U_{n+1}\|^{2p-2}\right].\end{aligned}$$

Moreover, since $\|\delta_n\| \leq C_m\|Z_n - m\|^2$ and since $\|V_n\|^2 \leq \left(1 + 2C^2c_\gamma^2\right)\|Z_n - m\|^2 + 2L_1\gamma_n^2$, let

$$\begin{aligned}(\star) &:= 2^{p-1}\gamma_n\mathbb{E}\left[\|Z_n - m\|\,\|\delta_n\|\,\|V_n\|^{2p-2}\right] \\ &\leq 2^{2p-3}C_m\left(1 + 2C^2c_\gamma^2\right)^{p-1}\gamma_n\mathbb{E}\left[\|Z_n - m\|^{2p+1}\right] + 2^{3p-4}L_1^{p-1}C_m\gamma_n^{2p-1}\mathbb{E}\left[\|Z_n - m\|^3\right] \\ &\leq \frac{2^{4p-6}C_m^2\left(1 + 2C^2c_\gamma^2\right)^{2p-2}}{c'}\gamma_n\mathbb{E}\left[\|Z_n - m\|^{2p+2}\right] + \frac{1}{4}c'\gamma_n\mathbb{E}\left[\|Z_n - m\|^{2p}\right] \\ &+ 2^{3p-5}C_mL_1^{p-1}\gamma_n^{2p-1}\mathbb{E}\left[\|Z_n - m\|^2\right] + 2^{3p-5}C_mL_1^{p-1}\gamma_n^{2p-1}\mathbb{E}\left[\|Z_n - m\|^4\right].\end{aligned}$$

Then, since $p \geq 2$, there are positive constants B_1, B_2 such that

$$\begin{aligned}2^{p-1}\gamma_n\mathbb{E}\left[\|Z_n - m\|\,\|\delta_n\|\,\|V_n\|^{2p-2}\right] \leq \left(\frac{c'}{4}\gamma_n + B_1\gamma_n^2\right)\mathbb{E}\left[\|Z_n - m\|^{2p}\right] \\ + B_2\gamma_n\mathbb{E}\left[\|Z_n - m\|^{2p+2}\right] + O\left(\frac{1}{n^{(p+1)\alpha}}\right).\end{aligned}$$

In the same way, since $\|\delta_n\| \leq C_m \|Z_n - m\|^2$, and applying inequality (D.4), let

$$\begin{aligned}(\star\star) &:= 2^{p-1}\gamma_n^{2p-1}\mathbb{E}\left[\|Z_n - m\| \|\delta_n\| \|U_{n+1}\|^{2p-2}\right] \\ &\leq 2^{3p-5}C_m\gamma_n^{2p-1}\mathbb{E}\left[\mathbb{E}\left[f(X_{n+1})^{2p-2}|\mathcal{F}_n\right]\|Z_n - m\|^2\right] + 2^{3p-5}C^{2p-2}C_m\gamma_n^{2p-1}\mathbb{E}\left[\|Z_n - m\|^{2p}\right] \\ &+ 2^{3p-5}C_m\gamma_n^{2p-1}\mathbb{E}\left[\mathbb{E}\left[f(X_{n+1})^{2p-2}|\mathcal{F}_n\right]\|Z_n - m\|^4\right] \\ &+ 2^{3p-5}C^{2p-2}C_m\gamma_n^{2p-1}\mathbb{E}\left[\|Z_n - m\|^{2p+2}\right].\end{aligned}$$

Since $p \geq 2$, applying inequality (D.10), there are positive constant A_1, A_2, A_3 such that

$$\begin{aligned}2\gamma_n\mathbb{E}\left[\|Z_n - m\| \|\delta_n\| \left(\|V_n\|^2 + \gamma_n^2\|U_{n+1}\|^2\right)^{p-1}\right] &\leq \left(\frac{c'}{4}\gamma_n + A_1\gamma_n^2\right)\mathbb{E}\left[\|Z_n - m\|^{2p}\right] \\ &+ A_2\gamma_n\mathbb{E}\left[\|Z_n - m\|^{2p+2}\right] + \frac{A_3}{n^{(p+1)\alpha}}.\end{aligned} \tag{D.20}$$

In a similar way, since (ξ_n) is a martingale differences sequence adapted to the filtration $(\mathcal{F}_n)$ and since Z_n is $\mathcal{F}_n$-measurable, let

$$\begin{aligned}(*)' &:= 4(p-1)\gamma_n^2\mathbb{E}\left[\|Z_n - m\| \|\delta_n\| \langle Z_n - m, \xi_{n+1}\rangle \left(\|V_n\|^2 + \gamma_n^2\|U_{n+1}\|^2\right)^{p-2}\right] \\ &\leq 4(p-1)\sum_{k=1}^{p-2}\binom{p-2}{k}\gamma_n^{2k+2}\mathbb{E}\left[\|Z_n - m\| \|\delta_n\| \langle Z_n - m, \xi_{n+1}\rangle \|V_n\|^{2(p-2-k)} \|U_{n+1}\|^{2k}\right].\end{aligned}$$

Note that this term is equal to 0 if $p = 2$. Moreover, since $\|\delta_n\| \leq C_m \|Z_n - m\|^2$, applying Cauchy-Schwarz's inequality,

$$\begin{aligned}\|Z_n - m\| \|\delta_n\| |\langle Z_n - m, \xi_{n+1}\rangle| &\leq C_m \|Z_n - m\|^4 \|\xi_{n+1}\| \\ &\leq \frac{C_m}{2}\|Z_n - m\|^4 \left(1 + \|\xi_{n+1}\|^2\right).\end{aligned}$$

Then,

$$\begin{aligned}(*)' &\leq 2C_m(p-1)\sum_{k=1}^{p-2}\binom{p-2}{k}\gamma_n^{2k+2}\mathbb{E}\left[\|V_n\|^{2(p-2-k)}\|Z_n - m\|^4\|U_{n+1}\|^{2k}\right] \\ &+ 2C_m(p-1)\sum_{k=1}^{p-2}\binom{p-2}{k}\gamma_n^{2k+2}\mathbb{E}\left[\|V_n\|^{2(p-2-k)}\|Z_n - m\|^4\|U_{n+1}\|^{2k}\|\xi_{n+1}\|^2\right].\end{aligned}$$

Thus, applying inequalities (D.10), (D.4) and (D.5), one can check that there are positive

constants A_1', A_2', A_3' such that

$$(*)' \leq A_1'\gamma_n^2 \mathbb{E}\left[\|Z_n - m\|^{2p}\right] + A_2'\gamma_n^2 \mathbb{E}\left[\|Z_n - m\|^{2p+2}\right] + \frac{A_3'}{n^{(p+1)\alpha}}. \tag{D.21}$$

Finally, with similar calculus, let

$$\begin{aligned}
(*)'' :&= 2\gamma_n \mathbb{E}\left[\|Z_n - m\| \|\delta_n\| \sum_{k=2}^{p-1} \binom{p-1}{k} 2^k \gamma_n^k \|Z_n - m\|^k \|\xi_{n+1}\|^k \left(\|V_n\|^2 + \gamma_n^2 \|U_{n+1}\|^2\right)^{p-1-k}\right] \\
&\leq C_m \sum_{k=2}^{p-1} \binom{p-1}{k} 2^{k-1} \gamma_n^{k+1} \mathbb{E}\left[\|Z_n - m\|^{2k+2} \left(\|V_n\|^2 + \gamma_n^2 \|U_{n+1}\|^2\right)^{p-1-k}\right] \\
&+ C_m \sum_{k=2}^{p-1} \binom{p-1}{k} 2^{k-1} \gamma_n^{k+1} \mathbb{E}\left[\|Z_n - m\|^2 \|\xi_{n+1}\|^{2k} \left(\|V_n\|^2 + \gamma_n^2 \|U_{n+1}\|^2\right)^{p-1-k}\right] \\
&+ C_m \sum_{k=2}^{p-1} \binom{p-1}{k} 2^{k-1} \gamma_n^{k+1} \mathbb{E}\left[\|Z_n - m\|^{2k+4} \left(\|V_n\|^2 + \gamma_n^2 \|U_{n+1}\|^2\right)^{p-1-k}\right] \\
&+ C_m \sum_{k=2}^{p-1} \binom{p-1}{k} 2^{k-1} \gamma_n^{k+1} \mathbb{E}\left[\|Z_n - m\|^4 \|\xi_{n+1}\|^{2k} \left(\|V_n\|^2 + \gamma_n^2 \|U_{n+1}\|^2\right)^{p-1-k}\right].
\end{aligned}$$

Thus, applying inequalities (D.10), (D.4) and (D.5), one can check that there are positive constants A_0'', A_1'', A_2'' such that

$$(*)'' \leq A_0''\gamma_n^2 \mathbb{E}\left[\|Z_n - m\|^{2p}\right] + A_1''\gamma_n^2 \mathbb{E}\left[\|Z_n - m\|^{2p+2}\right] + \frac{A_2''}{n^{(p+1)\alpha}}. \tag{D.22}$$

Finally, applying inequalities (D.20) to (D.22), there are positive constants B_0', B_1', B_2' such that

$$\begin{aligned}
2\gamma_n \mathbb{E}\left[\|Z_n - m\| \|\delta_n\| \|Z_{n+1} - m\|^{2p-2}\right] \leq &\left(\frac{1}{4}c'\gamma_n + B_0'\gamma_n^2\right) \mathbb{E}\left[\|Z_n - m\|^{2p}\right] \\
&+ B_1'\gamma_n \mathbb{E}\left[\|Z_n - m\|^{2p+2}\right] + \frac{B_2'}{n^{(p+1)\alpha}}.
\end{aligned}$$

Bounding $\gamma_n^2 \mathbb{E}\left[\|U_{n+1}\|^2 \|Z_{n+1} - m\|^{2p-2}\right]$ First, since $\|V_n\|^2 \leq \left(1 + 2c_\gamma^2 C^2\right) \|Z_n - m\|^2 + 2L_1\gamma_n^2$, let

$$\begin{aligned}
(\star) &:= \gamma_n^2 \mathbb{E}\left[\|U_{n+1}\|^2 \left(\|V_n\|^2 + \gamma_n^2 \|U_{n+1}\|^2\right)^{p-1}\right] \\
&\leq 3^{p-2} \left(1 + 2C^2 c_\gamma^2\right)^{p-1} \gamma_n^2 \mathbb{E}\left[\|U_{n+1}\|^2 \|Z_n - m\|^{2p-2}\right] + 3^{p-2}\gamma_n^{2p} \mathbb{E}\left[\|U_{n+1}\|^{2p}\right] \\
&+ 3^{p-2} 2^{p-1} L_1^{p-1} \gamma_n^{2p} \mathbb{E}\left[\|U_{n+1}\|^2\right].
\end{aligned}$$

Thus, applying inequalities (D.10), (D.4) and (D.5), there are positive constants A_0, A_1 such that

$$\gamma_n^2 \mathbb{E}\left[\|U_{n+1}\|^2 \left(\|V_n\|^2 + \gamma_n^2 \|U_{n+1}\|^2\right)^{p-1}\right] \leq A_0 \gamma_n^2 \mathbb{E}\left[\|Z_n - m\|^{2p}\right] + \frac{A_1}{n^{(p+1)\alpha}}. \tag{D.23}$$

In the same way, let

$$\begin{aligned}(*) :&= \left|2(p-1)\gamma_n^3 \mathbb{E}\left[\|U_{n+1}\|^2 \langle Z_n - m, \xi_{n+1}\rangle \left(\|V_n\|^2 + \gamma_n^2 \|U_{n+1}\|^2\right)^{p-2}\right]\right| \\ &\leq (p-1)\gamma_n^3 \mathbb{E}\left[\left(\|U_{n+1}\|^2 \|Z_n - m\|^2 + \|U_{n+1}\|^2 \|\xi_{n+1}\|^2\right)\left(\|V_n\|^2 + \gamma_n^2 \|U_{n+1}\|^2\right)^{p-2}\right].\end{aligned}$$

Thus, applying inequalities (D.10), (D.4) and (D.5), one can check that there are positive constants A_0', A_1' such that

$$\begin{aligned}\left|2(p-1)\gamma_n^3 \mathbb{E}\left[\|U_{n+1}\|^2 \langle Z_n - m, \xi_{n+1}\rangle \left(\|V_n\|^2 + \gamma_n^2 \|U_{n+1}\|^2\right)^{p-2}\right]\right| &\leq A_0' \gamma_n^2 \mathbb{E}\left[\|Z_n - m\|^{2p}\right] \\ &\quad + \frac{A_1'}{n^{(p+1)\alpha}}.\end{aligned} \tag{D.24}$$

Finally, let

$$\begin{aligned}(*)' :&= \sum_{k=2}^{p-1} \binom{p-1}{k} \gamma_n^{k+2} \mathbb{E}\left[\|U_{n+1}\|^2 \|Z_n - m\|^k \|\xi_{n+1}\|^k \left(\|V_n\|^2 + \gamma_n^2 \|U_{n+1}\|^2\right)^{p-1-k}\right] \\ &\leq \frac{1}{2}\sum_{k=2}^{p-1} \binom{p-1}{k} \gamma_n^{k+2} \mathbb{E}\left[\|U_{n+1}\|^2 \left(\|Z_n - m\|^{2k} + \|\xi_{n+1}\|^{2k}\right)\left(\|V_n\|^2 + \gamma_n^2 \|U_{n+1}\|^2\right)^{p-1-k}\right].\end{aligned}$$

Applying inequalities (D.10), (D.4) and (D.5), there are positive constants A_0'', A_1'' such that

$$(*)' \leq A_0'' \gamma_n^2 \mathbb{E}\left[\|Z_n - m\|^{2p}\right] + \frac{A_1''}{n^{(p+1)\alpha}}. \tag{D.25}$$

Thus, applying inequalities (D.24) and (D.25), there are positive constants B_0'', B_1'' such that

$$\gamma_n^2 \mathbb{E}\left[\|U_{n+1}\|^2 \|Z_{n+1} - m\|^{2p-2}\right] \leq B_0'' \gamma_n^2 \mathbb{E}\left[\|Z_n - m\|^{2p}\right] + \frac{B_1''}{n^{(p+1)\alpha}}.$$

Bounding $2\gamma_n \mathbb{E}\left[\langle Z_n - m, \xi_{n+1}\rangle \|Z_{n+1} - m\|^{2p-2}\right]$

First, since (ξ_n) is a martingale differences sequence adapted to the filtration $(\mathcal{F}_n)$, let

$$\begin{aligned}(*) :&= 2\gamma_n \mathbb{E}\left[\langle \xi_{n+1}, Z_n - m\rangle \left(\|V_n\|^2 + \gamma_n^2 \|U_{n+1}\|^2\right)^{p-1}\right] \\ &= 2\sum_{k=1}^{p-1} \binom{p-1}{k} \gamma_n^{2k+1} \mathbb{E}\left[\langle \xi_{n+1}, Z_n - m\rangle \|V_n\|^{2(p-1-k)} \|U_{n+1}\|^{2k}\right] \\ &\leq \sum_{k=1}^{p-1} \binom{p-1}{k} \gamma_n^{2k+1} \mathbb{E}\left[\left(\|\xi_{n+1}\|^2 + \|Z_n - m\|^2\right) \|V_n\|^{2(p-1-k)} \|U_{n+1}\|^{2k}\right].\end{aligned}$$

Thus, applying inequalities (D.10), (D.4) and (D.5), one can check that there are positive constants A_0, A_1 such that

$$2\gamma_n \mathbb{E}\left[\langle \xi_{n+1}, V_n\rangle \left(\|V_n\|^2 + \gamma_n^2 \|U_{n+1}\|^2\right)^{p-1}\right] \leq A_0 \gamma_n^2 \mathbb{E}\left[\|Z_n - m\|^{2p}\right] + \frac{A_1}{n^{(p+1)\alpha}}. \quad \text{(D.26)}$$

In the same way, applying Cauchy-Schwarz's inequality, let

$$\begin{aligned}(*)' :&= 4(p-1)\gamma_n^2 \mathbb{E}\left[\langle Z_n - m, \xi_{n+1}\rangle^2 \left(\|V_n\|^2 + \gamma_n^2 \|U_{n+1}\|^2\right)^{p-2}\right] \\ &\leq 2^{p-1}(p-1)\gamma_n^2 \mathbb{E}\left[\|Z_n - m\|^2 \|\xi_{n+1}\|^2 \left(\|V_n\|^{2p-4} + \gamma_n^{2p-4} \|U_{n+1}\|^{2p-4}\right)\right]\end{aligned}$$

Thus, since $p \geq 2$, applying inequalities (D.10), (D.4) and (D.5), one can check that there are positive constants A_0', A_1' such that

$$4(p-1)\gamma_n^2 \mathbb{E}\left[\langle Z_n - m, \xi_{n+1}\rangle^2 \left(\|V_n\|^2 + \gamma_n^2 \|U_{n+1}\|^2\right)^{p-2}\right] \leq A_0' \gamma_n^2 \mathbb{E}\left[\|Z_n - m\|^{2p}\right] + \frac{A_1'}{n^{(p+1)\alpha}}. \quad \text{(D.27)}$$

Finally, let

$$\begin{aligned}(*)'' :&= 2\sum_{k=2}^{p-1} \binom{p-1}{k} \gamma_n^{k+1} \mathbb{E}\left[\langle Z_n - m, \xi_{n+1}\rangle \|Z_n - m\|^k \|\xi_{n+1}\|^k \left(\|V_n\|^2 + \gamma_n^2 \|U_{n+1}\|^2\right)^{p-k}\right] \\ &\leq \sum_{k=2}^{p-1} \binom{p-1}{k} 2^{p-k-2} \gamma_n^{k+1} \\ &\quad \mathbb{E}\left[\left(\|Z_n - m\|^2 + \|\xi_{n+1}\|^2\right) \left(\|Z_n - m\|^{2k} + \|\xi_{n+1}\|^{2k}\right) \left(\|V_n\|^{2p-2k} + \gamma_n^{2p-2k} \|U_{n+1}\|^{2p-2k}\right)\right]\end{aligned}$$

Thus, applying inequalities (D.10), (D.4) and (D.5), one can check that there are positive constants A_0'', A_1'', A_2'' such that

$$(*)'' \leq A_0'' \gamma_n^2 \mathbb{E}\left[\|Z_n - m\|^{2p}\right] + A_1'' \gamma_n^2 \mathbb{E}\left[\|Z_n - m\|^{2p+2}\right] + \frac{A_2''}{n^{(p+1)\alpha}}. \quad \text{(D.28)}$$

Thus, applying inequalities (D.26) to (D.28), there are positive constants B_0''', B_1''', B_2''' such that

$$2\gamma_n \mathbb{E}\left[\langle Z_n - m, \xi_{n+1}\rangle \|Z_{n+1} - m\|^{2p-2}\right] \leq B_0''' \gamma_n^2 \mathbb{E}\left[\|Z_n - m\|^{2p}\right] + B_1''' \gamma_n^2 \mathbb{E}\left[\|Z_n - m\|^{2p+2}\right] + \frac{B_2'''}{n^{(p+1)\alpha}}.$$

Conclusion We have proved that there are positive constants c_0, C_1, C_2 such that for all $n \geq n'_\alpha$;

$$\mathbb{E}\left[\|Z_{n+1} - m\|^{2p}\right] \leq \left(1 - \frac{c'}{2}\gamma_n + c_0\gamma_n^2\right) \mathbb{E}\left[\|Z_n - m\|^{2p}\right] + C_1\gamma_n \mathbb{E}\left[\|Z_n - m\|^{2p+2}\right] + \frac{C_2}{n^{(p+1)\alpha}}.$$

Thus, there are a positive constant c and a rank $n_\alpha \geq n'_\alpha$ such that for all $n \geq n_\alpha$, $1 - \frac{c'}{2}\gamma_n + c_0\,\gamma_n^2 \leq 1 - c\gamma_n$, and in a particular case, for all $n \geq n_\alpha$,

$$\mathbb{E}\left[\|Z_{n+1} - m\|^{2p}\right] \leq (1 - c\gamma_n)\, \mathbb{E}\left[\|Z_n - m\|^{2p}\right] + C_1\gamma_n \mathbb{E}\left[\|Z_n - m\|^{2p+2}\right] + \frac{C_2}{n^{(p+1)\alpha}}. \quad \text{(D.29)}$$

□

Quatrième partie

Estimation des paramètres d'une distribution sphérique tronquée

Chapitre 8

An averaged projected Robbins-Monro algorithm for estimating the parameters of a truncated spherical distribution

Résumé

L'objectif de ce travail est de proposer un algorithme pour ajuster une sphère sur un nuage de point 3D bruité distribué autour d'une sphère complète, ou d'une sphère tronquée. Un algorithme de type back-fitting a été proposé par [BP14]) mais aucun résultat de convergence n'a été démontré pour le cas de la sphère tronquée. Un des soucis majeur est que la fonction que l'on voudrait minimiser n'est convexe que sur un sous espace. Pour pallier cet inconvénient, nous introduisons un algorithme de gradient stochastique projeté (voir Chapitre 1) et sa version moyennée qui permettent d'estimer le centre et le rayon de la sphère. On donne des résultats asymptotiques tels que la convergence presque sûre de ces algorithmes (Théorème 8.4.1) ainsi que la normalité asymptotique de l'algorithme moyenné (Théorème 8.4.4). De plus, quelques résultats non-asymptotiques sont donnés, tels que les vitesses de convergence en moyenne quadratique (Théorèmes 8.4.2 et 8.4.3). Quelques simulations montrent l'efficacité des algorithmes pour des donnée simulées, pour des échantillons de petite taille à taille moyenne.

This Chapter is based on a work with Bruno Portier ([GBP16]).

Abstract

The objective of this work is to propose a new algorithm to fit a sphere on a noisy 3D point cloud distributed around a complete or a truncated sphere. More precisely, we introduce a projected Robbins-Monro algorithm and its averaged version for estimating the center and the radius of the sphere. We give asymptotic results such as the almost sure convergence of these algorithms as well as the asymptotic normality of the averaged algorithm. Furthermore, some non-asymptotic results will be given, such as the rates of convergence in quadratic mean. Some numerical experiments show the efficiency of the proposed algorithm on simulated data for small to moderate sample sizes.

8.1 Introduction

Primitive shape extraction from data is a recurrent problem in many research fields such as archeology [Tho55], medicine [ZSW+07], mobile robotics [MNP08], motion capture [STG01] and computer vision [Rab06, LW14]. This process is of primary importance since it provides a high level information on the data structure.

First works focused on the case of 2D shapes (lines, circles), but recent technologies enable to work with three dimensional data. For instance, in computer vision, depth sensors provide 3D point clouds representing the scene in addition to usual color images. In this work, we are interested in the estimation of the center $\mu \in \mathbb{R}^3$ and the radius $r > 0$ of a sphere from a set of 3D noisy data. In practical applications, only a discrete set of noisy measurements is available. Moreover, sample points are usually located only near a portion of the spherical surface. Two kinds of problem can be distinguished : shape detection and shape fitting.

Shape detection consists in finding a given shape in the whole data without any prior knowledge on which observations belong to it. In that case, the data set may represent several objects of different nature and may therefore contain a high number of outliers. Two main methods are used in practise to solve this problem. The Hough transform [ANC13] performs a discretization of the parameter space. Each observation is associated to a set of parameters corresponding to all possible shapes that could explain the sample point. Then, a voting strategy is applied to select the parameter vectors of the detected shapes. The advantage of this method is that several instances of the shape can be detected. However, a large amount of memory is required to discretize the parameter space, especially in the case of three dimensional models. The RANSAC (RANdom SAmple Consensus) paradigm [FB81, SWK07] is a probabilistic method based on random sampling. Observations are randomly selected among the whole data set and candidate models are generated. Then, shapes can be detected thanks to an individual scoring scheme. The success of the method depends on a given probability related to the number of sampling and the fraction of points belonging to the shape.

The shape fitting problem assumes that all the data points belong to the shape. For example, spherical fitting techniques have been used in several domains such as industrial inspection [JJ98], GPS localization [BP12], robotics [VHSR05] and 3D modelling [TCL15]. Geometric and algebric methods have been proposed [Lan87, RTKD03, AS14] for parameters estimation. Moreover, let us note that fitting methods are generally applied for shape detection as a post-processing step in order to refine the parameters of the detected shapes [TCL15].

In a recent paper, Brazey and Portier [BP14] introduced a new spherical probability density function belonging to the family of elliptical distributions, and designed to model points spread near a spherical surface. This probability density function depends on three parameters, namely a center $\mu \in \mathbb{R}^3$, a radius $r > 0$ and a dispersion parameter $\sigma > 0$. In their paper, the model is formulated in a general form in $\mathbb{R}^d$. To estimate μ and r, a backfitting algorithm (see e.g. [BF85]) similar to the one used in [Lan87] is employed. A convergence result is given in the case of the complete sphere. However, no result is established in the case of a truncated sphere while simulations showed the efficiency of the algorithm.

The objective of this work is to propose a new algorithm to fit a sphere on a noisy 3D point cloud distributed around a complete or a truncated sphere. We shall assume that the observations are independent realizations of a random vector X defined as

$$X = \mu + r\, W\, U_\Omega, \tag{8.1}$$

where W is a positive real random variable such that $\mathbb{E}[W] = 1$, U_Ω is uniformly distributed on a measurable subset Ω of the unit sphere of $\mathbb{R}^3$, W and U_Ω are independent. Parameters $\mu \in \mathbb{R}^3$ and $r > 0$ are respectively the center and the radius of the sphere we are trying to adjust to the point cloud. Random variable W allows to model the fluctuations of points in the normal direction of the sphere. When Ω coincides with the complete sphere, then the distribution of X is spherical (see e.g. [Mui09]). Indeed, if we set $Y = (X - \mu)/r$, then the distribution of Y is rotationally invariant.

We are interested in estimating center μ and radius r. As $\|U_\Omega\| = 1$, we easily deduce from (8.1) that

$$\mu = \mathbb{E}\left[X - r\,\frac{(X-\mu)}{\|X-\mu\|}\right] \tag{8.2}$$

$$r = \mathbb{E}\left[\|X-\mu\|\right]. \tag{8.3}$$

It is clear that from these two equations, we cannot deduce explicit estimators of parameters μ and r using the method of moments since each parameter depends on the other. To overcome this problem, we can use a backfitting type algorithm (as in [BP14]) or introduce a recursive stochastic algorithm, which seems well-suited for this problem since equations (8.2) and (8.3) can also be derived from the local minimization of the following quadratic criteria

$$G(\mu, r) := \frac{1}{2}\,\mathbb{E}\left[(\|X-\mu\| - r)^2\right]. \tag{8.4}$$

Stochastic algorithms, and more precisely Robbins-Monro algorithms, are effective and fast methods (see e.g. [Duf97, KY03, RM51]). They do not need too much computational efforts and can easily be updated, which make of them good candidates to deal with big data for example. However, usual sufficient conditions to prove the convergence of this kind of algorithm are sometimes not satisfied and it is necessary to modify the basic algorithm. We can, for example, introduce a projected version of the Robbins-Monro algorithm which consists in keeping the usual estimators in a nice subspace with the help of a projection. Such an algorithm has been recently considered in [BF12] and [LJSB12].

In this paper, due to the non global convexity of function G, we estimate parameters μ and r using a projected Robbins-Monro algorithm. We also propose an averaged algorithm which consists in averaging the projected algorithm. In general, this averaged algorithm allows to improve the rate of convergence of the basic estimators, or to reduce the variance, or not to have to make a good choice of the step sequence, which can be as exhaustive as to estimate the parameters. It is widely used when having to deal with Robbins-Monro algorithms (see [PJ92]) or [Pel98] amoung others).

This paper is organized as follows. In Section 2, we specify the framework and assumptions. After a short explanation on the non-convergence of the Robbins-Monro algorithm, the projected algorithm and its averaged version are introduced in Section 3. Section 4 is concerned with the convergence results. Some simulation experiments are provided in Section 5, showing the efficiency of the algorithms. Proofs of the different results are postponed in a supplementary file.

8.2 Framework and assumptions

We consider in this paper a more general framework than the one described in the introduction. Let X be a random vector of $\mathbb{R}^d$ with $d \geq 2$. Let F denotes the distribution of X. We assume that X can be decomposed under the form

$$X = \mu + r\, W\, U_\Omega. \tag{8.5}$$

where $\mu \in \mathbb{R}^d$, $r > 0$, W is a positive real continuous random variable (and with a bounded density if $d = 2$), U_Ω is uniformly distributed on a measurable subset Ω of the unit sphere of $\mathbb{R}^d$. Moreover, let us suppose that W and U_Ω are independent.

Model (8.5) allows to model a point cloud of $\mathbb{R}^d$ spread around a complete or truncated sphere of center $\mu \in \mathbb{R}^d$ and radius $r > 0$. Random vector U_Ω defines the position of the

points on the sphere and random variable W defines the fluctuations in the normal direction of the sphere. As mentioned in the introduction, when Ω is the complete unit sphere, then the distribution of X is spherical.

When W satisfies the condition $\mathbb{E}[W] = 1$, the radius r is identifiable and can be directly estimated. Indeed, since $\|U_\Omega\| = 1$, then $\|X - \mu\| = rW$ and $\mathbb{E}[\|X - \mu\|] = r\,\mathbb{E}[W] = r$. However, this condition is sometimes not satisfied (as in [BP14]) and only $r^\star := r\,\mathbb{E}[W]$ can be estimated. Therefore, in what follows, we are interested in estimating $\theta := \left(\mu^T, r^\star\right)^T$, which will be denoted by $(\mu, r^\star)$ for the sake of simplicity.

We suppose from now that the following assumptions are fulfilled :

— **Assumption [A1].** The random vector X is not concentrated around μ :

$$\mathbb{E}\left[\|X - \mu\|^{-2}\right] < \infty.$$

— **Assumption [A2].** The random vector X admits a second moment :

$$\mathbb{E}\left[\|X - \mu\|^{2}\right] < \infty.$$

These assumptions ensure that the values of X are concentrated around the sphere and not around the center μ, without in addition too much dispersion. This framework totally corresponds to the real situation that we want to model. Moreover, using (8.5), Assumptions [A1] and [A2] reduce to assumptions on W. More precisely, [A1] reduces to $\mathbb{E}\left[W^{-2}\right] < \infty$ and [A2] to $\mathbb{E}\left[W^2\right] < \infty$.

Let us now introduce two examples of distribution allowing to model points spread around a complete sphere and satisfying assumptions [A1] and [A2].

Exemple 8.2.1. *Let us consider a random vector X taking values in $\mathbb{R}^d$ with a distribution absolutely continuous with respect to the Lebesgue measure, with a probability density function f_δ defined for all $\delta > 0$ by*

$$f_\delta(x) = \frac{C_d}{\|x - \mu\|^{d-1}} \mathbb{1}\{\|x - \mu\| / r \in [1 - \delta\,,\, 1 + \delta]\}, \tag{8.6}$$

where C_d is the normalization constant. Then, we can rewrite X under the form (8.5) with $U_\Omega = U$, $W \sim \mathcal{U}([1 - \delta, 1 + \delta])$ and $\mathbb{E}[W] = 1$ for any $\delta > 0$.

Exemple 8.2.2. *Let us consider the probability density function introduced in [BP14]. It is defined*

for any $x \in \mathbb{R}^d$ *by*

$$f(x) = K_d \exp\left(-\frac{1}{2\sigma^2}\left(\|x-\mu\| - r\right)^2\right), \tag{8.7}$$

where K_d *is the normalization constant. Then, a random vector* X *with probability density function* f *can be rewritten under the form (8.5), with* $\mathbb{E}[W] \neq 1$*, but* $\mathbb{E}[W]$ *is closed to* 1 *when the variance* σ *is negligible compared to the radius* r*.*

To obtain points distributed around a truncated sphere, it is sufficient to modify the previous densities by considering densities of the form $f_{\overline{\Omega}}(x) = C_{\overline{\Omega}} f(x) \mathbb{1}_{\left\{(x-\mu) \in \overline{\Omega}\right\}}$ where $\overline{\Omega}$ is the set of points of $\mathbb{R}^d$ whose polar coordinates are given by $(\rho, \theta_1, \ldots, \theta_{d-1} \in \mathbb{R}_+^* \times \Theta)$ where Θ defines the convex part Ω of the surface of the unit sphere of $\mathbb{R}^d$ we want to consider.

8.3 The algorithms

We present in this section two algorithms for estimating the unknown parameter θ which can be seen as a local minimizer (under conditions) of a function. Indeed, let us consider the function $G: \mathbb{R}^d \times \mathbb{R} \longrightarrow \mathbb{R}$ defined for all $y = (z, a) \in \mathbb{R}^d \times \mathbb{R}_+^*$ by

$$G(y) := \frac{1}{2}\mathbb{E}\left[(\|X-z\| - a)^2\right] = \frac{1}{2}\mathbb{E}\left[g(X, y)\right], \tag{8.8}$$

where we denote by g the function defined for any $x \in \mathbb{R}^d$ and $y = (z, a) \in \mathbb{R}^d \times \mathbb{R}_+^*$ by $g(x, y) := (\|x-z\| - a)^2$. The function G is Frechet-differentiable and we denote by Φ its gradient, which is defined for all $y = (z, a) \in \mathbb{R}^d \times \mathbb{R}_+^*$ by

$$\Phi(y) := \nabla G(y) = \mathbb{E}\left[\nabla_y g(X, y)\right] = \begin{pmatrix} z - \mathbb{E}[X] - a\,\mathbb{E}\left[\dfrac{z-X}{\|z-X\|}\right] \\ a - \mathbb{E}\left[\|z-X\|\right]. \end{pmatrix} \tag{8.9}$$

From (8.5) and definition of $\theta = (\mu, r^\star)$, we easily verify that $\nabla G(\theta) = 0$. Therefore, since θ is a local minimizer of function G (under assumptions) or a zero of ∇G, an idea could be to introduce a stochastic gradient algorithm for estimating θ.

8.3.1 The Robbins-Monro algorithm.

Let $(X_n)_{n\geq 1}$ be a sequence of independent and identically distributed random vectors of $\mathbb{R}^d$ following the same law as X and let $(\gamma_n)_{n\geq 1}$ be a decreasing sequence of positive real numbers satisfying the usual conditions

$$\sum_{n\geq 1} \gamma_n = \infty \quad \text{and} \quad \sum_{n\geq 1} \gamma_n^2 < \infty. \tag{8.10}$$

When the functional G is convex or verifies nice properties, a usual way to estimate the unknown parameter θ is to use the following recursive algorithm

$$\theta_{n+1} = \theta_n - \gamma_n \, \nabla_y g(X_{n+1}, \theta_n), \tag{8.11}$$

with θ_1 chosen arbitrarily bounded. The term $\nabla_y g\,(X_{n+1}, \theta_n)$ can be seen as an estimate of the gradient of G at θ_n, and the step sequence (γ_n) controls the convergence of the algorithm.

The convergence of such an algorithm is often established using the Robbins-Siegmund's theorem (see e.g. [Duf97]) and a sufficient condition to get it, is to verify that for any $y \in \mathbb{R}^d \times \mathbb{R}_+^*$, $\langle \Phi(y), y - \theta \rangle > 0$ where $\langle ., . \rangle$ denotes the usual inner product and $\|.\|$ the associated norm. However, we can show that this condition is only satisfied for y belonging to a subset of $\mathbb{R}^d \times \mathbb{R}_+^*$ to be specified. Thus, if at time $(n+1)$, the update of θ_n (using (8.11)) leaves this subset, then it does not necessarily converge. Therefore, we have to introduce a projected Robbins-Monro algorithm.

8.3.2 The Projected Robbins-Monro algorithm

Let $\mathcal{K}$ be a compact and convex subset of $\mathbb{R}^d \times \mathbb{R}_+^*$ containing $\theta = (\mu, r^*)$ and let $\pi : \mathbb{R}^d \times \mathbb{R}_+^* \longrightarrow \mathcal{K}$ be a projection satisfying

$$\begin{cases} \forall y, y' \in \mathbb{R}^d \times \mathbb{R}_+^*, \ \|\pi(y) - \pi(y')\| \leq \|y - y'\| \\ \forall y \notin \mathcal{K}, \ \pi(y) \in \partial\mathcal{K} \end{cases} \tag{8.12}$$

where $\partial\mathcal{K}$ is the frontier of $\mathcal{K}$. An example will be given later.

Then, we estimate θ using the following Projected Robbins-Monro algorithm (PRM), defined recursively by

$$\widehat{\theta}_{n+1} = \pi\left(\widehat{\theta}_n - \gamma_n \nabla_y g\left(X_{n+1}, \widehat{\theta}_n\right)\right), \tag{8.13}$$

where $\widehat{\theta}_1$ is arbitrarily chosen in $\mathcal{K}$, and (γ_n) is a decreasing sequence of positive real num-

bers satisfying (8.10).

Of course the choice of subset $\mathcal{K}$ and projector π is crucial. It is clear that if $\mathcal{K}$ is poorly chosen for a given projector, the convergence of the projected algorithm towards θ will be slower, even if from a theoretical point of view, we shall see in the next section dedicated to the theoretical results, that this algorithm is almost the same as the traditional Robbins-Monro algorithm since the updates of $\widehat{\theta}_n$, ie. the quantities $\left(\widehat{\theta}_n - \gamma_n \nabla_y g\left(X_{n+1}, \widehat{\theta}_n\right)\right)$, leave $\mathcal{K}$ only a finite number of times.

Let us now discuss the choice of $\mathcal{K}$ and π. The choice of $\mathcal{K}$ is directly related to the following assumption that we introduce to ensure the existence of a compact subset on which the scalar product $\langle \Phi(y), y - \theta \rangle$ is positive.

— **Assumption [A3].** There are two positive constants R_μ and R_r such that for all $y = (z, a) \in \overline{\mathcal{B}(\mu, R_\mu)} \times \overline{\mathcal{B}(r^*, R_r)}$,

$$\sup_{z \in \overline{\mathcal{B}(\mu, R_\mu)}} \lambda_{\max}\big(\Gamma(z)\big) < \frac{1 - \left\|\mathbb{E}\left[U_\Omega\right]\right\|^2 / A}{r^* + \frac{3}{2} R_r}, \tag{8.14}$$

with A such that $\left\|\mathbb{E}\left[U_\Omega\right]\right\|^2 < A < 1$, and $\lambda_{\max}(M)$ denotes the largest eigenvalue of matrix M, and

$$\Gamma(z) := \mathbb{E}\left[\frac{1}{\|X - z\|}\left(I_d - \frac{(X - z)(X - z)^T}{\|X - z\|^2}\right)\right].$$

Remark 8.3.1. The less the sphere is troncated, the more $\left\|\mathbb{E}\left[U_\Omega\right]\right\|$ is close to 0 and the constraints on R_μ and R_r are relaxed. In particular, when the sphere is complete, ie. $U_\Omega = U$ where U denotes the random vector uniformly distributed on the whole unit sphere of $\mathbb{R}^d$, then $\mathbb{E}\left[U_\Omega\right] = 0$ and Assumption [A3] reduces to

$$\sup_{z \in \overline{\mathcal{B}(\mu, R_\mu)}} \lambda_{\max}\big(\Gamma(z)\big) < \frac{1}{r^* + \frac{3}{2} R_r}.$$

The main consequence of Assumption [A3] is the following proposition which is one of the key point to establish the convergence of the PRM algorithm.

Proposition 8.3.1. *Assume that [A1] to [A3] hold. Then, there is a positive constant c such that for all $y \in \overline{\mathcal{B}(\mu, R_\mu)} \times \overline{\mathcal{B}(r^*, R_r)}$,*

$$\langle \Phi(y), y - \theta \rangle \geq c \left\|y - \theta\right\|^2.$$

Proof. The proof is given in Section E.1. □

Assumption [A3] is therefore crucial but only technical. It reflects the fact that the sphere is not too much truncated and that the points are not too far away from the sphere which corresponds to the real situations we want to model.
In a general framework, this technical assumption is difficult to verify since it requires to specify the distribution of X. In the case of distribution of Example 8.2.1 with $\delta < 1/10$, we can easily exhibit constant R_μ and R_r. Indeed taking $R_\mu = R_r = r^*/10$, then assumption [A3] holds. When the distribution of X is compactly supported with a support included in $[1-\delta, 1+\delta]$, it is fairly easy to find the constants provided that δ is small enough. It is quite more difficult when dealing with distribution of Example 8.2.2. Nevertheless, topological results can ensure that these constants exist.

From constants R_μ and R_r of Assumption [A3], it is then possible to simply define a projector π which satisfies condition (8.12). Indeed, let us set $\mathcal{K} = \mathcal{K}_\mu \times \mathcal{K}_r$ with $\mathcal{K}_\mu = \mathcal{B}(\mu, R_\mu)$ and $\mathcal{K}_r = \mathcal{B}(r^*, R_r)$, and define for any $y = (z, a) \in \mathbb{R}^d \times \mathbb{R}_+^*$ by $\pi(y) := \big(\pi_\mu(z), \pi_r(a)\big)$, with

$$\pi_\mu y(z) := \begin{cases} z & \text{if } z \in \mathcal{K}_\mu \\ \mu + R_\mu \dfrac{(z-\mu)}{\|z-\mu\|} & \text{otherwise} \end{cases}$$

and

$$\pi_r(a) := \begin{cases} a & \text{if } a \in \mathcal{K}_{r_0} \\ r + R_r \dfrac{(a-r^\star)}{|a-r^\star|} & \text{otherwise} \end{cases}$$

Such projector satisfies the requested conditions. However, it is clear that this projector can not be implemented since μ and r^* are unknown. We shall see in the simulation study how to overcome this problem.

We suppose from now that $\mathcal{K}$ is a compact and convex subset of $\overline{\mathcal{B}(\mu, R_\mu)} \times \overline{\mathcal{B}(r^*, R_r)}$ such that $\theta \in \mathcal{K}$, but $\theta \notin \partial\mathcal{K}$, where $\partial\mathcal{K}$ is the frontier of $\mathcal{K}$, i.e there is a positive constant $d_{\min}$ such that $\overline{\mathcal{B}(\theta, d_{\min})} \subset \mathcal{K}$.

8.3.3 The averaged algorithm

Averaging is a usual method to improve the rate of convergence of Robbins-Monro algorithms, or to reduce the variance, or finally not to have to make a good choice of the step sequence (see [PJ92]), but for the projected algorithms, this method is not widespread in the litterature. In this paper, we improve the estimation of θ by adding an averaging step to the PRM algorithm. Starting from the sequence $(\widehat{\theta}_n)_{n\geq 1}$ given by (8.13), we introduce for any

$n \geq 1$,

$$\overline{\theta}_n = \frac{1}{n} \sum_{k=1}^{n} \widehat{\theta}_k,$$

which can also be recursively defined by

$$\overline{\theta}_{n+1} = \overline{\theta}_n + \frac{1}{n+1} \left(\widehat{\theta}_{n+1} - \overline{\theta}_n \right), \quad \text{and} \quad \overline{\theta}_1 = \widehat{\theta}_1. \tag{8.15}$$

We shall see in the following two sections, the gain provided by this algorithm.

8.4 Convergence properties

We now give asymptotic properties of the algorithms. All the proofs are postponed in Section E.2. The following theorem gives the strong consistency of the PRM algorithm as well as properties on the number of times we really use the projection.

Theorem 8.4.1. *Let (X_n) be a sequence of iid random vectors following the same law as X. Assume that [A1] to [A3] hold, then*

$$\lim_{n \to \infty} \|\widehat{\theta}_n - \theta\| = 0 \quad a.s.$$

Moreover, the number of times the random vectors $\widehat{\theta}_n - \gamma_n \nabla_y g \left(X_{n+1}, \widehat{\theta}_n \right)$ do not belong to $\mathcal{K}$ is almost surely finite.

The following theorem gives the rate of convergence in quadratic mean and the L^p rates of convergence of the PRM algorithm (under conditions) as well as an upper bound of the probability that the random vector $\widehat{\theta}_n - \gamma_n \nabla_y g \left(X_{n+1}, \widehat{\theta}_n \right)$ does not belong to $\mathcal{K}$.

Theorem 8.4.2. *Let (X_n) be a sequence of iid random vectors following the same law as X. Assume that [A1] to [A3] hold and consider a step sequence (γ_n) of the form $\gamma_n = c_\gamma n^{-\alpha}$, with $c_\gamma > 0$ and $\alpha \in]1/2, 1[$. Then, there is a positive constant C_1 such that for all $n \geq 1$,*

$$\mathbb{E}\left[\left\| \widehat{\theta}_n - \theta \right\|^2 \right] \leq \frac{C_1}{n^\alpha}.$$

Moreover, for all positive integer p such that $\mathbb{E}\left[\|X - \mu\|^{2p} \right] < \infty$, there is a positive constant C_p such that for all $n \geq 1$,

$$\mathbb{E}\left[\left\| \widehat{\theta}_n - \theta \right\|^{2p} \right] \leq \frac{C_p}{n^{p\alpha}},$$

and for all $n \geq 1$,

$$\mathbb{P}\left[\widehat{\theta}_n - \gamma_n \nabla_y g \left(X_{n+1}, \widehat{\theta}_n \right) \notin \mathcal{K} \right] \leq \frac{C_p}{d_{\min}^{2p} \, n^{p\alpha}},$$

where $d_{\min} := \inf_{y \in \partial K} \{\|y - \theta\|\}$ *and* ∂K *is the frontier of* K.

We now focus on the asymptotic behavior of the averaged algorithm. First of all, applying Theorem 8.4.1 and Toeplitz's lemma for example, we easily obtain the strong consistency of the averaged estimator $\overline{\theta}_n$. Introducing the following assumption, we can specify its rate of convergence in quadratic mean as well as its asymptotic normality.

— **Assumption [A4].** The Hessian of G at $\theta = (\mu, r^*)$, denoted by Γ_θ and defined by

$$\Gamma_\theta := \begin{pmatrix} I_d - r^* \mathbb{E}\left[\frac{1}{\|X-\mu\|}\left(I_d - \frac{(X-\mu)\otimes(X-\mu)}{\|X-\mu\|^2}\right)\right] & \mathbb{E}\left[\frac{X-\mu}{\|X-\mu\|}\right] \\ \mathbb{E}\left[\frac{X-\mu}{\|X-\mu\|}\right]^T & 1 \end{pmatrix}$$

is a positive definite matrix.

Note that thanks to topological results, this assumption also implies Proposition 8.3.1 but is not useful to obtain the constants R_μ and R_r. Nevertheless, this assumption is crucial to establish the results of the two following theorems but it is satisfied as soon as the sphere is not too much truncated and the dispersion around the sphere not too important which corresponds to the real situations encountered. Using model (8.5), Γ_θ rewrites under the form

$$\Gamma_\theta = \begin{pmatrix} (I_d - \beta\,(I_d - \mathbb{E}\left[U_\Omega\, U_\Omega^T\right]) & \mathbb{E}\left[U_\Omega\right] \\ \mathbb{E}\left[U_\Omega^T\right] & 1 \end{pmatrix} \quad \text{with} \quad \beta = \mathbb{E}\left[W\right]\mathbb{E}\left[W^{-1}\right]. \tag{8.16}$$

When the sphere is complete, ie. $U_\Omega = U$, then $\mathbb{E}\left[U_\Omega\right] = 0$, $\mathbb{E}\left[U_\Omega\, U_\Omega^T\right] = (1/d) I_d$ and $\lambda_{\min}(\Gamma_\theta) > 0$ as soon as $\beta < d/(d-1)$. In the case of distribution of Example 8.2.1, we have $\beta = (\log(1+\delta) - \log(1-\delta))/(2\delta)$ and [A4] is satisfied as soon as δ is small enough. In the case of distribution of Example 8.2.2, [A4] is satisfied as soon as $r \gg \sigma$. When the sphere is not complete, we can easily show that a sufficient condition to ensure [A4] is $\lambda_{\min}\left(\mathbb{V}\mathrm{ar}\left[U_\Omega\right]\right) < 1 - 1/\beta$, where $\mathbb{V}\mathrm{ar}\left[U_\Omega\right]$ is the covariance matrix of the random variable U_Ω. In the case of the half sphere and $d = 3$, we have $\lambda_{\min}\left(\mathbb{V}\mathrm{ar}\left[U_\Omega\right]\right) = 1/12$ and Γ_θ is definite positive as soon as $\beta < 12/11$. This condition holds for distribution of Example 8.2.1 with $\delta < 0.4$ for instance, and distribution of Example 8.2.2 as soon as $r \gg \sigma$.

Theorem 8.4.3. *Let* (X_n) *be a sequence of iid random vectors following the same law as* X. *Assume that [A1] to [A4] hold and consider a step sequence* (γ_n) *of the form* $\gamma_n = c_\gamma n^{-\alpha}$, *with* $c_\gamma > 0$ *and* $\alpha \in]1/2, 1[$. *Moreover, suppose that* $\mathbb{E}[\|X-\mu\|^{12}] < \infty$. *Then there is a positive constant* C *such that for all* $n \geq 1$,

$$\mathbb{E}\left[\left\|\overline{\theta}_n - \theta\right\|^2\right] \leq \frac{C}{n}.$$

With respect to results of Theorem 8.4.2, we clearly improve the rate of convergence in quadratic mean. Note that the computed rate is the optimal one for such stochastic algorithms. We finally give a central limit theorem which can be useful to build confidence balls for the different parameters of the sphere.

Theorem 8.4.4. *Let (X_n) be a sequence of iid random vectors following the same law as X and let us choose the step sequence (γ_n) of the form $\gamma_n = c_\gamma n^{-\alpha}$, with $c_\gamma > 0$ and $\alpha \in]1/2, 1[$. Assume that [A1] to [A4] hold and suppose that $\mathbb{E}[\|X-\mu\|^{12}] < \infty$. Then $\left(\overline{\theta}_n\right)$ satisfies*

$$\sqrt{n}\left(\overline{\theta}_n - \theta\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}\left(0, \Gamma_\theta^{-1}\Sigma\,\Gamma_\theta^{-1}\right) \tag{8.17}$$

with

$$\Sigma := \mathbb{E}\left[\begin{pmatrix} \mu - X - r^*\dfrac{(\mu - X)}{\|\mu - X\|} \\ r^* - \|\mu - X\| \end{pmatrix}\begin{pmatrix} \mu - X - r^*\dfrac{(\mu - X)}{\|\mu - X\|} \\ r^* - \|\mu - X\| \end{pmatrix}^T\right]. \tag{8.18}$$

From result (8.17) of Theorem 8.4.4, we easily derive that

$$\sqrt{n}\,\Sigma^{-1/2}\Gamma_\theta\left(\overline{\theta}_n - \theta\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}\left(0, I_{d+1}\right). \tag{8.19}$$

Therefore, in order to build confidence balls or statistical tests for the parameters of the sphere, matrices Γ_θ and Σ must be estimated.
Let us decompose $\overline{\theta}_n$ under the form $(\overline{Z}_n, \overline{A}_n)$ where $\overline{Z}_n \in \mathbb{R}^d$ estimates the center μ and $\overline{A}_n \in \mathbb{R}_+^*$ the radius r^*, and let us denote $U_n := (X_n - \overline{Z}_n)/\left\|X_n - \overline{Z}_n\right\|$. Then we can estimate Γ_θ and Σ by $\widehat{\Gamma}_n$ and $\widehat{\Sigma}_n$ iteratively as follows

$$n\widehat{\Gamma}_n = (n-1)\widehat{\Gamma}_{n-1} + \begin{pmatrix} \left(1 - \frac{\overline{A}_n}{\|X_n - \overline{Z}_n\|}\right) I_d + \frac{\overline{A}_n}{\|X_n - \overline{Z}_n\|} U_n U_n^T & U_n \\ U_n^T & 1 \end{pmatrix},$$

$$n\widehat{\Sigma}_n = (n-1)\widehat{\Sigma}_{n-1} + \begin{pmatrix} X_n - \overline{Z}_n + \overline{A}_n U_n \\ \overline{A}_n - \|X_n - \overline{Z}_n\| \end{pmatrix}\begin{pmatrix} X_n - \overline{Z}_n + \overline{A}_n U_n \\ \overline{A}_n - \|X_n - \overline{Z}_n\| \end{pmatrix}^T,$$

where $\widehat{\Sigma}_1 = I_{d+1}$ and $\widehat{\Gamma}_1 = I_{d+1}$ to avoid usual problems of invertibility. It is not hard to show that $\widehat{\Gamma}_n$ and $\widehat{\Sigma}_n$ respectively converge to Γ_θ and Σ and then deduce that

$$Q_n := \sqrt{n}\,\widehat{\Sigma}_n^{-1/2}\widehat{\Gamma}_n\left(\overline{\theta}_n - \theta\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}\left(0, I_{d+1}\right). \tag{8.20}$$

The simulation study of the next section will illustrate the good approximation of the distribution of Q_n by the standard gaussian for moderate sample sizes.

8.5 Some experiments on simulated data

We study in this section the behavior of the PRM and averaged algorithms on simulated data in the case $d = 3$, for small to moderate sample sizes. This section first begins with the specification of the compact set involved in the definition of the PRM algorithm which is of course a crucial point. We then study the performance of the two algorithms in the case of the whole sphere with the distributions of Examples 8.2.1 and 8.2.2. Finally, we consider the case of the truncated sphere (a half-sphere) and we compare our strategy with the one proposed by [BP14].

In this simulation study, we shall always consider the same sphere defined by its center $\mu = (0, 0, 0)^T$ and its radius $r = 50$. In addition, to reduce sampling effects, our results are based on 200 samples of size n. Finally, let us mention that simulations were carried out using the statistical software R (see R Core Team, 2013).

8.5.1 Choice of the compact set and of the projection

We discuss here the crucial point of the choice of the compact set $\mathcal{K}$ and of the projection π involved in the definition of the PRM algorithm. The main problem is to find a compact set containing the unknown parameter θ. We propose to build a preliminary estimation of θ, using a geometric approach which consists in finding the center and the radius of a sphere of $\mathbb{R}^3$ from 4 non-coplanar distinct points. We denote by (μ_0, r_0) this initial estimate of θ. From this estimate, we define the compact set $\mathcal{K}$ by $\mathcal{K} := \mathcal{K}_{\mu_0} \times \mathcal{K}_{r_0}$ with $\mathcal{K}_{\mu_0} := \overline{\mathcal{B}(\mu_0, r_0/10)}$ and $\mathcal{K}_{r_0} := \overline{\mathcal{B}(r_0, r_0/10)}$, where the choice of the value $r_0/10$ for the radius of the balls is justified by the discussion about Assumption [A3] in Section 3.2. We then define the projector π as follows : for any $y = (z, a) \in \mathbb{R}^3 \times \mathbb{R}_+^*$, we set $\pi(y) := (\pi_{\mu_0}(z), \pi_{r_0}(a))$ with

$$\pi_{\mu_0} y(z) := \begin{cases} z & \text{if } z \in \mathcal{K}_{\mu_0} \\ \mu_0 + \dfrac{r_0}{10} \dfrac{(z - \mu_0)}{\|z - \mu_0\|} & \text{otherwise} \end{cases}$$

and

$$\pi_{r_0}(a) := \begin{cases} a & \text{if } a \in \mathcal{K}_{r_0} \\ r_0 + \dfrac{r_0}{10} \dfrac{(a - r_0)}{|a - r_0|} & \text{otherwise} \end{cases}$$

With this strategy, we can raisonnably hope that if our initial estimate is not too poor, then the true parameter belongs to $\mathcal{K}$ and the quadratic criteria G is convex on $\mathcal{K}$. We will see below that even if this preliminary estimation is rough, the true parameter belongs to $\mathcal{K}$ and the PRM algorithm improves the estimation of θ.

Let us now describe our strategy to obtain a preliminary estimation of the parameter $\theta = (\mu, r^\star)$. Since the data points are spread around the sphere, the estimation of the parameters from only one quadruplet of points is not robust to random fluctuations. In order to make the estimation more robust, we consider instead N quadruplets sampled with replacement from the first K points of the sample $X_1, \ldots, X_n$. For each quadruplet, we calculate the center of the sphere which passes through these four points, which gives a sequence of centers $(\widehat{\mu}_i)_{1 \leq i \leq N}$. The initial estimate of the center, denoted by μ_0, is then computed as the median point. Finally, we obtain an estimation of the radius by calculating the empirical mean of the sequence $(\|X_i - \mu_0\|)_{1 \leq i \leq 50}$.

A simulation study carried out for various values of K and N in the case of the whole and truncated sphere, shows that by taking $K = 50$ and $N = 200$, we obtain a preliminary estimation of θ sufficiently good to ensure that the compact $\mathcal{K}$ contains θ.

To close this section, let us mention that although the initial estimate is quite accurate, it is necessary to project the Robbins-Monro algorithm to ensure the convergence of the estimator. Indeed, taking a step sequence of the form $\gamma_n = c_\gamma n^{-\alpha}$, the results given in Table 8.1 show that for some values of c_γ and α, the parameter θ is poorly estimated by the Robbins-Monro algorithm, while the PRM algorithm (Table 8.2) is less sensitive to the step sequence choice.

		α				
		0.51	0.6	0.66	0.75	0.99
c_γ	1	0.27	0.14	0.09	0.06	0.23
	5	10^8	10^6	10^5	10^4	10^5
	10	10^{31}	10^{18}	10^{14}	10^{10}	10^6

TABLE 8.1 – Robbins-Monro algorithm. Errors in quadratic mean of the 200 estimations of the center μ for samples of size $n = 2000$ in the case of the distribution of Example 8.2.1.

		α				
		0.51	0.6	0.66	0.75	0.99
c_γ	1	0.28	0.15	0.09	0.05	0.24
	5	1.55	0.76	0.48	0.24	0.05
	10	3.22	1.35	0.94	0.43	0.08

TABLE 8.2 – PRM algorithm. Errors in quadratic mean of the 200 estimations of the center μ for samples of size $n = 2000$ in the case of the distribution of Example 8.2.1.

In the sequel of the simulation study, we take a step sequence of the form $\gamma_n := n^{-2/3}$

($\alpha = 2/3$ is often considered as the optimal choice in the literature).

8.5.2 Case of the whole sphere

In what follows, we are interested in the behavior of the PRM and averaged algorithms when samples are distributed on the whole sphere according to the distribution of Example 8.2.1 with $\delta = 0.1$.

Figure 8.1 shows that the accuracy of the estimations increases with the sample size. In particular, as expected, the PRM algorithm significantly improves the initial estimations of the center and the radius (see the first boxplots which correspond to the initial estimations). Moreover, as expected in the case of the "whole sphere", we can see that the three components of the center μ are estimated with the same accuracy.

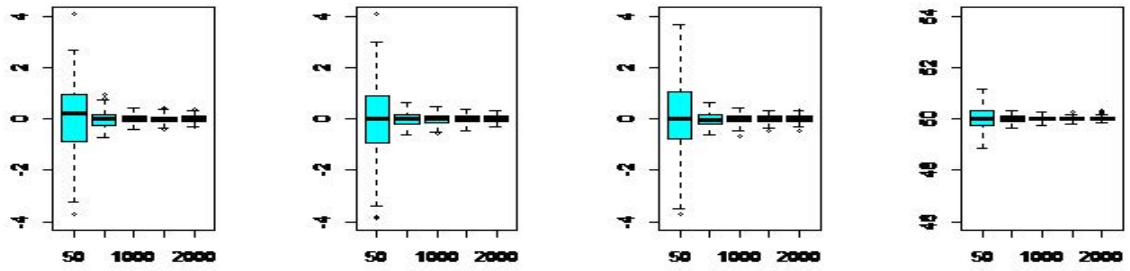

FIGURE 8.1 – Whole sphere with distribution of Example 8.2.1. From the left to the right, boxplots of estimates of μ_x, μ_y, μ_z and r obtained with the PRM algorithm for different sample sizes.

Let us now examine the gain provided by the use of the averaged algorithm. Figure 8.2 shows that for small sample sizes, the performances of the two algorithms are comparable, but when n is greater than 500, the averaged algorithm is more accurate than the PRM algorithm. We can even think that by forgetting the first estimates of the PRM algorithm, we improve the behavior of the averaged algoritm when the sample size is small.

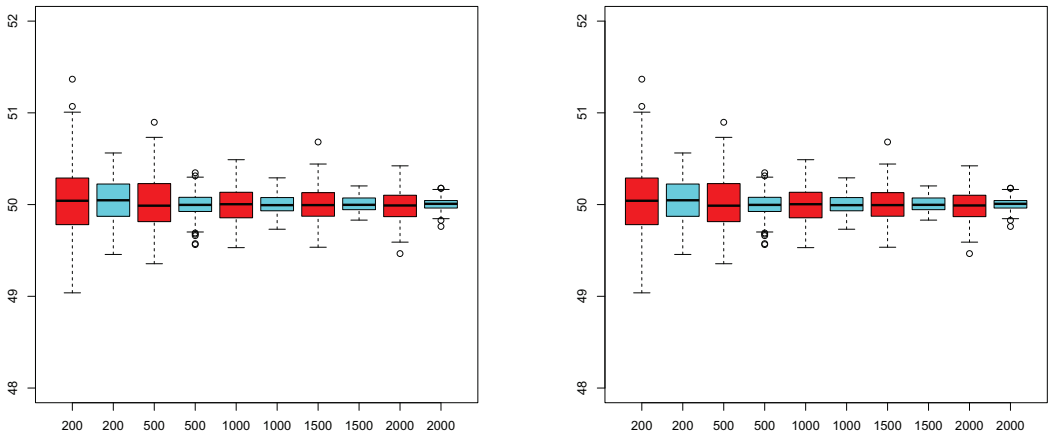

FIGURE 8.2 – Whole sphere with distribution of Example 8.2.1. Boxplots of estimates of μ_y (left) and r (right) obtained with the PRM algorithm (in red) and with the averaged algorithm (in blue) for different sample sizes.

Finally, let us study the quality of the Gaussian approximation of the distribution of Q_n for a moderate sample size. This point is crucial for building confidence intervals or statistical tests for the parameters of the sphere.

Figure 8.3 shows that this approximation is reasonable when $n = 2000$. Indeed, we can see that the estimated density of each component of Q_n is well superimposed with the density of the $\mathcal{N}(0,1)$. To validate these approximations, we perform a Kolmogorov-Smirnov test at level 5%. The test enables us to conclude that the normality is not rejected for each component of Q_n.

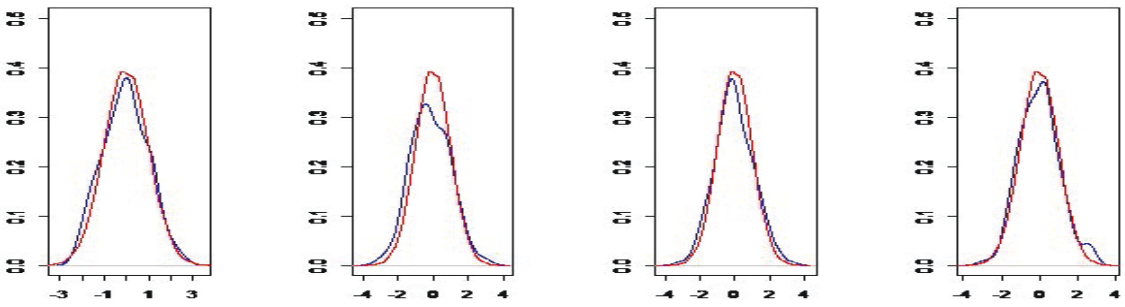

FIGURE 8.3 – From the left to the right, estimated densities of each components of Q_{2000} superimposed with the standard gaussian density.

8.5.3 Comparison with a backfitting-type algorithm in the case of a half-sphere

In this section, we compare the performances of the averaged algorithm with the ones of the backfitting algorithm introduced by [BP14]. In what follows, we consider samples coming from the distribution of Example 8.2.2, with $\sigma = 1$, in the case of the half sphere defined by the set of points whose y-component is positive.

Results obtained with the two algorithms are presented in Figure 8.4. We focus on parameter μ_y for the center since it is the more difficult to estimate. We can see that even if the backfitting (BF for short) algorithm is better than the averaged algorithm, the performances are globally good, which validates the use of our algorithm for estimating the parameters of a sphere from 3D-points distributed around a truncated sphere. Recall that convergence results are available for our algorithm in the case of the truncated sphere, contrary to the backfitting algorithm for which no theoretical result is available in that case.

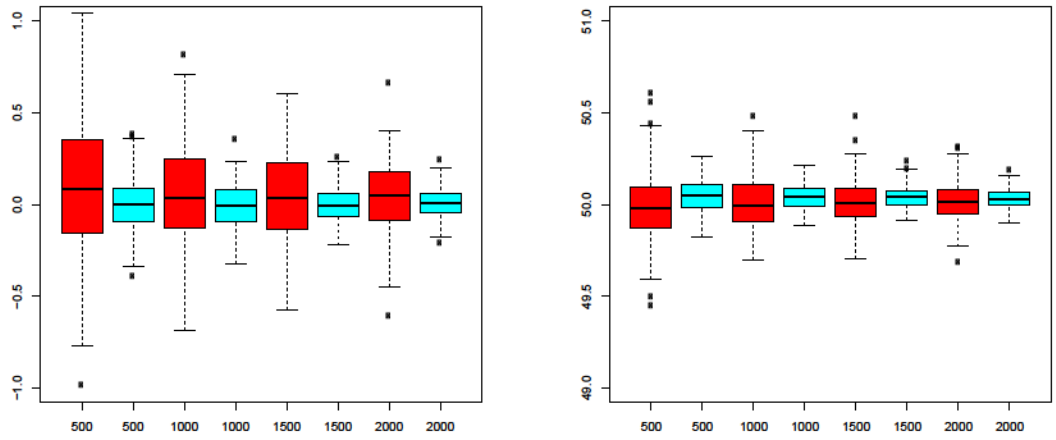

FIGURE 8.4 – Comparison of averaged and BF algorithms. Boxplots of the estimates of μ_y (on the left) and r (on the right), obtained with the BF algorithm (in blue) and with the averaged algorithm (in red) for the half sphere in the case of Example 8.2.2.

8.6 Conclusion

We presented in this work a new stochastic algorithm for estimating the center and the radius of a sphere from a sample of points spread around the sphere, the points being distributed around the complete sphere or only around a part of the sphere.

We shown on simulated data that this algorithm is efficient, less accurate than the back-fitting algorithm proposed in [BP14] but for which no convergence result is available for the case of the truncated sphere. Therefore, our main contribution is to have proposed an algorithm for which we have given asymptotic results such as its strong consistency and its asymptotic normality which can be useful to build confidence balls or statistical tests for example, as well as non asymptotic results such as the rates of convergence in quadratic mean.

A possible extension of this work could be to extend the obtained results to the case of the finite mixture model. This framework has been considered in [BP14] but no convergence result is established. Proposing a stochastic algorithm for estimating the different parameters of the model and obtaining convergence results would be a nice challenge.

Acknowledgements. The authors would like to thank Aurélien Vasseur for his contribution to the start of this work. We also would like to thank Peggy Cénac and Denis Brazey for their constructive remarks and for their careful reading of the manuscript that allowed to improve the presentation of this work. Finally, we would like to thank Nicolas Godichon for his help in the creation of Figure E.1.

Annexe E

An averaged projected Robbins-Monro algorithm for estimating the parameters of a truncated spherical distribution. Appendix

Résumé

Dans cette partie, nous commençons par donner des résultats de convexité sur la fonction dont l'on cherche un minimum local. De plus, on donne la preuve de la Proposition 8.3.1, qui assure que la fonction admet un unique minimum sur le compact sur lequel on projette l'algorithme. De plus, les preuves des Théorèmes 8.4.1 à 8.4.4, qui donnent les vitesses de convergence des algorithmes ainsi que la normalité asymptotique, sont données.

Abstract

In this part, we first give some convexity results on the function we would like to minimize. Moreover, we give the proof of Proposition 8.3.1, which ensure that the function admit an unique minimizer on the compact on each we project the algorithm. Moreover, the proofs of Theorems 8.4.1 to 8.4.4, which give the rate of convergence of the algorithms as well as the asymptotic normality of the averaged algorithm, are given.

E.1 Some convexity results and proof of proposition 8.3.1

The following lemma ensures that the Matrix in Assumption [A3] is well defined and that the Hessian of G exists for all $y \in \mathbb{R}^d \times \mathbb{R}$.

Lemma E.1.1. *Assume [A1] holds. If $d \geq 3$, there is a positive constant C such that for all $z \in \mathbb{R}^d$,*

$$\mathbb{E}\left[\frac{1}{\|X-z\|}\right] \leq C.$$

Moreover, suppose that W admits a bounded density, then for all $d \geq 2$, there is a positive constant C such that for all $z \in \mathbb{R}^d$,

$$\mathbb{E}\left[\frac{1}{\|X-z\|}\right] \leq C.$$

Note that for the sake of simplicity, we denote by the same way the two constants.

Proof of Lemma E.1.1. **Step 1 :** $d \geq 3$

By continuity and applying Assumption **[A1]**, there are positive constants ϵ, C' such that for all $z \in \mathcal{B}(\mu, \epsilon)$,

$$\mathbb{E}\left[\frac{1}{\|X-z\|}\right] \leq C'.$$

Moreover, let $z \in \mathbb{R}^d$ such that $\|z-\mu\| \geq \epsilon$, we have

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{\|X-z\|}\right] &= \int_0^{+\infty} \mathbb{P}\left[\|X-z\| \leq \frac{1}{t}\right] dt \\
&= \int_0^{M} \mathbb{P}\left[\|X-z\| \leq t^{-1}\right] dt + \int_M^{\infty} \mathbb{P}\left[\|X-z\| \leq t^{-1}\right] dt \\
&\leq M + \int_M^{\infty} \mathbb{P}\left[\|X-z\| \leq t^{-1}\right] dt,
\end{aligned}$$

with M positive and defined later. Moreover, let $t \geq M$,

$$\begin{aligned}
\mathbb{P}\left[\|X-z\| \leq t^{-1}\right] &= \mathbb{P}\left[\|\mu + rWU_\Omega - z\| \leq t^{-1}\right] \\
&\leq \mathbb{P}\left[-t^{-1} + \|z-\mu\| \leq rW \leq t^{-1} + \|z-\mu\|, (\mu + rWU_\Omega) \cap \mathcal{B}\left(z, t^{-1}\right) \neq \emptyset\right],
\end{aligned}$$

taking $M = \frac{2}{\epsilon}$. With previous condition on rW, calculating $\mathbb{P}\left[(\mu + rWU_\Omega) \cap \mathcal{B}\left(z, t^{-1}\right) \neq \emptyset\right]$ consists in measuring the intersection between a truncated sphere with radius bigger than $\epsilon/2$ with a ball of radius $\frac{1}{t}$, with $\frac{1}{t} \leq \frac{\epsilon}{2}$. This is smaller than the surface of the frontier of the ball (see the following figure).

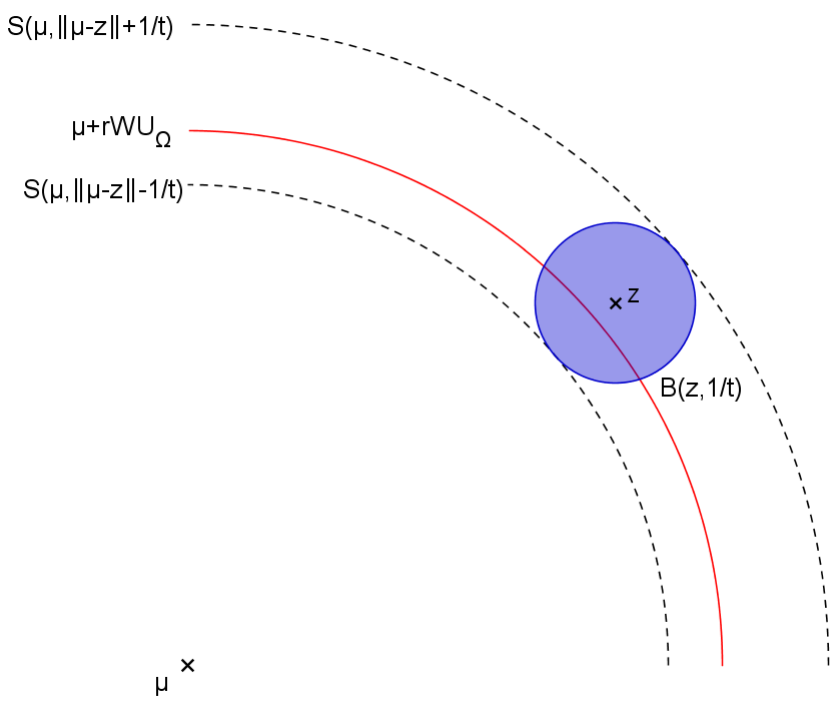

FIGURE E.1 – Intersection between a ball and a sphere

Thus, there is a positive constant k such that for all $t \geq M$,

$$\mathbb{P}\left[\|X - z\| \leq t^{-1}\right] \leq \frac{k}{t^{d-1}}. \tag{E.1}$$

Finally,

$$\begin{aligned}\mathbb{E}\left[\frac{1}{\|X - z\|}\right] &\leq \frac{2}{\epsilon} + \int_{\frac{2}{\epsilon}}^{+\infty} k\frac{1}{t^{d-1}}dt \\ &= \frac{2}{\epsilon} + k\frac{\epsilon^{d-2}}{2^{d-2}(d-2)}.\end{aligned}$$

We conclude the proof taking $C = \max\left\{C', \frac{2}{\epsilon} + k\frac{\epsilon^{d-2}}{2^{d-2}(d-2)}\right\}$.

Step 2 : $d = 2$ and W admits a bounded density

Let $f_{\max}$ be a bound of the density function of W. As in previous case, let $z \in \mathbb{R}^d$ such that

$\|z-\mu\| \geq \epsilon$,

$$\begin{aligned}\mathbb{P}\left[\|X-z\| \leq t^{-1}\right] &\leq \mathbb{P}\left[-t^{-1}+\|z-\mu\| \leq rW \leq t^{-1}+\|z-\mu\|, (\mu+rWU_\Omega)\cap\mathcal{B}\left(z,t^{-1}\right)\neq\emptyset\right]\\ &= \mathbb{P}\left[(\mu+rWU_\Omega)\cap\mathcal{B}\left(z,t^{-1}\right)\neq\emptyset \middle| -t^{-1}+\|z-\mu\| \leq rW \leq t^{-1}+\|z-\mu\|\right]\\ &\times \mathbb{P}\left[-t^{-1}+\|z-\mu\| \leq rW \leq t^{-1}+\|z-\mu\|\right].\end{aligned}$$

As in previous case, if $t \geq \frac{2}{\epsilon}$, there is a positive constant k such that for all $t \geq \frac{2}{\epsilon}$,

$$\mathbb{P}\left[(\mu+rWU_\Omega)\cap\mathcal{B}\left(z,t^{-1}\right)\neq\emptyset \middle| -t^{-1}+\|z-\mu\| \leq rW \leq t^{-1}+\|z-\mu\|\right] \leq kt^{-1}.$$

Moreover, since $f_{\max}$ is a bound of the density function of W,

$$\mathbb{P}\left[-\frac{1}{t}+\|z-\mu\| \leq rW \leq \frac{1}{t}+\|z-\mu\|\right] \leq \frac{2rf_{\max}}{t}$$

Thus, for all $t \geq \frac{2}{\epsilon}$,

$$\mathbb{P}\left[\|X-z\| \leq t^{-1}\right] \leq \frac{2rf_{\max}k}{t^2},$$

and in a particular case,

$$\mathbb{E}\left[\frac{1}{\|X-z\|}\right] \leq \frac{2}{\epsilon}+krf_{\max}\epsilon, \tag{E.2}$$

and one can conclude the proof taking $C = \max\left\{C', 2\epsilon^{-1}+krf_{\max}\epsilon\right\}$. □

Proof of Proposition 8.3.1. We want to show there is $c>0$ such that for any $y=(z,a) \in \overline{\mathcal{B}(\mu,\varepsilon_\mu)} \times \overline{\mathcal{B}(r^*,\varepsilon_r)}$, $P(y) := \langle y-\theta\,,\,\Phi(y)\rangle \geq c\,\|y-\theta\|$. We have

$$P(y) = P(z,a) = \left\langle \begin{pmatrix} z-\mu \\ a-r^* \end{pmatrix}, \begin{pmatrix} z-\mathbb{E}[X]-a\,\mathbb{E}\left[\frac{z-X}{\|z-X\|}\right] \\ a-\mathbb{E}[\|X-z\|] \end{pmatrix}\right\rangle. \tag{E.3}$$

For any $z \in \mathbb{R}^d$, let us set $F(z) := \mathbb{E}[\|X-z\|]$ and $f(z) := \mathbb{E}[(z-X)/\|z-X\|]$. Note that f is the gradient of F. Using (2.1), we deduce that $F(\mu)=r^*$, $f(\mu) = -\mathbb{E}[U_\Omega]$ and $\mathbb{E}[X] = \mu - r^\star f(\mu)$. Then, (E.3) can be rewritten as

$$\begin{aligned}P(y) &= \|z-\mu\|^2 + r^*\langle z-\mu, f(\mu)\rangle - a\langle z-\mu, f(z)\rangle + (a-r^*)^2 - (a-r^*)(F(z)-F(\mu))\\ &= \|z-\mu\|^2 - (a-r^\star)\langle z-\mu, f(\mu)\rangle - a\langle z-\mu, f(z)-f(\mu)\rangle + (a-r^\star)^2\\ &- (a-r^\star)(F(z)-F(\mu)).\end{aligned}$$

Moreover, using the following Taylor's expansions,

$$F(z) = F(\mu) + \langle z-\mu, f(\mu)\rangle + \frac{1}{2}(z-\mu)^T \nabla f(c)(z-\mu),$$
$$f(z) = f(\mu) + \left\langle \nabla f(c'), z-\mu \right\rangle,$$

with $c, c' \in [z, \mu]$. We get

$$\begin{aligned} P(y) = \|z-\mu\|^2 - 2(a-r^\star)\,\langle z-\mu, f(\mu)\rangle - a(z-\mu)^T \nabla f(c')(z-\mu) \\ + (a-r^*)^2 - \frac{1}{2}(a-r^\star)(z-\mu)^T \nabla f(c)(z-\mu) \end{aligned} \tag{E.4}$$

Now, remarking that for any positive constant A and real numbers x, y, we have $2xy \;\leq\; A\,x^2 \;+\; y^2/A$, we derive

$$\begin{aligned} P(y) \geq \|z-\mu\|^2 - A(a-r^*)^2 - \frac{1}{A}\,\|z-\mu\|^2\,\|f(\mu)\|^2 - a\;\|\nabla f(c)\|_{op}\,\|z-\mu\|^2 \\ + (a-r^*)^2 - \frac{1}{2}\,|a-r^\star|\,\left\|\nabla f(c')\right\|_{op}\,\|z-\mu\|^2. \end{aligned}$$

Let us denote by $\lambda_M = \sup_{z \in \overline{\mathcal{B}(\mu,\varepsilon_\mu)}} \lambda_{\max} \nabla f(z)$ and choose A such that

$$\|f(\mu)\|^2 \;=\; \|\mathbb{E}\,[U_\Omega]\|^2 \;<\; A \;<\; 1.$$

Then, for any $z \in \overline{\mathcal{B}(\mu, \varepsilon_\mu)}$ and $a \in \overline{\mathcal{B}(r^*, \varepsilon_r)}$, we have

$$P(y) \geq \left(1 - \frac{1}{A}\,\|f(\mu)\|^2 - (r^* + \frac{3}{2}\varepsilon_r)\,\lambda_M\right)\|z-\mu\|^2 + (1-A)\,(a-r^*)^2$$

Finally, using assumption **[A3]**, we close the proof.

□

In order to linearize the gradient in the decompositions of the PRM algorithm and get a nice decomposition of the averaged algorithm, we introduce the Hessian matrix of G, denoted, for all $y = (z, a) \in \mathbb{R}^d \times \mathbb{R}$, by $\Gamma_y : \mathbb{R}^d \times \mathbb{R} \longrightarrow \mathbb{R}^d \times \mathbb{R}$ and defined by :

$$\Gamma_y = \begin{pmatrix} I_d - a\mathbb{E}\left[\dfrac{1}{\|X-z\|}\left(I_d - \dfrac{(X-z)\otimes(X-z)}{\|X-z\|^2}\right)\right] & \mathbb{E}\left[\dfrac{X-z}{\|X-z\|}\right] \\ \mathbb{E}\left[\dfrac{X-z}{\|X-z\|}\right]^T & 1 \end{pmatrix}, \tag{E.5}$$

with, for all $z, z', z'' \in \mathbb{R}^d$, $z \otimes z'(z'') = \langle z, z'' \rangle z'$. Applying Lemma E.1.1, the Hessian matrix exists for all $y \in \mathbb{R}^{d+1}$.

Proposition E.1.1. *Suppose [A1] to [A3] hold, there is a positive constant C_θ such that for all $y \in \mathcal{K}$,*

$$\|\Phi(y) - \Gamma_\theta (y - \theta)\| \leq C_\theta \|y - \theta\|^2 .$$

Proof of Proposition E.1.1. Under Assumption [A1], by continuity, there are positive constants C', ϵ' such that for $z \in \mathcal{B}(\mu, \epsilon')$,

$$\mathbb{E}\left[\frac{1}{\|X - y\|^2}\right] \leq C'.$$

Moreover, note that for all $y \in \mathcal{K}$,

$$\Phi(y) = \int_0^1 \Gamma_{\theta + t(y-\theta)}(y - \theta) dt.$$

Thus, with analogous calculus to the ones in the proof of Lemma 5.1 in [CCGB15], one can check that there is a positive constant C'' such that for all $y \in \mathcal{B}(\theta, \epsilon') \cap \mathcal{K}$,

$$\|\Phi(y) - \Gamma_\theta\| \leq C'' \|y - \theta\|^2 .$$

Moreover, for all $y = (z, a) \in \mathcal{K}$ and $y' = (z', a') \in \mathbb{R}^d \times \mathbb{R}$,

$$\Gamma_y(y') = \begin{pmatrix} z' - y\mathbb{E}\left[\frac{1}{\|X-z\|}\left(z - \frac{\langle X-z, z'\rangle (X-z)}{\|X-z\|^2}\right)\right] + a'\mathbb{E}\left[\frac{X-z}{\|X-z\|}\right] \\ \mathbb{E}\left[\frac{\langle X-z, z'\rangle}{\|X-z\|}\right] + a' \end{pmatrix}.$$

Thus, applying Cauchy-Schwarz's inequality,

$$\begin{aligned} \|\Gamma_y(y')\|^2 &= \left\| z' - a\mathbb{E}\left[\frac{1}{\|X-z\|}\left(z' - \frac{\langle X-z, z'\rangle (X-z)}{\|X-z\|^2}\right)\right] + a'\mathbb{E}\left[\frac{X-z}{\|X-z\|}\right]\right\|^2 \\ &\quad + \left\|\mathbb{E}\left[\frac{\langle X-z, z'\rangle}{\|X-z\|}\right] + a'\right\|^2 \\ &\leq 3\|z'\|^2 + 3\|a\|^2 \|z'\|^2 \mathbb{E}\left[\frac{1}{\|X-z\|}\right]^2 + 3\|a'\|^2 + 2\|z'\|^2 + 2\|a'\|^2 \end{aligned}$$

Thus, applying Lemma E.1.1, there are positive constants A_1, A_2 such that

$$\|\Gamma_y(y')\| \leq A_1 \|y'\| + A_2 \|y\| \|y'\|$$

Note that since $\mathcal{K}$ is compact and convex, there is a positive constant $C_{\mathcal{K}}$ such that for all $y \in \mathcal{K}$ and $t \in [0,1]$, $\|\theta + t(y-\theta)\| \leq C_{\mathcal{K}}$, and in a particular case,

$$\left\|\Gamma_{\theta+(y-\theta)}(y-\theta)\right\| \leq (A_1 + A_2 C_{\mathcal{K}}) \|y-\theta\|.$$

Thus, for all $y \in \mathcal{K}$ such that $\|y-\theta\| \geq \epsilon'$,

$$\begin{aligned}\|\Phi(y) - \Gamma_\theta(y-\theta)\| &\leq \int_0^1 \left\|\Gamma_{\theta+t(y-\theta)}(y-\theta)\right\| dt \\ &\leq (A_1 + A_2 C_{\mathcal{K}}) \|y-\theta\| \\ &\leq \frac{1}{\epsilon'}(A_1 + A_2 C_{\mathcal{K}}) \|y-\theta\|^2.\end{aligned}$$

Thus, we conclude the proof taking $C_\theta = \max\left\{C'', \frac{1}{\epsilon'}(A_1 + A_2 C_{\mathcal{K}})\right\}$. □

E.2 Proof of Section 8.4

Proof of Theorem 8.4.1. Let us recall that there is a positive constant c such that for all $y \in \mathcal{K}$, $\langle \Phi(y), y-\theta\rangle \geq c\|y-\theta\|^2$. The aim is to use previous inequality and the fact that the projection is 1-lipschitz in order to get an upper bound of $\mathbb{E}\left[\left\|\widehat{\theta}_{n+1}-\theta\right\|^2 | \mathcal{F}_n\right]$ and apply Robbins-Siegmund theorem to get the almost sure convergence of the algorithm.

Almost sure convergence of the algorithm : Since π is 1-lipschitz,

$$\begin{aligned}\left\|\widehat{\theta}_{n+1}-\theta\right\|^2 &= \left\|\pi\left(\widehat{\theta}_n - \gamma_n \nabla_y g\left(X_{n+1}, \widehat{\theta}_n\right)\right) - \pi(\theta)\right\|^2 \\ &\leq \left\|\widehat{\theta}_n - \gamma_n \nabla_y g\left(X_{n+1}, \widehat{\theta}_n\right) - \theta\right\|^2 \\ &= \left\|\widehat{\theta}_n - \theta\right\|^2 - 2\gamma_n \left\langle \nabla_y g\left(X_{n+1}, \widehat{\theta}_n\right), \widehat{\theta}_n - \theta\right\rangle + \gamma_n^2 \left\|\nabla_y g\left(X_{n+1}, \widehat{\theta}_n\right)\right\|^2\end{aligned}$$

Thus, since $\widehat{\theta}_n$ is $\mathcal{F}_n$-measurable,

$$\begin{aligned}
\mathbb{E}\left[\left\|\widehat{\theta}_{n+1}-\theta\right\|^2|\mathcal{F}_n\right] &\leq \left\|\widehat{\theta}_n-\theta\right\|^2 - 2\gamma_n\left\langle\mathbb{E}\left[\nabla_y g\left(X_{n+1},\widehat{\theta}_n\right)\Big|\mathcal{F}_n\right],\widehat{\theta}_n-\theta\right\rangle \\
&\quad+\gamma_n^2\mathbb{E}\left[\left\|\nabla_y g\left(X_{n+1},\widehat{\theta}_n\right)\right\|^2\Big|\mathcal{F}_n\right] \\
&= \left\|\widehat{\theta}_n-\theta\right\|^2 - 2\gamma_n\left\langle\Phi(\widehat{\theta}_n),\widehat{\theta}_n-\theta\right\rangle+\gamma_n^2\mathbb{E}\left[\left\|\nabla_y g\left(X_{n+1},\widehat{\theta}_n\right)\right\|^2|\mathcal{F}_n\right] \\
&\leq \left\|\widehat{\theta}_n-\theta\right\|^2 - 2c\gamma_n\left\|\widehat{\theta}_n-\theta\right\|^2+\gamma_n^2\mathbb{E}\left[\left\|\nabla_y g\left(X_{n+1},\widehat{\theta}_n\right)\right\|^2|\mathcal{F}_n\right]
\end{aligned}$$

Moreover, let $\widehat{\theta}_n := (Z_n, A_n)$ with $Z_n \in \mathbb{R}^d$ and $A_n \in \mathbb{R}$, we have

$$\begin{aligned}
\mathbb{E}\left[\left\|\nabla_y g\left(X_{n+1},\widehat{\theta}_n\right)\right\|^2|\mathcal{F}_n\right] &= \mathbb{E}\left[\left\|Z_n - X_{n+1} - A_n\frac{Z_n-X_{n+1}}{\|Z_n-X_{n+1}\|}\right\|^2\Big|\mathcal{F}_n\right] \\
&\quad+\mathbb{E}\left[\left|A_n-\|Z_n-X_{n+1}\|\right|^2|\mathcal{F}_n\right] \\
&\leq 4\mathbb{E}\left[\|Z_n-X_{n+1}\|^2\big|\mathcal{F}_n\right]+4\left(A_n\right)^2 \\
&\leq 8\|Z_n-\mu\|^2+8(r^\star)^2+8\left(A_n-r^\star\right)^2+8\mathbb{E}\left[\|\mu-X_{n+1}\|^2|\mathcal{F}_n\right] \\
&= 8\left\|\widehat{\theta}_n-\theta\right\|^2+8(r^\star)^2+8r^2\mathbb{E}\left[W^2\right].
\end{aligned}$$

Let $M := 8(r^\star)^2+8r^2\mathbb{E}\left[W^2\right]$, we have

$$\mathbb{E}\left[\left\|\widehat{\theta}_n-\theta\right\|^2|\mathcal{F}_n\right] \leq \left(1+8\gamma_n^2\right)\left\|\widehat{\theta}_n-\theta\right\|^2 - 2c\gamma_n\left\|\widehat{\theta}_n-\theta\right\|^2+\gamma_n^2 M. \tag{E.6}$$

Applying Robbins-Siegmund's theorem (see [Duf97] for instance), $\left\|\widehat{\theta}_n-\theta\right\|^2$ converges almost surely to a finite random variable, and in a particular case,

$$\sum_{k=1}^{\infty}\gamma_k\left\|\widehat{\theta}_k-\theta\right\|^2 < +\infty.$$

Thus, since $\sum_{k\geq 1}\gamma_k = +\infty$,

$$\lim_{n\to+\infty}\left\|\widehat{\theta}_n-\theta\right\|^2 = 0 \quad a.s. \tag{E.7}$$

Number of times the projection is used

Let $N_n := \sum_{k=1}^n \mathbf{1}_{\left\{\widehat{\theta}_k-\gamma_k\nabla_y\left(X_{k+1},\widehat{\theta}_k\right)\notin\mathcal{K}\right\}}$. This sequence is non-decreasing, and suppose by contradiction that N_n goes to infinity. Thus, there is a subsequence (n_k) such that (N_{n_k})

is increasing, i.e for all $k \geq 1$, $\widehat{\theta}_{n_k} - \gamma_n \nabla_y g\left(X_{n_k+1}, \widehat{\theta}_{n_k}\right) \notin \mathcal{K}$, and in a particular case, $\widehat{\theta}_{n_k+1} \in \partial\mathcal{K}$, where $\partial\mathcal{K}$ is the frontier of $\mathcal{K}$. Let us recall that θ is in the interior of $\mathcal{K}$, i.e let $d_{\min} := \inf_{y\in\partial\mathcal{K}} \|\theta - y\|$, we have $d_{\min} > 0$. Thus,

$$\left\|\widehat{\theta}_{n_k+1} - \theta\right\| \geq d_{\min} \quad a.s,$$

and,

$$\lim_{k\to\infty} \left\|\widehat{\theta}_{n_k+1} - \theta\right\| = 0 \geq d_{min} > 0 \quad a.s,$$

which leads to a contradiction. □

Proof of Theorem 8.4.2. **Convergence in quadratic mean**

The aim is to obtain an induction relation for the quadratic mean error. Let us recall inequality (E.6),

$$\mathbb{E}\left[\left\|\widehat{\theta}_{n+1} - \theta\right\|^2 |\mathcal{F}_n\right] \leq \left(1 + 8\gamma_n^2\right) \left\|\widehat{\theta}_n - \theta\right\|^2 - 2c\gamma_n \left\|\widehat{\theta}_n - \theta\right\|^2 + \gamma_n^2 M.$$

Then we have

$$\mathbb{E}\left[\left\|\widehat{\theta}_{n+1} - \theta\right\|^2\right] \leq \left(1 - c\gamma_n + 8\gamma_n^2\right) \mathbb{E}\left[\left\|\widehat{\theta}_n - \theta\right\|^2\right] + M\gamma_n^2,$$

and one can conclude the proof with the help of an induction (see [GB15] for instance) or applying a lemma of stabilization (see [Duf96]).

L^p **rates of convergence**

Let $p \geq 2$, we now prove with the help of a strong induction that for all integer $p' \leq p$, there is a positive constant $C_{p'}$ such that for all $n \geq 1$,

$$\mathbb{E}\left[\left\|\widehat{\theta}_n - \theta\right\|^{2p'}\right] \leq \frac{C_{p'}}{n^{p'\alpha}}.$$

This inequality is already checked for $p' = 1$. Let $p' \geq 2$, we suppose from now that for all integer $k < p'$, there is a positive constant C_k such that for all $n \geq 1$,

$$\mathbb{E}\left[\left\|\widehat{\theta}_n - \theta\right\|^{2k}\right] \leq \frac{C_k}{n^{k\alpha}}.$$

We now search to give an induction relation for the $L^{2p'}$-error. Let us recall that

$$\left\|\widehat{\theta}_{n+1}-\theta\right\|^2 \leq \left\|\widehat{\theta}_n-\theta\right\|^2 - 2\gamma_n\left\langle \nabla_y g\left(X_{n+1},\widehat{\theta}_n\right),\widehat{\theta}_n-\theta\right\rangle + \gamma_n^2\left\|\nabla_y g\left(X_{n+1},\widehat{\theta}_n\right)\right\|^2.$$

We suppose from now that $\mathbb{E}\left[W^{2p}\right] < +\infty$ (and in a particular case, $\mathbb{E}\left[W^k\right] < +\infty$ for all integer $k \leq 2p$) and let $U_{n+1} := \nabla_y g\left(X_{n+1},\widehat{\theta}_n\right)$. We have

$$\begin{aligned}\left\|\widehat{\theta}_{n+1}-\theta\right\|^{2p'} \leq{}& \left(\left\|\widehat{\theta}_n-\theta\right\|^2+\gamma_n^2\left\|U_{n+1}\right\|^2\right)^{p'}\\ &-2p'\gamma_n\left\langle\widehat{\theta}_n-\theta,U_{n+1}\right\rangle\left(\left\|\widehat{\theta}_n-\theta\right\|^2+\gamma_n^2\left\|U_{n+1}\right\|^2\right)^{p'-1}\\ &+\sum_{k=2}^{p'}\binom{p'}{k}\gamma_n^k\left|\left\langle\widehat{\theta}_n-\theta,U_{n+1}\right\rangle\right|^k\left(\left\|\widehat{\theta}_n-\theta\right\|^2+\gamma_n^2\left\|U_{n+1}\right\|^2\right)^{p'-k}. \qquad \text{(E.8)}\end{aligned}$$

The aim is to bound each term on the right-hand side of previous inequality. In this purpose, we first need to introduce some technical inequalities.

$$\begin{aligned}\left\|U_{n+1}\right\|^2 &\leq 2\left\|\nabla_y g\left(X_{n+1},\widehat{\theta}_n\right)\right\|^2+2\mathbb{E}\left[\left\|\nabla_y g\left(X_{n+1},\widehat{\theta}_n\right)\right\|^2\middle|\mathcal{F}_n\right]\\ &\leq 16\left(2\left\|\widehat{\theta}_n-\theta\right\|^2+2\left(r^\star\right)^2+\left\|\mu-X_{n+1}\right\|^2+r^2\mathbb{E}\left[W^2\right]\right).\end{aligned}$$

Thus, applying Lemma A.1 in [GB15] for instance, for all integer $k \leq p'$,

$$\left\|U_{n+1}\right\|^{2k} \leq 4^{k-1}16^k\left(2^k\left\|\widehat{\theta}_n-\theta\right\|^{2k}+2^k\left(r^\star\right)^{2k}+\left\|X_{n+1}-\mu\right\|^{2k}+r^{2k}\left(\mathbb{E}\left[W^2\right]\right)^k\right).$$

In a particular case, since for all $k \leq p$, $\mathbb{E}\left[W^{2k}\right] < +\infty$, there are positive constants $A_{1,k}, A_{2,k}$ such that for all $n \geq 1$,

$$\begin{aligned}\mathbb{E}\left[\left\|U_{n+1}\right\|^{2k}\middle|\mathcal{F}_n\right] &\leq 4^{3k-1}\left(2^k\left\|\widehat{\theta}_n-\theta\right\|^{2k}+2^k\left(r^\star\right)^{2k}+r^{2k}\mathbb{E}\left[W^{2k}\right]+r^{2k}\left(\mathbb{E}\left[W^2\right]\right)^k\right)\\ &\leq A_{1,k}\left\|\widehat{\theta}_n-\theta\right\|^{2k}+A_{2,k}. \qquad \text{(E.9)}\end{aligned}$$

We can now bound the expectation of the three terms on the right-hand side of inequality

(E.8). First, since $\widehat{\theta}_n$ is $\mathcal{F}_n$- measurable, applying inequality (E.9), let

$$\begin{aligned}(*) &:= \mathbb{E}\left[\left(\left\|\widehat{\theta}_n - \theta\right\| + \gamma_n^2 \left\|U_{n+1}\right\|^2\right)^{p'}\right] \\ &= \mathbb{E}\left[\left\|\widehat{\theta}_n - \theta\right\|^{2p'}\right] + \sum_{k=1}^{p'} \binom{p'}{k} \gamma_n^{2k} \mathbb{E}\left[\left\|U_{n+1}\right\|^{2k} \left\|\widehat{\theta}_n - \theta\right\|^{2p'-2k}\right] \\ &\leq \mathbb{E}\left[\left\|\widehat{\theta}_n - \theta\right\|^{2p'}\right] + \sum_{k=1}^{p'} \binom{p'}{k} \gamma_n^{2k} \mathbb{E}\left[\left(A_{1,k}\left\|\widehat{\theta}_n - \theta\right\|^{2k} + A_{2,k}\right)\left\|\widehat{\theta}_n - \theta\right\|^{2p'-2k}\right]\end{aligned}$$

Let $B := \sum_{k=1}^{p'} c_\gamma^{2k-2} A_{1,k}$, using previous inequality and by induction,

$$\begin{aligned}(*) &\leq \left(1 + B\gamma_n^2\right) \mathbb{E}\left[\left\|\widehat{\theta}_n - \theta\right\|^{2p'}\right] + \sum_{k=1}^{p'} \binom{p'}{k} \gamma_n^{2k} A_{2,k} \mathbb{E}\left[\left\|\widehat{\theta}_n - \theta\right\|^{2p'-2k}\right] \\ &\leq \left(1 + B\gamma_n^2\right) \mathbb{E}\left[\left\|\widehat{\theta}_n - \theta\right\|^{2p'}\right] + \sum_{k=1}^{p'} \binom{p'}{k} c_\gamma^{2k} A_{2,k} \frac{C_k}{n^{(p'+k)\alpha}} \\ &\leq \left(1 + B\gamma_n^2\right) \mathbb{E}\left[\left\|\widehat{\theta}_n - \theta\right\|^{2p'}\right] + O\left(\gamma_n^{p'+1}\right). \end{aligned} \tag{E.10}$$

In the same way, applying Cauchy-Schwarz's inequality, let

$$\begin{aligned}(**) &:= -2p'\gamma_n \mathbb{E}\left[\left\langle \widehat{\theta}_n - \theta, U_{n+1}\right\rangle \left(\left\|\widehat{\theta}_n - \theta\right\|^2 + \gamma_n^2 \left\|U_{n+1}\right\|^2\right)^{p'-1}\right] \\ &\leq -2p'\gamma_n \mathbb{E}\left[\left\langle \widehat{\theta}_n - \theta, U_{n+1}\right\rangle \left\|\widehat{\theta}_n - \theta\right\|^{2p'-2}\right] \\ &+ 2p'\gamma_n \mathbb{E}\left[\left\|\widehat{\theta}_n - \theta\right\| \left\|U_{n+1}\right\| \sum_{k=1}^{p'-1} \binom{p'-1}{k} \gamma_n^{2k} \left\|U_{n+1}\right\|^{2k} \left\|\widehat{\theta}_n - \theta\right\|^{2p'-2k}\right].\end{aligned}$$

Moreover, since $\widehat{\theta}_n$ is $\mathcal{F}_n$-measurable, applying Proposition 8.3.1,

$$\begin{aligned}-2p'\gamma_n \mathbb{E}\left[\left\langle \widehat{\theta}_n - \theta, U_{n+1}\right\rangle \left\|\widehat{\theta}_n - \theta\right\|^{2p'-2}\right] &= -2p'\gamma_n \mathbb{E}\left[\left\langle \widehat{\theta}_n - \theta, \mathbb{E}\left[U_{n+1}|\mathcal{F}_n\right]\right\rangle \left\|\widehat{\theta}_n - \theta\right\|^{2p'-2}\right] \\ &= -2p'\gamma_n \mathbb{E}\left[\left\langle \widehat{\theta}_n - \theta, \Phi(\widehat{\theta}_n)\right\rangle \left\|\widehat{\theta}_n - \theta\right\|^{2p'-2}\right] \\ &\leq -2p'c\gamma_n \mathbb{E}\left[\left\|\widehat{\theta}_n - \theta\right\|^{2p'}\right].\end{aligned}$$

Moreover, since $2ab \leq a^2 + b^2$, let

$$\begin{aligned}
(**') &:= 2p'\gamma_n \mathbb{E}\left[\left\|\widehat{\theta}_n - \theta\right\| \|U_{n+1}\| \sum_{k=1}^{p'-1} \binom{p'-1}{k} \gamma_n^{2k} \|U_{n+1}\|^{2k} \left\|\widehat{\theta}_n - \theta\right\|^{2p'-2k}\right] \\
&\leq p'\gamma_n \mathbb{E}\left[\left(\left\|\widehat{\theta}_n - \theta\right\|^2 + \|U_{n+1}\|^2\right) \sum_{k=1}^{p'-1} \binom{p'-1}{k} \gamma_n^{2k} \|U_{n+1}\|^{2k} \left\|\widehat{\theta}_n - \theta\right\|^{2p'-2k}\right] \\
&\leq p'\gamma_n \sum_{k=1}^{p'-1} \binom{p'-1}{k} \gamma_n^{2k} \\
&\left(\mathbb{E}\left[\|U_{n+1}\|^{2k+2} \left\|\widehat{\theta}_n - \theta\right\|^{2p'-2k}\right] + \mathbb{E}\left[\|U_{n+1}\|^{2k} \left\|\widehat{\theta}_n - \theta\right\|^{2p'+2-2k}\right]\right).
\end{aligned}$$

With analogous calculus to the ones for inequality (E.10), one can check that there is a positive constant B' such that for all $n \geq 1$,

$$(**') \leq B'\gamma_n^2 \mathbb{E}\left[\left\|\widehat{\theta}_n - \theta\right\|^{2p'}\right] + O\left(\gamma_n^{(p'+1)\alpha}\right).$$

Thus,

$$\begin{aligned}
-2\gamma_n \mathbb{E}\left[\gamma_n \left\langle \widehat{\theta}_n - \theta, U_{n+1}\right\rangle \left(\left\|\widehat{\theta}_n - \theta\right\|^2 + \gamma_n^2 \|U_{n+1}\|^2\right)^{p'-1}\right] &\leq \left(-2cp'\gamma_n + B'\gamma_n^2\right) \mathbb{E}\left[\left\|\widehat{\theta}_n - \theta\right\|^{2p'}\right] \\
&\quad + O\left(\gamma_n^{(p'+1)\alpha}\right).
\end{aligned} \tag{E.11}$$

Finally, applying Lemma A.1 in [GB15] and since $|\langle a, b\rangle| \leq \frac{1}{2}\|a\|^2 + \frac{1}{2}\|b\|^2$, let

$$\begin{aligned}
(***) &:= \sum_{k=2}^{p'} \binom{p'}{k} \gamma_n^k \mathbb{E}\left[\left|\left\langle \widehat{\theta}_n - \theta, U_{n+1}\right\rangle\right|^k \left(\left\|\widehat{\theta}_n - \theta\right\|^2 + \gamma_n^2 \|U_{n+1}\|^2\right)^{p'-k}\right] \\
&\leq \sum_{k=2}^{p'} \binom{p'}{k} \gamma_n^k \mathbb{E}\left[\left(\frac{1}{2}\left\|\widehat{\theta}_n - \theta\right\|^2 + \frac{1}{2}\|U_{n+1}\|^2\right)^k \left(\left\|\widehat{\theta}_n - \theta\right\|^2 + \gamma_n^2 \|U_{n+1}\|^2\right)^{p'-k}\right] \\
&\leq \sum_{k=2}^{p'} \binom{p'}{k} 2^{p'-k-2} \gamma_n^k \\
&\mathbb{E}\left[\left(\left\|\widehat{\theta}_n - \theta\right\|^{2k} + \|U_{n+1}\|^{2k}\right) \left(\left\|\widehat{\theta}_n - \theta\right\|^{2p'-2k} + \gamma_n^{2p'-2k} \|U_{n+1}\|^{2p'-2k}\right)\right]
\end{aligned}$$

Thus, with analogous calculus to the ones for inequality (E.10), one can check that there is a

positive constant B'' such that for all $n \geq 1$,

$$\sum_{k=2}^{p'} \binom{p'}{k} \gamma_n^k \mathbb{E}\left[\left| \left\langle \widehat{\theta}_n - \theta, U_{n+1} \right\rangle \right|^k \left(\left\| \widehat{\theta}_n - \theta \right\|^2 + \gamma_n^2 \left\| U_{n+1} \right\|^2 \right)^{p'-k} \right] \leq B'' \gamma_n^2 \mathbb{E}\left[\left\| \widehat{\theta}_n - \theta \right\|^{2p'} \right] + O\left(\gamma_n^{p'+1} \right). \tag{E.12}$$

Finally, applying inequalities (E.10) to (E.12), there are positive constants B_1, B_2 such that for all $n \geq 1$,

$$\mathbb{E}\left[\left\| \widehat{\theta}_{n+1} - \theta \right\|^{2p'} \right] \leq \left(1 - 2p'c\gamma_n + B_1\gamma_n^2\right) \mathbb{E}\left[\left\| \widehat{\theta}_n - \theta \right\|^{2p'} \right] + B_2 \gamma_n^{p'+1}. \tag{E.13}$$

Thus, with the help of an induction on n or applying a lemma of stabilization (see [Duf96] for instance), one can check that there is a positive constant $C_{p'}$ such that for all $n \geq 1$,

$$\mathbb{E}\left[\left\| \widehat{\theta}_n - \theta \right\|^{2p'} \right] \leq \frac{C_{p'}}{n^{p'\alpha}},$$

which concludes the induction on p' and the proof.

Bounding $\mathbb{P}\left[\widehat{\theta}_n - \gamma_n \nabla_y g\left(X_{n+1}, \widehat{\theta}_n \right) \notin \mathcal{K} \right]$

Let us recall that $d_{\min} = \inf_{y \in \partial\mathcal{K}} \|y - \theta\| > 0$ and that if W admits a $2p$-th moment, there is a positive constant C_p such that for all $n \geq 1$, $\mathbb{E}\left[\left\| \widehat{\theta}_n - \theta \right\|^{2p} \right] \leq \frac{C_p}{n^{p\alpha}}$. Thus, for all $n \geq 1$,

$$\begin{aligned} \frac{C_p}{(n+1)^{p\alpha}} &\geq \mathbb{E}\left[\left\| \widehat{\theta}_{n+1} - \theta \right\|^{2p} \right] \\ &\geq \mathbb{E}\left[\left\| \widehat{\theta}_{n+1} - \theta \right\|^{2p} \mathbf{1}_{\left\{ \widehat{\theta}_n - \gamma_n \nabla_y g\left(X_{n+1}, \widehat{\theta}_n \right) \notin \mathcal{K} \right\}} \right] \\ &\geq d_{\min}^{2p} \mathbb{P}\left[\widehat{\theta}_n - \gamma_n \nabla_y g\left(X_{n+1}, \widehat{\theta}_n \right) \notin \mathcal{K} \right]. \end{aligned}$$

Finally,

$$\mathbb{P}\left[\widehat{\theta}_n - \gamma_n \nabla_y g\left(X_{n+1}, \widehat{\theta}_n \right) \notin \mathcal{K} \right] \leq \frac{C_p}{d_{\min}^{2p}} \frac{1}{(n+1)^{p\alpha}} \leq \frac{C_p}{d_{\min}^{2p}} \frac{1}{n^{p\alpha}}.$$

□

Proof of Theorem 8.4.3. The aim is, in a first time, to exhibit a nice decomposition of the averaged algorithm. In this purpose, let us introduce this new decomposition of the PRM algorithm

$$\widehat{\theta}_{n+1} - \theta = \widehat{\theta}_n - \theta - \gamma_n \Phi\left(\widehat{\theta}_n \right) + \gamma_n \xi_{n+1} + r_n, \tag{E.14}$$

with

$$\xi_{n+1} := -\nabla_y g\left(X_{n+1}, \widehat{\theta}_n\right) + \Phi\left(\widehat{\theta}_n\right),$$
$$r_n := \pi\left(\widehat{\theta}_n - \gamma_n \nabla_y g\left(X_{n+1}, \widehat{\theta}_n\right)\right) - \widehat{\theta}_n + \gamma_n \nabla_y g\left(X_{n+1}, \widehat{\theta}_n\right).$$

Remark that (ξ_n) is a sequence of martingale differences adapted to the filtration $(\mathcal{F}_n)$ and r_n is equal to 0 when $\widehat{\theta}_n - \gamma_n \nabla_y g\left(X_{n+1}, \widehat{\theta}_n\right) \in \mathcal{K}$. Moreover, linearizing the gradient, decomposition (E.14) can be written as

$$\widehat{\theta}_{n+1} - \theta = \left(I_{\mathbb{R}^d \times \mathbb{R}} - \gamma_n \Gamma_\theta\right)\left(\widehat{\theta}_n - \theta\right) + \gamma_n \xi_{n+1} - \gamma_n \delta_n + r_n, \tag{E.15}$$

where $\delta_n := \Phi\left(\widehat{\theta}_n\right) - \Gamma_\theta\left(\widehat{\theta}_n - \theta\right)$ is the remainder term in the Taylor's expansion of the gradient. This can also be decomposed as

$$\Gamma_\theta\left(\widehat{\theta}_n - \theta\right) = \frac{\widehat{\theta}_n - \theta}{\gamma_n} - \frac{\widehat{\theta}_{n+1} - \theta}{\gamma_n} - \delta_n + \frac{r_n}{\gamma_n} + \xi_{n+1}.$$

As in [Pel00], summing these equalities, applying Abel's transform and dividing by n,

$$\Gamma_\theta\left(\overline{\theta}_n - \theta\right) = \frac{1}{n}\left(\frac{\widehat{\theta}_1 - \theta}{\gamma_1} - \frac{\widehat{\theta}_{n+1} - \theta}{\gamma_n} + \sum_{k=2}^{n}\left(\frac{1}{\gamma_k} - \frac{1}{\gamma_{k-1}}\right)\left(\widehat{\theta}_k - \theta\right) - \sum_{k=1}^{n}\delta_k + \sum_{k=1}^{n}\frac{r_k}{\gamma_k}\right) + \frac{1}{n}\sum_{k=1}^{n}\xi_{k+1}. \tag{E.16}$$

We now give the rate of convergence in quadratic mean of each term using Theorem 8.4.2. In this purpose, let us recall the following technical lemma.

Lemma E.2.1 ([GB15]). *Let $Y_1, ..., Y_n$ be random variables taking values in a normed vector space such that for all positive constant q and for all $k \geq 1$, $\mathbb{E}\left[\|Y_k\|^q\right] < \infty$. Thus, for all constants $a_1, ..., a_n$ and for all integer p,*

$$\mathbb{E}\left[\left\|\sum_{k=1}^{n} a_k Y_k\right\|^p\right] \leq \left(\sum_{k=1}^{n} |a_k| \left(\mathbb{E}\left[\|Y_k\|^p\right]\right)^{\frac{1}{p}}\right)^p \tag{E.17}$$

The remainder terms : First, one can check that

$$\frac{1}{n^2}\mathbb{E}\left[\left\|\frac{\widehat{\theta}_1 - \theta}{\gamma_1}\right\|^2\right] = o\left(\frac{1}{n}\right). \tag{E.18}$$

In the same way, applying Theorem 8.4.2,

$$\begin{aligned}\frac{1}{n^2}\mathbb{E}\left[\left\|\frac{\widehat{\theta}_{n+1}-\theta}{\gamma_n}\right\|^2\right] &= \frac{1}{c_\gamma^2}\frac{1}{n^{2-2\alpha}}\mathbb{E}\left[\left\|\widehat{\theta}_{n+1}-\theta\right\|^2\right] \\ &\leq \frac{C_1}{c_\gamma^2}\frac{1}{n^{2-\alpha}} \\ &= o\left(\frac{1}{n}\right). \end{aligned} \tag{E.19}$$

Moreover, since $\gamma_k^{-1}-\gamma_{k-1}^{-1} \leq 2\alpha c_\gamma^{-1}k^{\alpha-1}$, applying Lemma E.2.1,

$$\begin{aligned}\frac{1}{n^2}\mathbb{E}\left[\left\|\sum_{k=2}^{n}\left(\frac{1}{\gamma_k}-\frac{1}{\gamma_{k-1}}\right)\left(\widehat{\theta}_k-\theta\right)\right\|^2\right] &\leq \frac{1}{n^2}\left(\sum_{k=2}^{n}\left(\frac{1}{\gamma_k}-\frac{1}{\gamma_{k-1}}\right)\sqrt{\mathbb{E}\left[\left\|\widehat{\theta}_k-\theta\right\|^2\right]}\right)^2 \\ &\leq \frac{4\alpha^2 c_\gamma^{-2}C_1}{n^2}\left(\sum_{k=2}^{n}\frac{1}{k^{1-\alpha/2}}\right)^2 \\ &= O\left(\frac{1}{n^{2-\alpha}}\right) \\ &= o\left(\frac{1}{n}\right). \end{aligned} \tag{E.20}$$

Thanks to Lemma E.1.1, there is a positive constant C_θ such that for all $n \geq 1$,

$$\|\delta_n\| \leq C_\theta\left\|\widehat{\theta}_n-\theta\right\|^2.$$

Thus, applying Lemma E.2.1 and Theorem 8.4.2, there is a positive constant C_2 such that

$$\begin{aligned}\frac{1}{n^2}\mathbb{E}\left[\left\|\sum_{k=1}^{n}\delta_k\right\|^2\right] &\leq \frac{1}{n^2}\left(\sum_{k=1}^{n}\sqrt{\mathbb{E}\left[\|\delta_k\|^2\right]}\right)^2 \\ &\leq \frac{C_\theta^2}{n^2}\left(\sum_{k=1}^{n}\sqrt{\mathbb{E}\left[\left\|\widehat{\theta}_k-\theta\right\|^4\right]}\right)^2 \\ &\leq \frac{C_\theta^2 C_2}{n^2}\left(\sum_{k=1}^{n}\frac{1}{k^\alpha}\right)^2 \\ &= O\left(\frac{1}{n^{2\alpha}}\right) \\ &= o\left(\frac{1}{n}\right). \end{aligned}$$

Let $U_{n+1} := \nabla_y g\left(X_{n+1}, \widehat{\theta}_n\right)$, note that if $\widehat{\theta}_n - \gamma_n U_{n+1} \in \mathcal{K}$, then $r_n = 0$. Thus, applying Lemma E.2.1 and Cauchy-Schwarz's inequality,

$$\begin{aligned}
\frac{1}{n^2}\mathbb{E}\left[\left\|\sum_{k=1}^{n}\frac{r_k}{\gamma_k}\right\|^2\right] &\leq \frac{1}{n^2}\left(\sum_{k=1}^{n}\frac{1}{\gamma_k}\sqrt{\mathbb{E}\left[\|r_k\|^2\right]}\right)^2 \\
&= \frac{1}{n^2}\left(\sum_{k=1}^{n}\frac{1}{\gamma_k}\sqrt{\mathbb{E}\left[\|r_k\|^2 \mathbf{1}_{\widehat{\theta}_k - \gamma_k U_{k+1} \notin \mathcal{K}}\right]}\right)^2 \\
&\leq \frac{1}{n^2}\left(\sum_{k=1}^{n}\frac{1}{\gamma_k}\left(\mathbb{E}\left[\|r_k\|^4\right]\right)^{\frac{1}{4}}\left(\mathbb{P}\left[\widehat{\theta}_k - \gamma_k U_{k+1} \notin \mathcal{K}\right]\right)^{\frac{1}{4}}\right)^2.
\end{aligned}$$

Moreover, since π is 1-lipschitz,

$$\begin{aligned}
\|r_n\|^4 &= \left\|\pi\left(\widehat{\theta}_n - \gamma_n U_{n+1}\right) - \theta + \theta - \widehat{\theta}_n + \gamma_n U_{n+1}\right\|^4 \\
&\leq \left(\left\|\pi\left(\widehat{\theta}_n - \gamma_n U_{n+1}\right) - \pi(\theta)\right\| + \left\|\widehat{\theta}_n - \gamma_n U_{n+1} - \theta\right\|\right)^4 \\
&\leq \left(2\left\|\widehat{\theta}_n - \theta - \gamma_n U_{n+1}\right\|\right)^4 \\
&\leq 2^7\left\|\widehat{\theta}_n - \theta\right\|^4 + 2^7\gamma_n^4\|U_{n+1}\|^2.
\end{aligned}$$

Thus, applying inequality (E.9), there are positive constants A_1, A_2 such that for all $n \geq 1$,

$$\mathbb{E}\left[\|r_n\|^4 \middle| \mathcal{F}_n\right] \leq A_1\left\|\widehat{\theta}_n - \theta\right\|^4 + A_2\gamma_n^4.$$

In a particular case, applying Theorem 8.4.2, there is a positive constant A_3 such that for all $n \geq 1$,

$$\mathbb{E}\left[\|r_n\|^4\right] \leq \frac{A_3}{n^{2\alpha}}.$$

Moreover, applying Theorem 8.4.2, there is a positive constant C_6 such that for all $n \geq 1$,

$$\mathbb{P}\left[\widehat{\theta}_n - \gamma_n U_{n+1} \notin \mathcal{K}\right] \leq \frac{C_6}{d_{\min}^{12} n^{6\alpha}}.$$

Then,

$$\begin{aligned}\frac{1}{n^2}\mathbb{E}\left[\left\|\sum_{k=1}^{n}\frac{r_k}{\gamma_k}\right\|^2\right] &\leq \frac{\sqrt{C_6 A_3}}{d_{\min}^6 c_\gamma n^2}\left(\sum_{k=1}^{n}\frac{1}{k^\alpha}\right)^2 \\ &= O\left(\frac{1}{n^{2\alpha}}\right) \\ &= o\left(\frac{1}{n}\right). \end{aligned} \tag{E.21}$$

The martingale term : Since (ξ_n) is a sequence of martingale differences adapted to the filtration $(\mathcal{F}_n)$,

$$\begin{aligned}\frac{1}{n^2}\mathbb{E}\left[\left\|\sum_{k=1}^{n}\xi_{k+1}\right\|^2\right] &= \frac{1}{n^2}\sum_{k=1}^{n}\mathbb{E}\left[\left\|\xi_{k+1}\right\|^2\right] + \frac{2}{n^2}\sum_{k=1}^{n}\sum_{k'=k+1}^{n}\mathbb{E}\left[\left\langle\xi_{k+1},\xi_{k'+1}\right\rangle\right] \\ &= \frac{1}{n^2}\sum_{k=1}^{n}\mathbb{E}\left[\left\|\xi_{k+1}\right\|^2\right] + \frac{2}{n^2}\sum_{k=1}^{n}\sum_{k'=k+1}^{n}\mathbb{E}\left[\left\langle\xi_{k+1},\mathbb{E}\left[\xi_{k'+1}|\mathcal{F}_{k'}\right]\right\rangle\right] \\ &= \frac{1}{n^2}\sum_{k=1}^{n}\mathbb{E}\left[\left\|\xi_{k+1}\right\|^2\right].\end{aligned}$$

Moreover,

$$\begin{aligned}\mathbb{E}\left[\left\|\xi_{n+1}\right\|^2\right] &= \mathbb{E}\left[\left\|U_{n+1}\right\|^2 - 2\mathbb{E}\left[\left\langle\mathbb{E}\left[U_{n+1}|\mathcal{F}_n\right],\Phi(\widehat{\theta}_n\right\rangle\right] + \mathbb{E}\left[\left\|\Phi(\widehat{\theta}_n\right)\right\|^2\right] \\ &= \mathbb{E}\left[\left\|U_{n+1}\right\|^2\right] - \mathbb{E}\left[\left\|\Phi(\widehat{\theta}_n)\right\|^2\right] \\ &\leq \mathbb{E}\left[\left\|U_{n+1}\right\|^2\right].\end{aligned}$$

Finally, applying inequality (E.9) and Theorem 8.4.2, there is a positive constant M such that

$$\begin{aligned}\mathbb{E}\left[\left\|\xi_{n+1}\right\|^2\right] &\leq A_{1,1}\mathbb{E}\left[\left\|\widehat{\theta}_n - \theta\right\|^2\right] + A_{2,1} \\ &\leq M.\end{aligned}$$

Then,

$$\frac{1}{n^2}\mathbb{E}\left[\left\|\sum_{k=1}^{n}\xi_{k+1}\right\|^2\right] \leq \frac{1}{n^2}\sum_{k=1}^{n}M = \frac{M}{n},$$

which concludes the proof. □

Proof of Theorem 8.4.4. Let us recall that the averaged algorithm can be written as follows

$$\sqrt{n}\Gamma_\theta\left(\overline{\theta}_n-\theta\right) = \frac{1}{\sqrt{n}}\left(\frac{\widehat{\theta}_1-\theta}{\gamma_1}-\frac{\widehat{\theta}_{n+1}-\theta}{\gamma_n}+\sum_{k=2}^{n}\left(\frac{1}{\gamma_k}-\frac{1}{\gamma_{k-1}}\right)\left(\widehat{\theta}_k-\theta\right)-\sum_{k=1}^{n}\delta_k+\sum_{k=1}^{n}\frac{r_k}{\gamma_k}\right) + \frac{1}{\sqrt{n}}\sum_{k=1}^{n}\xi_{k+1}. \tag{E.22}$$

We now prove that the first terms on the right-hand side of previous equality converge in probability to 0 and apply a Central Limit Theorem to the last one.

The remainder terms : Applying inequalities (E.18) to (E.21),

$$\begin{aligned}
\frac{1}{\sqrt{n}}\frac{\widehat{\theta}_1-\theta}{\gamma_1} &\xrightarrow[n\to\infty]{\mathbb{P}} 0,\\
\frac{1}{\sqrt{n}}\frac{\widehat{\theta}_{n+1}-\theta}{\gamma_n} &\xrightarrow[n\to\infty]{\mathbb{P}} 0,\\
\frac{1}{\sqrt{n}}\sum_{k=2}^{n}\left(\frac{1}{\gamma_k}-\frac{1}{\gamma_{k-1}}\right)\left(\widehat{\theta}_k-\theta\right) &\xrightarrow[n\to\infty]{\mathbb{P}} 0,\\
\frac{1}{\sqrt{n}}\sum_{k=1}^{n}\frac{r_k}{\gamma_k} &\xrightarrow[n\to\infty]{\mathbb{P}} 0.
\end{aligned}$$

The martingale term Let $\widehat{\theta}_n = (Z_n, A_n) \in \mathbb{R}^d \times \mathbb{R}$, then ξ_{n+1} can be written as $\xi_{n+1} =$

$\xi'_{n+1} + \epsilon_{n+1} + \epsilon'_{n+1}$, with

$$\xi'_{n+1} := \begin{pmatrix} \mu - X_{n+1} - \mathbb{E}\left[\mu - X_{n+1} | \mathcal{F}_n\right] - r^\star \left(\frac{\mu - X_{n+1}}{\|\mu - X_{n+1}\|} - \mathbb{E}\left[\frac{\mu - X_{n+1}}{\|\mu - X_{n+1}\|} \middle| \mathcal{F}_n \right] \right) \\ r^\star - \|\mu - X_{n+1}\| - r^\star + \mathbb{E}\left[\|\mu - X_{n+1}\| \middle| \mathcal{F}_n \right] \end{pmatrix},$$

$$\epsilon_{n+1} := - \begin{pmatrix} (A_n - r^\star) \left(\frac{Z_n - X_{n+1}}{\|Z_n - X_{n+1}\|} - \mathbb{E}\left[\frac{Z_n - X_{n+1}}{\|Z_n - X_{n+1}\|} \middle| \mathcal{F}_n \right] \right) \\ \|Z_n - X_{n+1}\| - \mathbb{E}\left[\|Z_n - X_{n+1}\| \, \mathcal{F}_n \right] - \|\mu - X_{n+1}\| + \mathbb{E}\left[\|\mu - X_{n+1}\| \middle| \mathcal{F}_n \right] \end{pmatrix},$$

$$\epsilon'_{n+1} := - \begin{pmatrix} r^\star \left(\frac{Z_n - X_{n+1}}{\|X_{n+1} - Z_n\|} - \frac{\mu - X_{n+1}}{\|\mu - X_{n+1}\|} - \mathbb{E}\left[\frac{Z_n - X_{n+1}}{\|X_{n+1} - Z_n\|} - \frac{\mu - X_{n+1}}{\|\mu - X_{n+1}\|} \middle| \mathcal{F}_n \right] \right) \\ 0 \end{pmatrix}.$$

Note that $(\xi_n), (\epsilon_n), (\epsilon'_n)$ are martingale differences sequences adapted to the filtration $(\mathcal{F}_n)$. Thus,

$$\begin{aligned} \frac{1}{n}\mathbb{E}\left[\left\| \sum_{k=1}^n \epsilon_{k+1} \right\|^2 \right] &= \frac{1}{n} \sum_{k=1}^n \mathbb{E}\left[\|\epsilon_{k+1}\|^2 \right] + \frac{2}{n} \sum_{k=1}^n \sum_{k'=k+1}^n \mathbb{E}\left[\langle \epsilon_{k+1}, \epsilon_{k'+1} \rangle \right] \\ &= \frac{1}{n} \sum_{k=1}^n \mathbb{E}\left[\|\epsilon_{k+1}\|^2 \right] + \frac{2}{n} \sum_{k=1}^n \sum_{k'=k+1}^n \mathbb{E}\left[\langle \epsilon_{k+1}, \mathbb{E}\left[\epsilon_{k'+1} | \mathcal{F}_{k'} \right] \rangle \right] \\ &= \frac{1}{n} \sum_{k=1}^n \mathbb{E}\left[\|\epsilon_{k+1}\|^2 \right]. \end{aligned}$$

Moreover,

$$\mathbb{E}\left[\|\epsilon_{n+1}\|^2 | \mathcal{F}_n \right]$$

$$= \mathbb{E}\left[\left\| \begin{pmatrix} (A_n - r^\star) \frac{Z_n - X_{n+1}}{\|Z_n - X_{n+1}\|} \\ \|Z_n - X_{n+1}\| - \|\mu - X_{n+1}\| \end{pmatrix} \right\|^2 \middle| \mathcal{F}_n \right] + \left\| \begin{pmatrix} (A_n - r^\star) \mathbb{E}\left[\frac{Z_n - X_{n+1}}{\|Z_n - X_{n+1}\|} \middle| \mathcal{F}_n \right] \\ \mathbb{E}\left[\|Z_n - X_{n+1}\| - \|\mu - X_{n+1}\| \, | \mathcal{F}_n \right] \end{pmatrix} \right\|^2$$

$$- 2\mathbb{E}\left[\left\langle \begin{pmatrix} (A_n - r^\star) \frac{Z_n - X_{n+1}}{\|Z_n - X_{n+1}\|} \\ \|Z_n - X_{n+1}\| - \|\mu - X_{n+1}\| \end{pmatrix}, \begin{pmatrix} (A_n - r^\star) \mathbb{E}\left[\frac{Z_n - X_{n+1}}{\|Z_n - X_{n+1}\|} \middle| \mathcal{F}_n \right] \\ \mathbb{E}\left[\|Z_n - X_{n+1}\| - \|\mu - X_{n+1}\| \, | \mathcal{F}_n \right] \end{pmatrix} \right\rangle \middle| \mathcal{F}_n \right].$$

Thus, one can check that

$$\begin{aligned}\mathbb{E}\left[\|\epsilon_{n+1}\|^2 |\mathcal{F}_n\right] &\leq \mathbb{E}\left[\left\|\begin{pmatrix} -(A_n - r^\star)\frac{Z_n - X_{n+1}}{\|Z_n - X_{n+1}\|} \\ \|Z_n - X_{n+1}\| - \|\mu - X_{n+1}\| \end{pmatrix}\right\|^2 \Big| \mathcal{F}_n\right] \\ &\leq \|A_n - r^\star\|^2 + \|Z_n - \mu\|^2 \\ &= \left\|\widehat{\theta}_n - \theta\right\|^2.\end{aligned}$$

Thus, applying Theorem 8.4.2,

$$\begin{aligned}\frac{1}{n}\mathbb{E}\left[\left\|\sum_{k=1}^n \epsilon_{k+1}\right\|^2\right] &= \frac{1}{n}\sum_{k=1}^n \mathbb{E}\left[\|\epsilon_{k+1}\|^2\right] \\ &\leq \frac{1}{n}\sum_{k=1}^n \mathbb{E}\left[\left\|\widehat{\theta}_n - \theta\right\|^2\right] \\ &\leq \frac{1}{n}\sum_{k=1}^n \frac{C'}{n^\alpha} \\ &= O\left(\frac{1}{n^\alpha}\right).\end{aligned}$$

As a particular case,

$$\frac{1}{\sqrt{n}}\sum_{k=1}^n \epsilon_{k+1} \xrightarrow[n\to\infty]{\mathbb{P}} 0.$$

Similarly, since $\left\|\frac{Z_n - X_{n+1}}{\|Z_n - X_{n+1}\|} - \frac{\mu - X_{n+1}}{\|\mu - X_{n+1}\|}\right\| \leq 2$,

$$\begin{aligned}\mathbb{E}\left[\left\|\epsilon'_{n+1}\right\|^2 |\mathcal{F}_n\right] &\leq \mathbb{E}\left[\left\|\begin{pmatrix} r^\star\left(\frac{Z_n - X_{n+1}}{\|X_{n+1} - Z_n\|} - \frac{\mu - X_{n+1}}{\|\mu - X_{n+1}\|}\right) \\ 0 \end{pmatrix}\right\|^2 \Big| \mathcal{F}_n\right] \\ &\leq 2\left(r^\star\right)^2 \mathbb{E}\left[\left\|\frac{Z_n - X_{n+1}}{\|X_{n+1} - Z_n\|} - \frac{\mu - X_{n+1}}{\|\mu - X_{n+1}\|}\right\| \Big| \mathcal{F}_n\right].\end{aligned}$$

This last term is closely related to the gradient of the function we need to minimize to get the geometric median (see [Kem87]) for example) and it is proved in [CCGB15] that since Lemma E.1.1 is verified, then

$$\mathbb{E}\left[\left\|\frac{Z_n - X_{n+1}}{\|X_{n+1} - Z_n\|} - \frac{\mu - X_{n+1}}{\|\mu - X_{n+1}\|}\right\| \Big| \mathcal{F}_n\right] \leq C\|Z_n - \mu\| \leq C\left\|\widehat{\theta}_n - \theta\right\|.$$

Thus,

$$\mathbb{E}\left[\left\|\epsilon_{n+1}\right\|^2 |\mathcal{F}_n\right] \leq 2C\left(r^\star\right)^2 \left\|\widehat{\theta}_n - \theta\right\|.$$

Finally, since (ϵ_n) is a sequence of martingale differences adapted to the filtration $(\mathcal{F}_n)$, applying Theorem 8.4.2 and Cauchy-Schwarz's inequality,

$$\begin{aligned}
\frac{1}{n}\mathbb{E}\left[\left\|\sum_{k=1}^{n}\epsilon'_{k+1}\right\|^2\right] &= \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}\left[\left\|\epsilon'_{k+1}\right\|^2\right] \\
&\leq 2C\left(r^\star\right)^2\frac{1}{n}\sum_{k=1}^{n}\mathbb{E}\left[\left\|\widehat{\theta}_n - \theta\right\|\right] \\
&\leq 2C\left(r^\star\right)^2\sqrt{C_1}\frac{1}{n}\sum_{k=1}^{n}\frac{1}{k^{\alpha/2}} \\
&= O\left(\frac{1}{n^{\alpha/2}}\right).
\end{aligned}$$

Note that with more assumptions on W, we could get a better rate but this one is sufficient. Indeed, thanks to previous inequality,

$$\frac{1}{\sqrt{n}}\sum_{k=1}^{n}\epsilon'_{k+1} \xrightarrow[n\to\infty]{\mathbb{P}} 0.$$

Finally, applying a Central Limit Theorem (see [Duf97]) for example), we have the convergence in law

$$\frac{1}{\sqrt{n}}\sum_{k=1}^{n}\xi'_{k+1} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}\left(0,\Sigma\right), \tag{E.23}$$

with

$$\Sigma := \mathbb{E}\left[\begin{pmatrix}\mu - X - r^\star\frac{\mu-X}{\|\mu-X\|}\\ r^\star - \|\mu - X\|\end{pmatrix}\otimes\begin{pmatrix}\mu - X - r^\star\frac{\mu-X}{\|\mu-X\|}\\ r^\star - \|\mu - X\|\end{pmatrix}\right],$$

which also can be written as

$$\Sigma = \mathbb{E}\left[\begin{pmatrix}r^\star U_\Omega - rWU_\Omega\\ r^\star - rW\end{pmatrix}\otimes\begin{pmatrix}r^\star U_\Omega - rWU_\Omega\\ r^\star - rW\end{pmatrix}\right].$$

Thus, we have the convergence in law

$$\sqrt{n}\Gamma_\theta\left(\overline{\theta}_n - \theta\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}\left(0,\Sigma\right),$$

and in a particular case,

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

□

Conclusion et perspectives

Dans ma thèse, je me suis d'abord concentré sur des algorithmes de gradient stochastiques moyennés pour estimer de manière rapide et efficace la médiane géométrique. Nous avons conforté l'intérêt de telles approches algorithmiques en établissant les vitesses de convergence L^p de ces estimateurs ([GB15]) et en construisant des boules de confiance non asymptotiques ([CCGB15]), à l'aide d'une nouvelle technique de démonstration basée sur une double récurrence. Nous avons ensuite introduit un algorithme de gradient stochastique et sa version moyennée pour l'estimation de la "Median Covariation Matrix", et grâce au travail effectué sur la médiane, nous avons obtenu les bonnes vitesses de convergence en moyenne quadratique ([CGB15]) de ces estimateurs. Ces différents travaux ont permis d'exhiber les mêmes vitesses de convergence presque sûre et L^p des algorithmes de Robbins-Monro et de leurs moyennés dans un cadre plus général (voir Chapitre 7).

Un premier prolongement serait de s'inspirer du cadre général introduit au Chapitre 7 pour donner des hypothèses garantissant la normalité asymptotique des estimateurs moyennés. L'obtention d'un tel résultat ainsi que d'un estimateur de la covariance du moyenné permettrait l'obtention de "boules" de confiance plus précises, par exemple, pour les estimateurs de la médiane. En effet, au Chapitre 4, nous avons obtenu des boules de confiance uniquement basées sur la plus petite valeur propre de la Hessienne de la fonction que l'on voulait minimiser, et donc sans prendre en compte les autres valeurs propres, ce qui, par conséquent, entraine une perte d'information. Cependant, nous avons vu que dans le cas où l'on traite de gros échantillons à valeurs dans des espaces de grande dimension, estimer la covariance peut être très coûteux en terme de temps de calcul. Pour contourner ce problème, il est notamment important d'éviter une inversion de matrice à chaque itération, comme le ferait un algorithme de Newton itératif, et il serait par conséquent intéressant d'introduire un algorithme récursif s'inspirant de celui introduit par Gahbiche et Pelletier ([GP00]). L'objectif est alors de donner des conditions suffisantes pour établir les vitesses de convergence presque sûre et en moyenne quadratique des estimateurs récursifs de la covariance.

J'aimerais également utiliser les idées introduites dans ma thèse pour estimer la médiane

conditionnelle à l'aide d'algorithmes de gradient stochastiques pondérés (voir [CCZ12]). Ces estimateurs consistent à introduire des poids contrôlés par un noyau de lissage ("kernel smoother") dans l'estimateur de la médiane étudié aux Chapitres 4 et 5. Ces poids, qui sont contrôlés par une fenêtre, permettent d'approcher l'espérance conditionnelle, et donc le gradient de la fonction que l'on veut minimiser (voir [Rév77] parmi d'autres). La normalité asymptotique de l'estimateur moyenné a déjà été établie par [CCZ12]. A l'aide des méthodes introduites dans cette thèse, on pourrait obtenir les vitesses de convergence presque sûre et L^p des algorithmes moyennés ainsi que des boules de confiance non-asymptotiques avant d'étendre ces résultats dans le cadre plus général du Chapitre 7.

Dans la dernière partie de ma thèse, j'ai utilisé des algorithmes de gradient projetés pour estimer les paramètres d'une loi sphérique tronquée. Une perspective serait, en s'inspirant du Chapitre 7, de donner un cadre général garantissant les bonnes vitesses de convergence presque sûre et L^p des algorithmes projetés et de leurs versions moyennées. Il serait également intéressant d'établir la normalité asymptotique des algorithmes moyennés et donner un estimateur récursif de la covariance.

De plus, des simulations (voir Figure 8.4) laissent penser que les estimateurs introduits au Chapitre 8 ont une moins bonne variance que l'algorithme de back-fitting introduit par [BP14]. Cependant, aucun résultat de convergence n'est établi pour cet algorithme dans le cas d'une sphère tronquée, et des simulations montrent qu'il peut diverger. Une perspective serait donc d'introduire deux nouveaux algorithmes : un algorithme de back-fitting projeté, et un algorithme de gradient projeté pour estimer dans un premier temps le centre de la sphère, et dans un deuxième temps le rayon (et non plus simultanément). Les tentatives de simulations sont dans ce sens très encourageantes.

A plus long terme, j'aimerais étudier la convergence des algorithmes de gradient stochastiques moyennés et pondérés introduits par [DR97], de la forme

$$\overline{Z}_{n,\beta} = \frac{1}{\sum_{k=1}^{n} k^{\beta}} \sum_{k=1}^{n} k^{\beta} Z_k,$$

avec $\beta \geq 0$, et $(Z_n)_{n\geq 1}$ une suite d'estimateurs obtenue à l'aide d'un algorithme de Robbins-Monro. L'intérêt de ces algorithmes est qu'ils permettent, à travers les pondérations, d'obtenir un algorithme avec une vitesse en moyenne quadratique de l'ordre de $\frac{1}{n}$ qui soit moins sensible aux mauvaises initialisations que l'algorithme moyenné. Cependant, ces algorithmes ont une plus grande variance asymptotique. Une idée serait donc, en "mixant" ces deux algorithmes, de donner un estimateur récursif qui soit moins sensible aux mauvaises initialisations pour des petites tailles d'échantillon, tout en conservant une variance asymptotique

optimale.

Enfin, en raison de l'évolution des méthodes permettant de recueillir des données, l'"estimation distribuée" est un outil de plus en plus utilisé en statistique (voir [BFL$^+$15] et [BZ14] parmi d'autres). Ce type de méthode consiste à utiliser plusieurs "machines" qui récoltent en parallèle et traitent indépendamment des données, avant de "centraliser" les résultats. Un objectif à long terme serait donc, à travers le cadre mis en place au Chapitre 7, d'étudier comment adapter l'algorithme de gradient et sa version moyennée à ce contexte, tout en conservant les mêmes vitesses de convergence.

Table des figures

Bibliographie

[ADPY12] Marc Arnaudon, Clément Dombry, Anthony Phan, and Le Yang. Stochastic algorithms for computing means of probability measures. *Stochastic Processes and their Applications*, 122(4) :1437–1455, 2012.

[ANC13] Anas Abuzaina, Mark S Nixon, and John N Carter. Sphere detection in kinect point clouds via the 3d hough transform. pages 290–297. Springer, 2013.

[AS14] Ali Al-Sharadqah. Further statistical analysis of circle fitting. *Electronic Journal of Statistics*, 8(2) :2741–2778, 2014.

[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.

[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.

[BDF13] Akshay Balsubramani, Sanjoy Dasgupta, and Yoav Freund. The fast convergence of incremental pca. In *Advances in Neural Information Processing Systems*, pages 3174–3182, 2013.

[BDM01] Nathalie Bartoli and Pierre Del Moral. Simulation et algorithmes stochastiques. *Cépaduès éditions*, page 5, 2001.

[BF85] Leo Breiman and Jerome H Friedman. Estimating optimal transformations for multiple regression and correlation. *Journal of the American statistical Association*, 80(391) :580–598, 1985.

[BF12] Bernard Bercu and Philippe Fraysse. A robbins–monro procedure for estimation in semiparametric regression models. *The Annals of Statistics*, 40(2) :666–693, 2012.

[BFL+15] Heather Battey, Jianqing Fan, Han Liu, Junwei Lu, and Ziwei Zhu. Distributed estimation and inference with statistical guarantees. *arXiv :1509.05457*, 2015.

[Bil13] Patrick Billingsley. *Convergence of probability measures*. John Wiley & Sons, 2013.

[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.

[BMP90] Albert Benveniste, Michel Métivier, and Pierre Priouret. *Adaptive Algorithms and Stochastic Approximations*, volume 22 of *Applications of Mathematics*. Springer-Verlag, New York, 1990.

[Bos00] Denis Bosq. *Linear processes in function spaces*, volume 149 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 2000. Theory and applications.

[Bot10] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.

[BP12] Amir Beck and Dror Pan. On the solution of the gps localization and circle fitting problems. *SIAM Journal on Optimization*, 22(1) :108–134, 2012.

[BP14] Denis Brazey and Bruno Portier. A new spherical mixture model for head detection in depth images. *SIAM Journal on Imaging Sciences*, 7(4) :2423–2447, 2014.

[BS15] Amir Beck and Shoham Sabach. Weiszfeld's method : old and new results. *J. Optim. Theory Appl.*, 164(1) :1–40, 2015.

[BSGV14] Enea G Bongiorno, Ernesto Salinelli, Aldo Goia, and Philippe Vieu. *Contributions in infinite-dimensional statistics and related topics*. Società Editrice Esculapio, 2014.

[BV04] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[BZ14] Gérard Biau and Ryad Zenine. Online asynchronous distributed regression. *arXiv :1407.4373*, 2014.

[Cad01] Benoît Cadre. Convergent estimators for the L_1-median of a Banach valued random variable. *Statistics*, 35(4) :509–521, 2001.

[CC14] Anirvan Chakraborty and Probal Chaudhuri. The spatial distribution in infinite dimensional spaces and related quantiles and depths. *The Annals of Statistics*, 42 :1203–1231, 2014.

[CCC10] Hervé Cardot, Peggy Cénac, and Mohamed Chaouch. Stochastic approximation to the multivariate and the functional median. In Y. Lechevallier and G. Saporta, editors, *Compstat 2010*, pages 421–428. Physica Verlag, Springer., 2010.

[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.

[CCZ12] Hervé Cardot, Peggy Cénac, and Pierre-André Zitt. Recursive estimation of the conditional geometric median in Hilbert spaces. *Electronic Journal of Statistics*, 6 :2535–2562, 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.

[CD15] Hervé Cardot and David Degras. Online principal components analysis : which algorithm to choose ? Technical report, Institut de Mathématiques de Bourgogne, France, 2015.

[CDPB09] Yixin Chen, Xin Dang, Hanxiang Peng, and Henry L Bart. Outlier detection with the kernelized spatial depth function. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(2) :288–305, 2009.

[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.

[CGB15] Hervé Cardot and Antoine Godichon-Baggioni. Fast estimation of the median covariation matrix with application to online robust principal components analysis. *arXiv :1504.02852*, 2015.

[CGER07] Jean Cupidon, David Gilliam, Randall Eubank, and Frits Ruymgaart. The delta method for analytic functions of random operators with application to functional data. *Bernoulli*, 13 :1179–1194, 2007.

[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.

[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.

[Cue88] Antonio Cuevas. Qualitative robustness in abstract inference. *Journal of statistical planning and inference*, 18(3) :277–289, 1988.

[Cue14] Antonio Cuevas. A partial overview of the theory of statistics with functional data. *Journal of Statistical Planning and Inference*, 147 :1–23, 2014.

[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.

[DJ92] Bernard Delyon and Anatoli Juditsky. Stochastic optimization with averaging of trajectories. *Stochastics : An International Journal of Probability and Stochastic Processes*, 39(2-3) :107–118, 1992.

[DJ93] Bernard Delyon and Anatoli Juditsky. Accelerated stochastic approximation. *SIAM Journal on Optimization*, 3(4) :868–881, 1993.

[DPR82] Jacques Dauxois, Alain Pousse, and Yves Romain. Asymptotic theory for principal components analysis of a random vector function : some applications to statistical inference. *Journal of Multivariate Analysis*, 12 :136–154, 1982.

[DR97] Jürgen Dippon and Joachim Renz. Weighted means in stochastic approximation of minima. *SIAM Journal on Control and Optimization*, 35(5) :1811–1827, 1997.

[Duf96] Marie Duflo. *Algorithmes stochastiques*. Springer Berlin, 1996.

[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.

[FB81] Martin Fischler and Robert Bolles. Random sample consensus : a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6) :381–395, 1981.

[FFC12] Heinrich Fritz, Peter Filzmoser, and Christophe Croux. A comparison of algorithms for the multivariate L_1-median. *Comput. Stat.*, 27 :393–410, 2012.

[FV06] Frédéric Ferraty and Philippe Vieu. *Nonparametric functional data analysis : theory and practice*. Springer Science & Business Media, 2006.

[FVJ09] Thomas Fletcher, Suresh Venkatasubramanian, and Sarang Joshi. The geometric median on riemannian manifolds with application to robust atlas estimation. *NeuroImage*, 45(1) :S143–S152, 2009.

[GB15] 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*, 2015.

[GBP16] Antoine Godichon-Baggioni and Bruno Portier. An averaged projected robbins-monro algorithm for estimating the parameters of a truncated spherical distribution. *arXiv preprint arXiv :1606.04276*, 2016.

[Gee00] Sara A Geer. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.

[Ger08] Daniel Gervini. Robust functional estimation using the median and spherical principal components. *Biometrika*, 95(3) :587–600, 2008.

[GP00] Mouna Gahbiche and Mariane Pelletier. On the estimation of the asymptotic covariance matrix for the averaged robbins–monro algorithm. *Comptes Rendus de l'Académie des Sciences-Series I-Mathematics*, 331(3) :255–260, 2000.

[Hal48] J. B. S. Haldane. Note on the median of a multivariate distribution. *Biometrika*, 35(3-4) :414–417, 1948.

[Ham71] Frank R Hampel. A general qualitative definition of robustness. *The Annals of Mathematical Statistics*, pages 1887–1896, 1971.

[HP06] Marc Hallin and Davy Paindaveine. Semiparametrically efficient rank-based inference for shape. i. optimal rank-based tests for sphericity. *The Annals of Statistics*, 34(6) :2707–2756, 2006.

[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.

[Hub64] Peter Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1) :73–101, 1964.

[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.

[JJ98] Bernard C Jiang and SJ Jiang. Machine vision based inspection of oil seals. *Journal of manufacturing systems*, 17(3) :159–166, 1998.

[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.

[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.

[KP12] David Kraus and Victor M. Panaretos. Dispersion operators and resistant second-order functional data analysis. *Biometrika*, 99 :813–832, 2012.

[KSZ12] Volker Krätschmer, Alexander Schied, and Henryk Zähle. Qualitative and infinitesimal robustness of tail-dependent statistical functionals. *Journal of Multivariate Analysis*, 103(1) :35–47, 2012.

[Kuh73] Harold W Kuhn. A note on Fermat's problem. *Mathematical programming*, 4(1) :98–107, 1973.

[KY03] Harold J Kushner and George Yin. *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer Science & Business Media, 2003.

[Lan87] UM Landau. Estimation of a circular arc center and its radius. *Computer Vision, Graphics, and Image Processing*, 38(3) :317–326, 1987.

[LJSB12] Simon Lacoste-Julien, Mark Schmidt, and Francis Bach. A simpler approach to obtaining an O (1/t) convergence rate for the projected stochastic subgradient method. *arXiv preprint arXiv :1212.2002*, 2012.

[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.

[LW14] Jin Liu and Zhong-ke Wu. An adaptive approach for primitive shape extraction from point clouds. *Optik-International Journal for Light and Electron Optics*, 125(9) :2000–2008, 2014.

[Min14] Stanislav Minsker. Geometric median and robust estimation in Banach spaces. *Bernoulli, to appear*, 2014.

[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.

[MNO10] Jyrki Möttönen, Klaus Nordhausen, and Hannu Oja. Asymptotic theory of the spatial median. In *Nonparametrics and Robustness in Modern Statistical Inference and Time Series Analysis : A Festschrift in honor of Professor Jana Jurečková*, volume 7, pages 182–193. IMS Collection, 2010.

[MNP08] Daniel A Martins, António JR Neves, and Armando J Pinho. Real-time generic ball recognition in robocup domain. In *Proc. of the 3rd International Workshop on Intelligent Robotics, IROBOT*, pages 37–48, 2008.

[MP06] Abdelkader Mokkadem and Mariane Pelletier. Convergence rate and averaging of nonlinear two-time-scale stochastic approximation algorithms. *Ann. Appl. Probab.*, 16(3) :1671–1702, 2006.

[MSLD14] Stanislav Minsker, Sanvesh Srivastava, Lizhen Lin, and David Dunson. Scalable and robust bayesian inference via the median posterior. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1656–1664, 2014.

[Mui09] Robb J Muirhead. *Aspects of multivariate statistical theory*, volume 197. John Wiley & Sons, 2009.

[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.

[NNY94] Yurii Nesterov, Arkadii Nemirovskii, and Yinyu Ye. *Interior-point polynomial algorithms in convex programming*, volume 13. SIAM, 1994.

[ON85] Hannu Oja and Ahti Niinimaa. Asymptotic properties of the generalized median in the case of multivariate normality. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 372–377, 1985.

[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.

[Pet95] Valentin V Petrov. Limit theorems of probability theory. sequences of independent random variables, vol. 4 of. *Oxford Studies in Probability*, 1995.

[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.

[R D10] R Development Core Team. *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2010. ISBN 3-900051-07-0.

[Rab06] Tahir Rabbani. Automatic reconstruction of industrial installations using point clouds and images. Technical report, NCG Nederlandse Commissie voor Geodesie Netherlands Geodetic Commission, 2006.

[Rév77] Pál Révész. How to apply the method of stochastic approximation in the nonparametric estimation of a regression function 1. *Statistics : A Journal of Theoretical and Applied Statistics*, 8(1) :119–126, 1977.

[RM51] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[Roc15] Ralph Tyrell Rockafellar. *Convex analysis*. Princeton university press, 2015.

[RS85] Herbert Robbins and David Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In *Herbert Robbins Selected Papers*, pages 111–135. Springer, 1985.

[RS05] James O. Ramsay and Bernard W. Silverman. *Functional Data Analysis*. Springer, New York, second edition, 2005.

[RTKD03] Corneliu Rusu, Marius Tico, Pauli Kuosmanen, and Edward J Delp. Classical geometrical approach to circle fitting—review and new developments. *Journal of Electronic Imaging*, 12(1) :179–193, 2003.

[Rud14] Mark Rudelson. Recent developments in non-asymptotic theory of random matrices. *Modern Aspects of Random Matrix Theory*, 72 :83, 2014.

[Rup88] David Ruppert. Efficient estimations from a slowly convergent robbins-monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988.

[RV10] Mark Rudelson and Roman Vershynin. Non-asymptotic theory of random matrices : extreme singular values. In *Proceedings of the International Congress of Mathematicians. Volume III*, pages 1576–1602. Hindustan Book Agency, New Delhi, 2010.

[RvD99] Peter Rousseeuw and Katrien van Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41 :212–223, 1999.

[Ser06] Robert Serfling. Depth functions in nonparametric multivariate inference. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, 72 :1, 2006.

[Sma90] Christopher G. Small. A survey of multidimensional medians. *International Statistical Review / Revue Internationale de Statistique*, 58(3) :263–277, 1990.

[Sma96] Florentin Smarandache. *Collected Papers, Vol. I*, volume 1. Infinite Study, 1996.

[SR05] Bernard Silverman and James Ramsay. *Functional Data Analysis*. Springer, 2005.

[STG01] M Shahid Shafiq, S Turgut Tümer, and H Cenk Güler. Marker detection and trajectory generation algorithms for a multicamera based gait analysis system. *Mechatronics*, 11(4) :409–437, 2001.

[SWK07] Ruwen Schnabel, Roland Wahl, and Reinhard Klein. Efficient ransac for point-cloud shape detection. In *Computer graphics forum*, volume 26, pages 214–226. Wiley Online Library, 2007.

[TCL15] Trung-Thien Tran, Van-Toan Cao, and Denis Laurendeau. Extraction of cylinders and estimation of their parameters from point clouds. *Computers & Graphics*, 46 :345–357, 2015.

[Tho55] Alexander Thom. A statistical examination of the megalithic sites in britain. *Journal of the Royal Statistical Society. Series A (General)*, pages 275–295, 1955.

[TKO12] Sara Taskinen, Inge Koch, and Hannu Oja. Robustifying principal components analysis with spatial sign vectors. *Statist. and Probability Letters*, 82 :765–774, 2012.

[TY14] Pierre Tarrès and Yuan Yao. Online learning as stochastic approximation of regularization paths : optimality and almost-sure convergence. *IEEE Trans. Inform. Theory*, 60 :5716–5735, 2014.

[VHSR05] Felix Von Hundelshausen, Michael Schreiber, and Raúl Rojas. A constructive feature detection approach for robotic vision. In *RoboCup 2004 : Robot Soccer World Cup VIII*, pages 72–83. Springer, 2005.

[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.

[Wal77] Harro Walk. An invariance principle for the robbins-monro process in a hilbert space. *Probability Theory and Related Fields*, 39(2) :135–150, 1977.

[Wei37a] 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.

[Wei37b] Endre Weiszfeld. Sur le point pour lequel la somme des distances de n points donnés est minimum. *Tohoku Math. J*, 43(355-386) :2, 1937.

[WF29] Alfred Weber and Carl Joachim Friedrich. Alfred weber's theory of the location of industries. 1929.

[Woo72] Michael Woodroofe. Normal approximation and large deviations for the Robbins-Monro process. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, 21 :329–338, 1972.

[WZH03] Juyan Weng, Yilu Zhang, and Wey-Shiuan Hwang. Candid covariance-free incremental principal component analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25 :1034–1040, 2003.

[Yan10] Le Yang. Riemannian median and its estimation. *LMS Journal of Computation and Mathematics*, 13 :461–479, 2010.

[Zäh16] Henryk Zähle. A definition of qualitative robustness for general point estimators, and examples. *Journal of Multivariate Analysis*, 143 :12–31, 2016.

[ZSW+07] Xiangwei Zhang, Jonathan Stockel, Matthias Wolf, Pascal Cathier, Geoffrey McLennan, Eric A Hoffman, and Milan Sonka. A new method for spherical object detection and its application to computer aided detection of pulmonary nodules in ct images. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2007*, pages 842–849. Springer, 2007.