Stochastic approximation in Hilbert spaces

Aymeric Dieuleveut

THÈSE DE DOCTORAT

de l'Université de recherche Paris Sciences Lettres
PSL Research University

Préparée à l'École normale supérieure

Stochastic Approximation in Hilbert Spaces

Approximation Stochastique dans les Espaces de Hilbert

École doctorale n°386

ÉCOLE DOCTORALE DE SCIENCES MATHÉMATIQUES DE PARIS CENTRE

Spécialité MATHÉMATIQUES

COMPOSITION DU JURY :

M. Arnak Dalalyan
ENSAE / CREST, Paris,
Rapporteur

M. Lorenzo Rosasco
MIT, University of Genova,
Rapporteur

M. Francis Bach
Inria Paris, DI ENS
Directeur de thèse

M. Stéphane Boucheron
DMA ENS,
Président du Jury

M. François Glineur
École polytechnique de Louvain,
Examinateur

Soutenue par Aymeric DIEULEVEUT
le 28.09.2016

Dirigée par **Francis BACH**

ENS ÉCOLE NORMALE SUPÉRIEURE | PSL RESEARCH UNIVERSITY PARIS

Abstract

The goal of supervised machine learning is to infer relationships between a phenomenon one seeks to predict and "explanatory" variables. To that end, multiple occurrences of the phenomenon are observed, from which a prediction rule is constructed. The last two decades have witnessed the apparition of very large data-sets, both in terms of the number of observations (e.g., in image analysis) and in terms of the number of explanatory variables (e.g., in genetics). This has raised two challenges: first, avoiding the pitfall of over-fitting, especially when the number of explanatory variables is much higher than the number of observations; and second, dealing with the computational constraints, such as when the mere resolution of a linear system becomes a difficulty of its own.

Algorithms that take their roots in stochastic approximation methods tackle both of these difficulties simultaneously: these stochastic methods dramatically reduce the computational cost, without degrading the quality of the proposed prediction rule, and they can naturally avoid over-fitting. As a consequence, the core of this thesis will be the study of stochastic gradient methods.

The popular parametric methods give predictors which are linear functions of a set of explanatory variables. However, they often result in an imprecise approximation of the underlying statistical structure. In the non-parametric setting, which is paramount in this thesis, this restriction is lifted. The class of functions from which the predictor is proposed depends on the observations. In practice, these methods have multiple purposes, and are essential for learning with non-vectorial data, which can be mapped onto a vector in a functional space using a positive definite kernel. This allows to use algorithms designed for vectorial data, but requires the analysis to be made in the non-parametric associated space: the reproducing kernel Hilbert space. Moreover, the analysis of non-parametric regression also sheds some light on the parametric setting when the number of predictors is much larger than the number of observations.

The first contribution of this thesis is to provide a detailed analysis of stochastic approximation in the non-parametric setting, precisely in reproducing kernel Hilbert spaces. This analysis proves optimal convergence rates for the averaged stochastic gradient descent algorithm. As we take special care in using minimal assumptions, it applies to numerous situations, and covers both the settings in which the number of observations is known *a priori*, and situations in which the learning algorithm works in an on-line fashion.

The second contribution is an algorithm based on acceleration, which converges at optimal speed, both from the optimization point of view and from the statistical one. In the non-parametric setting, this can improve the convergence rate up to optimality, even in particular regimes for which the first algorithm remains sub-optimal.

Finally, the third contribution of the thesis consists in an extension of the framework beyond the least-square loss. The stochastic gradient descent algorithm is analyzed as a Markov chain. This point of view leads to an intuitive and insightful interpretation, that outlines the differences between the quadratic setting and the more general setting. A simple method resulting in provable improvements in the convergence is then proposed.

Keywords: stochastic approximation, convex optimization, supervised learning, non-parametric estimation, reproducing kernel Hilbert spaces.

Résumé

Le but de l'apprentissage supervisé est d'inférer des relations entre un phénomène que l'on souhaite prédire et des variables "explicatives". À cette fin, on dispose d'observations de multiples réalisations du phénomène, à partir desquelles on propose une règle de prédiction. L'émergence récente de sources de données à très grande échelle, tant par le nombre d'observations effectuées (en analyse d'image, par exemple) que par le grand nombre de variables explicatives (en génétique), a fait émerger deux difficultés : d'une part, il devient difficile d'éviter l'écueil du sur-apprentissage lorsque le nombre de variables explicatives est très supérieur au nombre d'observations; d'autre part, l'aspect algorithmique devient déterminant, car la seule résolution d'un système linéaire dans les espaces en jeu peut devenir une difficulté majeure.

Des algorithmes issus des méthodes d'approximation stochastique proposent une réponse simultanée à ces deux difficultés : l'utilisation d'une méthode stochastique réduit drastiquement le coût algorithmique, sans dégrader la qualité de la règle de prédiction proposée, en évitant naturellement le sur-apprentissage. En particulier, le coeur de cette thèse portera sur les méthodes de gradient stochastique.

Les très populaires méthodes paramétriques proposent comme prédictions des fonctions linéaires d'un ensemble choisi de variables explicatives. Cependant, ces méthodes aboutissent souvent à une approximation imprécise de la structure statistique sous-jacente. Dans le cadre non-paramétrique, qui est un des thèmes centraux de cette thèse, la restriction aux prédicteurs linéaires est levée. La classe de fonctions dans laquelle le prédicteur est construit dépend elle-même des observations. En pratique, les méthodes non-paramétriques sont cruciales pour diverses applications, en particulier pour l'analyse de données non vectorielles, qui peuvent être associées à un vecteur dans un espace fonctionnel via l'utilisation d'un noyau défini positif. Cela autorise l'utilisation d'algorithmes associés à des données vectorielles, mais exige une compréhension de ces algorithmes dans l'espace non-paramétrique associé: l'espace à noyau reproduisant. Par ailleurs, l'analyse de l'estimation non-paramétrique fournit également un éclairage révélateur sur le cadre paramétrique, lorsque le nombre de prédicteurs surpasse largement le nombre d'observations.

La première contribution de cette thèse consiste en une analyse détaillée de l'approximation stochastique dans le cadre non-paramétrique, en particulier dans le cadre des espaces à noyaux reproduisants. Cette analyse permet d'obtenir des taux de convergence optimaux pour l'algorithme de descente de gradient stochastique moyennée. L'analyse proposée s'applique à de nombreux cadres, et une attention particulière est portée à l'utilisation d'hypothèses minimales, ainsi qu'à l'étude des cadres où le nombre d'observations est connu à l'avance, ou peut évoluer.

La seconde contribution est de proposer un algorithme, basé sur un principe d'accélération, qui converge à une vitesse optimale, tant du point de vue de l'optimisation que du point de vue statistique. Cela permet, dans le cadre non-paramétrique, d'améliorer

la convergence jusqu'au taux optimal, dans certains régimes pour lesquels le premier algorithme analysé restait sous-optimal.

Enfin, la troisième contribution de la thèse consiste en l'extension du cadre étudié au delà de la perte des moindres carrés : l'algorithme de descente de gradient stochastique est analysé comme une chaine de Markov. Cette approche résulte en une interprétation intuitive, et souligne les différences entre le cadre quadratique et le cadre général. Une méthode simple permettant d'améliorer substantiellement la convergence est également proposée.

Mots-clés : approximation stochastique, optimisation convexe, apprentissage supervisé, estimation non-paramétrique, espaces de Hilbert à noyaux reproduisants.

Remerciements

En premier lieu, je souhaite exprimer mon immense reconnaissance à mon directeur de thèse, Francis Bach. Francis, travailler avec toi a été une chance incroyable, et m'a énormément appris sur de nombreux plans. J'ai apprécié ta vision des problèmes, ton enthousiasme, ta gentillesse, sans oublier le cadre incroyable dans lequel tu nous permets de travailler.

Arnak Dalalyan et Lorenzo Rosasco ont accepté de rapporter cette thèse, je les en remercie vivement. Merci aussi à François Glineur et Stéphane Boucheron d'avoir accepté de faire partie de mon jury. Présenter mes travaux devant vous tous est un grand honneur.

Je souhaiterais exprimer ma gratitude à Martin Wainwright, qui en m'accueillant six mois à Berkeley l'année dernière m'a permis de replonger dans la beauté des statistiques, de découvrir parmi bien d'autres choses, la joie de la localisation (statistique), les délices des cafés, et la Californie.

Je souhaite également remercier les professeurs qui m'ont encadré depuis l'ENS et ont fait porter mon choix sur l'apprentissage statistique, en particulier Gérard Biau qui m'a initié aux statistiques et était mon tuteur, Gilles Stoltz, qui supervisait mon mémoire de première année et co-encadrait avec Francis, Sylvain Arlot, Guillaume Obozinski et Olivier Catoni le passionnant cours d'apprentissage statistique. Enfin l'ensemble des professeurs d'Orsay, en particulier Pascal Massart, Elizabeth Gassiat, Christophe Giraud, Vincent Rivoirard, dont les cours, les conseils et les expériences sont d'une grande richesse.

Je voudrais aussi remercier Martin Jaggi de m'accueillir à l'EPFL: je suis impatient de découvrir et travailler sur de nombreux autres sujets.

J'ai eu la chance de collaborer avec Nicolas Flammarion, et Alain Durmus. Travailler avec vous fut source d'enseignements, de surprises, et de résultats. J'espère qu'il y aura d'autres occasions !

Par ailleurs, cette thèse ne serait pas telle qu'elle est sans certaines relectures de Vincent R. et Damien S., les remarques anglo-saxonnes de Dominique, les inestimables contributions graphiques de Jordane, ou les relectures délicieusement pointilleuses de Daphné et Rémi. Je vous en remercie tous particulièrement.

Travailler au sein de l'équipe Sierra fut un grand plaisir. Je voudrais remercier tous ceux avec qui j'ai partagé ces années. En particulier, j'ai une pensée pour mes acolytes de place d'It', Piotr Bojanowski et Sesh Kumar, comme ceux de gare de Lyon, Nicolas Flammarion et Damien Garreau. Nicolas et Damien, réfléchir et discuter avec vous, partager tant de cafés, et avoir votre soutien dans les instants d'hésitation ou de doute, fut un honneur, une chance, et une joie. J'ai une pensée pour tous les autres membres du labo avec qui j'ai partagé de superbes moments, Jean-Baptiste Alayrac, Nicolas Boumal, Guilhem Chéron, Théophile Dalens, Rémy Degenne, Vincent Delaitre, Christophe Dupuy, Fajwel Fogel, Rémi Leblond, Pierre Gaillard, Pascal Germain, Edouard Grave, Gauthier Gidel, Robert Gower, Yana Hasson, Vadim Kantorov, Rémi Lajugie, Loic Landrieu, Maxime Oquab, Julia Peyre,

Anastasia Podosinnikova, Antoine Recanati, Vincent Roulet, Damien Scieur, Guillaume Seguin, Nino Shervashidze, et Gül Varol.

I was also most lucky to be surrounded by amazing people in Berkeley, and I am grateful to Reinhard Heckel and Fanny Yang, and also Ahmed El Alaoui, Vipul Gupta, Sang Min Han, Vidya Muthukumar, Orhan Ocal, Ashwin Pananjady, Aaditya Ramdas, Dominik Rothenhäusler Ludwig Schmidt, Nihar Shah, Ilan Shomorony and Yuting Wei for welcoming me and/or sharing unforgettable moments together. I would also like to send a special thank to my fantastic roommates Alex and Karl.

Sur un plan plus personnel, cette thèse a été ponctuée de nombreuses aventures parallèles, parmi lesquelles une me tient particulièrement à coeur, en un lieu de création et de dépassement bien différent, dans les collines du massif central. Je voudrais remercier mes amis qui y ont pris part, en particulier Clémentine, Juliette, Adrien, Hugo, Vincent V., Thibault, Jad, Vincent B., et tous ceux cités par ailleurs qui se reconnaîtront. Merci pour ces moments de rire, d'épuisement, ou de craquages.

Merci à Irène de m'avoir accueilli et supporté en de nombreuses occasions ces dernières années. Merci à Iryna pour les soirées interminables.

Merci à toute ma famille, pour leur indéfectible soutien. À Yvonne, qui est une perle de bienveillance. À Papa, qui en nous initiant aux subtilités *pures et pires* de la logique, fit sûrement naître un certain goût des paradoxes. À Maman, qui a construit notre bonheur. À Daphné, de qui j'ai suivi le chemin bien trop longtemps pour que ce soit un hasard. À Anouk, dont le courage est incroyable. À Floriane, qui ne cesse de me surprendre.

À Jordane, qui illumine ma vie.

Contents

Contributions and thesis outline 1

1 Introduction 3
- 1.1 Statistical Learning . . . 5
- 1.2 Convex optimization . . . 12
- 1.3 Stochastic approximation . . . 16
- 1.4 Non-parametric regression in reproducing kernel Hilbert spaces . . . 28

2 Non-parametric Stochastic Approximation with Large Step-sizes 39
- 2.1 Introduction . . . 41
- 2.2 Learning with positive-definite kernels . . . 43
- 2.3 Stochastic approximation in Hilbert spaces . . . 50
- 2.4 Links with existing results . . . 57
- 2.5 Experiments on artificial data . . . 62
- 2.6 Conclusion . . . 65

A Appendix to Non-parametric Stochastic Approximation with Large Step-sizes 68
- A.1 Minimal assumptions . . . 68
- A.2 Sketch of the proofs . . . 71
- A.3 Reproducing kernel Hilbert spaces . . . 73
- A.4 Proofs . . . 83

3 Faster Convergence Rates for Least-Squares Regression 113
- 3.1 Introduction . . . 115
- 3.2 Least-Squares Regression . . . 117
- 3.3 Averaged Stochastic Gradient Descent . . . 121
- 3.4 Accelerated Stochastic Averaged Gradient Descent . . . 123
- 3.5 Tighter Dimension-independent Convergence Rates . . . 125
- 3.6 Rates of Convergence for Kernel Regression . . . 126
- 3.7 Conclusion . . . 131

B Appendix to Faster Convergence Rates for Least-Squares Regression 132
- B.1 Proofs of Section 3.3 . . . 132
- B.2 Proof of Theorem 3.2 . . . 134
- B.3 Convergence of Accelerated Averaged Stochastic Gradient Descent . . . 141
- B.4 Tighter bounds . . . 151
- B.5 Technical Lemmas . . . 153

4 Bridging the Gap between Constant Step Size SGD and Markov Chains **158**
4.1 Introduction . . . 160
4.2 Main results . . . 162
4.3 Detailed analysis . . . 165
4.4 Experiments . . . 171
4.5 Conclusion . . . 171

C Appendix to Bridging the Gap between SGD and Markov Chains **173**
C.1 Generalities on convex and strongly convex functions . . . 173
C.2 Results on the Markov chain defined by SGD . . . 175
C.3 Further properties of the Markov chain $(\theta_k^{(\gamma)})_{k\geqslant 0}$ . . . 187
C.4 Regularity of the gradient flow and estimates on Poisson solution . . . 190
C.5 Proof of Theorem 4.6 . . . 193

5 Conclusion and Future Work **196**
5.1 Summary of the thesis . . . 196
5.2 Perspectives . . . 197

Bibliography **199**

List of Figures **210**

List of Tables **211**

Contributions and thesis outline

Chapter 1: In this opening Chapter, we describe the key areas that come into play in the following chapters. More specifically, we first introduce the general setting of supervised machine learning, namely its statistical and computational goals. We then present how these goals can be achieved using convex optimization and/or stochastic approximation. Finally, we describe the non-parametric estimation framework, which will be of interest in Chapters 2 and 3.

Chapter 2: We consider the random-design least-squares regression problem within a reproducing kernel Hilbert space. We consider the least-mean-squares algorithm, and detail the benefits of using sufficiently large step-sizes together with averaging, without regularization. Our analysis is based on two assumptions: on the smoothness of the optimal prediction function and on the eigenvalue decay of the covariance operators of the reproducing kernel Hilbert space. We prove that the convergence rate of the algorithm depends on two factors: the speed at which initial conditions are forgotten, and the influence of the noise. We describe how both of these factors behave, which leads to an optimal choice for the learning rate. For this choice, we get optimal non-asymptotic rates of convergence, in both the finite horizon setting and the online setting, over a variety of regimes.

We furthermore give minimal assumptions, regarding the input space and the distributions, for our results to hold. We finally compare our results to existing work, both theoretically and empirically.

While the least-mean-squares algorithm with averaging is able to achieve the optimal rate of convergence in many situations, it remains sub-optimal for some difficult problems (typically when the hypothesis space is too small and/or the regression function most irregular). In such a situation, the speed at which the initial conditions are forgotten always dominates the convergence rate. This was one of the key motivations for Chapter 3.

Chapter 3: In this Chapter, we propose a new algorithm based on averaged *accelerated* regularized gradient descent to minimize a quadratic objective function whose gradients are accessible through a stochastic oracle. For least-squares regression, we show that in the parametric regime, this algorithm improves on the previous one (without acceleration). Indeed, it improves the speed at which the initial conditions are forgotten, up to the optimal rate $O(n^{-2})$ for a first order algorithm, while preserving the statistically optimal dependence $O(d/n)$ on the noise and dimension d of the problem. In the non-parametric regime, disregarding computational limits, this allows us to recover the statistical performance for a wider class of regimes than in Chapter 2 (though in practice, the algorithm cannot always be computed as it relies on the knowledge of the covariance operator).

We also propose a simplified analysis of the averaged algorithm considered in Chapter 2, by means of an additional regularization, which does not change the convergence rate.

A crucial aspect of Chapters 2 and 3 was the use of a quadratic objective function. In Chapter 4, we relax this assumption and consider a more general smooth and strongly-convex objective function.

Chapter 4: In this Chapter, we consider the averaged stochastic gradient descent with a constant learning rate, in order to minimize a strongly-convex and smooth objective function. While this behaves optimally in the quadratic case, it does not even converge to the global optimum in the general setting. We propose a detailed analysis of the different factors influencing this convergence. More precisely, we provide an explicit expansion of the moments of the averaged stochastic gradient descent iterates, that outlines the dependence on initial conditions, the effect of noise and the effect of the non-decaying step-size. We also use a simple trick from numerical analysis, Richardson-Romberg extrapolation, to substantially improve the convergence. We support these results both with theoretical and empirical results.

To conduct such an analysis, we use tools from Markov chain theory to analyze stochastic gradient descent. This allows for an intuitive understanding of the behavior of averaged stochastic gradient descent in the general case.

Chapter 5: This Chapter concludes the thesis by summarizing our contributions and describing possible extensions.

Publications related to this manuscript are listed bellow:

- **Chapter 2** is based on *Non-parametric Stochastic Approximation with Large Step-sizes*, A. Dieuleveut and F. Bach, published in the Annals of Statistics (Dieuleveut and Bach, 2016).
- **Chapter 3** is based on *Harder, Better, Faster, Stronger Convergence Rates for Least-squares Regression*, A. Dieuleveut, N. Flammarion and F. Bach, accepted for publication in Journal of Machine Learning Research (Dieuleveut et al., 2016).
- **Chapter 4** is based on *Bridging the Gap between Constant Step Size Stochastic Gradient Descent and Markov Chains*, A. Dieuleveut, A. Durmus, F. Bach (Dieuleveut et al., 2017).

1

Introduction

This introduction describes the key areas that interplay in the following chapters, namely supervised machine learning (Section 1.1), convex optimization (Section 1.2), stochastic approximation (Section 1.3) and non-parametric estimation, especially with reproducing kernel Hilbert spaces (Section 1.4).

Chapters 2, 3 and 4 will use tools from these different settings. The main sources of influence are summarized in Table 1.1.

	Chapter 2	Chapter 3	Chapter 4
Supervised Machine Learning	✓	✓	✓
Convex Optimization		✓	
Stochastic Approximation	✓	✓	✓
Non-parametric estimation	✓	✓	

Table 1.1: Schematic interplay of supervised machine learning, convex optimization, stochastic approximation and non-parametric estimation in the main chapters of this thesis.

Contents

1.1 Statistical Learning 5
1.1.1 Supervised machine learning 5
1.1.2 Observations and Empirical Risk Minimization 5
1.1.3 Linear predictors: the parametric setting 7
1.1.4 Statistical point of view on least-squares and logistic regression . . . 7
1.1.5 Risk decomposition: approximation and estimation errors 8
1.1.6 Upper bounds on the estimation error 8
1.1.7 Minimax rates of convergence 11
1.1.8 Computational cost of ERM 12
1.2 Convex optimization 12
1.2.1 Assumptions 12
1.2.2 Gradient methods 13
1.2.3 Accelerated gradient descent 13
1.2.4 Lower complexity bounds 14
1.3 Stochastic approximation 16
1.3.1 Convergence of the last iterate 17
1.3.2 Polyak-Ruppert averaging 17
1.3.3 Stochastic gradient descent. 18
1.3.4 Application to machine learning and optimization. 18
1.3.5 Assumptions on the noise 20
1.3.6 Non-asymptotic results: stochastic approximation for minimizing convex functions 22
1.3.7 Non-asymptotic results: stochastic approximation for minimizing smooth convex functions 23
1.3.8 Non-asymptotic results: stochastic approximation for least-squares regression and logistic regression 25
1.4 Non-parametric regression in reproducing kernel Hilbert spaces 28
1.4.1 Reproducing kernel Hilbert spaces 30
1.4.2 Examples 32
1.4.3 Least-squares regression in RKHS 33
1.4.4 Consequences in finite dimension 37
1.4.5 Computations in RKHS 37

1.1 Statistical Learning

1.1.1 Supervised machine learning

In *supervised machine learning* (Vapnik, 1995; Hastie et al., 2001; Shalev-Shwartz and Ben-David, 2014) one aims to predict an outcome $Y \in \mathcal{Y}$, based on some feature(s) $X \in \mathcal{X}$ that are supposed to have some influence on this outcome. The set $\mathcal{Y}$, which describes the outcome, can be either quantitative ($\mathcal{Y} \subset \mathbb{R}$), or categorical ($\mathcal{Y}$ is a finite set, typically $\{-1, 1\}$ if there are two possible categories). This leads to the two most important tasks of supervised learning:

- *Regression*, when one predicts a quantitative outcome.
- *Classification*, when one predicts a categorical outcome, with $\mathcal{Y} = \{-1, 1\}$.

A *predictor* is defined as a (measurable) function $f : \mathcal{X} \to \mathcal{Y}$, and the set of possible predictors is denoted $\mathcal{M}(\mathcal{X}, \mathcal{Y})$. Variables X and Y are modeled as random variables following a joint distribution denoted ρ, and we denote ρ_X the marginal distribution of X.

To measure the quality of a predictor, we introduce a loss function

$$\ell : (\mathcal{X} \times \mathcal{Y}) \times \mathcal{M}(\mathcal{X}, \mathcal{Y}) \to \mathbb{R}_+,$$

such that $\ell((X, Y), f)$ is small when $f(X)$ is a good prediction of Y. The *risk*, or *generalization error*, of a predictor f is the averaged loss under the distribution of observations $R(f) := \mathbb{E}_{(X,Y)\sim\rho}[\ell((X, Y), f)]$. Our general goal is to find a predictor minimizing the risk, *i.e.*,

$$\operatorname*{arg\,min}_{f \in \mathcal{M}(\mathcal{X},\mathcal{Y})} R(f) .$$

The most classical loss functions are the following:

- For regression, the squared loss: $\ell((X, Y), f) = \frac{1}{2}(Y - f(X))^2$.
- For classification, the binary loss $\ell((X, Y), f) = \mathbb{1}_{Y \neq sign(f(X))}$. However this loss is often replaced by surrogates that allow easier computations of predictors, for example the logistic loss $\ell((X, Y), f) = \log(1 + \exp(-Yf(X)))$, or the hinge loss, $\ell((X, Y), f) = \max\{0, 1 - Yf(X)\}$.

The optimal predictor f_ρ, that minimizes the risk, is called the *Bayes predictor*, when it exists. While it may have a closed form depending on ρ (e.g., for regression with the square loss, $f_\rho(X) = \mathbb{E}_\rho[Y|X]$), as the distribution ρ is unknown, the Bayes predictor cannot be directly computed in practice and needs to be approximated.

In most situations, only weak assumptions are made on the distribution ρ; we refer to this setting as the "distribution free" approach (Györfi et al., 2002). Unlike in parametric statistics, where a model of the distribution is chosen, we here generally only assume that the marginal laws have first order moments (typically 2 or 4 moments are necessary).

1.1.2 Observations and Empirical Risk Minimization

In practice, in order to build a predictor, n *observations* $(x_k, y_k)_{k\in[\![1;n]\!]} \in (\mathcal{X} \times \mathcal{Y})^n$ are used as training examples[1]. They correspond to n independent and identically distributed

[1] we use the notation $[\![1; n]\!] := [1; n] \cap \mathbb{N}$.

examples of possible inputs and outputs, with $n \in \mathbb{N}$. The intuition is that, given multiple input/output pairs of observations, one can infer a prediction rule that generalizes to any input.

A *learning rule* $\mathcal{A}$ (or *statistical learning algorithm*) is thus naturally defined as a measurable function that maps the set of observations to an *estimator* $\hat{f}$: it is a function $\mathcal{A} : \cup_{n=1}^{\infty}(\mathcal{X} \times \mathcal{Y})^n \to \mathcal{M}(\mathcal{X}, \mathcal{Y})$. As we describe our observations as random variables, the predictor $\hat{f} = \mathcal{A}((x_k, y_k)_{k\in[\![1;n]\!]})$ is a random variable in $\mathcal{M}(\mathcal{X}, \mathcal{Y})$, and its risk $R(\hat{f})$ is a random variable in $\mathbb{R}_+$. One of our main goals is to find learning rules that minimize the expectation (under the law of the observations) of the risk of the predictor: $\mathbb{E}_{(x_k,y_k)_{k\in[\![1;n]\!]}\sim\rho^{\otimes n}}\left[R(\mathcal{A}((x_k, y_k)_{k\in[\![1;n]\!]}))\right]$ (simply denoted $\mathbb{E}[R(\hat{f})]$ from now on). While one could also try to control this quantity with high probability, we mainly propose results in expectation in this thesis. Extensions to high probability bounds could be the subject of future work.

The most common learning rule consists in proposing a predictor that behaves well on the observed data. This method is known as *empirical risk minimization* (ERM): as we cannot access the risk function itself, we minimize instead the averaged loss on the observed points, defined as the *empirical risk*, or *training error*, $R_n(f) = n^{-1}\sum_{k=1}^{n}\ell((x_k, y_k), f)$. The learning rule is thus:

$$\mathcal{A}((x_k, y_k)_{k\in[\![1;n]\!]}) \quad = \quad \underset{f\in\mathcal{M}(\mathcal{X},\mathcal{Y})}{\arg\min}\ R_n(f)\ .$$

This idea historically dates back to Legendre (1805) and Gauss (1809), who introduced independently the least-squares principle, which corresponds to empirical risk minimization for linear predictors and the square loss.

Even though this rule seems intuitive, it is still questionable: choosing a predictor that behaves well on observed points does not guarantee that it has small generalization error. On the contrary, in most situations, a predictor minimizing the empirical risk has poor generalization performance as it only fits the observed sample. This phenomenon is known as "overfitting". Indeed, any function g such that $g(x_k) = y_k$ for any k has minimal (null) empirical risk, while its generalization error may be arbitrarily large. To avoid such a pitfall, one needs to avoid selecting functions with such pathological behavior. This is the purpose of *regularization* (Hastie et al., 2001), which can take several forms. The most popular approaches consist in either restricting the class of functions over which the empirical risk is minimized, or adding a penalization term that artificially increases the risk of the undesirable functions. This means considering the following predictors:

- **Constrained formulation:**
$$\hat{f}_{\mathcal{F}} = \underset{f\in\mathcal{F}}{\arg\min}\, R_n(f),$$
where $\mathcal{F}$ is a set of functions, called the *hypothesis space*.

- **Penalized formulation:**
$$\hat{f}_{\lambda} = \underset{f\in\mathcal{M}(X,Y)}{\arg\min}\ (R_n(f) + \lambda\ \mathrm{pen}(f))\ ,$$
where $\mathrm{pen} : \mathcal{M}(X, Y) \to \mathbb{R}_+$ is a penalty on some functions (generally the least regular ones), and $\lambda > 0$.

Beyond the choice of the regularization, two keys challenges appear: computing the estimator and analyzing its statistical properties. The cost of computing such estimators may be prohibitive. Interestingly, some estimators designed to ensure a lower computational cost naturally regularize the problem. This is the case of some *gradient methods* and *stochastic gradient methods*. These methods will be introduced later in Section 1.2 and 1.3; and stochastic gradient methods will be the key ingredient of the main three chapters of the thesis. Stochastic gradient descent will be used as an alternative to ERM. To facilitate the comparison, we first describe some statistical properties of ERM, in the particular case of *linear predictors*. We introduce this particular setting in Section 1.1.3 and relate it to statistical models in Section 1.1.4; we then describe how the risk can be upper bounded in Section 1.1.5 and 1.1.6, and show how the upper bounds can be considered as optimal in Section 1.1.7.

1.1.3 Linear predictors: the parametric setting

In an important particular case, referred to as the *parametric regime*, we look for estimators in a finite-dimensional space parameterized by vectors in $\mathbb{R}^d$, for some dimension $d \in \mathbb{N}^*$. The space $\mathbb{R}^d$ is embedded with the Euclidean norm and the associated inner product $\langle \cdot, \cdot \rangle$. We consider a function $\Phi : \mathcal{X} \to \mathbb{R}^d$: for any $x \in \mathcal{X}$, $\Phi(x)$ is a finite-dimensional vector containing features of x. We here choose for our hypothesis space the set of *linear predictors* of these features: $f_\theta : x \mapsto \langle \Phi(x), \theta \rangle$. That is, we minimize the empirical risk over the set of functions $\{f_\theta, \theta \in \mathbb{R}^d\}$. In such a setting, with a slight abuse in notation, the loss can be written $\ell(\langle \Phi(X), \theta \rangle, Y) = \ell((X, Y), f_\theta)$, and the risk can be denoted $R(\theta) = R(f_\theta)$.

This particular case is of major importance as many practitioners rely on such linear predictors (Neter et al., 1996; Seber and Lee, 2012). Of course, it is not assumed that the regression function f_ρ truly belongs to the class of linear functions. If it does, we are in the *well-specified setting*. Also note that Φ is not necessarily linear, and can be learned itself, for example using deep learning techniques (Le Cun et al., 2015). Yet, in this Chapter, Φ is considered as known.

1.1.4 Statistical point of view on least-squares and logistic regression

It is worth pointing out that for a variety of losses, the ERM framework can be understood as a particular case of maximum likelihood estimation, with a well-suited statistical model. In statistics, a *model* is a set of possible joint distributions on (X, Y). We give here two classical examples.

Gaussian linear regression statistical model: we first consider the statistical model $\{p_\theta(X, Y),\ \theta \in \mathbb{R}^d\}$, where $p_\theta(Y|X) = \mathcal{N}(\langle \theta, \Phi(X) \rangle, \sigma^2)$, *i.e.*, the law of Y knowing X is a normal law with variance $\sigma^2 > 0$. The maximum likelihood estimator is then also the ordinary least-squares estimator, that minimizes the empirical square loss.

Logistic regression statistical model: we now consider the following statistical model: $\{p_\theta(X, Y),\ \theta \in \mathbb{R}^d\}$, where $p_\theta(Y|X) = \mathcal{B}\left(\frac{\exp\langle \theta, \Phi(X) \rangle}{1+\exp\langle \theta, \Phi(X) \rangle}\right)$ is a Bernoulli law. Then the maximum likelihood estimator is also the empirical risk minimizer for the logistic loss.

The properties of maximum likelihood estimators have been widely studied (Van der Vaart, 1998). As we intend to address a more general situation, without assuming a particular statistical model, we will not build on such results. Indeed, in most situations, we

have neither (sub-) Gaussian assumption on the noise in our approach, nor the assumption that the model is well-specified.

1.1.5 Risk decomposition: approximation and estimation errors

We consider here, for simplicity, the constrained version. In fact, the penalized version is in most situations "equivalent" to the constrained one[2], but the latter has a simpler geometrical interpretation, as illustrated in Figure 1.1.

Let $\hat{f}_{\mathcal{F}} \in \arg\min_{f\in\mathcal{F}} R_n(f)$, $f_{\mathcal{F}} \in \arg\min_{f\in\mathcal{F}} R(f)$ be the predictors that minimize the empirical and generalization risk over $\mathcal{F}$, and recall that f_ρ minimizes R. The excess of generalization error of our estimator can be decomposed in two terms:

$$R(\hat{f}_{\mathcal{F}}) - R(f_\rho) = \underbrace{R(\hat{f}_{\mathcal{F}}) - R(f_{\mathcal{F}})}_{\text{estimation error}} + \underbrace{R(f_{\mathcal{F}}) - R(f_\rho)}_{\text{approximation error}} . \tag{1.1}$$

This error decomposition is illustrated in Figure 1.1. The approximation error is due to the fact that $\mathcal{F}$ is smaller than $\mathcal{M}(X,Y)$. It decreases as $\mathcal{F}$ increases. The estimation error is linked to the fact that $\hat{f}$ minimizes the empirical risk and not the true risk. It decreases as the number of observations increase and as the size of $\mathcal{F}$ decreases. Moreover, we have the following upper bound on the estimation error:

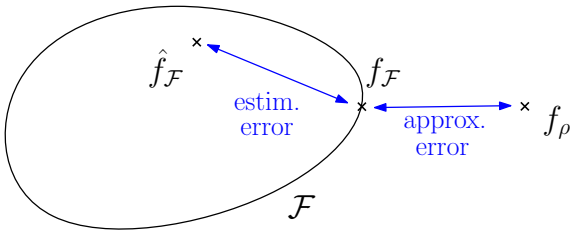


Figure 1.1: Risk decomposition

$$\begin{aligned} R(\hat{f}_{\mathcal{F}}) - R(f_{\mathcal{F}}) &= R(\hat{f}_{\mathcal{F}}) - R_n(\hat{f}_{\mathcal{F}}) + \underbrace{R_n(\hat{f}_{\mathcal{F}}) - R_n(f_{\mathcal{F}})}_{\leqslant 0} + R_n(f_{\mathcal{F}}) - R(f_{\mathcal{F}}) \\ &\leqslant R(\hat{f}_{\mathcal{F}}) - R_n(\hat{f}_{\mathcal{F}}) + R_n(f_{\mathcal{F}}) - R(f_{\mathcal{F}}) \\ &\leqslant 2\sup_{f\in\mathcal{F}} |R_n(f) - R(f)| . \end{aligned}$$

As a consequence, one can derive bounds on the estimation error from uniform bounds on the function $R_n - R$. Analyzing deviations between empirical quantities and their averages is one of the key problems studied in empirical process theory (van der Vaart and Wellner, 2000; Van der Vaart and Wellner, 2007). We develop these methods in the next section.

1.1.6 Upper bounds on the estimation error

General case

In order to control uniformly the function $R_n - R$, we introduce the Rademacher complexity of the class of functions $\{(X,Y) \mapsto \ell((X,Y),f), f \in \mathcal{F}\}$:

$$\mathcal{R}_n = \mathbb{E}\left[\sup_{f\in\mathcal{F}} \left(\frac{1}{n}\sum_{k=1}^{n} \varepsilon_k \ell((x_k,y_k),f)\right)\right] ,$$

where the ε_k, $k \in [\![1;n]\!]$ are i.i.d. Rademacher variables ($\mathbb{P}(\varepsilon_k = 1) = \mathbb{P}(\varepsilon_k = -1) = 1/2$) independent from $(x_k,y_k)_{k\in[\![1;n]\!]}$; and the expectation is taken with respect to both the distribution of observations and the randomness of $(\varepsilon_k)_{k\in[\![1;n]\!]}$.

[2] especially, if the class $\mathcal{F}$ in the constrained case is the set of predictors f such that $\mathbf{pen}(f) \leqslant C$ for some C, then the Lagrangian of the constrained version is equivalent to the penalized problem (Rockafellar, 1970).

A symmetrization argument shows that $\mathbb{E}[\sup_{f\in\mathcal{F}}|R_n(f)-R(f)|]\leqslant 2\mathcal{R}_n$. For simple classes of functions, it is possible to control the Rademacher complexity, which immediately yields upper bounds on the estimation error. For example, in the parametric setting, if we consider almost surely (a.s.) bounded inputs such that $\|\Phi(X)\|\leqslant R$, a G-Lipschitz loss function, and minimize the empirical risk over $\{\theta \text{ s.t. } \|\theta\|\leqslant D\}$, then using Ledoux-Talagrand inequality, one has $\mathcal{R}_n\leqslant\frac{RGD}{\sqrt{n}}$ (Ledoux and Talagrand, 1991).

While this $O(1/\sqrt{n})$ bound shows that our estimator is weakly consistent ($\mathbb{E}[R(\hat{\theta})-R(\theta_*)]\overset{n\to\infty}{\to}0$), this rate of convergence is pessimistic in several settings. For example, Sridharan et al. (2008); Boucheron and Massart (2011) proved that for a μ-strongly convex loss, one has $R(\hat{\theta})-R(\theta_*)=O(1/(n\mu))$. In this thesis, we aim at proving rates faster than $O(1/\sqrt{n})$. In order to derive these faster rates of convergence, two approaches have mainly been used: performing a direct and explicit calculation, or extending and refining the framework presented above, considering *localized* Rademacher complexities (Koltchinskii, 2001, 2006; Bartlett et al., 2005) to obtain a sharper bound, without using a uniform upper bound on the empirical process over the whole set $\mathcal{F}$. In the next paragraph, we address the case of linear regression with the square loss, where all computations are made explicitly.

Linear least-squares regression

In this section, we show that the excess risk (defined as $R(\cdot)-\inf_\theta R(\theta)$) is of order $\frac{\sigma^2 d}{n}$ for the empirical risk minimizer for linear least-squares regression. We have $\Phi(X)\in\mathbb{R}^d$, and $R(\theta)=\frac{1}{2}\mathbb{E}\left[(\langle\theta,\Phi(X)\rangle-Y)^2\right]$. We denote $\Sigma=\mathbb{E}[\Phi(X)\Phi(X)^\top]$ the covariance matrix. The optimal predictor θ_* satisfies the first order condition:

$$R'(\theta_*)=2\mathbb{E}\left[(Y-\langle\Phi(X),\theta_*\rangle)\Phi(X)\right]=0, \tag{1.2}$$

i.e., $\Sigma\theta_*=\mathbb{E}[Y\Phi(X)]$. Therefore, the best linear predictor exists and is unique if Σ is an invertible matrix. We then have the following excess risk decomposition:

$$\begin{aligned} R(\theta) &= \frac{1}{2}\mathbb{E}\left[(\langle\theta,\Phi(X)\rangle-Y)^2\right]=\frac{1}{2}\mathbb{E}\left[(\langle\theta-\theta_*,\Phi(X)\rangle)^2\right]+R(\theta_*)\\ &= \frac{1}{2}(\theta-\theta_*)^\top\mathbb{E}[\Phi(X)\Phi(X)^\top](\theta-\theta_*)^\top+R(\theta_*)\,, \end{aligned}$$

where $\mathbb{E}[(Y-\langle\Phi(X),\theta_*\rangle)\langle\theta-\theta_*,\Phi(X)\rangle]=0$ is due to Equation (1.2). The function R is thus *quadratic*, and the excess risk is written:

$$R(\theta)-R(\theta_*) = \frac{1}{2}\left\|\Sigma^{1/2}(\theta-\theta_*)\right\|^2. \tag{1.3}$$

We denote $\mathbf{\Phi}\in\mathbb{R}^{n\times d}$ the feature matrix, whose k-th row contains the feature vectors $\Phi(x_k)$ for $k\in[\![1;n]\!]$. The empirical risk is thus

$$R_n(\theta)=\frac{1}{2}\sum_{k=1}^n(y_k-\langle\Phi(x_k),\theta\rangle)^2=\frac{1}{2}\left\|\boldsymbol{Y}-\mathbf{\Phi}\theta\right\|^2\,,$$

with $\boldsymbol{Y}=(y_k)_{k\in[\![1;n]\!]}\in\mathbb{R}^n$. We consider the ordinary least-squares estimator (OLS):

$$\hat{\theta}=\operatorname*{arg\,min}_{\theta\in\mathbb{R}^d}R_n(\theta)=(\mathbf{\Phi}^\top\mathbf{\Phi})^{-1}\mathbf{\Phi}^\top\boldsymbol{Y}\,, \tag{E}$$

assuming $\mathbf{\Phi}^\top\mathbf{\Phi}$ to be invertible.

Fixed design analysis. In the *fixed design setting*, the covariates $(x_k)_{k\in[\![1;n]\!]}$ are considered to be deterministic: only the outputs $(y_k)_{k\in[\![1;n]\!]}$ are treated as random. As a consequence, the covariance matrix $\mathbf{\Phi}$ is fixed and known, which simplifies the analysis of standard estimators, including the OLS estimator. Indeed, Equation (1.2) gives that $\theta_* = (\mathbf{\Phi}^\top\mathbf{\Phi})^{-1}\mathbf{\Phi}^\top\mathbb{E}[\boldsymbol{Y}]$ and $\hat{\theta} = (\mathbf{\Phi}^\top\mathbf{\Phi})^{-1}\mathbf{\Phi}^\top\boldsymbol{Y}$. Denoting $\boldsymbol{\varepsilon} = \boldsymbol{Y} - E[\boldsymbol{Y}] \in \mathbb{R}^n$ and using the fact that $\mathbf{\Phi}(\mathbf{\Phi}^\top\mathbf{\Phi})^{-1}\mathbf{\Phi}^\top$ is a projector, we have:

$$R(\hat{\theta}) - R(\theta_*) \;=\; \frac{1}{n}\left\|\mathbf{\Phi}(\hat{\theta}-\theta_*)\right\|^2 = \frac{1}{n}\left\|\mathbf{\Phi}(\mathbf{\Phi}^\top\mathbf{\Phi})^{-1}\mathbf{\Phi}^\top\boldsymbol{\varepsilon}\right\|^2 = \mathrm{tr}((\mathbf{\Phi}^\top\mathbf{\Phi})^{-1}\mathbf{\Phi}^\top\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top)\,.$$

Finally, if $\mathbb{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top] = \sigma^2\,\mathrm{Id}$, we conclude that $\mathbb{E}[R(\hat{\theta}) - R(\theta_*)] = \frac{\sigma^2\mathbf{rank}(\mathbf{\Phi})}{n}$. More generally, if $\mathbb{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top] \preccurlyeq \sigma^2\,\mathrm{Id}$, where $\preccurlyeq$ is the natural order between positive definite matrices (by definition $A \preccurlyeq B$ if $B - A$ is non-negative), then $\mathbb{E}[R(\hat{\theta}) - R(\theta_*)] \leqslant \frac{\sigma^2 d}{n}$.

Random design analysis. The fixed design does not directly address the "out of sample" error (the error made at input points that were not present in the training set), which may be the most important in practice. The *random design setting*, in which both X and Y are random, is thus the most relevant. Consistency and asymptotic behavior of the OLS estimator are consequences of the general study of *M-estimators* (see, for example, chapter 5 in Van der Vaart, 1998). As $\hat{\theta}$ is such an estimator, it is consistent ($\hat{\theta}$ converges in probability to θ_*) and asymptotically normal, as $\sqrt{n}(\hat{\theta}-\theta_*) \xrightarrow{d} \mathcal{N}(0, \Sigma^{-1}C\Sigma^{-1})$, with[3] $C := \mathbb{E}[((\langle\theta_*, \Phi(X)\rangle - Y)\Phi(X))^{\otimes 2}]$. Such a limit distribution is optimal, in the sense that the estimator is *asymptotically efficient*: an estimator is asymptotically efficient if it converges to the *asymptotic lower bound* (see, for example, Chapter 8 in Van der Vaart, 1998). For parametric models, under some regularity assumptions, the asymptotic lower bound is a normal distribution with mean 0 and covariance the inverse of the Fisher information matrix. Asymptotic efficiency can be understood as an "asymptotic Cramer-Rao bound" (the Cramer-Rao bound gives a lower bound on the variance of an *unbiased* estimator). However, this does not do justice to the depth and complexity of results on asymptotic efficiency.

Moreover, $n(R(\hat{\theta}) - R(\theta_*)) = n\|\Sigma^{1/2}(\hat{\theta}-\theta_*)\|^2$ converges in distribution to a Wishart distribution $W(\Sigma^{-1/2}C\Sigma^{-1/2})$ (with one degree of freedom), a natural generalization of a χ^2 distribution when the variance is not the identity (Wishart, 1928). The noise is said to be *structured* when $C \preccurlyeq \sigma^2\Sigma$ for some $\sigma^2 > 0$. In such a situation, $W(\Sigma^{-1/2}C\Sigma^{-1/2})$ is dominated in probability[4] by $\sigma^2\chi^2(d)$, which has expectation $\sigma^2 d$. We thus have an asymptotic upper bound on $\mathbb{E}[R(\hat{\theta})] - R(\theta_*)$.

However, we are mainly interested in *non-asymptotic bounds*. Under a kurtosis condition (there exists $C > 0$ such that for all $\theta \in \mathbb{R}^d$, $\mathbb{E}[\langle\theta, \Phi(X)\rangle^4] \leqslant C\mathbb{E}^2[\langle\theta, \Phi(X)\rangle^2]$), a uniform one-sided law holds: with high probability, for any $n \geqslant d$, the generalization error is uniformly upper bounded by the in-sample error (Raskutti et al., 2014; Mendelson, 2014). This results in a non-asymptotic bound on the generalization error. A tighter error decomposition has been proposed by Hsu et al. (2014). Overall, for parametric least-squares, the excess of generalization error is of order $O\left(\frac{\sigma^2 d}{n}\right)$ (Lecué and Mendelson, 2016). The next section will give elements to show that this rate is optimal for least-squares regression.

[3] we denote $v^{\otimes 2} = vv^\top$.

[4] for two distributions μ, ν with support included in $\mathbb{R}$, μ dominates ν if the cumulative distribution function of μ is below the one of ν.

1.1.7 Minimax rates of convergence

In this introduction, and throughout the thesis, we derive upper convergence rates for estimation procedures. If the analysis is not tight, the upper bound may sometimes not reflect the actual behavior of an estimator. Moreover, the estimator itself may not be the best one. In order to fully understand the behavior of the estimator, and whether a better estimator exists, we may ask the two following questions:

- Given a learning rule, can we prove a lower bound on (the expectation of) its excess risk?
- Can we prove that given n observations, any learning rule has a lower bounded excess risk?

These two questions are quite different, as the first one is specific to a procedure or an algorithm, while the second one embraces all possible learning rules, regardless of their other properties (complexity, storage, etc.).

In this paragraph, we answer the second question for parametric least-squares. Such lower bounds are extremely insightful as they describe the optimal statistical performance that can be expected. Consequently, if we can prove that an estimator converges at a rate which matches the lower bound, we know that the upper bound actually reflects its behavior, and that no estimator would perform better. In Section 1.4, we will explain how these results can be extended to the non-parametric setting.

This approach takes its roots in the seminal work of Shannon (1948, 1949) on information theory. We define the minimax risk associated to a statistical model $\{P_\theta, \theta \in \Theta\}$ (Massart, 2007; Tsybakov, 2008):

$$\mathcal{R}_n^* := \inf_{\hat{\theta}} \sup_{\theta_* \in \Theta} \mathbb{E}_{P_{\theta_*}} \left[\mathrm{dist}\left(\hat{\theta}, \theta_*\right) \right] ,$$

where the infimum is taken over all estimators (*i.e.*, over all learning rules), $\mathbb{E}_{P_{\theta_*}}$ stands for the averaging over the observation's law P_{θ_*}, and dist is a distance on Θ. While in the parametric case we could also consider the squared euclidean distance $\mathrm{dist}(\hat{\theta}, \theta_*) = \|\hat{\theta} - \theta_*\|^2$, throughout this document, we mainly focus on prediction errors for the square loss. We thus state results on the distance which corresponds to the excess risk of a predictor (see Equation (1.3)): $\mathrm{dist}(\hat{\theta} - \theta_*) = \|\Sigma^{1/2}(\hat{\theta} - \theta_*)\|^2$.

We call ψ_n an *optimal rate of convergence* if there exist positive constants m, M such that

$$m \leqslant \liminf_{n\to\infty} \mathcal{R}_n^* \psi_n \leqslant \limsup_{n\to\infty} \mathcal{R}_n^* \psi_n \leqslant M \ . \tag{1.4}$$

Moreover, we say that an estimator $\hat{\theta}_n$ converges at the optimal statistical rate if $\sup_{\theta\in\Theta} \mathbb{E}_\theta[\|\hat{\theta}_n - \theta\|^2] \leqslant C\psi_n$ for some $C > 0$.

Lower bound for linear least-squares regression. In the parametric linear regression setting with fixed design, under a Gaussian model for the noise, with homoscedastic variance σ^2, the optimal rate is $\psi_n = \frac{\sigma^2 d}{n}$ (Massart, 2007). Similarly, in the random design setting, the rate is also $\frac{\sigma^2 d}{n}$ (Tsybakov, 2003).

It thus appears that ERM for least squares regression achieves the optimal rate of convergence. Yet, computing the empirical risk minimizer can be difficult.

1.1.8 Computational cost of ERM

In machine learning, dimension d often becomes very large. Therefore, finding the empirical risk minimizer, *i.e.*, solving the linear system (Equation (E)), can be prohibitive. In the next two sections, we describe the main tools (mainly stochastic approximation and convex optimization) used to build estimators than can be computed more efficiently than the ERM. Bottou and Bousquet (2008) underlined two key insights which shed light on how to approach this task: first, the true goal is to minimize the generalization error, thus using the empirical risk is not necessary; second, as no estimator can converge faster than the statistical rate, it is un-necessary to solve optimization problems beyond the statistical level.

1.2 Convex optimization

As the problem we address is expressed as a minimization problem, which is convex if the loss is a.s. convex in θ, we here recall a few results from convex optimization. We introduce gradient descent in Section 1.2.2, and a modification relying on a second order system, accelerated gradient descent in Section 1.2.3. We then briefly describe the framework used to show that these rates are optimal in Section 1.2.4.

1.2.1 Assumptions

We denote $\mathcal{C}^p(\mathbb{R}^d)$ the set of p times continuously differentiable functions from $\mathbb{R}^d$ into $\mathbb{R}$. For $f \in \mathcal{C}^p(\mathbb{R}^d)$, we denote $f^{(n)}$ the n-th differential of f. In particular, when f is $\mathcal{C}^1$, we denote f' its gradient. By definition, a function $f \in \mathcal{C}^1(\mathbb{R}^d)$ is *convex* if for any $\theta, \eta \in \mathbb{R}^d$ we have:

$$f(\eta) \geqslant f(\theta) + \langle f'(\theta), \eta - \theta \rangle \, .$$

Moreover, f is L-*smooth* if its gradient is L-Lipschitz, *i.e.*, if there exists a constant $L > 0$, such that for any $\theta, \eta \in \mathbb{R}^d$ we have:

$$\|f'(\eta) - f'(\theta)\| \leqslant L\|\eta - \theta\| \, .$$

Finally, f is μ-*strongly convex* if there exists a constant $\mu > 0$, such that for any $\theta, \eta \in \mathbb{R}^d$ we have:

$$f(\eta) \geqslant f(\theta) + \langle f'(\theta), \eta - \theta \rangle + \frac{\mu}{2}\|\theta - \eta\|^2 \, .$$

Note that if f is $\mathcal{C}^2$, then it is convex if and only if for all $\theta \in \mathbb{R}^d$, the hessian matrix $f''(\theta)$ at θ satisfies $f''(\theta) \succcurlyeq 0$, strongly convex if and only if for all $\theta \in \mathbb{R}^d$, $f''(\theta) \succcurlyeq \mu \, \mathrm{Id}$, and L-smooth if and only if for all $\theta \in \mathbb{R}^d$, $f''(\theta) \preccurlyeq L \, \mathrm{Id}$. Following Nesterov (2004), we denote $\mathcal{F}_L^1(\mathbb{R}^d)$ the subset of $\mathcal{C}^1(\mathbb{R}^d)$ of convex L-smooth functions and $\mathcal{S}_{\mu,L}^1(\mathbb{R}^d)$ the subset of $\mathcal{C}^1(\mathbb{R}^d)$ of μ-strongly convex and L-smooth functions.

Convex-smooth functions satisfy the following inequality, which is central for the analysis of stochastic gradient descent: for all $\theta, \eta \in \mathbb{R}^d$,

$$\|f'(\theta) - f'(\eta)\|^2 \leqslant L \langle \theta - \eta, f'(\theta) - f'(\eta) \rangle \, , \tag{1.5}$$

This is the so called co-coercivity Lemma (Nesterov, 2004): it strengthens the Lipschitz gradient inequality when the function is also convex.

1.2.2 Gradient methods

For a convex function f in $\mathcal{F}^1_L(\mathbb{R}^d)$, we consider the following optimization problem:

$$\mathcal{P}_f := \min_{\theta \in \mathbb{R}^d} f(\theta) \ . \tag{1.6}$$

For simplicity, we assume that f has a unique minimum θ_* (this always holds when f is in $\mathcal{S}^1_{\mu,L}(\mathbb{R}^d)$).

The simplest first-order algorithm is *gradient descent* (GD). We start from some θ_0 and compute the sequence $(\theta_n)_{n\in\mathbb{N}^*}$ defined recursively by:

$$\theta_n = \theta_{n-1} - \gamma_n f'(\theta_{n-1}) \ . \tag{1.7}$$

This is a simple incremental algorithm, which updates the current iterate moving in the opposite direction of the gradient, as illustrated in Figure 1.2. At each step, we only access or use a gradient, thus the computational cost is much lower than with Newton's methods, which use second derivatives of the function (Boyd and Vandenberghe, 2004). The sequence $(\gamma_n)_{n\in\mathbb{N}}$ is called the sequence of *step sizes* or *learning rates*. The choice of the learning rate is fundamental, and has been one of the most studied questions.

For a smooth function, the simplest choice is to use a constant learning rate: indeed, choosing $\gamma_n = \frac{1}{L}$ for all $n \in \mathbb{N}^*$, ensures convergence for all smooth convex functions, with a faster rate if the function is strongly convex. More precisely, we have the following proposition:

Proposition 1.1 (Convergence of gradient descent (Nesterov, 2004))**.** *Let $f \in \mathcal{F}^1_L(\mathbb{R}^d)$, and $\gamma_n = \frac{1}{L}$, for all $n \in \mathbb{N}^*$. The gradient method (1.7) generates a sequence $(\theta_n)_{n\in\mathbb{N}}$ satisfying*

$$f(\theta_n) - f(\theta_*) \leqslant \frac{2L \left\|\theta_0 - \theta_*\right\|^2}{n+4} \ .$$

Moreover, if $f \in \mathcal{S}^1_{\mu,L}(\mathbb{R}^d)$, we also have, with $\gamma_n = \frac{1}{L}$ for all $n \in \mathbb{N}^*$:

$$f(\theta_n) - f(\theta_*) \leqslant \left(1 - \frac{\mu}{L}\right)^n \left(f(\theta_0) - f(\theta_*)\right) \ ,$$

and a slightly more powerful result if $\gamma_n = \frac{2}{L+\mu}$ for all $n \in \mathbb{N}^*$:

$$f(\theta_n) - f(\theta_*) \leqslant \left(1 - \frac{2\mu}{\mu + L}\right)^{2n} \frac{L}{2} \left\|\theta_0 - \theta_*\right\|^2 \ .$$

The first two equations show that the choice $\gamma = 1/L$ is very powerful, as the algorithm then adapts to the difficulty of the problem: a unique algorithm works for all convex functions, but the convergence is faster if the function is strongly convex. Moreover, this choice of step sizes does not require the knowledge of μ. The last equation, on the contrary, holds for a step size which depends on μ, and while convergence is asymptotically faster, it does not show convergence if μ goes to 0.

1.2.3 Accelerated gradient descent

Nesterov (1983) proposed an improvement of the gradient method, able to achieve faster rates of convergence: *accelerated gradient descent* (AGD). This algorithm takes the following generic form: starting form some θ_0 and $\eta_0 = \theta_0$, for any $n \in \mathbb{N}$, for some sequences of step size $(\gamma_n)_{n\in\mathbb{N}^*}$, and momentum $(\delta_n)_{n\in\mathbb{N}^*}$,

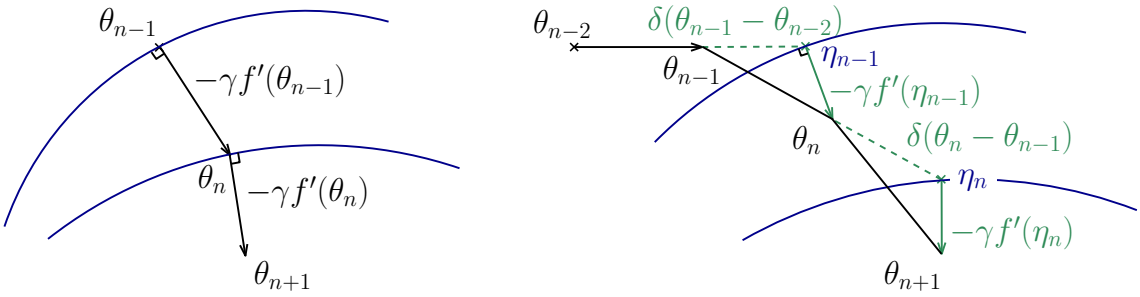


Figure 1.2: Gradient descent and Accelerated Gradient Descent
Left: Gradient descent. *Right:* Accelerated Gradient descent. The blue lines are the level lines of the objective function f.

$$\begin{cases} \theta_n &= \eta_{n-1} - \gamma_n f'(\eta_{n-1}) \\ \eta_n &= \theta_n + \delta_n(\theta_n - \theta_{n-1}) \ . \end{cases} \tag{1.8}$$

We thus compute two updates, one being a normal gradient update, the second one being an extrapolation from the two previous points, an "acceleration" proportional to the momentum coefficient δ_n, with $0 \leqslant \delta_n \leqslant 1$. See Figure 1.2.

Proposition 1.2 (Convergence of accelerated gradient descent)**.** *Let $f \in \mathcal{F}_L^1(\mathbb{R}^d)$, and for all $n \in \mathbb{N}^*$, let $\gamma_n = \frac{1}{L}$, and $\delta_n = \frac{n-1}{n+2}$. The accelerated gradient method (1.8) generates a sequence (θ_n) satisfying:*

$$f(\theta_n) - f(\theta_*) \leqslant \frac{2L \left\|\theta_0 - \theta_*\right\|^2}{(n+1)^2} \ .$$

Moreover, if $f \in \mathcal{S}_{\mu,L}^1(\mathbb{R}^d)$, and for all $n \in \mathbb{N}^*$, $\gamma_n = \frac{1}{L}$, $\delta_n = \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$, we have:

$$f(\theta_n) - f(\theta_*) \leqslant \left(1 - \sqrt{\frac{\mu}{L}}\right)^n \frac{L+\mu}{2} \left\|\theta_0 - \theta_*\right\|^2 \ .$$

Acceleration with a momentum term was originally proposed by Nesterov (1983). The result given here is from Nesterov (2004) for the strongly convex case, and from Schmidt et al. (2011) for the convex case with a simple sequence of steps (a summary and proofs of these results can be found for example in Bubeck (2015)). Acceleration has since received large attention, with several approaches to explain the improvement in the rate: the idea of adding a momentum terms dates back to the heavy ball algorithm by Polyak (1964), but several other interpretations have been proposed: Allen-Zhu and Orecchia (2017) viewed AGD as a linear coupling of gradient descent and mirror descent, Bubeck et al. (2015) proposed a simple geometric reason for the possibility of acceleration, while Su et al. (2014) described it as the discretization of a certain second-order ODE.

1.2.4 Lower complexity bounds

These results raise again a very natural question: can we do better? To address such a question, one first needs to define its precise meaning: what is the problem we are trying to solve, what type of information is our method allowed to access, and how do we measure the cost of a method. To do so, we introduce the *black box optimization* framework (Nemirovsky and Yudin, 1983; Nesterov, 2004; Juditsky and Nemirovski, 2011).

This framework formalizes the analysis of optimization complexity. We consider a class of problems (here, a class of functions), and a set of available methods. For a problem $\mathcal{P}_f$ (as in Equation (1.6)), we define the *performance* of a method as the amount of computation needed to *solve* the problem: to *solve* a problem is to find an ε-approximate solution. The performance of a method on a class of problems is the worst case performance of the method on any problem of the class. Here, we consider two classes of problems: respectively $\mathcal{F}_L^1(\mathbb{R}^d)$ and $\mathcal{S}_{\mu,L}^1(\mathbb{R}^d)$ for fixed $L, \mu > 0$.

To define the set of available methods, we describe the information the method accesses in terms of calls to an *oracle*. An *oracle* is a routine that answers question asked by the method. Typically, for $p \geqslant 0$, a p-th order oracle provides the value of the function at a point, together with the first p derivatives of the function at that point. We consider here the set of *first-order incremental methods*. Such a method incrementally performs oracle calls, and updates its current prediction with respect to the available information. We also restrict to methods that build estimators in the linear span of the gradients, which holds for the majority of practical methods.

The amount of computation can be either defined as the number of oracle calls, or as the total number of arithmetic operations. Here, we consider only the number of iterations as it allows to directly compare the convergence rates of the different methods. However, to guarantee a fair comparison, we need to check that the cost of the updates (given the answer of the oracle), is the same for the different methods.

We have the following results on the lower complexity bounds. We consider $\mu, L > 0$:

Proposition 1.3 (Lower complexity bounds for $\mathcal{F}_L^1(\mathbb{R}^d)$ and $\mathcal{S}_{\mu,L}^1(\mathbb{R}^d)$ (Nesterov, 2004))**.** ***Convex case:*** *For any $n \in \mathbb{N}$ such that $1 \leqslant n \leqslant \frac{d-1}{2}$, for any $\theta_0 \in \mathbb{R}^d$, there exists a function $f \in \mathcal{F}_L^1(\mathbb{R}^d)$, such that for any first-order method providing θ_n after n iterations, we have:*

$$f(\theta_n) - f(\theta_*) \geqslant \frac{3L \left\|\theta_0 - \theta_*\right\|^2}{32(n+1)^2} .$$

Strongly-convex case: *Let ℓ_2 be the set of squared summable sequences, embedded with the norm $\|(x_k)_{k\in\mathbb{N}}\|_{\ell_2}^2 := \sum_{k=0}^{\infty} x_k^2$. For any $n \in \mathbb{N}$, for any $\theta_0 \in \ell_2$, there exists a function $f \in \mathcal{S}_{\mu,L}(\ell_2)$, such that for any first-order method providing θ_n after n iterations, we have:*

$$f(\theta_n) - f(\theta_*) \geqslant \frac{\mu}{2} \left\|\theta_n - \theta_*\right\|_{\ell_2}^2 \geqslant \frac{\mu}{2} \left(1 - \sqrt{\frac{L}{\mu}}\right)^{2n} \left\|\theta_0 - \theta_*\right\|_{\ell_2}^2 .$$

In the strongly convex case, as ℓ_2 is an infinite-dimensional space, there is no restriction on the number of iterations as in the non-strongly convex case.

To prove such results, Nesterov (2004) proposed a simple quadratic function, for which no algorithm can converge faster than the described rate. Note that in the first case, the result only holds for the first $n \leqslant \frac{d}{2}$ iterations. Rates are summarized in the following table:

	$\mathcal{F}_L^1(\mathbb{R}^d)$	$\mathcal{S}_{\mu,L}^1(\mathbb{R}^d)$
GD	$\dfrac{L \left\|\theta_0 - \theta_*\right\|^2}{n}$	$\left(1 - \dfrac{\mu}{L}\right)^n (f(\theta_0) - f(\theta_*))$
AGD	$\dfrac{L \left\|\theta_0 - \theta_*\right\|^2}{n^2}$	$\left(1 - \sqrt{\dfrac{\mu}{L}}\right)^n L \left\|\theta_0 - \theta_*\right\|^2$

In machine learning, convex optimization can be used to optimize the empirical risk. Specifically, instead of looking for exact minimizer of the empirical loss, we can compute

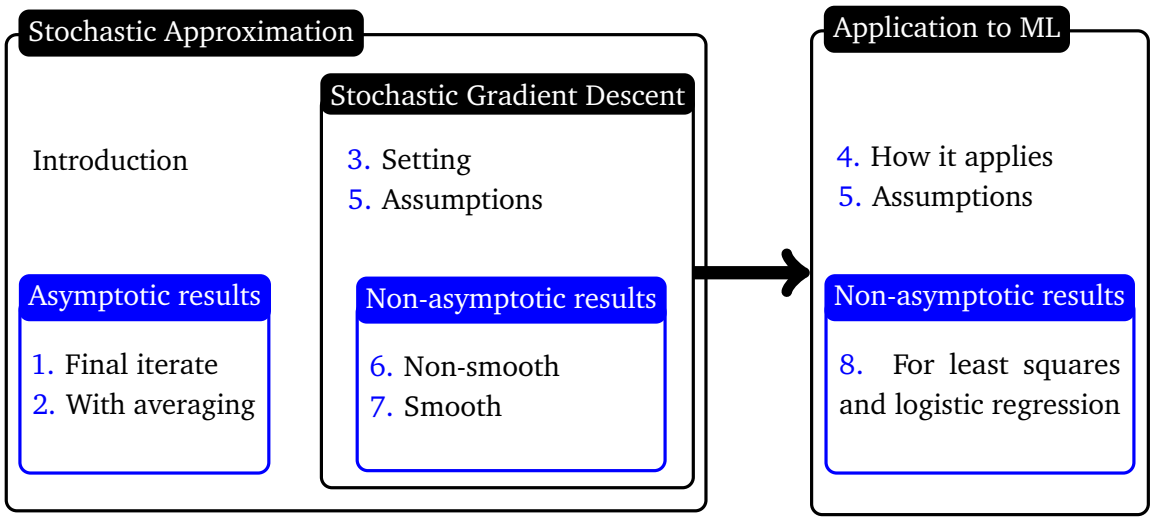


Table 1.2: Organization of the Section 1.3. All numbers are references to subsections of Section 1.3. Stochastic gradient descent is a particular case of stochastic approximation, and that be used in machine learning.

an iterative sequence of estimators, using GD or AGD on the empirical risk; the cost of each iteration is then of order $O(nd)$ per iteration, in order to compute the gradient of the average of n functions. This generates an additional *optimization error*, but it appears that after n iterations for GD (or $\sqrt{n}$ iterations for AGD), this additional error is of the same order as the statistical error. However, as it will appear in the following part, using randomized gradient methods further reduces the complexity.

1.3 Stochastic approximation

Robbins and Monro (1951) introduced *stochastic approximation* (SA) as the following iterative sequence, for a function $h : \mathbb{R}^d \to \mathbb{R}^d$, for any $n \in \mathbb{N}^*$:

$$\theta_n = \theta_{n-1} - \gamma_n(h(\theta_{n-1}) + \varepsilon_n) \ ,$$

where $(\varepsilon_n)_{n\in\mathbb{N}^*}$ are random variables corresponding to a noise and $(\gamma_n)_{n\in\mathbb{N}^*}$ is a positive deterministic sequence of step sizes. We assume that there exists a filtration[5] $(\mathcal{F}_n)_{n\in\mathbb{N}}$ such that θ_n is $\mathcal{F}_n$-measurable, and that the noise has 0 mean given past information, *i.e.*, $\mathbb{E}\left[\varepsilon_n|\mathcal{F}_{n-1}\right] = 0$. The original work of Robbins and Monro was motivated by the problem of finding a root of a continuous function, when the function is not completely known, but noisy evaluations of the function at any desired point are available.

In machine learning, stochastic approximation is used to find the minimum of functions by searching for roots of their gradients (we consider $h = f'$). However, stochastic approximation goes beyond minimization problems, and can be applied without convexity. Notably, it has been used in wireless communications, repeated games, decision problems in economics, amongst others (Benaim and Hirsch, 1999; Marcet and Sargent, 1989). Overall, it resulted in a tremendous amount of both theoretical work and practical applications (Kushner and Yin, 2003; Benveniste et al., 2012).

In this section, we describe several results of the literature. In Section 1.3.1 and Section 1.3.2, we review asymptotic results on stochastic approximation, respectively

[5] an increasing sequence of σ-algebras (Billingsley, 2008).

for the last iterate and using an averaging scheme on the iterates. We describe how stochastic approximation is used for optimization, introducing stochastic gradient descent in Section 1.3.3, then how it applies to machine learning in Section 1.3.4. We discuss assumptions that can be made on the sequence of noise in Section 1.3.5. Finally, we describe results for stochastic optimization in Section 1.3.6 and Section 1.3.7, and in some special case of machine learning in Section 1.3.8. The structure of this chapter is summarized in Table 1.2.

1.3.1 Convergence of the last iterate

Theoretical results include the proof of the convergence to a root of h (in probability, or almost surely), the analysis of the convergence speed, and the asymptotical behavior, depending on the learning rate $(\gamma_n)_{n\in\mathbb{N}^*}$ (Duflo, 1997).

The convergence almost always depends on two aspects: the possibility of forgetting the initialization choice, and the robustness to noise. As a consequence, traditional learning rates satisfy two properties: $\sum_{n=1}^{\infty}\gamma_n = \infty$, which ensures that the initial condition will be forgotten, and $\sum_{n=1}^{\infty}\gamma_n^2 < \infty$, which limits the influence of the noise. As it will appear later on, this second condition can be relaxed.

Traditional analysis requires the introduction of a Lyapunov function, *i.e.*, a differentiable smooth function $V : \mathbb{R}^d \to \mathbb{R}$ such that for any $\theta \in \mathbb{R}^d$, $\|h(\theta)\|^2 \leqslant C(1 + V(\theta))$ and $\langle h(\theta), V'(\theta)\rangle \geqslant \mu \|V'(\theta)\|^2$. In the context of machine learning, for μ-strongly convex and L-smooth risks, the function V such that $V(\theta) := R(\theta) - R(\theta_*)$ is a Lyapunov function. If such a Lyapunov function exists, under simple assumptions on the noise (typically, $\mathbb{E}[\|\varepsilon_n\|^2|\mathcal{F}_{n-1}] \leqslant \sigma^2$), the convergence of $V(\theta_n)$ to 0 is guaranteed; this convergence holds in expectation if $\sum_{n=1}^{\infty}\gamma_n = \infty$ and $\gamma_n \overset{n\to\infty}{\to} 0$, and almost surely if one also has $\sum_{n=1}^{\infty}\gamma_n^2 = \infty$ (Robbins and Siegmund, 1985). These properties suggested to use decaying sequences of step size, typically scaling as $\gamma_n \propto n^{-\zeta}$ for $\zeta \in]\frac{1}{2}; 1]$. The convergence can then be extended to non-random noise with vanishing magnitudes (Duflo, 1997; Schmidt et al., 2011).

In regards to asymptotical results, Fabian (1968) showed that for $\gamma_n = \gamma_0 n^{-1}$, with $\gamma_0 \geqslant \mu^{-1}$, the sequence $\sqrt{n}(\theta_n - \theta_*)$ is asymptotically normal. This is a powerful result but unfortunately, the variance of the chain is potentially large (scaling as μ^{-2}), sensitive to ill conditioning ($\mu \to 0$), and the proposed choice of initial step size depends on the unknown μ.

1.3.2 Polyak-Ruppert averaging

This difficulty of selecting an appropriate step size was partially tackled in a fundamental paper by Polyak and Juditsky (1992), generalizing one-dimensional results from Ruppert (1988). The *Polyak-Ruppert averaging* consists in considering the sequence of averaged iterates:

$$\bar{\theta}_n := \frac{1}{n+1}\sum_{k=0}^{n}\theta_k \, .$$

Loosely speaking, they showed that for a wider range of decaying step sizes, this averaged sequence converges to its limit at optimal rate: it underlines that one can use larger step sizes than n^{-1} and benefit from the fact that the off-line averaging naturally reduces the

higher noise induced by a larger step size. In addition, this averaged sequence can be computed online, as for any $n \in \mathbb{N}$

$$\bar{\theta}_n = \frac{1}{n+1}\theta_n + \frac{n}{n+1}\bar{\theta}_{n-1}.$$

More precisely, Polyak and Juditsky (1992) showed that for any sequence of step sizes $\gamma_n = \gamma_0 n^{-\zeta}$, with $\zeta \in]\frac{1}{2}; 1[$, the sequence $\sqrt{n}(\bar{\theta}_n - \theta_*)$ converges in distribution to a normal law, with a variance that *does not* depend on (γ_n), and is asymptotically efficient[6] (Van der Vaart, 1998). This result led to numerous extensions (Chen, 1993; Delyon and Juditsky, 1992; Kushner and Yang, 1993; Yin, 1991).

Deriving *non-asymptotic convergence rates* is more challenging. In the following sections, we summarize a few important results in the context of stochastic optimization, first without strong convexity, then with strong convexity. But before, we describe how stochastic approximation is used for optimization.

1.3.3 Stochastic gradient descent.

To optimize a convex function f, one searches for a root of its gradient f'. Therefore, we consider the stochastic gradient descent algorithm (SGD), for $k \in \mathbb{N}^*$:

$$\theta_k \quad = \quad \theta_{k-1} - \gamma_k f'_k(\theta_{k-1})$$

with $f'_k(\theta_{k-1})$ being an unbiased estimate of the gradient, *i.e.*,

$$\mathbb{E}\left[f'_k(\theta_{k-1})|\mathcal{F}_{k-1}\right] = f'(\theta_{k-1}), \tag{1.9}$$

for a filtration $(\mathcal{F}_k)_{k\in\mathbb{N}}$, such that θ_k is $\mathcal{F}_k$-measurable. Thus with $(\varepsilon_k)_{k\in\mathbb{N}^*}$ the sequence of noise *functions*, such that for all $k \in \mathbb{N}^*$, $\varepsilon_k = f' - f'_k$, the recursion is written:

$$\theta_k \quad = \quad \theta_{k-1} - \gamma_k(f'(\theta_{k-1}) + \varepsilon_k(\theta_{k-1})) \ .$$

The noise functions satisfy:

$$\mathbb{E}\left[\varepsilon_k(\theta_{k-1})|\mathcal{F}_{k-1}\right] = 0. \tag{1.10}$$

As it is defined, f'_k is not necessarily the gradient of a function f_k, but just an arbitrary notation for the noisy gradient, as $f'_k = f' + \varepsilon_k$. However, this notation is convenient as there may exist functions f_k such that $f'_k = (f_k)'$.

In a slightly different setting, the assumption that the noise is random and is a 0 mean random variable is removed, and methods are analyzed for deterministic "small errors", a context occurring if one only observes an inexact oracle on the gradient (Devolder et al., 2014). We do not consider such a situation, as the machine learning setting naturally fits into the framework of Equation 1.9, as described in the next paragraph.

1.3.4 Application to machine learning and optimization.

In supervised machine learning, the function we seek to minimize is either the generalization error, or the training loss for ERM. Using i.i.d. observations, both of these tasks can be

[6] definition was given in Section 1.1.6.

addressed with SGD. For any $k \in [\![1; n]\!]$, we define the loss on observation k as the function f_k defined by

$$f_k : \theta \mapsto \ell(\langle \theta, \Phi(x_k) \rangle, y_k). \tag{1.11}$$

We use these losses to build stochastic gradients. Note that we could also use a batch (or mini-batch) of observations as each step (Cotter et al., 2011; Dekel et al., 2012; Jain et al., 2016).

For the empirical error: Recall that

$$R_n(\theta) = \frac{1}{n} \sum_{k=1}^{n} \ell(\langle \theta, \Phi(x_k) \rangle, y_k).$$

We consider the filtration $(\mathcal{F}_k)_{k \in \mathbb{N}}$, with

$$\mathcal{F}_k = \sigma((x_i, y_i)_{i \in [\![1;n]\!]}, (I_i)_{i \in [\![1;k]\!]}) \ ,$$

where for any step $k \in \mathbb{N}^*$, $I_k \sim \mathcal{U}[\![1; n]\!]$ is an index uniformly sampled over $[\![1; n]\!]$. We emphasize that, for any $k \in \mathbb{N}$, all the observations are $\mathcal{F}_k$-measurable. Then $f'_{I_k}(\theta_{k-1}) = \ell'(\langle \theta_{k-1}, \Phi(x_{I_k}) \rangle, y_{I_k})$ is an unbiased gradient of the *empirical* risk R_n:

$$\mathbb{E}[f'_{I_k}(\theta_{k-1}) | \mathcal{F}_{k-1}] = R'_n(\theta_{k-1}).$$

With this setting, SGD can be used to minimize the empirical risk. However, our goal remains to minimize the generalization error; once the empirical loss is minimized, this requires an additional control on $R_n - R$.

For the generalization error: Recall that

$$R(\theta) = \mathbb{E}[\ell(\langle \theta, \Phi(X) \rangle), Y].$$

We consider the filtration $(\mathcal{F}_k)_{k \in [\![0;n]\!]}$, with

$$\mathcal{F}_k = \sigma((x_i, y_i)_{i \in [\![1;k]\!]}) \ ,$$

where for any $k \in [\![1; n]\!]$, a new point (x_k, y_k), *independent* of θ_{k-1} has been added to the $\mathcal{F}_k$. Then $f'_k(\theta_{k-1}) = \ell'(\langle \theta_{k-1}, \Phi(x_k) \rangle, y_k)$ is in an unbiased gradient of the *true* risk R. Indeed,

$$\mathbb{E}[f'_k(\theta_{k-1}) | \mathcal{F}_{k-1}] = R'(\theta_{k-1}). \tag{1.12}$$

This is a very powerful method, which emphasizes that stochastic approximation is far more than a simple optimization tool. It directly minimizes the true risk, which is an un-known function. As a consequence, it avoids the need for regularization, which is only meant to avoid converging to a minimum of the empirical risk that would poorly generalize. In other words, as the algorithm converges to a minimum of the true risk, it *cannot over-fit*. The only constraint is that one can only perform a *single pass* through the data-set.

In the case of least-squares regression, this algorithm is called *least-mean-squares* (LMS).

Smoothness and strong convexity conditions in machine learning applications. In order to understand how the results described for stochastic approximation apply to the true risk R or the empirical risk R_n, it is necessary to check which assumptions these functions satisfy.

First, most usual loss functions ℓ are almost surely convex[7] in θ (*i.e.*, for a.s. any (X, Y), $\theta \mapsto \ell(\theta, (X, Y))$ is convex). By integration, R is convex.

Moreover, if the loss is twice differentiable, R is also twice differentiable, and for all $\theta \in \mathbb{R}^d$, $R''(\theta) = \mathbb{E}\left[\ell''(\langle \theta, \Phi(X)\rangle, Y)\Phi(X)\Phi(X)^\top\right]$. Thus if ℓ is also L_ℓ-smooth in its first variable (this is the case for the logistic loss and the square loss, but not for the hinge loss), R is smooth if $\mathbb{E}[\|\Phi(X)\|^2] \leqslant r^2$. Indeed, we then have

$$R''(\theta) \preccurlyeq L_\ell \mathbb{E}\left[\Phi(X)\Phi(X)^\top\right] \preccurlyeq L_\ell r^2 \operatorname{Id} .$$

Similarly, if ℓ is twice differentiable and μ_ℓ-strongly convex and if $\Phi(X)\Phi(X)^\top$ is invertible, then R is also strongly convex, as

$$R''(\theta) \succcurlyeq \mu_\ell \mathbb{E}\left[\Phi(X)\Phi(X)^\top\right] \succcurlyeq \mu_\ell \lambda_{\min}(\Phi(X)\Phi(X)^\top) \operatorname{Id} .$$

However, unless a regularization is added to force strong convexity, $\lambda_{\min}(\Phi(X)\Phi(X)^\top)$ may be arbitrarily small. Indeed, if the feature vector has finite second order moment, which is a common assumption, then the smallest eigenvalue covariance $\lambda_{\min}(\Phi(X)\Phi(X)^\top)$ satisfies

$$\lambda_{\min}\left(\Phi(X)\Phi(X)^\top\right) \leqslant \frac{\operatorname{tr}(\mathbb{E}\left[\Phi(X)\Phi(X)^\top\right])}{d} = \frac{\mathbb{E}\left[\|\Phi(X)\|^2\right]}{d},$$

thus it is generally very close to 0 when dimension d is large.

In the following Section, we introduce supplementary assumptions on the noise functions, and discuss their validity in the machine learning context as we go along.

1.3.5 Assumptions on the noise

In order to obtain more precise results, several types of assumptions can be made on the noise function $(\varepsilon_k)_{k\in\mathbb{N}^*}$ or equivalently when they exist on the function $(f_k)_{k\in\mathbb{N}^*}$. We still assume that we observe *unbiased* gradients, satisfying Equation (1.10). Moreover, in the context of this thesis, we only consider i.i.d. noise functions:

H1 (I.i.d. noise functions). *The sequence $(\varepsilon_k)_{k\in\mathbb{N}^*}$ is i.i.d.*

This is a reasonably weak assumption, which always applies in the context of machine learning described above, since the observations are i.i.d. and we use one observation for each gradient. Dependent (or non identically distributed) observations are studied in the context of *online learning* (Cesa-Bianchi and Lugosi, 2006). This situation is relevant in practice, as the observations may come from times series, or even from an adversary who tries to degrade the performance of the learner (this adversarial situation naturally relates to game theory). In such a situation, there is no common distribution for all the observations. It is then impossible to define a generalization risk, thus one seeks to minimize another function, called the *regret* instead. Analysis of gradient methods for online learning has led to numerous studies (Hazan et al., 2007; Bubeck and Cesa-Bianchi,

[7] apart from the 0-1 loss in classification, which is, for this reason, generally replaced by a convex surrogate.

2012; Shalev-Shwartz, 2011). Interestingly, this theory also brings ways to analyze the performance of *multiple passes* stochastic gradient descent (Shamir, 2016).

Even though the noise functions $(\varepsilon_k)_{k\in\mathbb{N}^*}$ are i.i.d., the sequence of noise actually suffered $(\varepsilon_k(\theta_{k-1}))_{k\in\mathbb{N}^*}$ is generally *not i.i.d.* , since the iterate θ_{k-1} depends on the previous noise functions. To overcome this issue when it is necessary, we may use the following assumption, generally described as the *"semi-stochastic" setting*.

H2 (Semi-Stochastic). *For any $k \in \mathbb{N}^*$[8], the noise function ε_k is constant: $\varepsilon_k(\theta)$ does not depend on θ.*

Under Assumption **H**2, the sequence $(\varepsilon_k(\theta_{k-1}))_{k\in\mathbb{N}^*}$ is $i.i.d.$.

More generally, any noise can de decomposed as an additive noise plus a more general noise:

$$\varepsilon_k(\theta_{k-1}) = \underbrace{\varepsilon_k(\theta_{k-1}) - \varepsilon_k(\theta_*)}_{\text{General noise}} + \underbrace{\varepsilon_k(\theta_*)}_{\text{Additive noise}} . \tag{1.13}$$

As the additive noise does not depend on θ_k, the noise satisfies Assumption **H**2 if and only if the "general noise" is null. For least-mean-squares, since $f_k(\theta_{k-1}) = \frac{1}{2}(\langle\Phi(x_k), \theta_{k-1}\rangle - y_k)^2$ decomposition (1.13) can be written as:

$$\begin{aligned}
\varepsilon_k(\theta_{k-1}) &= f_k'(\theta_{k-1}) - f'(\theta_{k-1}) \\
&= (\langle\Phi(x_k), \theta_{k-1}\rangle - y_k)\Phi(x_k) - \mathbb{E}\left[(\langle\Phi(X), \theta_{k-1}\rangle - Y)\Phi(X)\right] \\
&= \underbrace{\left(\Phi(x_k)\Phi(x_k)^\top - \mathbb{E}\left[\Phi(X)\Phi(X)^\top\right]\right)(\theta_{k-1} - \theta_*)}_{\text{Multiplicative noise}} + \underbrace{(\langle\Phi(x_k), \theta_*\rangle - y_k)\Phi(x_k)}_{\text{Additive noise}},
\end{aligned}$$

using that $\mathbb{E}\left[\Phi(X)\Phi(X)^\top\right]\theta_* = \mathbb{E}[Y\Phi(X)]$ (Equation (1.2)). Due to its particular multiplicative structure, the general part for least mean squares is called the *multiplicative noise*. Proving results for general noises is harder than for additive noise functions: in Chapter 3, some of the results for acceleration only hold under Assumption **H**2.

However, this assumption is not always valid (e.g., for LMS), making the theorems that rely on it somehow limited. One can often use less restrictive assumptions, which aim to control the general part of the noise:

H3 (Lipschitz noise). *For any $k \in \mathbb{N}^*$, the noise function ε_k is a.s. L_∞-Lipschitz. In particular, a.s.,*

$$\|\varepsilon_k(\theta_{k-1}) - \varepsilon_k(\theta_*)\| \leqslant L\|\theta_{k-1} - \theta_*\| .$$

Assumption **H**3 will often be seen as a consequence of the following assumption.

H4 (f_k a.s. convex smooth). *There exist functions $(f_k)_{k\in\mathbb{N}^*}$, such that for any $k \in \mathbb{N}^*$, $(f_k)' = f_k'$. Moreover, f_k is a.s. convex and L_∞-smooth, thus satisfies (Eq. (1.5)):*

$$\left\|f_k'(\theta_{k-1}) - f_k'(\theta_*)\right\|^2 \leqslant L_\infty \left\langle f_k'(\theta_{k-1}) - f_k'(\theta_*), \theta_{k-1} - \theta_*\right\rangle . \tag{1.14}$$

[8] since the noise functions are i.i.d., Assumptions **H**2-6 hold for any $k \in \mathbb{N}^*$ if and only if they hold for $k = 1$. Assumptions could be made only on ε_1, but equivalently holds for any k.

This assumption can be extended to the setting where the noisy gradients f'_k do not come from the derivation of a function f_k. The assumption is then that the noisy gradients are a.s. L_∞-co-coercive, which is, by definition, Equation 1.14 (Zhu and Marcotte, 1996), see Chapter 4 for details. Finally, a slightly weaker version only assumes this inequality 1.14 in mean

H5 (f_k a.s. convex smooth in QM). *There exist functions $(f_k)_{k\in\mathbb{N}^*}$, such that for any $k \in \mathbb{N}^*$, $(f_k)' = f'_k$. Moreover, f_k is a.s. convex and L_2-smooth in* quadratic mean, *i.e.,*

$$\mathbb{E}\left[\left\|f'_k(\theta_{k-1}) - f'_k(\theta_*)\right\|^2 |\mathcal{F}_{k-1}\right] \leqslant L_2 \left\langle f'(\theta_{k-1}) - f'(\theta_*), \theta_{k-1} - \theta_* \right\rangle .$$

In the machine learning context, the existence and a.s. convexity of functions f_k is natural, since the loss measured on one observation (Equation (1.11)) is a.s. convex. Thus, Assumption **H**4 is generally true for bounded inputs ($\|\Phi(X)\| \leqslant R$, ρ_X-a.s.) and Assumption **H**5 holds if $\mathbb{E}_{\rho_X}[\|\Phi(X)\|^2] \leqslant R^2$. Moreover, Assumptions **H**4-5 only make sense if the risk is L-smooth, since they imply its smoothness[9]. The constants involved (if taken "optimally"), verify $L \leqslant L_2 \leqslant L_\infty$, which can typically influence the quality of the result (see discussion in Agarwal and Bottou, 2015).

In the non-smooth context, since assumptions **H**4 and **H**5 cannot hold, it is generally assumed that the noise is a.s. bounded (Bach and Moulines, 2011):

H6 (Bounded noise). *There exists B, such that, for all $k \in \mathbb{N}^*$, a.s. , $\|\varepsilon_k(\theta_{k-1})\| \leqslant B$.*

This assumption can be true in machine learning, e.g., when iterates are a.s. bounded (for example, constrained to live in a ball), assumption **H**3 is satisfied, and the additive part of the noise is a.s. bounded.

Assumptions made on the noise are summarized in the following table:

Chapter 2	Chapter 3	Chapter 4
H1, **H**5	**H**1, **H**2 or **H**5	**H**1, **H**4

Apart from these assumptions, the following assumption is also made in most parts of this thesis.

Structured noise. The noise is *structured* if the additive part of the noise $\varepsilon_k(\theta_*)$ satisfies $\mathbb{E}[\varepsilon_k(\theta_*)\varepsilon_k(\theta_*)^\top] \leqslant \sigma^2\Sigma$, for some $\sigma^2 > 0$. For LMS, as $\varepsilon_k(\theta_*) = (\langle\Phi(x_k), \theta_*\rangle - y_k)\Phi(x_k)$, this assumption is true for example if $(\langle\Phi(x_k), \theta_*\rangle - y_k)^2 \leqslant \sigma^2$ almost surely, or if the model is well-specified, (e.g., $y_k = \langle\theta_*, \Phi(x_k)\rangle + \xi_k$, with $(\xi_k)_{k\in[\![1;n]\!]}$ i.i.d. of variance σ^2 and independent of $\Phi(x_k)$). This is a crucial assumption: convergence rates are dramatically different if the noise is un-structured. For example Lan (2012) proved an optimal convergence rate of order $O(1/\sqrt{n})$ for un-structured noise.

1.3.6 Non-asymptotic results: stochastic approximation for minimizing convex functions

For convex functions (without strong convexity), although the sequence of iterates does not always converge, it is possible to prove bounds on $f(\bar{\theta}_n) - f(\theta_*)$, which shows convergence of the function values . For example, for $\gamma_n \propto n^{-1/2}$, Zhang (2004) showed that function values converge at speed $n^{-1/2}$.

[9] for any $k \in \mathbb{N}^*$, for any $\eta, \theta \in \mathbb{R}^d$, $\|R'(\theta) - R'(\eta)\|^2 = \|\mathbb{E}[f'_k(\theta) - f'_k(\eta)]\|^2 \leqslant \mathbb{E}[\|f'_k(\theta) - f'_k(\eta)\|^2] \leqslant \operatorname{ess\,sup} \|f'_k(\theta) - f'_k(\eta)\|^2$.

In the strongly convex case (but without smoothness), the upper bound on the function values scales as $(n\mu)^{-1}$, for decaying step sizes $\gamma_n = 2/(\mu(n+1))$, and a non uniform averaging or tail averaging scheme (Lacoste-Julien et al., 2012; Rakhlin et al., 2011); while with Polyak-Ruppert averaging, a $\log$ factor is lost in the worst case.

Note that both of these rates are optimal, respectively for the class of convex functions and the class of strongly convex functions: this optimality is meant in the sense that no algorithm querying n *stochastic first-order oracles* can achieve a better rate of convergence. Analysis dates back to Nemirovsky and Yudin (1983) and was nicely summarized and extended by Agarwal et al. (2012). It brings together tools from optimization (esp. first-order oracle models), information theory and statistics: the problem is shown to be as hard as an estimation problem, and classical tools from statistics are used, such as Fano's inequality.

1.3.7 Non-asymptotic results: stochastic approximation for minimizing smooth convex functions

The smoothness assumption does not change the minimax convergence rate in the convex case: Flammarion and Bach (2015) showed $\Omega(n^{-1/2})$ lower bound for a smooth (even quadratic) function (with unstructured noise). Following the intuition coming from the asymptotic rate by Polyak and Juditsky (1992), Bach and Moulines (2011) proposed a non-asymptotic analysis for *smooth strongly convex functions*, showing upper bounds for any decaying step size $\gamma_n \propto n^{-\zeta}$, with $\zeta \in [0,5;1]$. Precisely, the averaged iterate $\bar{\theta}_n$ converges in quadratic mean to θ_*: $\mathbb{E}[\|\bar{\theta}_n - \theta_*\|^2] = O((n\mu)^{-1})$, and the function values decay at asymptotic rate $O(n^{-1})$ if $\zeta \in]\frac{1}{2};1[$. Moreover, for logistic regression, Bach (2014) proved that for function values, with $\gamma_n \propto n^{-1/2}$, averaged SGD achieves the rate $O((n\mu)^{-1})$[10]. This uses the additional property that logistic regression is self-concordant.

In such situations, the step size $\gamma_n \propto n^{-1/2}$ is adaptive to strong convexity in the smooth case. This *single procedure* indeed achieves the optimal rate of convergence in both situations, without depending on the strong convexity parameter. This was not the case in the non-smooth case, where one needed to use much smaller steps (scaling as n^{-1}) to get the optimal rate in the strongly convex case.

To understand differences between quadratic functions and other smooth-strongly convex function, we briefly describe the proof technique in the following paragraph.

Proof technique for averaged iterate.

General case: In the general setting, proofs for the averaged iterate rely on an expansion of $f_k'(\theta_{k-1})$: for $k \in [\![1;n]\!]$, we have $\gamma_k(f'(\theta_{k-1}) + \varepsilon_k(\theta_{k-1})) = \theta_{k-1} - \theta_k$. Using a first order Taylor expansion of $f'(\theta_{k-1})$, as $f''(\theta_*)(\theta_{k-1} - \theta_*) + O(\|\theta_{k-1} - \theta_*\|^2$, we get (Polyak and Juditsky, 1992):

$$\gamma_k f''(\theta_*)(\theta_{k-1} - \theta_*) = \theta_{k-1} - \theta_k - \gamma_k \varepsilon_k(\theta_{k-1}) - \gamma_k O\left(\|\theta_{k-1} - \theta_*\|^2\right) .$$

Averaging over k from 1 to n then yields:

$$f''(\theta_*)(\bar{\theta}_n - \theta_*) = \frac{1}{n}\sum_{k=1}^{n} \frac{\theta_{k-1} - \theta_k}{\gamma_k} - \frac{1}{n}\sum_{k=1}^{n} \varepsilon_k(\theta_{k-1}) - \frac{1}{n}\sum_{k=1}^{n} O\left(\|\theta_{k-1} - \theta_*\|^2\right). \qquad \text{(D)}$$

[10] here, μ is the strong convexity constant at the optimum, as the problem is not globally strongly convex.

In this decomposition, the first term (on the right-hand part of the inequality) corresponds to the speed at which the initial conditions are forgotten, the second one is a variance term, and the third one a residual term (which can be upper bounded using convergence in higher order moments on the chain). In the quadratic case, as $f'(\theta_{k-1}) = f''(\theta_*)(\theta_{k-1} - \theta_*)$, the residual term is removed.

Quadratic case: In the quadratic case, the proof does not necessarily rely on Equation (D): we instead use the fact that $\bar{\theta}_n - \theta_*$ can be written as a linear transform of the initial distance $\theta_0 - \theta_*$ and of the noise sequence; we then analyze the behavior of the linear operator.

Decaying *vs.* constant learning rates. With n observations, the practitioner can follow two strategies concerning the learning rate: either use decaying steps $\gamma_k = \frac{1}{\sqrt{k}}$ for $k \in [\![1; n]\!]$, or use constant steps $\gamma_k = \frac{1}{\sqrt{n}}$ for $k \in [\![1; n]\!]$. This formally corresponds to two different regimes:

- In the *online-setting*, a.k.a. *"any-time"*, or *decaying step-size*, between steps 1 and n, we use a sequence of step sizes $(\gamma_k)_{k \in [\![1;n]\!]}$, which is a subsequence of a universal sequence $(\gamma_k)_{k \in \mathbb{N}^*}$. This is simple to use in practice: if a first estimator θ_{n_1} has been computed with n_1 observations, and n_2 additional observations are opportunely provided, a new estimator can be computed by just making n_2 additional steps starting from θ_{n_1}, with step size $(\gamma_{n_1+k})_{k \in [\![1;n_2]\!]}$. It is thus not necessary to know in advance how many iterations will be made.

- In the *finite-horizon* setting, frequently used for the purpose of analysis (Bach, 2014; Ying and Pontil, 2008), we use a constant learning rate, which often depends on the number of iterations we plan to make, assumed to be known and fixed. Formally, for a sequence $(\Gamma_n)_{n \in \mathbb{N}^*}$, n being the number of iterations, we choose $\gamma_k = \Gamma_n$, for any $k \in [\![1; n]\!]$. In such a setting, changing the horizon from n_1 to $n_1 + n_2$ implies to recompute the entire sequence with a new (constant) learning rate $\Gamma_{n_1+n_2}$, instead of Γ_{n_1}.

Using *doubling tricks* allows to pass from constant steps to varying steps (Hazan and Kale, 2011), but is not fully satisfactory as it results in the definition of "epochs", which create discontinuities in the performance.

In most situations, up to constants or logarithmic terms, the performance is the same for both regimes, when using the "same decay", *i.e.*, if $(\gamma_k)_{k \in \mathbb{N}^*} = (\Gamma_k)_{k \in \mathbb{N}^*}$: one intuition is that the convergence often mainly depends on the sum of the steps[11], and if $\gamma_k = k^{-\zeta}$ for all k, then $\sum_{k=1}^n \gamma_k$ and $n\Gamma_n$ are both of order $\Theta(n^{1-\zeta})$.

However, this heuristical argument admits at least one noticeable exception: in the smooth strongly convex case, for averaged SGD with $\gamma_n = \Gamma_n \propto n^{-\zeta}$, $1/2 < \zeta < 1$, only the online version reaches the asymptotically optimal rate $O(n^{-1})$. The bias is indeed of order $O(n^{-2}\gamma_n^{-1}\mu^{-3})$ in the online case, but of order $O(n^{-2}\Gamma_n^{-2}\mu^{-2})$ in the finite-horizon case. Technically, the difference appears in Equation (D): in the finite-horizon case, the first term is exactly $n^{-1}\Gamma_n^{-1}(\theta_0 - \theta_*)$, while in the strongly convex case, using Abel's summation formula, we have that this first term can be upper bounded by $O(n^{-1}{\gamma_n}^{-\frac{1}{2}}\mu^{-1})$.

[11] for example, in the convex case, with gradients a.s. bounded by B, we have that $f(\bar{\theta}_n) - f(\theta_*) \leqslant \frac{\|\theta_n\|^2}{n\gamma_n} + \frac{B^2 \sum_{k=1}^n \gamma_k}{n}$.

While this could tend to discredit constant learning rates, the difference vanishes without strong convexity: in Chapter 2, we prove similar rates of convergence for both situations. In Chapters 3 and 4, we only consider constant learning rates. This is summarized in the following tabular.

	Chapter 2	Chapter 3	Chapter 4
Finite horizon	✓	✓	✓
Online	✓		

1.3.8 Non-asymptotic results: stochastic approximation for least-squares regression and logistic regression

Equation (D) also emphasizes that the algorithm may behave substantially better if the third derivative of the function is null. For quadratic functions, the asymptotically dominant term in the convergence rate is of order $O(n^{-1})$ for any decaying step-size scaling as $n^{-\zeta}$, $\zeta \in]0;1[$ (Bach and Moulines, 2011). This allows to consider much larger learning rates than $n^{-1/2}$. Bach and Moulines (2013) built on Györfi and Walk (1996) to exploit this idea with a constant step γ, that does not depend on the number of observations. They showed the following result:

Theorem 1.4. *Consider the averaged least mean squares algorithm, with structured noise (such that $\mathbb{E}[\varepsilon_k(\theta_*)\varepsilon_k(\theta_*)^\top] \leqslant \sigma^2\Sigma$). Writing $r^2 = E_{\rho_X}[\|\Phi(X)\|^2]$, and using $\gamma \leqslant \frac{1}{2r^2}$, we have, for any $n \in \mathbb{N}^*$,*

$$f(\bar{\theta}_n) - f(\theta_*) \leqslant 4\frac{\sigma^2 d}{n} + 2\frac{\|\theta_0 - \theta_*\|^2}{\gamma n} \,. \tag{F}$$

Therefore, it is possible to obtain a non-asymptotic rate $O(n^{-1})$ without dependence on the strong convexity constant. This bound decomposes into a *variance term*, $\frac{\sigma^2 d}{n}$ that matches the statistical lower bound described in Section 1.1.7, and a *bias term* corresponding to the speed at which initial conditions are forgotten. This bound leads to the choice of the largest possible step size, $\gamma = \frac{1}{2r^2}$.

One can also get this fast convergence rate for logistic regression. The main idea is to use an algorithm built in two different steps: first, averaged SGD with $\gamma_n \propto n^{-1/2}$ which gives a first estimator $\tilde{\theta}_n$; then, n steps of averaged LMS with constant step-size $\frac{1}{2r^2}$ for the quadratic approximation of f around $\tilde{\theta}_n$ (Bach and Moulines, 2013).

Theorem 1.4 is the cornerstone of this thesis: it indeed allows for several extensions which are the starting points for the three next chapters.

Robustness to the lack of strong convexity. In the non-parametric setting, which will be introduced in detail in the Section 1.4, problems are typically never strongly convex if they are not regularized. The fact that the convergence rate in Equation (F) does not depend on the strong convexity constant opens the door to its analysis in infinite dimension. We address this question in Chapter 2.

Forgetting initial conditions. The speed at which initial conditions are forgotten (n^{-1}) does not impact the order of magnitude of the bound but sometimes has an important influence in practice, as it can be the leading term during the first iterations. Moreover, in

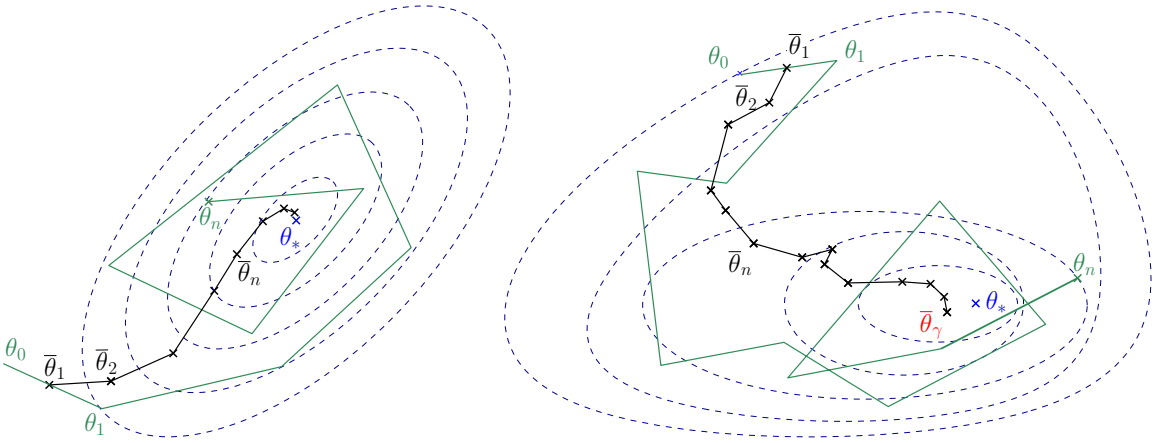


Figure 1.3: Stochastic Gradient Descent with constant learning rate. Dashed lines are the level lines of the objective function f, green points correspond to the main recursion, and black to the averaged one. *Left:* Quadratic case, the limit is the optimal point. *Right:* General case, the limit is a different point.

comparison to the rates of deterministic optimization summarized in Section 1.2, a gap appears between the optimal rate for non strongly convex problems $O(n^{-2})$ and the rate in Equation (F). We bridge this gap in Chapter 3.

Markov chain interpretation. Interestingly, SGD with constant learning rate γ is an *homogeneous Markov* chain (Meyn and Tweedie, 1993): the distribution of θ_n only depends on the distribution of θ_{n-1}, and the way this distribution evolves does not change with time. Ergodic theorems for Markov chains then show that the averaged iterate $\bar{\theta}_n$ almost surely converges to a point $\bar{\theta}_\gamma$, and that it satisfies a central limit theorem: $\sqrt{n}(\bar{\theta}_n - \bar{\theta}_\gamma) \xrightarrow{d} \mathcal{N}(0, V)$ for a certain variance matrix V. This gives a simple intuition on Theorem 1.4: $n\mathbb{E}\|\bar{\theta}_n - \bar{\theta}_\gamma\|^2$ converges to a constant, therefore $f(\bar{\theta}_n) - f(\bar{\theta}_\gamma)$ converges to zero at rate n^{-1}. Moreover, $(\theta_n)_{n\in\mathbb{N}}$ converges to a limit distribution π_γ, such that $\bar{\theta}_\gamma = \mathbb{E}_{\pi_\gamma}[\theta]$. As this limit distribution is stable (if $\theta_{n-1} \sim \pi_\gamma$ then $\theta_n \sim \pi_\gamma$), if $\theta_0 \sim \pi_\gamma$, then $\theta_1 = \theta_0 - \gamma f_1'(\theta_0) \sim \pi_\gamma$ and taking expectations on both sides thus yields

$$\mathbb{E}_{\pi_\gamma}\left[f_1'(\theta)\right] = 0.$$

In least-squares regression, as the function is quadratic, the gradients are linear functions and we get $\mathbb{E}_{\pi_\gamma}\left[f_1'(\theta)\right] = \Sigma(\bar{\theta}_\gamma - \theta_*) = 0$, *i.e.*, $\bar{\theta}_\gamma = \theta_*$ if Σ invertible. The limit of the averaged stochastic gradient is thus the optimal point, and $f(\bar{\theta}_n) - f(\theta_*) = O(n^{-1})$.

On the contrary, if the risk is not quadratic, the averaged recursion converges to a limit which is not the optimal point. This is illustrated in Figure 1.3. We develop and use the Markov chain approach in Chapter 4 to improve the convergence rate with large step sizes in the non quadratic case.

Rates of convergence for stochastic approximation are summarized in Table 1.3.

Smoothness	**Strong Convexity**	**Step size**	**Averaging**	**Upper bound on $f(\theta_n) - f(\theta_*)$**	**Reference**
Non smooth	Convex Lip.	$\gamma_n \propto n^{-1/2}$	PR[a] av.	$n^{-1/2}$	Zhang (2004)
Non smooth	Strongly Convex	$\gamma_n \propto (n\mu)^{-1}$	Tail av.	$n^{-1}\mu^{-1}$	Rakhlin et al. (2011)
Smooth	Convex	$\gamma_n \propto n^{-1/2}$	PR av.	$n^{-1/2}$	Bach and Moulines (2011)
⋮	Strongly Convex	$\gamma_n \propto n^{-1/2}$	PR av.	$n^{-1}\mu^{-2}$	
⋮	Strongly Convex	$\gamma_n \propto n^{-\zeta}, \zeta \in]0.5, 1[$	PR av.	$n^{-1} + n^{-\tau_{1,\zeta}}\mu^{-\tau_{2,\zeta}}$ [b]	
⋮	Logistic	$\gamma_n \propto n^{-1/2}$	PR av.	$n^{-1}\mu^{-1}$	Bach (2014)
⋮	Quadratic	$\gamma = 1/2R^2$	PR av.	$\sigma^2 d n^{-1}$	Bach and Moulines (2013)
Smooth	Self Concordant + kurtosis	2 epochs procedure		$O(n^{-1})$	

Table 1.3: Summary of rates for Stochastic Approximation.

[a] Polyak Ruppert averaging, *i.e.*, uniform averaging.

[b] $\tau_{1,\zeta} > 1$, $\tau_{2,\zeta} > 2\tau_{1,\zeta} - 1 > 1$.

1.4 Non-parametric regression in reproducing kernel Hilbert spaces

Parametric models often only provide an imprecise approximation of the underlying statistical structure: relationships between inputs and outputs are rarely linear (or even linear functions of features), and the approximation error (the excess risk of the best linear predictor) remains non negligible. Looking for an estimator in an infinite-dimensional space can substantially reduce the approximation error. In this section, we introduce the framework of non-parametric regression. Non-parametric statistical estimation dates back to the 50s and goes beyond regression, especially with density estimation (Rosenblatt, 1956; Parzen, 1962); a nice introduction is provided by Tsybakov (2008).

Analyzing non-parametric regression is paramount both for applications that use infinite-dimensional feature spaces, and to gain intuition on the behavior of algorithms in large dimension $d \gg n$ (see Section 1.4.4).

Random design non-parametric regression

In non-parametric regression, we consider the regression problem presented in Section 1.1, but allow the estimator to live in a broader class $\mathcal{F}$ of functions from $\mathcal{X}$ to $\mathcal{Y}$ than in the parametric setting. We aim to find a predictor solving the following problem:

$$\inf_{f \in \mathcal{F}} R(f).$$

For the square loss, this is called *non-parametric least-squares*:

$$\inf_{f \in \mathcal{F}} \frac{1}{2} \mathbb{E}_{(X,Y)\sim\rho} \left[(f(X) - Y)^2 \right].$$

Recall that if the marginal distribution ρ_Y has a second order moment, the Bayes predictor has a simple expression: $f_\rho(X) = \mathbb{E}[Y|X]$. We still do not assume that this predictor belongs to the class $\mathcal{F}$. However, as this class is much larger than a class of linear functions, the approximation error $\inf_{f\in\mathcal{F}} R(f) - R(f_\rho)$ can be much smaller than in the parametric case, and is even always 0 if the class $\mathcal{F}$ is large enough. Moreover, the excess risk can be written as the $L^2_{\rho_X}$-norm of the difference

$$R(f) - R(f_\rho) = \frac{1}{2} \mathbb{E}_{X\sim\rho_X} \left[(f(X) - f_\rho(X))^2 \right] = \|f - f_\rho\|^2_{L^2_{\rho_X}}, \tag{1.15}$$

where $L^2_{\rho_X}$ is the space of squared integrable functions with respect to ρ_X.

Non-parametric statistics: minimax rates

Under a well-defined statistical model, assuming that $y_k = f_*(x_k) + \varepsilon_k$, with $f_* \in \mathcal{F}$, and $(\varepsilon_k)_{k\in[\![1;n]\!]}$ i.i.d. Gaussian variables, it is possible to derive upper and lower bounds for least-squares regression.

At first sight, considering the minimax bound $\Theta\left(\frac{\sigma^2 d}{n}\right)$ in finite dimension, it might seem pointless to expect anything in the non-parametric regime, with $d = \infty$. This difficulty is circumvented by making some extra assumptions on the class of distributions followed by the observations. Recalling Section 1.1.7, the minimax risk is here defined as:

$$\inf_{\hat{f}} \sup_{f_*\in\mathcal{F}} \mathbb{E}_{P_{f_*}} \left[\left\| \hat{f} - f_* \right\|^2_{L^2_{\rho_X}} \right] \tag{1.16}$$

for a non-parametric class $\mathcal{F}$ being an infinite-dimensional space, and the associated non-parametric model $\{P_f, f \in \mathcal{F}\}$. For example, $\mathcal{F}$ can be a Sobolev or Hölder class (Tsybakov, 2008), and the family of distributions a Gaussian model, *i.e.*, for some $f_* \in \mathcal{F}$,

$$P_{f_*}(Y|X) \sim \mathcal{N}(f_*(X), \sigma^2).$$

For $\delta \in \mathbb{N}^*$, the Sobolev space $\mathbb{W}^\delta[0;1]$ of order δ is defined as the class of real valued functions on $[0;1]$, δ-times differentiable, such that $f^{(\delta)}$ Lebesgue-integrable, $f^{(\delta-1)}$ is absolutely continuous, and $f(0) = \cdots = f^{(\delta-1)}(0) = 0$. This space is embedded with the inner product $\langle f, g\rangle = \int_0^1 f^{(\delta)}(x)g^{(\delta)}(x)\mathrm{d}x$. Minimax rates for Sobolev classes (and Hölder classes) with Gaussian noise were proved by Ibragimov and Has' Minskii (1979, 1982), and Stone (1982). Upper convergence bounds were also proved by Nemirovski et al. (1983, 1984, 1985). These lines of work show that under suitable assumptions on the model, if the class of functions containing the Bayes predictor is known, then the minimizer of the empirical risk over that set has optimal error rate in $L^2_{\rho_X}$ norm. For Sobolev spaces of order δ, this minimax rate is $n^{\frac{-2\delta}{2\delta+1}}$.

In practice however, the class in which the regression function lives is unknown (and the noise is not necessarily Gaussian). The choice of the hypothesis space and learning algorithm are left to the learner, with the following constraints: the hypothesis space should be large enough for the approximation error to vanish, the estimator should be computable, and the combination of the hypothesis space and learning rule should avoid over-fitting.

Example of estimators

Several learning rules have been proposed; among others, noticeable examples include

- *Local regressors*, that estimate the unknown value at a point x by a mean of the observed values at points $x_k, k \in [\![1;n]\!]$ which are "close" to x in a sense. They include Nadaraya-Watson estimator (Nadaraya, 1964; Watson, 1964), locally polynomial estimators (Stone, 1977) and nearest neighbor regression (Altman, 1992).
- *Shape constrained regressors*, for which the hypothesis space is defined by some criteria — typically monotonic (Brunk, 1955, 1970) and/or convex functions (Hildreth, 1954; Hanson and Pledger, 1976).
- *Estimators in reproducing kernel Hilbert spaces* (Wahba, 1990; Schölkopf and Smola, 2002), which will be the setting for Chapters 2 and 3, and thus the main focus of the rest of this Section.

Overall, non-parametric regression has been an extremely active topic over the last 60 years; the books by Györfi et al. (2002) and Wasserman (2006) give a complete overview and contain most necessary references. In this thesis, we consider estimators built in a *reproducing kernel Hilbert space* (RKHS): we introduce this spaces in Section 1.4.1, see that they have good statistical properties (Section 1.4.3), and computational properties (Section 1.4.5). We describe a few settings in which RKHSs are useful in Section 1.4.2, and touch on how it is insightful for parametric regression in Section 1.4.4.

1.4.1 Reproducing kernel Hilbert spaces

Hilbert spaces of functions (complete[12] linear spaces with an inner product) are well suited as hypothesis spaces for regression: as they can have infinite dimension, they can include a sufficiently large class of functions (for example, to have small approximation error), and they enjoy a geometric structure similar to ordinary Euclidean spaces (for example, one can define projections). A particular class of such function-based Hilbert spaces are those defined with respect to a reproducing kernel; these spaces are known as *reproducing kernel Hilbert spaces*.

There are several definitions and points of view on reproducing kernel Hilbert spaces. In particular, they can be defined together with a reproducing kernel (Definition 1.5), more abstractly without even introducing any kernel function (Definition 1.6), or on the contrary, *implicitly* (without describing the space itself) with a positive definite kernel (Definition 1.7 and Theorem 1.8).

Definition 1.5. *Consider a set $\mathcal{X}$ and $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$ a Hilbert space of real valued functions on $\mathcal{X}$, with inner product $\langle , \rangle_{\mathcal{H}}$. The function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a reproducing kernel of $\mathcal{H}$ if:*

- *For any $x \in \mathcal{X}$, $\mathcal{H}$ contains the function K_x, defined by:*

$$\begin{aligned} K_x : \mathcal{X} &\to \mathbb{R} \\ y &\mapsto K(x, y). \end{aligned}$$

- *For any $x \in \mathcal{X}$ and $f \in \mathcal{H}$, the* reproducing property *holds:*

$$\langle K_x, f \rangle_{\mathcal{H}} = f(x). \tag{1.17}$$

If a reproducing kernel exists, then $\mathcal{H}$ is called a reproducing kernel Hilbert space (RKHS) (associated with K).

The reproducing property (1.17) allows to compute inner products in the Hilbert space as function evaluations. Moreover, one can show that the reproducing kernel of an RKHS is unique, and that distinct RKHS have distinct reproducing kernels.

From the point of view of Riesz's representation theorem, the reproducing property states that K_x is the representer of the evaluation functional $L_x : \mathcal{H} \to \mathbb{R}$ which maps f to $f(x)$. If $\mathcal{H}$ is an RKHS, this application is clearly continuous (by Cauchy-Schwarz inequality $|f(x)| = |\langle K_x, f \rangle|_{\mathcal{H}} \leqslant \|K_x\|_{\mathcal{H}} \|f\|_{\mathcal{H}}$). Conversely, an equivalent definition of RKHS holds:

Definition 1.6. *An RKHS is a Hilbert space of real valued functions on $\mathcal{X}$ such that for each $x \in X$, the evaluation functional L_x is continuous.*

In other words, if all evaluation functionals are continuous, the space is an RKHS and there exists a reproducing kernel associated to the space. Interestingly, this definition does not require the introduction of the kernel function K. For example, Definition 1.6 implies that the Sobolev space $\mathbb{W}^{\delta}[0; 1]$ is a reproducing kernel Hilbert space[13] (Wahba, 1990). However, the general expression of the kernel function for Sobolev spaces is complicated (though for $\delta = 1$, the kernel is simply $K(x, y) = \min(x, y)$, for $x, y \in [0; 1]^2$).

While these first two definitions give a central role to the set $\mathcal{H}$, RKHS can also be defined implicitly. To do so, we introduce positive semi-definite kernels.

[12] w.r.t. the norm defined by their inner product.

[13] this can be seen using a Taylor expansion and Cauchy-Schwarz inequality.

Definition 1.7. *A* positive semi-definite (PSD) kernel K *is a symmetric function* $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$*, such that, for any* $p \in \mathbb{N}^*$ *and any* $(x_i)_{i \in [\![1;p]\!]} \in \mathcal{X}^p$*, the corresponding* kernel matrix $\boldsymbol{K} \in \mathbb{R}^{p \times p}$ *(such that, for all* $i, j \in [\![1;p]\!]^2$*,* $\boldsymbol{K}_{i,j} = K(x_i, x_j)$*), is positive semi-definite.*

For a reproducing kernel K, Equation (1.17) implies that

$$K(x_1, x_2) = K_{x_1}(x_2) = \langle K_{x_1}, K_{x_2} \rangle_{\mathcal{H}} .$$

As a consequence, a reproducing kernel is always a PSD kernel, as for any $a \in \mathbb{R}^p$ and any $(x_i)_{i \in [\![1;p]\!]} \in \mathcal{X}^p$,

$$\sum_{i,j=1}^{p} a_i a_j K(x_i, x_j) = \left\langle \sum_{i=1}^{p} a_i K_{x_i}, \sum_{j=1}^{p} a_j K_{x_j} \right\rangle_{\mathcal{H}} .$$

This simple property has an "extended converse": the following theorem states that for any PSD kernel, there exists an RKHS such that the kernel is its reproducing kernel. This result is known as Moore-Aronszajn theorem (Aronszajn, 1950):

Theorem 1.8. *For any PSD kernel* K*, there exists a unique reproducing kernel Hilbert space* $(\mathcal{H}, \langle , \rangle_{\mathcal{H}})$ *with reproducing kernel* K*.*

Here, the RKHS and its inner product are not explicitly described, and may be difficult to determine, but is fortunately un-necessary: using the reproducing property, inner products can be computed ignoring the precise definitions of both the space and the inner product.

These different definitions correspond to two trends in understanding (and using) kernels. In the "first trend", the focus is on the class of function $\mathcal{H}$, and the kernel satisfying the reproducing property serves as a tool. In the "second trend", the kernel K is the main focus of attention, together with its possible applications, while little interest is given to the space $\mathcal{H}$ itself. Historically, these two trends both emerged at the beginning of the 20^{th} century. The "first trend" results from the introduction of reproducing kernels by Zaremba (1907, 1909) for certain classes of functions. It was later on extended by Bergmann (1922) who considered classes of harmonic and analytical functions, describing the associated reproducing kernel. The second approach, for which the kernel function is central and space $\mathcal{H}$ secondary, takes its roots in the study of positive definite kernels. Those were introduced by Mercer (1909), building on the works of Hilbert (1904). Theorem 1.8 was stated in the general case by Aronszajn (1950), who attributes it to Moore (1916, 1935). The seminal paper by Aronszajn (1950) remains a most interesting reference, containing a detailed presentation of the results above, together with precise historical references.

The success of RKHSs for non-parametric regression is due to the combination of the following three facts:

- they are spaces of functions, thus naturally used as hypothesis spaces in which to build predictors,
- they have a Hilbert structure, *i.e.*, a linear space, for which numerous methods have already been designed
- finally, it is possible to compute inner products even though the space is infinite-dimensional (one can therefore treat non-parametric estimation in the same algebraic framework as parametric regression). This last property is the most useful in practice.

In other words, any algorithm designed for finite-dimensional vectors, and that can be expressed only in terms of pairwise inner products, can be applied to infinite-dimensional vectors in the feature space of a PSD kernel. Each inner product evaluation is then replaced by a kernel evaluation.

Let us emphasize that kernel methods allow us to separate the representation of the inputs (mapping points into an RKHS) from the algorithmic aspect and its analysis (proposing a generic algorithm in an RKHS and analyzing it). We now give a few examples, first of possible kernels, then of typical application settings.

1.4.2 Examples

PSD kernels. Numerous examples of positive semi-definite kernel exist. If the space $\mathcal{X}$ is $\mathbb{R}^d$, they include:

- The *Linear kernel*, and *polynomial kernels:* respectively $K(x,y) = \langle x,y\rangle_{\mathbb{R}^d}$ and $K(x,y) = \langle x,y\rangle_{\mathbb{R}^d}^m$ for some $m \in \mathbb{N}^*$. They lead to finite-dimensional reproducing kernel Hilbert spaces (with dimension $D \geqslant d$).
- The *Gaussian kernel*: $K(x,y) = \exp(-\frac{1}{2\sigma^2}\|x-y\|^2)$. It leads to an infinite-dimensional reproducing kernel Hilbert space. More generally, *radial basis function (RBF) kernels* are kernels that can be written as $K(x,y) = h(\text{dist}(x,y))$ for a function h and a distance dist.

Note that one of the key advantages of PSD kernels is that they can be defined on *non-vectorial data*: for example, on text (Lodhi et al., 2002), on (biological) sequences (Jaakkola et al., 1999; Leslie et al., 2002), on images (Harchaoui and Bach, 2007), on graphs (Borgwardt and Kriegel, 2005), or on measures (Cuturi et al., 2005), among many others.

Possible settings. The analysis we propose applies to several contexts, corresponding to different types of applications.

i) Learning with non-vectorial data. A positive definite kernel can be used to map non-vectorial inputs into a linear space. Indeed, the *feature map* $x \mapsto K_x$ associates to any input x a *feature vector*, which is an element of an RKHS. In such a situation, the kernel function is sometimes imposed by the setting, as for some particular applications, only few kernels have been designed. Here, the space $\mathcal{H}$ is not described, in the spirit of the "second trend" described above.

ii) Non-parametric regression for real valued inputs. Assume that $\mathcal{X} \subset \mathbb{R}$, and that we look for a predictor f in a set of functions having some regularity (typically, a Sobolev space of order δ, where δ can be chosen *arbitrarily*). These applications adopt the point of view of the "first trend", the primary focus being the choice of the space, which is often left to the user.

iii) Linear regression for $d \gg n$. Using a linear kernel in dimension d, parametric regression becomes a special case of RKHS regression. Results proved in the RKHS setting thus apply in finite dimension. This is useful in many applications for which linear estimation is used, but the number of features d is much larger than n. Indeed, the analysis in the

RKHS is inherently designed to address infinite-dimensional spaces: it relies on bounds and guarantees that do not depend on the dimension (these quantities are referred to as being "dimensionless"). As a consequence, it therefore provides some natural insight on the finite "large-dimensional" setting.

We now give a precise description of least-squares regression in the kernel setting.

1.4.3 Least-squares regression in RKHS

In a reproducing kernel Hilbert space, least-squares regression minimization is $\inf_{f\in\mathcal{H}} R(f)$, where $R(f)$ can be written using the reproducing property:

$$R(f) = \frac{1}{2}\mathbb{E}_{(X,Y)\sim\rho}\left[(\langle K_X, f\rangle - Y)^2\right].$$

Similarly, the empirical risk minimization is then written $\inf_{f\in\mathcal{H}} R_n(f)$, with

$$R_n(f) = \frac{1}{2n}\sum_{k=1}^{n}(\langle K_{x_k}, f\rangle - y_k)^2.$$

Tikhonov regularization

The most popular regularization is Tikhonov regularization (a.k.a. ridge regression), where we add the following penalty term:

$$\inf_{f\in\mathcal{H}} \frac{1}{2n}\sum_{k=1}^{n}(\langle K_{x_k}, f\rangle - y_k)^2 + \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2.$$

The *representer theorem* (Kimeldorf and Wahba, 1970) states that there exists a function minimizing the (penalized) empirical risk over $\mathcal{H}$, and that this minimum can be chosen in the space $\mathcal{H}_{1,n} := \text{span}\{K_{x_k}, k \in [\![1;n]\!]\}$. Using the Hilbertian structure, the space $\mathcal{H}$ can indeed be decomposed as the orthogonal sum $\mathcal{H}_{1,n} \overset{\perp}{\oplus} \mathcal{H}_{1,n}^{\perp}$. Note that, using the reproducing property, for any function $f^{\perp} \in \mathcal{H}_{1,n}^{\perp}$, for any $k \in [\![1;n]\!]$, $f^{\perp}(x_k) = \left\langle f^{\perp}, K_{x_k}\right\rangle_{\mathcal{H}} = 0$. Therefore, a function in the orthogonal space does not change predicted values on the observations. The empirical risk of a function is thus the same as the empirical risk of its projection on $\mathcal{H}_{1,n}$; and the infimum over $\mathcal{H}$ is thus larger than (thus equal to) the minimum over $\mathcal{H}_{1,n}$[14]. Moreover, when a penalization $\lambda\|f\|_{\mathcal{H}}^2$, for $\lambda > 0$ is added, the penalized risk of a function becomes larger than the one of its projection (strictly larger if the function is not in $\mathcal{H}_{1,n}$, as $\|f\|_{\mathcal{H}}^2 = \|p_{\mathcal{H}_{1,n}}(f)\|_{\mathcal{H}}^2 + \|f^{\perp}\|_{\mathcal{H}}^2$). Thus, there exists a unique minimizer to the penalized empirical risk, and this minimizer lies in $\mathcal{H}_{1,n}$.

In other words, the problem can be rewritten as an equivalent finite-dimensional minimization problem:

$$\min_{\alpha\in\mathbb{R}^n} \frac{1}{2n}\sum_{k=1}^{n}\|\boldsymbol{K}\alpha - \boldsymbol{Y}\|^2 + \frac{\lambda}{2}\alpha^{\top}\boldsymbol{K}\alpha, \tag{E'}$$

for $f = \sum_{k=1}^{n}\alpha_k K_{x_k}$, with $\boldsymbol{K}$ the kernel matrix and $\boldsymbol{Y} = (y_k)_{k\in[\![1;n]\!]}$. If $\lambda > 0$, the problem (E') has a unique solution $\hat{\alpha} = (\boldsymbol{K} + n\lambda I)^{-1}\boldsymbol{Y}$. This estimator can be explicitly computed,

[14] the infimum on $\mathcal{H}_{1,n}$ is attained as it is a quadratic function over a finite-dimensional space.

but this cost can be prohibitive, this will be discussed in Section 1.4.5. We now turn to the analysis of the risk of such an estimator.

In the end of this Section, we assume that the approximation error is null, *i.e.*, $\inf_{f\in\mathcal{H}} R(f) = R(f_\rho)$. This is true for any $f_\rho \in \mathcal{H}$ if $\mathcal{H}$ is dense in $L^2_{\rho_X}$ (with respect to the $L^2_{\rho_X}$-norm). We discuss this property in Chapter 2. Recalling the risk decomposition (1.1), the excess risk is then only the estimation error $R(\hat{f}) - R(f_\rho)$.

We recall that in the fixed design setting (see Section 1.1.6), the input points $(x_k)_{k\in[\![1;n]\!]}$ are considered as fixed, and the only randomness is in the distribution of $(y_k)_{k\in[\![1;n]\!]}$. We now give the error decomposition in the fixed design setting.

Fixed design error analysis. In the fixed design setting, the excess risk is

$$R(\hat{f}) - R(f_\rho) = \frac{1}{2n}\sum_{k=1}^{n} (\hat{f}(x_k) - E[y_k])^2 = \frac{1}{2n}\|\boldsymbol{K}(\boldsymbol{K}+n\lambda I)^{-1}\boldsymbol{Y} - \mathbb{E}[\boldsymbol{Y}]\|^2,$$

since the prediction vector $(\hat{f}(x_k))_{k\in[\![1;n]\!]}$ is equal to $\boldsymbol{K}(\boldsymbol{K}+n\lambda I)^{-1}\boldsymbol{Y}$. Following classical literature (Gu, 2002; Wahba, 1990), it leads to the following in-sample decomposition error:

$$\begin{aligned} R(\hat{f}) - R(f_\rho) &= \frac{1}{2n}\left\|\boldsymbol{K}(\boldsymbol{K}+n\lambda I)^{-1}\boldsymbol{Y} - \mathbb{E}[\boldsymbol{Y}]\right\|^2 \\ \mathbb{E}\left[R(\hat{f}) - R(f_\rho)\right] &= n\lambda^2 \left\|(\boldsymbol{K}+n\lambda I)^{-1}\,\mathbb{E}[\boldsymbol{Y}]\right\|^2 + \frac{1}{2n}\mathbb{E}\left\|\boldsymbol{K}(\boldsymbol{K}+n\lambda I)^{-1}\boldsymbol{\varepsilon}\right\|^2 \\ &= n\lambda^2\,\mathbb{E}[\boldsymbol{Y}]^\top(\boldsymbol{K}+n\lambda I)^{-2}\,\mathbb{E}[\boldsymbol{Y}] + \frac{1}{2n}\operatorname{tr}(\boldsymbol{K}^2(\boldsymbol{K}+n\lambda I)^{-2}C)\,, \end{aligned} \quad (1.18)$$

with $\boldsymbol{\varepsilon} = \boldsymbol{Y} - \mathbb{E}[\boldsymbol{Y}]$ and $C = \mathbb{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top]$. The two terms in Equation (1.18) are respectively a bias and a variance term: the bias increases with λ, while the variance decreases: intuitively, a non 0 regularization induces a bias, but reduces the variance. This leads to an optimal choice of λ if we can estimate the quantities involved.

The variance strongly depends on the eigenvalue decay of the kernel matrix $\boldsymbol{K}$: we denote[15] $(\hat{\mu}_i)_{i\in[\![1;n]\!]}$ the eigenvalues of the renormalized kernel matrix $n^{-1}\boldsymbol{K}$. The variance depends on the quantity $\operatorname{tr}(\boldsymbol{K}^2(\boldsymbol{K}+n\lambda I)^{-2}) = \sum_{i=1}^n \left(\frac{\hat{\mu}_i}{\hat{\mu}_i+\lambda}\right)^2$, which is known as the *degrees of freedom* (Hastie et al., 2001), and plays the role of an implicit dimension which depends on the spectrum of $\boldsymbol{K}$. If the RKHS has finite dimension D, then it is smaller than D, but it is still finite for most infinite-dimensional kernel spaces, since the eigenvalues of $\boldsymbol{K}$ are summable under weak assumptions (see Chapter 2 for details, and Figure 1.5 for two examples). As for the bias term, it depends on both the spectrum and the decomposition of the vector $\mathbb{E}[\boldsymbol{Y}]$ on the eigenvectors of $\boldsymbol{K}$.

While the random design analysis is slightly more complicated, it involves similar quantities, and relies on two fundamental assumptions that we present hereafter.

Random design error analysis. We introduce the covariance operator $T : \mathcal{H} \to \mathcal{H}$, such that for any $f \in \mathcal{H}$, $T(f) = \mathbb{E}_{\rho_X}[f(X)K_X]$[16]. Informally, this is an extension of the finite-dimensional covariance matrix to the non-parametric setting. This operator is crucial as

[15] we use the notation $\hat{\mu}$ because the eigenvalues of $n^{-1}\boldsymbol{K}$ are also the eigenvalues of the empirical covariance matrix $n^{-1}\sum_{k=1}^n K_{x_k}K_{x_k}^\top$ which is the empirical version of the covariance operator T defined below, whose eigenvalues are denoted $(\mu_i)_{i\in\mathbb{N}}$.

[16] Descriptions of how such expectations are defined will be given in Chapter 2.

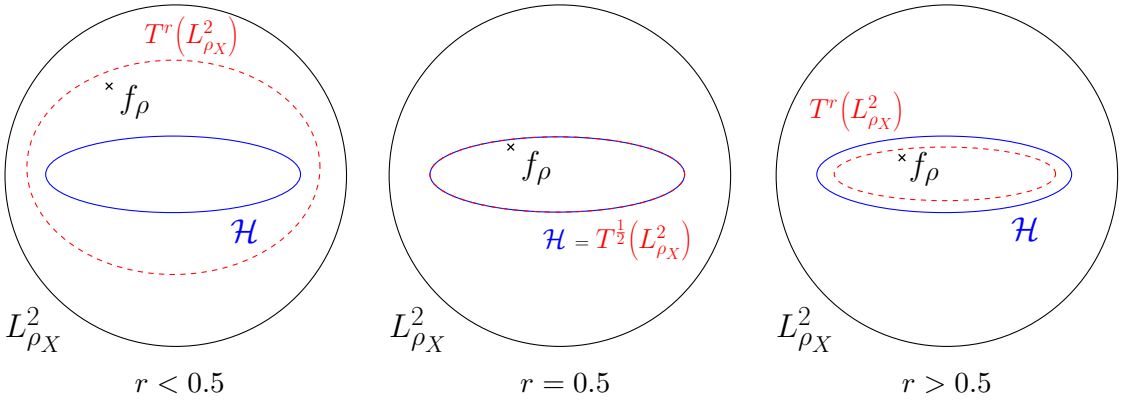


Figure 1.4: Schematic representation of the source condition **H**7: for different f_ρ, there exists an $r > 0$ such that **H**7 is satisfied.

it "identifies" the space in which we work: indeed, elements of the reproducing kernel space $\mathcal{H}$ can be decomposed as combinations of the eigenvectors of T, with conditions on the coefficients characterized by the eigenvalues of T; this is the purpose of Mercer's theorem (Aronszajn, 1950). Formally, we have the following result:

Theorem 1.9 (Mercer's theorem)**.** *Assume $\mathcal{X}$ compact, K continuous. Then the covariance operator T has a family of eigenvectors $\{\phi_i, i \in \mathbb{N}\}$, forming an an Hilbertian basis of $\mathcal{H}$, with associated eigenvalues $(\mu_i)_{i\in\mathbb{N}}$. Moreover the feature functions can be decomposed on the eigen-basis: for any $x \in \mathcal{X}$,*

$$K_x = \sum_{i=0}^{\infty} \mu_i \phi_i(x) \phi_i \ ,$$

where the convergence is absolute[17]*. Elements of the space $\mathcal{H}$ can then be decomposed over eigenvalues:*

$$\mathcal{H} = \left\{ f \in L^2_{\rho_X} : f = \sum_{i=0}^{\infty} a_i \phi_i, \ \textit{s.t.}, \ \sum_{i=0}^{\infty} \frac{a_i^2}{\mu_i} < \infty \right\}. \tag{1.19}$$

In Chapter 2 we propose an extension of this result under weaker assumptions, in particular *without* topological assumption on $\mathcal{X}$ (see Section A.1.3).

Decomposing any function on the eigenbasis $\{\phi_i, i \in \mathbb{N}\}$, we can define T^r, the r-th power of T, for any $r \in [0; 1]$. These operators are necessary to define the two following assumptions, which describe respectively the smoothness of the function and the size of the kernel, and strongly influence the performance of random design kernel regression:

H7 (Source Condition)**.** $f_\rho \in T^r\left(L^2_{\rho_X}\right)$ *for some* $r \geqslant 0$.

H8 (Capacity Condition)**.** *We sort the sequence $(\mu_i)_{i\in I}$ of non-zero eigenvalues of the operator T in decreasing order. We assume that $\mu_i \leqslant \frac{s^2}{i^\alpha}$ for some $\alpha > 1$ (so that $\mathrm{tr}(T) < \infty$), and some $s \in \mathbb{R}_+$.*

The first condition **H**7 should be understood as a regularity condition. The sequence of spaces $T^r\left(L^2_{\rho_X}\right)$ is a decreasing sequence of subspaces of $L^2_{\rho_X}$; for $r \geqslant \frac{1}{2}$, it implies

[17] more precisely, the convergence is in the sense of Bochner, which generalizes the absolute convergence for Banach spaces.

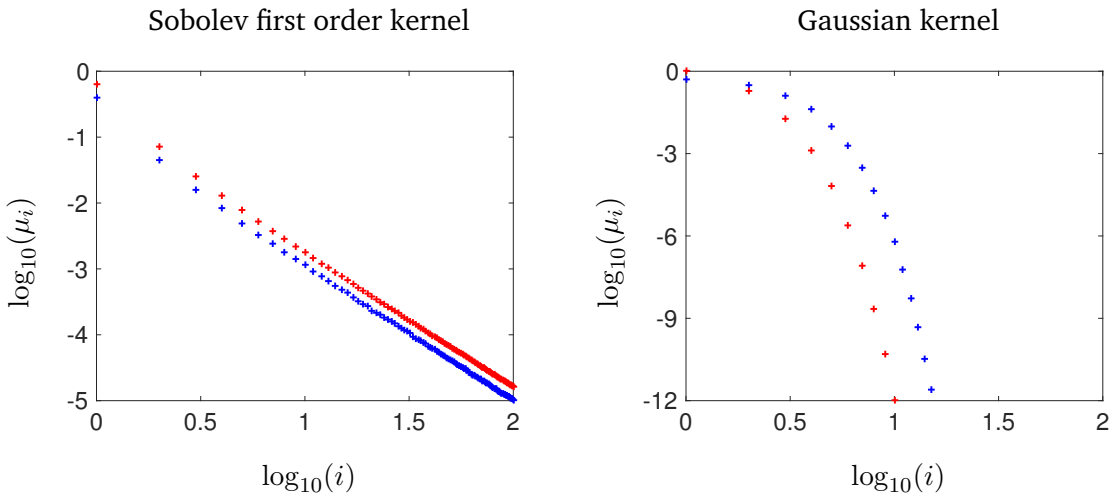


Figure 1.5: Eigenvalue decay of the empirical covariance operator (blue) (resp. population covariance operator (red)). *Left:* $\min$ kernel with $\rho_X = \mathcal{U}[0;1]$, inducing the first order Sobolev space. *Right:* Gaussian kernel with $\rho_X = \mathcal{U}[-1;1]$. The capacity condition is satisfied for any $\alpha \leqslant 2$ for the $\min$ kernel, and for any α for the Gaussian kernel.

that the objective function f_ρ truly lies in $\mathcal{H}$. A corollary of Theorem 1.9 is indeed that $T^{1/2}\left(L^2_{\rho_X}\right) = \mathcal{H}$. As r gets bigger, the assumption gets stronger. This condition is illustrated in Figure 1.4.

On the other hand, the capacity condition **H**8 describes the size of the kernel space. Again, considering Mercer's theorem (in particular Equation (1.19)), we see that for a fixed eigenbasis, the faster the eigenvalue decay is, the smaller $\mathcal{H}$ is. It is valid in practice for classical kernels, see Figure 1.5.

These assumptions are discussed in detail in Chapter 2, and have been used under multiple variants in the literature. Let us now summarize some of the important contributions to their study.

With Tikhonov regularization. Starting with the work of Smale and Cucker (2001), De Vito et al. (2005) proposed an approach of Tikhonov regularization under source conditions. Refined bounds for Tikhonov regularization were then successively proved by Smale and Zhou (2007). Rates with both source conditions and capacity conditions were proved by Zhang (2004), and optimal rates under both conditions provided by Caponnetto and De Vito (2007); Steinwart et al. (2009). Minimax rates under capacity conditions were also proved by (Raskutti et al., 2014).

With other regularizations. Using Tikhonov's regularization is not always necessary, and other algorithms have also been studied. Notably, analysis of batch gradient descent for ERM (where early stopping is used as a regularization) were provided by Yao et al. (2007) and Rosasco et al. (2014) under source conditions, Blanchard and Krämer (2010) for the conjugate gradient algorithm, and Raskutti et al. (2014). Raskutti et al. (2014) provided statistical minimax rates depending on the capacity condition parameter (under a Gaussian noise model, and valid for any algorithm), together with a data-dependent stopping rule for gradient descent. Interestingly, most classical regularizations (Tikhonov, Landweber

(early stopping GD), spectral cut-off) can be analyzed under a single framework (Bauer et al., 2007; Caponnetto and Yao, 2006).

The analysis of online methods in this context, in particular SGD (which naturally regularizes, as explained in Section 1.3), was initiated by Smale and Yao (2006), and then refined by Ying and Pontil (2008) and Tarrès and Yao (2014). Detailed description of these approaches is given in Chapter 2, where we consider stochastic approximation algorithm for least-squares regression in an RKHS. A refined analysis under both capacity condition and source condition was recently proposed by Lin and Rosasco (2016), under a unifying approach analyzing all mini-batches sizes, notably recovering both multiple passes SGD and early stopping as special cases.

Overall, the optimal rate, attained by some of the above papers and in Chapter 2, is of order $O\left(n^{\frac{-2r\alpha}{2r\alpha+1}}\right)$.

1.4.4 Consequences in finite dimension

Our analysis offers interesting insight to the analysis of the finite-dimensional setting. "*Conditions*" in infinite dimension translate to "*quantities*" in finite dimension: in infinite-dimension, conditions are either *satisfied* or *not satisfied*, thus their importance is clear; in finite dimension, the meaningful quantities are only "small" or "large". This contributes to blurring the understanding of their importance.

In other words, in finite dimension, the rate $\frac{\sigma^2 d}{n}$ is optimal, but this optimality corresponds to the worst case with respect to the distribution ρ_X. For any *fixed* ρ_X, and a fixed n, the ratio $\frac{\sigma^2 d}{n}$ might be a very pessimistic bound on the error we would like to reach.

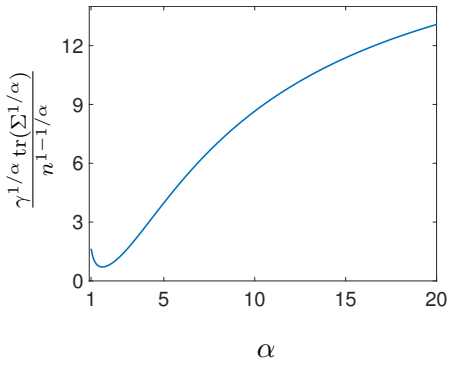


Figure 1.6: Upper bound on the variance term as a function of α. $d = 2000$, $n = 100$, eigenvalues decaying as $(i^{-2})_{i\in[\![1;d]\!]}$.

For the un-regularized ERM, the estimation error is *equal* to $\frac{\sigma^2 d}{n}$ independently of the distribution ρ_X (see the analysis of fixed design in 1.1.6). Thus if $d \gg n$, this error is "large" . Still, other algorithms may behave "well"; in particular, algorithms based on stochastic approximation can have a much smaller error than $\frac{\sigma^2 d}{n}$. In Chapter 2, we will prove that for averaged SGD with constant learning rate γ, in finite dimension, the variance term is upper bounded by $\frac{\gamma^{1/\alpha}\operatorname{tr}(\Sigma^{1/\alpha})}{n^{1-1/\alpha}}$, for any $\alpha \geqslant 1$. When $\alpha \to \infty$, we recover the bound $\frac{\sigma^2 d}{n}$. Considering the worst case over any matrix Σ, the best bound is $\frac{\sigma^2 d}{n}$; asymptotically, the best bound is $\frac{\sigma^2 d}{n}$; but for a fixed Σ, and n, this quantity can be much smaller than $\frac{\sigma^2 d}{n}$; the infimum in α of the bound may be achieved for a finite "small" α. We illustrate this phenomenon by plotting the bound as a function of the parameter α in Figure 1.6: the minimum quantity is obtained for an $\alpha \ll \infty$.

1.4.5 Computations in RKHS

Surprisingly, even though the space is infinite-dimensional, most estimators (based on Tikhonov regularization, batch gradient descent, or stochastic gradient descent) can be computed exactly: indeed, associated estimators can be decomposed as combinations of

the feature vectors $\{K_{x_k}, k \in [\![1; n]\!]\}$, and we can compute inner products of such functions as function evaluations using the reproducing property, a property known as the *Kernel trick*.

The algorithms based on Tikhonov regularization have complexity $O(n^3)$ to compute the empirical risk minimizer. Solving the linear system (Equation (E')) can be done either by using Cholesky decomposition or conjugate gradient, both with this complexity (Golub and Van Loan, 1996). Batch gradient methods have complexity $O(n^2)$ per iteration, resulting in a final complexity $O(Tn^2)$ after T iterations, and stochastic gradient methods have complexity $O(n)$ per iteration, resulting in a final complexity $O(n^2)$ after one single pass on observations.

It is interesting to note that dimension reduction techniques can be used to reduce these complexities: roughly speaking, there exists an implicit dimension d_n such that a projection on a space of dimension d_n still allows to get the optimal statistical rate. To do so, the two main methods are column sampling (Bach, 2012; El Alaoui and Mahoney, 2014; Lin and Rosasco, 2016), and random features (Rahimi and Recht, 2007, 2008; Rudi et al., 2016). Together with stochastic methods, these methods allow to derive optimal statistical rates for algorithms of overall complexity $O(nd_n)$.

2

Non-parametric Stochastic Approximation with Large Step-sizes

We consider the random-design least-squares regression problem within the reproducing kernel Hilbert space (RKHS) framework. Given a stream of independent and identically distributed input/output data, we aim to learn a regression function within an RKHS $\mathcal{H}$, even if the optimal predictor (*i.e.*, the conditional expectation) is not in $\mathcal{H}$. In a stochastic approximation framework where the estimator is updated after each observation, we show that the averaged unregularized least-mean-square algorithm (a form of stochastic gradient descent), given a sufficient large step-size, attains optimal rates of convergence for a variety of regimes for the smoothnesses of the optimal prediction function and the functions in $\mathcal{H}$.

This chapter is based on our work *Non-parametric Stochastic Approximation with Large Step-size*, A. Dieuleveut and F.Bach, published in the Annals of Statistics, 2016.

Contents

2.1 Introduction 41
2.2 Learning with positive-definite kernels 43
2.2.1 Reproducing kernel Hilbert spaces 43
2.2.2 Random variables 43
2.2.3 Minimization problem 44
2.2.4 Covariance operator 45
2.2.5 Minimal assumptions 47
2.2.6 Examples 48
2.2.7 Convergence rates 49
2.3 Stochastic approximation in Hilbert spaces 50
2.3.1 Regularization and linear systems 51
2.3.2 Stochastic approximation 52
2.3.3 Extra regularity assumptions 53
2.3.4 Main results (finite horizon) 53
2.3.5 Online setting 56
2.4 Links with existing results 57
2.4.1 Euclidean spaces 57
2.4.2 Optimal rates of estimation 59
2.4.3 Regularized stochastic approximation 59
2.4.4 Unregularized stochastic approximation 60
2.4.5 Summary of results 60
2.5 Experiments on artificial data 62
2.5.1 Splines on the circle 62
2.5.2 Experimental set-up 63
2.5.3 Optimal learning rate for our algorithm 63
2.5.4 Comparison to competing algorithms 64
2.6 Conclusion 65

2.1 Introduction

Positive-definite-kernel-based methods such as the support vector machine or kernel ridge regression are now widely used in many areas of science of engineering. They were first developed within the statistics community for non-parametric regression using splines, Sobolev spaces, and more generally reproducing kernel Hilbert spaces (see, e.g., Wahba (1990)). Within the machine learning community, they were extended in several interesting ways (see, e.g., Schölkopf and Smola (2002); Shawe-Taylor and Cristianini (2004)): (a) other problems were tackled using positive-definite kernels beyond regression problems, through the "kernelization" of classical unsupervised learning methods such as principal component analysis, canonical correlation analysis, or K-means, (b) efficient algorithms based on convex optimization have emerged, in particular for large sample sizes, and (c) kernels for non-vectorial data have been designed for objects like strings, graphs, measures, etc. A key feature is that they allow the separation of the representation problem (designing good kernels for non-vectorial data) and the algorithmic/theoretical problems (given a kernel, how to design, run efficiently and analyze estimation algorithms).

The theoretical analysis of non-parametric least-squares regression within the RKHS framework is well understood. In particular, regression on input data in $\mathbb{R}^d$, $d \geqslant 1$, and so-called *Mercer kernels* (continuous kernels over a compact set) that lead to dense subspaces of the space of square-integrable functions and non-parametric estimation (Tsybakov, 2008), has been widely studied in the last decade starting with the works of Smale and Cucker (2001, 2002) and being further refined (De Vito et al., 2005; Smale and Zhou, 2007) up to optimal rates (Caponnetto and De Vito, 2007; Steinwart et al., 2009; Bach, 2012) for Tikhonov regularization (batch iterative methods were for their part studied in (Blanchard and Krämer, 2010; Raskutti et al., 2014)). However, the kernel framework goes beyond Mercer kernels and non-parametric regression; indeed, kernels on non-vectorial data provide examples where the usual topological assumptions may not be natural, such as sequences, graphs and measures. Moreover, even finite-dimensional Hilbert spaces may need a more refined analysis when the dimension of the Hilbert space is much larger than the number of observations: for example, in modern text and web applications, linear predictions are performed with a large number of covariates which are equal to zero with high probability. The sparsity of the representation allows to reduce significantly the complexity of traditional optimization procedures; however, the finite-dimensional analysis which ignores the spectral structure of the data often leads to trivial guarantees because the number of covariates far exceeds the number of observations, while the analysis we carry out is meaningful (note that in these contexts sparsity of the underlying estimator is typically not a relevant assumption). In this chapter, we consider minimal assumptions regarding the input space and the distributions, so that our non-asymptotic results may be applied to all the cases mentioned above.

In practice, estimation algorithms based on regularized empirical risk minimization (e.g., penalized least-squares) face two challenges: (a) using the correct regularization parameter and (b) finding an approximate solution of the convex optimization problems. In this chapter, we consider these two problems jointly by following a stochastic approximation framework formulated directly in the RKHS, in which each observation is used only once and overfitting is avoided by making only a single pass through the data–a form of *early stopping*, which has been considered in other statistical frameworks such as boosting (Zhang and Yu, 2005). While this framework has been considered before (Rosasco et al., 2014;

Ying and Pontil, 2008; Tarrès and Yao, 2014), the algorithms that are considered either (a) require two sequences of hyper-parameters (the step-size in stochastic gradient descent and a regularization parameter) or (b) do not always attain the optimal rates of convergence for estimating the regression function. In this chapter, we aim to remove simultaneously these two limitations.

Traditional online stochastic approximation algorithms, as introduced by Robbins and Monro (Robbins and Monro, 1951), lead in finite-dimensional learning problems (e.g., parametric least-squares regression) to stochastic gradient descent methods with step-sizes decreasing with the number of observations n, which are typically proportional to $n^{-\zeta}$, with ζ between $1/2$ and 1. Short step-sizes ($\zeta = 1$) are adapted to well-conditioned problems (low dimension, low correlations between covariates), while longer step-sizes ($\zeta = 1/2$) are adapted to ill-conditioned problems (high dimension, high correlations) but with a worse convergence rate—see, e.g., Shalev-Shwartz (2011); Bach and Moulines (2011) and references therein. More recently Bach and Moulines (2013) showed that constant step-sizes *with averaging* could lead to the best possible convergence rate in Euclidean spaces (*i.e.*, in finite dimensions). In this chapter, we show that using longer step-sizes with averaging also brings benefits to Hilbert space settings needed for non-parametric regression.

With our analysis, based on positive definite kernels, under assumptions on both the objective function and the covariance operator of the RKHS, we derive improved rates of convergence (Caponnetto and De Vito, 2007), in both the finite horizon setting where the number of observations is known in advance and our bounds hold for the last iterate (with exact constants), and the online setting where our bounds hold for each iterate (asymptotic results only). It leads to an explicit choice of the step-sizes (which play the role of the regularization parameters) which may be used in stochastic gradient descent, depending on the number of training examples we want to use and on the assumptions we make.

In this chapter, we make the following contributions:

- We review in Section 2.2 a general though simple algebraic framework for least-squares regression in RKHS, which encompasses all commonly encountered situations. This framework however makes *unnecessary topological assumptions*, which we relax in Section 2.2.5 (with details in App. A.1).

- We characterize in Section 2.3 the convergence rate of averaged least-mean-squares (LMS) and show how the proper set-up of the step-size leads to optimal convergence rates (as they were proved in Caponnetto and De Vito (2007)), extending results from finite-dimensional (Bach and Moulines, 2013) to infinite-dimensional settings. The problem we solve here was stated as an open problem by Rosasco et al. (2014) and Ying and Pontil (2008). Moreover, our results apply as well in the usual finite-dimensional setting of parametric least-squares regression, showing adaptivity of our estimator to the spectral decay of the covariance matrix of the covariates (see Section 2.4.1).

- We compare our new results with existing work, both in terms of rates of convergence in Section 2.4, and with simulations on synthetic spline smoothing in Section 2.5.

Complete proofs are given in Chapter A. More precisely, minimal assumptions are presented in Section A.1 and sketches of the proofs are given in Section A.2. Then detailled proofs are given in Section A.3, and Section A.4.

2.2 Learning with positive-definite kernels

In this chapter, we consider a general random design regression problem, where observations (x_i, y_i) are independent and identically distributed (i.i.d.) random variables in $\mathcal{X} \times \mathcal{Y}$ drawn from a probability measure ρ on $\mathcal{X} \times \mathcal{Y}$. The set $\mathcal{X}$ may be any set equipped with a measure; moreover we consider for simplicity $\mathcal{Y} = \mathbb{R}$ and we measure the risk of a function $g : \mathcal{X} \to \mathbb{R}$, by the mean square error, that is, $\varepsilon(g) := \mathbb{E}_\rho\left[(g(X) - Y)^2\right]$.

The function g that minimizes $\varepsilon(g)$ over all measurable functions is known to be the conditional expectation, that is, $g_\rho(X) = \mathbb{E}[Y|X]$. In this chapter we consider formulations where our estimates lie in a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ with positive definite kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$.

2.2.1 Reproducing kernel Hilbert spaces

Throughout this section, we make the following assumption:

(A1) $\mathcal{X}$ is a compact topological space and $\mathcal{H}$ is an RKHS associated with a continuous kernel K on the set $\mathcal{X}$.

RKHSs are well-studied Hilbert spaces which are particularly adapted to regression problems (see, e.g., Berlinet and Thomas-Agnan (2004); Wahba (1990)). They satisfy the following properties:

1. $(\mathcal{H}, \langle\cdot,\cdot\rangle_\mathcal{H})$ is a separable Hilbert space of functions: $\mathcal{H} \subset \mathbb{R}^\mathcal{X}$.
2. $\mathcal{H}$ contains all functions $K_x : t \mapsto K(x,t)$, for all x in $\mathcal{X}$.
3. For any $x \in \mathcal{X}$ and $f \in \mathcal{H}$, the reproducing property holds:

$$f(x) = \langle f, K_x\rangle_\mathcal{H}.$$

The reproducing property allows to treat non-parametric estimation in the same algebraic framework as parametric regression. The Hilbert space $\mathcal{H}$ is totally characterized by the positive definite kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, which simply needs to be a symmetric function on $\mathcal{X} \times \mathcal{X}$ such that for any finite family of points $(x_i)_{i\in I}$ in $\mathcal{X}$, the $|I|\times|I|$-matrix of kernel evaluations is positive semi-definite. We provide examples in Section 2.2.6. For simplicity, we have here made the assumption that K is a Mercer kernel, that is, $\mathcal{X}$ is a compact set and $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is continuous. See Section 2.2.5 for an extension without topological assumptions.

2.2.2 Random variables

In this chapter, we consider a set $\mathcal{X}$ and $\mathcal{Y} \subset \mathbb{R}$ and a distribution ρ on $\mathcal{X} \times \mathcal{Y}$. We denote by ρ_X the marginal law on the space $\mathcal{X}$ and by $\rho_{Y|X=x}$ the conditional probability measure on Y given $x \in \mathcal{X}$. We may use the notations $\mathbb{E}[f(X)]$ or $\mathbb{E}_{\rho_X}[f(\cdot)]$ for $\int_\mathcal{X} f(x)d\rho_X(x)$. Beyond the moment conditions stated below, we will always make the assumption that the space $L^2_{\rho_X}$ of square ρ_X-integrable functions defined below is separable (this is the case in most interesting situations; see Thomson et al. (2000) for more details). Since we will assume that ρ_X has full support[1], we will make the usual simplifying identification of

[1] that is, the smallest closed space of measure 1 in the topological space $\mathcal{X}$ is $\mathcal{X}$ itself.

functions and their equivalence classes (based on equality up to a zero-measure set). We denote by $\|\cdot\|_{L^2_{\rho_X}}$ the norm:

$$\|f\|^2_{L^2_{\rho_X}} = \int_{\mathcal{X}} |f(x)|^2 d\rho_X(x).$$

The space $L^2_{\rho_X}$ is then a Hilbert space with norm $\|\cdot\|_{L^2_{\rho_X}}$.

Throughout this section, we make the following simple assumption regarding finiteness of moments:

(A2) $R^2 := \sup_{x\in\mathcal{X}} K(x,x)$ and $\mathbb{E}[Y^2]$ are finite; ρ_X has full support in $\mathcal{X}$.

Note that under these assumptions, any function in $\mathcal{H}$ is in $L^2_{\rho_X}$; however this inclusion is strict in most interesting situations.

2.2.3 Minimization problem

We are interested in minimizing the prediction error $\varepsilon(f)$ of a function f defined in Section 2.2. As we are looking for a function with a low prediction error in the particular function space $\mathcal{H}$, we aim to minimize $\varepsilon(f)$ over $f \in \mathcal{H}$. We have for $f \in L^2_{\rho_X}$:

$$\begin{aligned} \varepsilon(f) &= \|f\|^2_{L^2_{\rho_X}} - 2\left\langle f, \int_{\mathcal{Y}} y d\rho_{Y|X=\cdot}(y)\right\rangle_{L^2_{\rho_X}} + \mathbb{E}[Y^2] \\ &= \|f\|^2_{L^2_{\rho_X}} - 2\left\langle f, \mathbb{E}\left[Y|X=\cdot\right]\right\rangle_{L^2_{\rho_X}} + \mathbb{E}[Y^2]. \end{aligned} \tag{2.1}$$

A minimizer g of $\varepsilon(g)$ over $L^2_{\rho_X}$ is known to be such that $g(X) = \mathbb{E}[Y|X]$. Such a function is generally referred to as the regression function, and denoted g_ρ as it only depends on ρ. It is moreover unique (as an element of $L^2_{\rho_X}$). An important property of the prediction error is that the excess risk may be expressed as a squared distance to g_ρ, *i.e.*,

$$\forall f \in L^2_{\rho_X}, \qquad \varepsilon(f) - \varepsilon(g_\rho) = \|f - g_\rho\|^2_{L^2_{\rho_X}}. \tag{2.2}$$

A key feature of our analysis is that we only considered $\|f - g_\rho\|^2_{L^2_{\rho_X}}$ as a measure of performance and do not consider convergences in stricter norms (which are not true in general). *This allows us to neither assume that g_ρ is in $\mathcal{H}$ nor that $\mathcal{H}$ is dense in $L^2_{\rho_X}$.* We thus need to define a notion of the best estimator in $\mathcal{H}$. We first define the closure $\overline{F}$ (with respect to $\|\cdot\|_{L^2_{\rho_X}}$) of any set $F \subset L^2_{\rho_X}$ as the set of limits in $L^2_{\rho_X}$ of sequences in F. The space $\overline{\mathcal{H}}$ is a closed and convex subset in $L^2_{\rho_X}$. We can thus define $g_{\mathcal{H}} = \arg\min_{f\in\overline{\mathcal{H}}} \varepsilon(g)$, as the orthogonal projection of g_ρ on $\overline{\mathcal{H}}$, using the existence of the projection on any closed convex set in a Hilbert space. See Proposition A.1 in Section A.1 for details. Of course we do not have $g_{\mathcal{H}} \in \mathcal{H}$, that is *the infimum in $\mathcal{H}$ is in general not attained.*

Estimation from n i.i.d. observations builds a sequence $(g_n)_{n\in\mathbb{N}}$ in $\mathcal{H}$. We will prove under suitable conditions that such an estimator satisfies weak consistency, that is g_n ends up predicting as well as $g_{\mathcal{H}}$:

$$\mathbb{E}\left[\varepsilon(g_n) - \varepsilon(g_{\mathcal{H}})\right] \xrightarrow{n\to\infty} 0 \;\Leftrightarrow\; \|g_n - g_{\mathcal{H}}\|_{\mathcal{L}^2_{\rho_X}} \xrightarrow{n\to\infty} 0.$$

Seen as a function of $f \in \mathcal{H}$, our loss function ε is not coercive (*i.e.*, not strongly convex), as our covariance operator (see definition below) Σ has no minimal strictly positive eigenvalue (the sequence of eigenvalues decreases to zero). As a consequence, even if $g_{\mathcal{H}} \in \mathcal{H}$, g_n may not converge to $g_{\mathcal{H}}$ in $\mathcal{H}$, and *when $g_{\mathcal{H}} \notin \mathcal{H}$, we shall even have* $\|g_n\|_{\mathcal{H}} \to \infty$.

2.2.4 Covariance operator

We now define the *covariance operator* for the space $\mathcal{H}$ and probability distribution ρ_X. The spectral properties of such an operator have appeared to be a key point to characterize the convergence rates of estimators (Smale and Cucker, 2001; Smale and Zhou, 2007; Caponnetto and De Vito, 2007).

We implicitly define (via Riesz' representation theorem) a linear operator $\Sigma : \mathcal{H} \to \mathcal{H}$ through

$$\forall (f,g) \in \mathcal{H}^2, \quad \langle f, \Sigma g \rangle_{\mathcal{H}} = \mathbb{E}\left[f(X)g(X)\right] = \int_{\mathcal{X}} f(x)g(x)d\rho_X(x).$$

This operator is the *covariance operator* (defined on the Hilbert space $\mathcal{H}$). Using the reproducing property, we have:

$$\Sigma = \mathbb{E}\left[K_X \otimes K_X\right],$$

where for any elements $g, h \in \mathcal{H}$, we denote by $g \otimes h$ the operator from $\mathcal{H}$ to $\mathcal{H}$ defined as:

$$g \otimes h \ : f \mapsto \langle f, h \rangle_{\mathcal{H}} \, g.$$

Note that this expectation is formally defined as a Bochner expectation (an extension of Lebesgue integration theory to Banach spaces, see Mikusinski and Weiss (2014)), in $\mathcal{L}(\mathcal{H})$ the set of endomorphisms of $\mathcal{H}$.

In finite dimension, *i.e.*, $\mathcal{H} = \mathbb{R}^d$, for $g, h \in \mathbb{R}^d$, $g \otimes h$ may be identified to a rank-one matrix, that is, $g \otimes h = gh^\top = \left((g_i h_j)_{1 \leqslant i,j \leqslant d}\right) \in \mathbb{R}^{d \times d}$ as for any f, $(gh^\top)f = g(h^\top f) = \langle f, h \rangle_{\mathcal{H}} g$. In other words, $g \otimes h$ is a linear operator, whose image is included in Vect(g), the linear space spanned by g. Thus in finite dimension, Σ is the usual (non-centered) covariance matrix.

We have defined the covariance operator on the Hilbert space $\mathcal{H}$. If $f \in \mathcal{H}$, we have for all $z \in \mathcal{X}$, using the reproducing property:

$$\mathbb{E}[f(X)K(X,z)] = \mathbb{E}[f(X)K_z(X)] = \langle K_z, \Sigma f \rangle_{\mathcal{H}} = (\Sigma f)(z),$$

which shows that the operator Σ may be extended to any square-integrable function $f \in L^2_{\rho_X}$. In the following, we extend such an operator as an endomorphism T from $L^2_{\rho_X}$ to $L^2_{\rho_X}$.

Definition 2.1 (Extended covariance operator)**.** *Assume* ***(A1-2)****. We define the operator T as follows:*

$$\begin{array}{rccc} T: & L^2_{\rho_X} & \to & L^2_{\rho_X} \\ & g & \mapsto & \displaystyle\int_{\mathcal{X}} g(t)\ K_t \ d\rho_{\mathcal{X}}(t), \end{array}$$

so that for any $z \in \mathcal{X}$, $T(g)(z) = \displaystyle\int_{\mathcal{X}} g(x)\ K(x,z)\ d\rho_{\mathcal{X}}(t) = \mathbb{E}[g(X)K(X,z)]$.

From the discussion above, if $f \in \mathcal{H} \subset L^2_{\rho_X}$, then $Tf = \Sigma f$. We give here some of the most important properties of T. The operator T (which is an endomorphism of the separable Hilbert space $L^2_{\rho_X}$) may be reduced in some Hilbertian eigenbasis of $L^2_{\rho_X}$. It allows us to define the power of such an operator T^r, which will be used to quantify the regularity of the function $g_{\mathcal{H}}$. See proof in Section A.3.2, Proposition A.19.

Proposition 2.2 (Eigen-decomposition of T)**.** *Assume **(A1-2)**. T is a bounded self-adjoint semi-definite positive operator on $L^2_{\rho_X}$, which is trace-class. There exists a Hilbertian eigenbasis $(\phi_i)_{i\in I}$ of the orthogonal supplement S of the null space $\mathrm{Ker}(T)$, with summable strictly positive eigenvalues $(\mu_i)_{i\in I}$. That is:*

- $\forall i \in I,\ T\phi_i = \mu_i\phi_i$, $(\mu_i)_{i\in I}$ *strictly positive such that* $\sum_{i\in I}\mu_i < \infty$.
- $L^2_{\rho_X} = \mathrm{Ker}(T) \overset{\perp}{\oplus} S$, *that is,* $L^2_{\rho_X}$ *is the orthogonal direct sum of* $\mathrm{Ker}(T)$ *and* S.

When the space S has finite dimension, then I has finite cardinality, while in general I is countable. Moreover, the null space $\mathrm{Ker}(T)$ may be either reduced to $\{0\}$ (this is the more classical setting and such an assumption is often made), finite-dimensional (for example when the kernel has zero mean, thus constant functions are in S) or infinite-dimensional (e.g., when the kernel space only consists in even functions, the whole space of odd functions is in S).

Moreover, the linear operator T allows to relate $L^2_{\rho_X}$ and $\mathcal{H}$ in a very precise way. For example, when $g \in \mathcal{H}$, we immediately have $Tg = \Sigma g \in \mathcal{H}$ and $\langle g, Tg\rangle_{\mathcal{H}} = \mathbb{E}g(X)^2 = \|g\|^2_{L^2_{\rho_X}}$. As we formally state in the following propositions, this essentially means that $T^{1/2}$ will be an isometry from $L^2_{\rho_X}$ to $\mathcal{H}$. We first show that the linear operator T happens to have an image included in $\mathcal{H}$, and that the eigenbasis of T in $L^2_{\rho_X}$ may also be seen as eigenbasis of Σ in $\mathcal{H}$ (See proof in Section A.3.2, Proposition A.18):

Proposition 2.3 (Decomposition of Σ)**.** *Assume **(A1-2)**. $\Sigma : \mathcal{H} \to \mathcal{H}$ is injective. The image of T is included in $\mathcal{H}$: $Im(T) \subset \mathcal{H}$. Moreover, for any $i \in I$, $\phi_i = \frac{1}{\mu_i}T\phi_i \in \mathcal{H}$, thus $\left(\mu_i^{1/2}\phi_i\right)_{i\in I}$ is an orthonormal eigen-system of Σ and an Hilbertian basis of $\mathcal{H}$, i.e., for any i in I, $\Sigma\phi_i = \mu_i\phi_i$.*

This proposition will be generalized under relaxed assumptions (in particular as Σ will no more be injective, see Section 2.2.5 and Section A.1).

We may now define all powers T^r (they are always well defined because the sequence of eigenvalues is upper-bounded):

Definition 2.4 (Powers of T)**.** *We define, for any $r \geqslant 0$, $T^r : L^2_{\rho_X} \to L^2_{\rho_X}$, for any $h \in \mathrm{Ker}(T)$ and $(a_i)_{i\in I}$ such that $\sum_{i\in I} a_i^2 < \infty$, through: $T^r\left(h + \sum_{i\in I} a_i\phi_i\right) = \sum_{i\in I} a_i\mu_i^r\phi_i$. Moreover, for any $r > 0$, T^r may be defined as a bijection from S into $Im(T^r)$. We may thus define its unique inverse $T^{-r} :\ Im(T^r) \to S$.*

The following proposition is a consequence of Mercer's theorem (Smale and Cucker, 2001; Aronszajn, 1950). It describes how the space $\mathcal{H}$ is related to the image of operator $T^{1/2}$.

Proposition 2.5 (Isometry for Mercer kernels)**.** *Under assumptions **(A1,2)**, $\mathcal{H} = T^{1/2}\left(L^2_{\rho_X}\right)$ and $T^{1/2} : S \to \mathcal{H}$ is an isometrical isomorphism.*

The proposition has the following consequences:

Corollary 2.6. *Assume **(A1, A2)**:*

- *For any $r \geqslant 1/2$, $T^r(S) \subset \mathcal{H}$, because $T^r(S) \subset T^{1/2}(S)$, that is, with large enough powers r, the image of T^r is in the Hilbert space.*

– $\forall r > 0$, $\overline{T^r(L^2_{\rho_X})} = S = \overline{T^{1/2}(L^2_{\rho_X})} = \overline{\mathcal{H}}$, *because (a)* $T^{1/2}(L^2_{\rho_X}) = \mathcal{H}$ *and (b) for any* $r > 0$, $\overline{T^r(L^2_{\rho_X})} = S$. *In other words, elements of* $\overline{\mathcal{H}}$ *(on which our minimization problem attains its minimum), may seen as limits (in* $L^2_{\rho_X}$*) of elements of* $T^r(L^2_{\rho_X})$*, for any* $r > 0$.

– $\mathcal{H}$ is dense in $L^2_{\rho_X}$ if and only if T is injective (which is equivalent to $\ker(T) = \{0\}$)

The sequence of spaces $\{T^r(L^2_{\rho_X})\}_{r>0}$ is thus a decreasing (when r is increasing) sequence of subspaces of $L^2_{\rho_X}$ such that any of them is dense in $\overline{\mathcal{H}}$, and $T^r(L^2_{\rho_X}) \subset \mathcal{H}$ if and only if $r \geqslant 1/2$.

In the following, the regularity of the function $g_{\mathcal{H}}$ will be characterized by the fact that $g_{\mathcal{H}}$ belongs to the space $T^r(L^2_{\rho_X})$ (and not only to its closure), for a specific $r > 0$ (see Section 2.2.7). This space may be described depending on the eigenvalues and eigenvectors as

$$T^r(L^2_{\rho_X}) = \left\{ \sum_{i=1}^{\infty} b_i \phi_i \text{ such that } \sum_{i=1}^{\infty} \frac{b_i^2}{\mu_i^{2r}} < \infty \right\}.$$

We may thus see the spaces $T^r(L^2_{\rho_X})$ as spaces of sequences with various decay conditions.

2.2.5 Minimal assumptions

In this section, we describe under which "minimal" assumptions the analysis holds. We prove that the set $\mathcal{X}$ may only be assumed to be equipped with a measure, the kernel K may only be assumed to have bounded expectation $\mathbb{E}_\rho K(X, X)$ and the output Y may only be assumed to have finite variance. That is:

(A1') $\mathcal{H}$ is a separable RKHS associated with kernel K on the set $\mathcal{X}$.

(A2') $\mathbb{E}\left[K(X, X)\right]$ and $\mathbb{E}[Y^2]$ are finite.

In this section, we have to distinguish the set of square ρ_X-integrable functions $\mathcal{L}^2_{\rho_X}$ and its quotient $L^2_{\rho_X}$ that makes it a separable Hilbert space. We define p the projection from $\mathcal{L}^2_{\rho_X}$ into $L^2_{\rho_X}$ (precise definitions are given in Section A.1). Indeed it is no more possible to identify the space $\mathcal{H}$, which is a subset of $\mathcal{L}^2_{\rho_X}$, and its canonical projection $p(\mathcal{H})$ in $L^2_{\rho_X}$.

Minimality: The separability assumption is necessary to be able to expand any element as an infinite sum, using a countable orthonormal family (this assumption is satisfied in almost all cases, for instance it is simple as soon as $\mathcal{X}$ admits a topology for which it is separable and functions in $\mathcal{H}$ are continuous, see Berlinet and Thomas-Agnan (2004) for more details). Note that we do not make any topological assumptions regarding the set $\mathcal{X}$. We only assume that it is equipped with a probability measure.

Assumption **(A2')** is needed to ensure that every function in $\mathcal{H}$ is square-integrable, that is, $\mathbb{E}[K(X, X)] < \infty$ if and only if $\mathcal{H} \subset \mathcal{L}^2_{\rho_X}$; for example, for $f = K_z$, $z \in \mathcal{X}$, $\|K_z\|^2_{L^2_{\rho_X}} = \mathbb{E}[K(X, z)^2] \leqslant K(z, z)\mathbb{E}K(X, X)$ (see more details in the Section A.3, Proposition A.7).

Our assumptions are sufficient to analyze the minimization of $\varepsilon(f)$ with respect to $f \in \mathcal{H}$ and seem to allow the widest generality.

Comparison: These assumptions will include the previous setting, but also recover measures without full support (e.g., when the data lives in a small subspace of the whole space) and kernels on discrete objects (with non-finite cardinality).

Moreover, **(A1'), (A2')** are strictly weaker than **(A1)**, **(A2)**. In previous work, **(A2')** was sometimes replaced by the stronger assumptions $\sup_{x \in \mathcal{X}} K(x, x) < \infty$ (Rosasco et al.,

2014; Ying and Pontil, 2008; Tarrès and Yao, 2014) and $|Y|$ bounded (Rosasco et al., 2014; Tarrès and Yao, 2014). Note that in functional analysis, the weaker hypothesis $\int_{\mathcal{X}\times\mathcal{X}} k(x,x')^2 d\rho_X(x) d\rho_X(x') < \infty$ is often used (Brezis, 1983), but it is not adapted to the statistical setting.

Main differences: The main difference here is that we cannot identify $\mathcal{H}$ and $p(\mathcal{H})$: there may exist functions $f \in \mathcal{H} \setminus \{0\}$ such that $\|f\|_{\mathcal{L}^2_{\rho_X}} = 0$. This may for example occur if the support of ρ_X is strictly included in $\mathcal{X}$, and f is zero on this support, but not identically zero. See the Section A.3.5 for more details.

As a consequence, Σ is no more injective and we do not have $\mathrm{Im}(T^{1/2}) = \mathcal{H}$ anymore. We thus denote $\mathscr{S}$ an orthogonal supplement of the null space $\mathrm{Ker}(\Sigma)$. As we also need to be careful not to confuse $\mathcal{L}^2_{\rho_X}$ and $L^2_{\rho_X}$, we define an extension $\mathcal{T}$ of Σ from $\mathcal{L}^2_{\rho_X}$ into $\mathcal{H}$, then $T = p \circ \mathcal{T}$. We can define for $r \geqslant 1/2$ the power operator $\mathcal{T}^r$ of $\mathcal{T}$ (from $L^2_{\rho_X}$ into $\mathcal{H}$), see App. A.1 for details.

Conclusion: Our problem has the same behaviour under such assumptions. Proposition 2.2 remains unchanged. Decompositions in Prop. 2.3 and Corollary 2.6 must be slightly adapted (see Proposition A.3 and Corollary A.5 in Section A.1 for details). Finally, Proposition 2.5 is generalized by the next proposition, which states that $p(\mathscr{S}) = p(\mathcal{H})$ and thus S and $p(\mathcal{H})$ are isomorphic (see proof in Section A.3.2, Proposition A.19):

Proposition 2.7 (Isometry between supplements)**.** *$\mathcal{T}^{1/2} : S \to \mathscr{S}$ is an isometry. Moreover, $\mathrm{Im}(T^{1/2}) = p(\mathcal{H})$ and $T^{1/2} : S \to p(\mathcal{H})$ is an isomorphism.*

We can also derive a version of Mercer's theorem, which does not make anymore assumptions that are required for defining RKHSs. As we will not use it in this article, this proposition is only given in Section A.1.

Convergence results: In all convergence results stated below, *assumptions **(A1, A2)** may be replaced by assumptions **(A1', A2')***.

2.2.6 Examples

The property $\overline{\mathcal{H}} = S$, stated after Proposition 2.5, is important to understand what the space $\overline{\mathcal{H}}$ is, as we are minimizing over this closed and convex set. As a consequence the space $\mathcal{H}$ is dense in $L^2_{\rho_X}$ if and only if T is injective (or equivalently, $\mathrm{Ker}(T) = \{0\} \Leftrightarrow S = L^2_{\rho_X}$). We detail below a few classical situations in which different configurations for the "inclusion" $\mathcal{H} \subset \overline{\mathcal{H}} \subset L^2_{\rho_X}$ appear:

1. **Finite-dimensional setting with linear kernel**: in finite dimension, with $\mathcal{X} = \mathbb{R}^d$ and $K(x,y) = x^\top y$, we have $\mathcal{H} = \mathbb{R}^d$, with the scalar product in $\langle u, v\rangle_{\mathcal{H}} = \sum_{i=1}^d u_i v_i$. This corresponds to usual parametric least-squares regression. If the support of ρ_X has non-empty interior, then $\overline{\mathcal{H}} = \mathcal{H}$: $g_{\mathcal{H}}$ is the best linear estimator. Moreover, we have $\mathcal{H} = \overline{\mathcal{H}} \subsetneq L^2_{\rho_X}$: indeed $\mathrm{Ker}(T)$ is the set of functions such that $\mathbb{E}Xf(X) = 0$ (which is a large space).

2. **Translation-invariant kernels**: for instance the Gaussian kernel over $\mathcal{X} = \mathbb{R}^d$, with X following a distribution with full support in $\mathbb{R}^d$: in such a situation we have $\mathcal{H} \subsetneq \overline{\mathcal{H}} = L^2_{\rho_X}$. This last equality holds more generally for all universal kernels, which include all kernels of the form $K(x,y) = q(x-y)$ where q has a summable strictly positive Fourier transform (Micchelli et al., 2006; Sriperumbudur et al., 2011). These kernels are exactly the kernels such that T is an injective endomorphism of $L^2_{\rho_X}$.

3. **Splines over the circle:** When $X \sim \mathcal{U}[0;1]$ and $\mathcal{H}$ is the set of periodic m-times weakly differentiable functions (see Section 2.5), we have in general $\mathcal{H} \subsetneq \overline{\mathcal{H}} \subsetneq L^2_{\rho_X}$. In such a case, $\ker(T) = \operatorname{span}(x \mapsto 1) = \{x \mapsto c, c \in \mathbb{R}\}$, and $\overline{\mathcal{H}} \oplus \operatorname{span}(x \mapsto 1) = L^2_{\rho_X}$. This means we can approximate a function in $L^2_{\rho_X}$ by functions in $\mathcal{H}$ if and only if this function has zero-mean.

Many examples and more details may be found in Shawe-Taylor and Cristianini (2004); Aronszajn (1950); Vert (2014). In particular, kernels on non-vectorial objects may be defined (e.g., sequences, graphs or measures).

2.2.7 Convergence rates

In order to be able to establish rates of convergence in this infinite-dimensional setting, we have to make assumptions on the objective function and on the covariance operator eigenvalues. In order to account for all cases (finite and infinite dimensions), we now consider eigenvalues ordered in *non-increasing* order, that is, we assume that the set I is either $\{1, \dots, d\}$ if the underlying space is d-dimensional or $\mathbb{N}^*$ if the underlying space has infinite dimension.

(A3) We denote $(\mu_i)_{i \in I}$ the sequence of non-zero eigenvalues of the operator T, in decreasing order. We assume $\mu_i \leqslant \frac{s^2}{i^\alpha}$ for some $\alpha > 1$ (so that $\operatorname{tr}(T) < \infty$), with $s \in \mathbb{R}_+$.

(A4) $g_{\mathcal{H}} \in T^r\left(L^2_{\rho_X}\right)$ with $r \geqslant 0$, and as a consequence $\|T^{-r}(g_{\mathcal{H}})\|_{L^2_{\rho_X}} < \infty$.

We chose such assumptions in order to make the comparison with the existing literature as easy as possible, for example Caponnetto and De Vito (2007); Ying and Pontil (2008). However, some other assumptions may be found as in Bach (2012); Hsu et al. (2014).

Dependence on α and r. The two parameters r and α intuitively parametrize the strengths of our assumptions:

- In assumption **(A3)** a bigger α makes the assumption stronger: it means the reproducing kernel Hilbert space is smaller, that is if **(A3)** holds with some constant α, then it also holds for any $\alpha' < \alpha$. Moreover, if T is reduced in the Hilbertian basis $(\phi_i)_i$ of $L^2_{\rho_X}$, we have an effective search space $S = \{\sum_{i=1}^\infty b_i \phi_i / \sum_{i=1}^\infty \frac{b_i^2}{\mu_i} < \infty\}$: the smaller the eigenvalues, the smaller the space. Note that since $\operatorname{tr}(T)$ is finite, **(A3)** is always true for $\alpha = 1$. This assumption is generally referred to as the *capacity condition*.

- In assumption **(A4)**, for a fixed α, a bigger r makes the assumption stronger, that is the function $g_{\mathcal{H}}$ is actually smoother. Indeed, considering that **(A4)** may be rewritten $g_{\mathcal{H}} \in T^r(L^2_{\rho_X})$ and for any $r < r'$, $T^{r'}(L^2_{\rho_X}) \subset T^r(L^2_{\rho_X})$. In other words, $\{T^r\left(L^2_{\rho_X}\right)\}_{r \geq 0}$ are decreasing (r growing) subspaces of $L^2_{\rho_X}$.

  For $r = 1/2$, $T^{1/2}(L^2_{\rho_X}) = \mathcal{H}$; moreover, for $r \geqslant 1/2$, our best approximation function $g_{\mathcal{H}} \in \overline{\mathcal{H}}$ is in fact in $\mathcal{H}$, that is the optimization problem in the RKHS $\mathcal{H}$ is attained by a function of finite norm. However for $r < 1/2$ it is not attained. This assumption is generally referred to as the *source condition*.

– Furthermore, it is worth pointing the stronger assumption which is often used in the finite dimensional context, namely $\operatorname{tr}\left(\Sigma^{1/\alpha}\right) = \sum_{i\in I} \mu_i^{1/\alpha}$ finite. It turns out that this is a stronger assumption, indeed, since we have assumed that the eigenvalues (μ_i) are arranged in non-increasing order, if $\operatorname{tr}\left(\Sigma^{1/\alpha}\right)$ is finite, then **(A3)** is satisfied for $s^2 = \left[2\operatorname{tr}\left(\Sigma^{1/\alpha}\right)\right]^{\alpha}$. Such an assumption appears for example in Corollary 2.15.

Related assumptions. The assumptions **(A3)** and **(A4)** are adapted to our theoretical results, but some stricter assumptions are often used, that make comparison with existing work more direct. For comparison purposes, we will also use:

(a3) For any $i \in I = \mathbb{N}$, $u^2 \leqslant i^{\alpha}\mu_i \leqslant s^2$ for some $\alpha > 1$ and $u, s \in \mathbb{R}_+$.

(a4) We assume the coordinates $(\nu_i)_{i\in\mathbb{N}}$ of $g_{\mathcal{H}} \in L^2_{\rho_X}$ in the eigenbasis $(\phi_i)_{i\in\mathbb{N}}$ (for $\|.\|_{L^2_{\rho_X}}$) of T are such that $\nu_i i^{\delta/2} \leqslant W$, for some $\delta > 1$ and $W \in \mathbb{R}_+$ (so that $\|g_{\mathcal{H}}\|_{L^2_{\rho_X}} < \infty$).

Assumption **(a3)** directly imposes that the eigenvalues of T decay at rate $i^{-\alpha}$ (which imposes that there are infinitely many), and thus implies **(A3)**. Together, assumptions **(a3)**(α) and **(a4)**(δ), imply assumptions **(A3)**(α) and **(A4)**(r), with any r such that $\delta > 1 + 2\alpha r$ (and assumption **(A4)**(r) does not stand if $\delta < 1 + 2\alpha r$). Indeed, we have

$$\|T^{-r} g_{\mathcal{H}}\|^2_{L^2_{\rho_X}} = \sum_{i\in\mathbb{N}} \nu_i^2 \mu_i^{-2r} = \Theta\left(\frac{W^2}{u^{4r}} \sum_{i\in\mathbb{N}} i^{-\delta+2\alpha r}\right),$$

which is finite if and only if $2\alpha r + 1 < \delta$. Thus, the supremum element of the set of r such that **(A4)** holds is such that $\delta = 1 + 2\alpha r$. Thus, when comparing assumptions **(A3-4)** and **(a3-4)**, we will often make the identification above, that is, $\delta = 1 + 2\alpha r$.

The main advantage of the new assumptions is their interpretation when the basis $(\phi_i)_{i\in I}$ is common for several RKHSs (such as the Fourier basis for splines, see Section 2.5): **(a4)** describes the decrease of the coordinates of the best function $g_{\mathcal{H}} \in L^2_{\rho_X}$ *independently of the chosen RKHS*. Thus, the parameter δ characterizes the prediction function, while the parameter α characterizes the RKHS.

2.3 Stochastic approximation in Hilbert spaces

In this section, we consider estimating a prediction function $g \in \mathcal{H}$ from observed data, and we make the following assumption:

(A5) For $n \geqslant 1$, the random variables $(x_n, y_n) \in \mathcal{X} \times \mathbb{R}$ are independent and identically distributed with distribution ρ.

Our goal is to estimate a function $g \in \mathcal{H}$ from data, such that $\varepsilon(g) = \mathbb{E}(Y - g(X))^2$ is as small as possible. As shown in Section 2.2, this is equivalent to minimizing $\|g - g_{\mathcal{H}}\|^2_{L^2_{\rho_X}}$. Among others, two generic approaches to define an estimator are by regularization or by stochastic approximation (and combinations thereof). See also approaches by early-stopped gradient descent on the empirical risk in Yao et al. (2007).

2.3.1 Regularization and linear systems

Given n observations, regularized empirical risk minimization corresponds to minimizing with respect to $g \in \mathcal{H}$ the following objective function:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - g(x_i))^2 + \lambda\|g\|_{\mathcal{H}}^2.$$

Although the problem is formulated in a potentially infinite-dimensional Hilbert space, through the classical representer theorem (Schölkopf and Smola, 2002; Shawe-Taylor and Cristianini, 2004; Kimeldorf and Wahba, 1971), the unique (if $\lambda > 0$) optimal solution may be expressed as $\hat{g} = \sum_{i=1}^{n} a_i K_{x_i}$, and $a \in \mathbb{R}^n$ may be obtained by solving the linear system $(\mathbf{K} + n\lambda I)a = \mathbf{y}$, where $\mathbf{K} \in \mathbb{R}^{n\times n}$ is the kernel matrix, a.k.a. the Gram matrix, composed of pairwise kernel evaluations $\mathbf{K}_{ij} = K(x_i, x_j)$, $i, j = 1, \ldots, n$, and $\mathbf{y}$ is the n-dimensional vector of all n responses y_i, $i = 1, \ldots, n$.

The running-time complexity to obtain $a \in \mathbb{R}^n$ is typically $O(n^3)$ if no assumptions are made, but several algorithms may be used to lower the complexity and obtain an approximate solution, such as conjugate gradient (Golub and Van Loan, 1996) or column sampling (a.k.a. Nyström method) (Mahoney, 2011; Williams and Seeger, 2001; Bach, 2012).

In terms of convergence rates, assumptions **(a3-4)** allow to obtain convergence rates that decompose $\varepsilon(\hat{g}) - \varepsilon(g_{\mathcal{H}}) = \|\hat{g} - g_{\mathcal{H}}\|_{L^2_{\rho_X}}^2$ as the sum of two asymptotic terms (Caponnetto and De Vito, 2007; Hsu et al., 2014; Bach, 2012):

- *Variance term*: $O(\sigma^2 n^{-1}\lambda^{-1/\alpha})$, which is decreasing with λ, where σ^2 characterizes the noise variance, for example, in the homoscedastic case (i.i.d. additive noise), the marginal variance of the noise; see assumption **(A6)** for the detailed assumption that we need in our stochastic approximation context.

- *Bias term*: $O(\lambda^{\min\{(\delta-1)/\alpha, 2\}})$, which is increasing with λ. Note that the corresponding r from assumptions **(A3-4)** is $r = (\delta - 1)/2\alpha$, and the bias term becomes proportional to $\lambda^{\min\{2r,2\}}$.

There are then two regimes:

- *Optimal predictions*: If $r < 1$, then the optimal value of λ (that minimizes the sum of two terms and makes them asymptotically equivalent) is proportional to $n^{-\alpha/(2r\alpha+1)} = n^{-\alpha/\delta}$ and the excess prediction error $\|\hat{g} - g_{\mathcal{H}}\|_{L^2_{\rho_X}}^2 = O(n^{-2\alpha r/(2\alpha r+1)}) = O(n^{-1+1/\delta})$, and the resulting procedure is then "optimal" in terms of estimation of $g_{\mathcal{H}}$ in $L^2_{\rho_X}$ (see Section 2.4 for details).

- *Saturation*: If $r \geqslant 1$, where the optimal value of λ (that minimizes the sum of two terms and makes them equivalent) is proportional to $n^{-\alpha/(2\alpha+1)}$, and the excess prediction error is less than $O(n^{-2\alpha/(2\alpha+1)})$, which is suboptimal. Although assumption **(A4)** is valid for a larger r, the rate is the same than if $r = 1$.

In this chapter, we consider a stochastic approximation framework with improved running-time complexity and similar theoretical behavior as regularized empirical risk minimization, with the advantages of (a) needing a single pass through the data and (b) simple assumptions.

2.3.2 Stochastic approximation

Using the reproducing property, we have for any $g \in \mathcal{H}$, $\varepsilon(g) = \mathbb{E}(Y - g(X))^2 = \mathbb{E}(Y - \langle g, K_X\rangle_{\mathcal{H}})^2$, with gradient (defined with respect to the dot-product in $\mathcal{H}$) $\nabla\varepsilon(g) = -2\mathbb{E}\big[(Y - \langle g, K_X\rangle_{\mathcal{H}})K_X\big]$.

Thus, for each pair of observations (x_n, y_n), we have $\nabla\varepsilon(g) = -2\mathbb{E}\big[(y_n - \langle g, K_{x_n}\rangle_{\mathcal{H}})K_{x_n}\big]$, and thus, the quantity $(\langle g, K_{x_n}\rangle_{\mathcal{H}} - y_n)K_{x_n} = (g(x_n) - y_n)K_{x_n}$ is an *unbiased stochastic (half) gradient*. We thus consider the stochastic gradient recursion, in the Hilbert space $\mathcal{H}$, started from a function $g_0 \in \mathcal{H}$ (taken to be zero in the following):

$$g_n = g_{n-1} - \gamma_n\big[\langle g_{n-1}, K_{x_n}\rangle_{\mathcal{H}} - y_n\big]K_{x_n} = g_{n-1} - \gamma_n\big[g_{n-1}(x_n) - y_n\big]K_{x_n},$$

where γ_n is the *step-size*.

We may also apply the recursion using representants. Indeed, if $g_0 = 0$, which we now assume, then for any $n \geqslant 1$,

$$g_n = \sum_{i=1}^{n} a_i K_{x_i},$$

with the following recursion on the sequence $(a_n)_{n\geqslant 1}$:

$$a_n = -\gamma_n(g_{n-1}(x_n) - y_n) = -\gamma_n\left(\sum_{i=1}^{n-1} a_i K(x_n, x_i) - y_n\right).$$

We also output the averaged iterate defined as

$$\overline{g}_n = \frac{1}{n+1}\sum_{k=0}^{n} g_k = \frac{1}{n+1}\sum_{k=1}^{n}\Big(\sum_{j=1}^{k} a_j K_{x_j}\Big). \tag{2.3}$$

Running-time complexity. To compute $\overline{g}_n$, we need n steps of stochastic approximation. The running time complexity is $O(i)$ for iteration i—if we assume that kernel evaluations are $O(1)$, and thus $O(n^2)$ after n steps. This is a serious limitation for practical applications. Several authors have considered expanding g_n on a subset of all (K_{x_i}), which allows to bring down the complexity of each iteration and obtain an overall linear complexity $O(n)$ (Dekel et al., 2005; Bordes et al., 2005), but this comes at the expense of not obtaining the sharp generalization errors that we obtain in this chapter. Note that when studying regularized least-squares problem (*i.e.*, adding a penalization term), one has to update all the coefficients $(a_i)_{1\leqslant i\leqslant n}$ at step n, while in our situation, only a_n is computed at step n.

Relationship to previous works. Similar algorithms have been studied before (Rosasco et al., 2014; Ying and Pontil, 2008; Kivinen et al., 2004; Yao, 2006; Zhang, 2004), under various forms. Especially, in Tarrès and Yao (2014); Kivinen et al. (2004); Yao (2006); Zhang (2004) a regularization term is added to the loss function (thus considering the following problem: $\arg\min_{f\in\mathcal{H}} \varepsilon(f) + \lambda||f||_K^2$). In Rosasco et al. (2014); Ying and Pontil (2008), neither regularization nor averaging procedure are considered, but in the second case, multiple passes through the data are considered. In Zhang (2004), a non-regularized averaged procedure equivalent to ours is considered. However, the step-sizes γ_n which are proposed, as well as the corresponding analysis, are different. Our step-sizes are larger and our analysis uses more directly the underlying linear algebra to obtain better rates (while the proof of Zhang (2004) is applicable to all smooth losses).

Step-sizes. We are mainly interested in two different types of step-sizes (a.k.a. *learning rates*): the sequence $(\gamma_i)_{1\leqslant i\leqslant n}$ may be either:

1. a subsequence of a universal sequence $(\gamma_i)_{i\in\mathbb{N}}$, we refer to this situation as the *"online setting"*. Our bounds then hold for any of the iterates.

2. a sequence of the type $\gamma_i = \Gamma(n)$ for $i \leqslant n$, which will be referred to as the *"finite horizon setting"*: in this situation the number of samples is assumed to be known and fixed and we chose a constant step-size which may depend on this number. Our bound then holds only for the last iterate.

In practice it is important to have an online procedure, to be able to deal with huge amounts of data (potentially infinite). However, the analysis is easier in the "finite horizon" setting. Some *doubling tricks* allow to pass to varying steps (Hazan and Kale, 2011), but it is not fully satisfactory in practice as it creates jumps at every n which is a power of two.

2.3.3 Extra regularity assumptions

We denote by $\Xi = (Y - g_{\mathcal{H}}(X))K_X$ the residual, a random element of $\mathcal{H}$. We have $\mathbb{E}[\Xi] = 0$ but in general we do not have $\mathbb{E}[\Xi|X] = 0$ (unless the model of homoscedastic regression is well specified). We make the following extra assumption:

(A6) There exists $\sigma > 0$ such that $\mathbb{E}[\Xi \otimes \Xi] \preccurlyeq \sigma^2\Sigma$, where $\preccurlyeq$ denotes the order between self-adjoint operators.

In other words, for any $f \in \mathcal{H}$, we have $\mathbb{E}[(Y - g_{\mathcal{H}}(X))^2 f(X)^2] \leqslant \sigma^2\mathbb{E}[f(X)^2]$.

In the well specified homoscedastic case, we have that $(Y - g_{\mathcal{H}}(X))$ is independent of X and with $\sigma^2 = \mathbb{E}\left[(Y - g_{\mathcal{H}}(X))^2\right]$, $\mathbb{E}[\Xi|X] = \sigma^2\Sigma$ is clear: the constant σ^2 in the first part of our assumption characterizes the noise amplitude. Moreover when $|Y - g_{\mathcal{H}}(X)|$ is a.s. bounded by σ^2, we have **(A6)**.

We first present the results in the *finite horizon* setting in Section 2.3.4 before turning to the *online* setting in Section 2.3.5.

2.3.4 Main results (finite horizon)

We can first get some guarantee on the consistency of our estimator, for any small enough constant step-size:

Theorem 2.8. *Assume **(A1-6)**, then for any constant choice $\gamma_n = \gamma_0 < \frac{1}{2R^2}$, the prediction error of $\bar{g}_n$ converges in expectation to the one of $g_{\mathcal{H}}$, that is:*

$$\mathbb{E}\left[\varepsilon\left(\bar{g}_n\right) - \varepsilon(g_{\mathcal{H}})\right] = \mathbb{E}\|\bar{g}_n - g_{\mathcal{H}}\|^2_{L^2_{\rho_X}} \xrightarrow{n\to\infty} 0. \tag{2.4}$$

The expectation is considered with respect to the distribution of the sample $(x_i, y_i)_{1\leqslant i\leqslant n}$, as in all the following theorems (note that $\|\bar{g}_n - g_{\mathcal{H}}\|^2_{L^2_{\rho_X}}$ is itself a different expectation with respect to the law ρ_X).

Theorem 2.8 means that for the simplest choice of the learning rate as a constant, our estimator tends to perform as well as the best estimator in the class $\mathcal{H}$. Note that in general, the convergence in $\mathcal{H}$ is meaningless if $r < 1/2$. The following results will state some assertions on the speed of such a convergence; our main result, in terms of generality is the following:

Theorem 2.9 (Complete bound, γ constant, finite horizon)**.** *Assume* ***(A1-6)*** *and* $\gamma_i = \gamma = \Gamma(n)$*, for* $1 \leqslant i \leqslant n$*. If* $\gamma R^2 \leqslant 1/4$*, then*

$$\mathbb{E}\|\bar{g}_n - g_{\mathcal{H}}\|^2_{L^2_{\rho_X}} \leqslant \frac{4\sigma^2}{n}\left(1 + (s^2\gamma n)^{\frac{1}{\alpha}}\right) + 4(1 + q_{n,\gamma,s,r})\,\frac{\|T^{-r} g_{\mathcal{H}}\|^2_{L^2_{\rho_X}}}{\gamma^{2r} n^{2\min\{r,1\}}};$$

where $q_{n,\gamma,s,r} := (R^{2\alpha}\gamma^{1+\alpha} n s^2)^{\frac{2r-1}{\alpha}}$ *if* $r \geqslant \frac{1}{2}$ *and* $q_{n,\gamma,s,r} := 0$ *otherwise is a residual quantity.*

We can make the following observations:

- **Proof**: Theorem 2.8 is directly derived from Theorem 2.9, which is proved in Section A.4.3: we derive for our algorithm a new error decomposition and bound the different sources of error via algebraic calculations. More precisely, following the proof in Euclidean space Bach and Moulines (2013), we first analyze (in Section A.4.2) a closely related recursion (we replace $K_{x_n} \otimes K_{x_n}$ by its expectation Σ, and we thus refer to it as a semi-stochastic version of our algorithm):

  $$g_n = g_{n-1} - \gamma_n(y_n K_{x_n} - \Sigma g_{n-1}).$$

  It (a) leads to an easy computation of the main bias/variance terms of our result, (b) will be used to derive our main result by bounding the drifts between our algorithm and its semi-stochastic version. A more detailed sketch of the proof is given in Section A.2.

- **Bias/variance interpretation**: The two main terms have a simple interpretation. The first one is a variance term, which shows the effect of the noise σ^2 on the error. It is bigger when σ gets bigger, and moreover it also gets bigger when γ is growing (bigger steps mean more variance). As for the second term, it is a bias term, which accounts for the distance of the initial choice (the null function in general) to the objective function. As a consequence, it is smaller when we make bigger steps.

- **Assumption (A4)**: Our assumption **(A4)** for $r > 1$ is stronger than for $r = 1$ but we do not improve the bound. Indeed the bias term (see comments below) cannot decrease faster than $O(n^{-2})$: this phenomenon in known as saturation (Engl et al., 1996). To improve our results with $r > 1$ it may be interesting to consider another type of averaging. In the following, $r < 1$ shall be considered as the main and most interesting case.

- **Relationship to regularized empirical risk minimization**: Our bound ends up being very similar to bounds for regularized empirical risk minimization, with the identification $\lambda = \frac{1}{\gamma n}$. It is thus no surprise that once we optimize for the value of γ, we recover the same rates of convergence. Note that in order to obtain convergence, we require that the step-size γ is bounded, which corresponds to an equivalent λ which has to be lower-bounded by $1/n$.

- **Finite horizon**: Once again, this theorem holds in the finite horizon setting. That is we first choose the number of samples we are going to use, then the learning rate as a constant. It allows us to chose γ as a function of n, in order to balance the main terms in the error bound. The trade-off must be understood as follows: a bigger γ increases the effect of the noise, but a smaller one makes it harder to forget the initial condition.

We may now deduce the following corollaries, with specific optimized values of γ:

Corollary 2.10 (Optimal constant γ)**.** *Assume **(A1-6)** and a constant step-size $\gamma_i = \gamma = \Gamma(n)$, for $1 \leqslant i \leqslant n$:*

1. *If $\frac{\alpha-1}{2\alpha} < r$ and $\Gamma(n) = \gamma_0\, n^{\frac{-2\alpha\min\{r,1\}-1+\alpha}{2\alpha\min\{r,1\}+1}}$, $\gamma_0 R^2 \leqslant 1/4$, we have:*

$$\mathbb{E}\left(\|\bar{g}_n - g_{\mathcal{H}}\|^2_{L^2_{\rho_X}}\right) \leqslant A\, n^{-\frac{2\alpha\min\{r,1\}}{2\alpha\min\{r,1\}+1}}. \tag{2.5}$$

   with $A = 4\left(1 + (\gamma_0 s^2)^{\frac{1}{\alpha}}\right)\sigma^2 + \frac{4(1+o(1))}{\gamma_0^{2r}}||L_K^{-r} g_{\mathcal{H}}||^2_{L^2_{\rho_X}}$.

2. *If $0 < r < \frac{\alpha-1}{2\alpha}$, with $\Gamma(n) = \gamma_0$ is constant, $\gamma_0 R^2 \leqslant 1/4$, we have:*

$$\mathbb{E}\left(\|\bar{g}_n - g_{\mathcal{H}}\|^2_{L^2_{\rho_X}}\right) \leqslant A\, n^{-2r}, \tag{2.6}$$

   with the same constant A.

We can make the following observations:

- **Limit conditions:** Assumption **(A4)**, gives us some kind of "position" of the objective function with respect to our reproducing kernel Hilbert space. If $r \geqslant 1/2$ then $g_{\mathcal{H}} \in \mathcal{H}$. That means the regression function truly lies in the space in which we are looking for an approximation. However, it is neither necessary to get the convergence result, which holds for any $r > 0$, nor to get the optimal rate (see definition in Section 2.4.2), which is also true for $\frac{\alpha-1}{2\alpha} < r < 1$.

- **Evolution with r and α:** As it has been noticed above, a bigger α or r would be a stronger assumption. It is thus natural to get a rate which improves with a bigger α or r: the function $(\alpha, r) \mapsto \frac{2\alpha r}{2\alpha r+1}$ is increasing in both parameters.

- The quantity $o(1)$ in Equation (2.5) stands for $(\gamma_0 s^2 n^{-2\alpha^2 r+1})^{\frac{2r-1}{\alpha}}$ if $r \geqslant 1/2$ (0 otherwise) and is a quantity which decays to 0.

- **Different regions:** in Figure 2.1(a), we plot in the plan of coordinates α, δ (with $\delta = 2\alpha r + 1$) our limit conditions concerning our assumptions, that is, $r = 1 \Leftrightarrow \delta = 2\alpha + 1$ and $\frac{\alpha-1}{2\alpha} = r \Leftrightarrow \alpha = \delta$. The region between the two green lines is the region for which the optimal rate of estimation is reached. The magenta dashed lines stands for $r = 1/2$, which has appeared to be meaningless in our context.

  The region $\alpha \geqslant \delta \Leftrightarrow \frac{\alpha-1}{2\alpha} > r$ corresponds to a situation where regularized empirical risk minimization would still be optimal, but with a regularization parameter λ that decays faster than $1/n$, and thus, our corresponding step-size $\gamma = 1/(n\lambda)$ would not be bounded as a function of n. We thus saturate our step-size to a constant and the generalization error is dominated by the bias term.

  The region $\alpha \leqslant (\delta-1)/2 \Leftrightarrow r > 1$ corresponds to a situation where regularized empirical risk minimization reaches a saturating behaviour. In our stochastic approximation context, the variance term dominates.

2.3.5 Online setting

We now consider the second case when the sequence of step-sizes does not depend on the number of samples we want to use (online setting).

The computations are more tedious in such a situation so that we will only state asymptotic theorems in order to understand the similarities and differences between the finite horizon setting and the online setting, especially in terms of limit conditions.

Theorem 2.11 (Complete bound, $(\gamma_n)_n$ online)**.** *Assume **(A1-6)**, assume for any i, $\gamma_i = \frac{\gamma_0}{i^\zeta}$, $\gamma_0 R^2 \leqslant 1/2$:*

- *If $0 < r(1-\zeta) < 1$, if $0 < \zeta < \frac{1}{2}$ then*

$$\mathbb{E}\|\bar{g}_n - g_{\mathcal{H}}\|^2_{L^2_{\rho_X}} = O\left(\frac{\sigma^2 (s^2\gamma_n)^{\frac{1}{\alpha}}}{n^{1-\frac{1}{\alpha}}}\right) + O\left(\frac{||L_K^{-r} g_{\mathcal{H}}||^2_{L^2_{\rho_X}}}{(n\gamma_n)^{2r}}\right). \tag{2.7}$$

- *If $0 < r(1-\zeta) < 1$, $\frac{1}{2} < \zeta$*

$$\mathbb{E}\|\bar{g}_n - g_{\mathcal{H}}\|^2_{L^2_{\rho_X}} = O\left(\frac{\sigma^2 (s^2\gamma_n)^{\frac{1}{\alpha}}}{n^{1-\frac{1}{\alpha}}} \frac{1}{n\gamma_n^2}\right) + O\left(\frac{||L_K^{-r} g_{\mathcal{H}}||^2_{L^2_{\rho_X}}}{(n\gamma_n)^{2r}}\right). \tag{2.8}$$

The constants in the $O(\cdot)$ notations only depend on γ_0 and α.

Theorem 2.11 is proved in Section A.4.4. In the first case, the main bias and variance terms are the same as in the finite horizon setting, and so is the optimal choice of ζ. However in the second case, the variance term behaviour changes: it does not decrease anymore when ζ increases beyond $1/2$. Indeed, in such a case our constant averaging procedure puts too much weight on the first iterates, thus we do not improve the variance bound by making the learning rate decrease faster. Other type of averaging, as proposed for example in Lacoste-Julien et al. (2012), could help to improve the bound.

Moreover, the constraint $\zeta < 1/2$, to avoid saturation, changes a bit the regions where we get the optimal rate (see Figure 2.1(b)), and we have the following corollary:

Corollary 2.12 (Optimal decreasing γ_n)**.** *Assume **(A1-6)** (in this corollary, $O(\cdot)$ stands for a constant depending on $\alpha, ||L_K^{-r} g_{\mathcal{H}}||_{L^2_{\rho_X}}, s, \sigma^2, \gamma_0$ and universal constants):*

1. *If $\frac{\alpha-1}{2\alpha} < r < \frac{2\alpha-1}{2\alpha}$, with $\gamma_n = \gamma_0 n^{\frac{-2\alpha r-1+\alpha}{2\alpha r+1}}$ for any $n \geqslant 1$ we get the rate:*

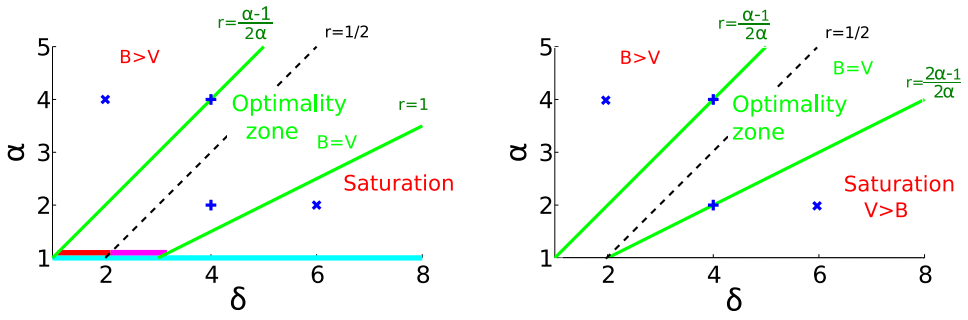
$$\mathbb{E}\|\bar{g}_n - g_{\mathcal{H}}\|^2_{L^2_{\rho_X}} = O\left(n^{-\frac{2\alpha r}{2\alpha r+1}}\right). \tag{2.9}$$

2. *If $\frac{2\alpha-1}{2\alpha} < r$, with $\gamma_n = \gamma_0 n^{-1/2}$ for any $n \geqslant 1$, we get the rate:*

$$\mathbb{E}\|\bar{g}_n - g_{\mathcal{H}}\|^2_{L^2_{\rho_X}} = O\left(n^{-\frac{2\alpha-1}{2\alpha}}\right). \tag{2.10}$$

3. *If $0 < r < \frac{\alpha-1}{2\alpha}$, with $\gamma_n = \gamma_0$ for any $n \geqslant 1$, we get the rate given in (2.6). Indeed the choice of a constant learning rate naturally results in an online procedure.*

This corollary is directly derived from Theorem 2.11, balancing the two main terms. The only difference with the finite horizon setting is the shrinkage of the optimality region as the condition $r < 1$ is replaced by $r < \frac{2\alpha-1}{2\alpha} < 1$ (see Figure 2.1(b)). In the next section, we relate our results to existing work.



(a) Finite Horizon (b) Online

Figure 2.1: Behaviour of convergence rates: (left) finite horizon and (right) online setting. We describe in the (α, δ) plan (with $\delta = 2\alpha r + 1$) the different optimality regions: between the two green lines, we achieve the optimal rate. On the left plot the red (respectively magenta and cyan) lines are the regions for which Zhang (2004) (respectively Tarrès and Yao (2014) and Ying and Pontil (2008)) proved to achieve the overall optimal rate (which may only be the case if $\alpha = 1$). The four blue points match the coordinates of the four couples (α, δ) that will be used in our simulations: they are spread over the different optimality regions.

2.4 Links with existing results

In this section, we relate our results from the previous section to existing results.

2.4.1 Euclidean spaces

Recently, Bach and Moulines (2013) showed that for least-squares regression, averaged stochastic gradient descent achieved a rate of $O(1/n)$, in a finite-dimensional Hilbert space (Euclidean space), under the same assumptions as above (except the first one of course), which is replaced by:

(A1-f) $\mathcal{H}$ is a d-dimensional Euclidean space.

They showed the following result:

Proposition 2.13 (Finite-dimensions (Bach and Moulines, 2013))**.** *Assume **(A1-f), (A2-6)**. Then for* $\gamma = \frac{1}{4R^2}$,

$$\mathbb{E}\left[\varepsilon\left(\overline{g}_n\right) - \varepsilon(g_{\mathcal{H}})\right] \leqslant \frac{4}{n}\left[\sigma\sqrt{d} + R\|g_{\mathcal{H}}\|_{\mathcal{H}}\right]^2. \tag{2.11}$$

We show that we can deduce such a result from Theorem 2.9 (and even with comparable constants). Indeed under **(A1-f)** we have:

- If $\mathbb{E}\left[||x_n||^2\right] \leqslant R^2$ then $\Sigma \preccurlyeq R^2 I$ and **(A3)** is true for any $\alpha \geq 1$ with $s^2 = R^2 d^\alpha$. Indeed $\lambda_i \leqslant R^2$ if $i \leqslant d$ and $\lambda_i = 0$ if $i > d+1$ so that for any $\alpha > 1, i \in \mathbb{N}^*$, $\lambda_i \leqslant R^2 \frac{d^\alpha}{i^\alpha}$.
- As we are in a finite-dimensional space **(A4)** is true for $r = 1/2$ as $||T^{-1/2} g_{\mathcal{H}}||^2_{\mathcal{L}^2_{\rho_X}} = ||g_{\mathcal{H}}||^2_{\mathcal{H}}$.

Under such remarks, the following corollary may be deduced from Theorem 2.9:

Corollary 2.14. *Assume **(A1-f), (A2-6)**, then for any $\alpha > 1$, with $\gamma R^2 \leqslant 1/4$:*

$$\mathbb{E}\|\bar{g}_n - g_{\mathcal{H}}\|^2_{L^2_{\rho_X}} \leqslant \frac{4\sigma^2}{n}\left(1 + (R^2\gamma d^\alpha n)^{\frac{1}{\alpha}}\right) + 4\frac{\|g_{\mathcal{H}}\|^2_{\mathcal{H}}}{n\gamma}.$$

So that, when $\alpha \to \infty$,

$$\mathbb{E}\left[\varepsilon\left(\overline{g}_n\right) - \varepsilon(g_{\mathcal{H}})\right] \leqslant \frac{4}{n}\left(\sigma\sqrt{d} + R\|g_{\mathcal{H}}\|_{\mathcal{H}}\frac{1}{\sqrt{\gamma R^2}}\right)^2.$$

This bound is easily comparable to (2.11) and shows that our more general analysis has not lost too much. Moreover our learning rate is proportional to $n^{\frac{-1}{2\alpha+1}}$ with $r = 1/2$, so tends to behave like a constant when $\alpha \to \infty$, which recovers the constant step set-up from Bach and Moulines (2013).

Moreover, the result can be extended in the following more general corollary of our Theorem 2.9:

Corollary 2.15. *Assume **(A1-f), (A2-6)**, and $||\Sigma^{-q} g_{\mathcal{H}}||^2_{\mathcal{H}} = ||\Sigma^{-(q+1/2)} g_{\mathcal{H}}||^2_{L^2_{\rho_X}} < \infty$, for some $q \in [-1/2; 1/2]$, then:*

$$\mathbb{E}\left[\varepsilon\left(\overline{g}_n\right) - \varepsilon(g_{\mathcal{H}})\right] \leqslant 16\frac{\sigma^2 \operatorname{tr}(\Sigma^{1/\alpha})(\gamma n)^{1/\alpha}}{n} + 8R^{4(q+1/2)}\frac{||\Sigma^{-q} g_{\mathcal{H}}||^2_{\mathcal{H}}}{(n\gamma R^2)^{2(q+1/2)}}.$$

Such a result is derived from Theorem 2.9 and with the stronger assumption $\operatorname{tr}(\Sigma^{1/\alpha}) < \infty$ clearly satisfied in finite dimension, and with $r = q + 1/2$. Note that the result above is true for all values of $\alpha \geqslant 1$ and all $q \geqslant -1/2$ (for the ones with infinite $||\Sigma^{-(q+1/2)} g_{\mathcal{H}}||^2_{L^2_{\rho_X}}$, the statement is trivial). This shows that we may take the infimum over all possible $\alpha \leqslant 1$ and $q \geqslant 0$, showing adaptivity of the estimator to the spectral decay of Σ and the smoothness of the optimal prediction function $g_{\mathcal{H}}$.

Thus with $\alpha \to \infty$, we obtain:

Corollary 2.16. *Assume **(A1-f), (A2-6)**, and $||\Sigma^{-q} g_{\mathcal{H}}||^2_{\mathcal{H}} = ||\Sigma^{-(q+1/2)} g_{\mathcal{H}}||^2_{L^2_{\rho_X}} < \infty$, for some $q \in [-1/2; 1/2]$, then:*

$$\mathbb{E}\left[\varepsilon\left(\overline{g}_n\right) - \varepsilon(g_{\mathcal{H}})\right] \leqslant 16\frac{\sigma^2 d}{n} + 8R^{4(q+1/2)}\frac{||\Sigma^{-q} g_*||^2_{\mathcal{H}}}{(n\gamma R^2)^{2(q+1/2)}}.$$

When $q = 1/2$, we get the following bound:

$$\mathbb{E}\left[\varepsilon\left(\overline{g}_n\right) - \varepsilon(g_{\mathcal{H}})\right] \leqslant 16\frac{\sigma^2 d}{n} + 8R^4\frac{\|\Sigma^{-1/2} g_{\mathcal{H}}\|^2}{(\gamma R^2)^2 n^2}. \tag{2.12}$$

which means that in finite dimension, the initial conditions are asymptotically forgotten at speed $1/n^2$. Moreover, we can make the following remarks:

- The constants 16 and 8 come from the upper bounds $(a+b)^2 \leqslant 2(a^2+b^2)$ and $1 + 1/\sqrt{d} \leqslant 2$ and are thus non optimal.
- We can also derive from Corollary 2.15, with $\alpha = 1$, $q = 0$, and $\gamma \propto n^{-1/2}$, we recover the rate $O(n^{-1/2})$ (where the constant does not depend on the dimension d of the Euclidean space). Such a rate was described, e.g., in Nemirovski et al. (2009).

Note that linking our work to the finite-dimensional setting is made using the fact that our assumption **(A3)** is true for any $\alpha > 1$.

2.4.2 Optimal rates of estimation

In some situations, our stochastic approximation framework leads to "optimal" rates of prediction in the following sense. In (Caponnetto and De Vito, 2007, Theorem 2) a minimax lower bound was given: let $\mathcal{P}(\alpha, r)$ $(\alpha > 1, r \in [1/2, 1])$ be the set of all probability measures ρ on $\mathcal{X} \times \mathcal{Y}$, such that:

- $|y| \leqslant M_\rho$ almost surely,
- $T^{-r} g_\rho \in L^2_{\rho_X}$,
- the eigenvalues $(\mu_j)_{j \in \mathbb{N}}$ arranged in a non increasing order, are subject to the decay $\mu_j = O(j^{-\alpha})$.

Then the following minimax lower rate holds:

$$\liminf_{n \to \infty} \inf_{g_n} \sup_{\rho \in \mathcal{P}(b,r)} \mathbb{P}\left\{ \varepsilon(g_n) - \varepsilon(g_\rho) > C n^{-2r\alpha/(2r\alpha+1)} \right\} = 1,$$

for some constant $C > 0$ where the infimum in the middle is taken over all algorithms as a map $((x_i, y_i)_{1 \leqslant i \leqslant n}) \mapsto g_n \in \mathcal{H}$.

When making assumptions **(a3-4)**, the assumptions regarding the prediction problem (*i.e.*, the optimal function g_ρ) are summarized in the decay of the components of g_ρ in an orthonormal basis, characterized by the constant δ. Here, the minimax rate of estimation (see, e.g., Johnstone (1994)) is $O(n^{-1+1/\delta})$ which is the same as $O(n^{-2r\alpha/(2r\alpha+1)})$ with the identification $\delta = 2\alpha r + 1$.

That means the rate we get is optimal for $\frac{\alpha-1}{2\alpha} < r < 1$ in the finite horizon setting, and for $\frac{\alpha-1}{2\alpha} < r < \frac{2\alpha-1}{2\alpha}$ in the online setting. This is the region between the two green lines on Figure 2.1.

2.4.3 Regularized stochastic approximation

It is interesting to link our results to what has been done in Yao (2006) and Tarrès and Yao (2014) in the case of regularized least-mean-squares, so that the recursion is written:

$$g_n \quad = \quad g_{n-1} - \gamma_n \left((g_{n-1}(x_n) - y_n) K_{x_n} + \lambda_n g_{n-1} \right)$$

with $(g_{n-1}(x_n) - y_n) K_{x_n} + \lambda_n g_{n-1}$ an unbiased gradient of $\frac{1}{2}\mathbb{E}_\rho\left[(g(x) - y)^2\right] + \frac{\lambda_n}{2}||g||^2$. In Tarrès and Yao (2014) the following result is proved (*Remark 2.8* following *Theorem C*):

Theorem 2.17 (Regularized, non averaged stochastic gradient(Tarrès and Yao, 2014))**.** *Assume that $T^{-r} g_\rho \in L^2_{\rho_X}$ for some $r \in [1/2, 1]$. Assume the kernel is bounded and $\mathcal{Y}$ compact. Then with probability at least $1 - \kappa$, for all $t \in \mathbb{N}$,*

$$\varepsilon(g_n) - \varepsilon(g_\rho) \leqslant O_\kappa \left(n^{-2r/(2r+1)} \right).$$

Where O_κ stands for a constant which depends on κ.

No assumption is made on the covariance operator beyond being trace class, but only on $\|T^{-r} g_\rho\|_{L^2_{\rho_X}}$ (thus no assumption **(A3)**). A few remarks may be made:

1. They get almost-sure convergence, when we only get convergence in expectation. We could perhaps derive a.s. convergence by considering moment bounds in order to be able to derive convergence in high probability and to use Borel-Cantelli lemma.
2. They only assume $\frac{1}{2} \leqslant r \leqslant 1$, which means that they assume the regression function to lie in the RKHS.

2.4.4 Unregularized stochastic approximation

In Ying and Pontil (2008), Ying and Pontil studied the same unregularized problem as we consider, under assumption **(A4)**. They obtain the same rates as above ($n^{-2r/(2r+1)}\log(n)$) in both online case (with $0 \leqslant r \leqslant \frac{1}{2}$) and finite horizon setting ($0 < r$).

They led as an open problem to improve bounds with some additional information on some decay of the eigenvalues of T, a question which is answered here.

Moreover, Zhang (2004) also studies stochastic gradient descent algorithms in an unregularized setting, also with averaging. As described in Ying and Pontil (2008), his result is stated in the linear kernel setting but may be extended to kernels satisfying $\sup_{x\in\mathcal{X}} K(x,x) \leqslant R^2$. Ying and Pontil derive from Theorem 5.2 in Zhang (2004) the following proposition:

Proposition 2.18 (Short step-sizes (Zhang, 2004))**.** *Assume we consider the algorithm defined in Section 2.3.2 and output $\bar{g}_n$ defined by equation (2.3). Assume the kernel K satisfies $\sup_{x\in\mathcal{X}} K(x,x) \leqslant R^2$. Finally assume g_ρ satisfies assumption **(A4)** with $0 < r < 1/2$. Then in the finite horizon setting, with $\Gamma(n) = \frac{1}{4R^2} n^{-\frac{2r}{2r+1}}$, we have:*

$$\mathbb{E}\left[\varepsilon\left(\bar{g}_n\right) - \varepsilon(g_{\mathcal{H}})\right] = O\left(n^{-\frac{2r}{2r+1}}\right).$$

Moreover, note that we may derive their result from Corollary 2.10. Indeed, using $\Gamma(n) = \gamma_0 n^{\frac{-2r}{2r+1}}$, we get a bias term which is of order $n^{\frac{-2r}{2r+1}}$ and a variance term of order $n^{-1+\frac{1}{2r\alpha+\alpha}}$ which is smaller. Our analysis thus recovers their convergence rate with their step-size. Note that this step-size is significantly smaller than ours, and that the resulting bound is worse (but their result holds in more general settings than least-squares). See more details in Section 2.4.5.

2.4.5 Summary of results

All three algorithms are variants of the following:

$$\begin{aligned} g_0 &= 0 \\ \forall n \geqslant 1, \quad g_n &= (1-\lambda_n) g_{n-1} - \gamma_n (y_n - g_{n-1}(x_n)) K_{x_n}. \end{aligned}$$

But they are studied under different settings, concerning regularization, averaging, assumptions: we sum up in Table 2.1 the settings of each of these studies. For each of them, we consider the finite horizon settings, where results are generally better.

Algorithm type	Ass. **(A3)**	Ass. **(A4)**	γ_n	λ_n	Rate	Conditions
This chapter	yes	yes	1	0	n^{-2r}	$r < \frac{\alpha-1}{2\alpha}$
This chapter	yes	yes	$n^{-\frac{2\alpha r+1-\alpha}{2\alpha r+1}}$	0	$n^{\frac{-2\alpha r}{2\alpha r+1}}$	$\frac{\alpha-1}{2\alpha} < r < 1$
This chapter	yes	yes	$n^{-\frac{\alpha+1}{2\alpha+1}}$	0	$n^{\frac{-2\alpha}{2\alpha+1}}$	$r > 1$
Zhang (2004)	no	yes	$n^{\frac{-2r}{2r+1}}$	0	$n^{\frac{-2r}{2r+1}}$	$0 \leqslant r \leqslant \frac{1}{2}$
Tarrès and Yao (2014)	no	yes	$n^{\frac{-2r}{2r+1}}$	$n^{\frac{-1}{2r+1}}$	$n^{\frac{-2r}{2r+1}}$	$\frac{1}{2} \leqslant r \leqslant 1$
Ying and Pontil (2008)	no	yes	$n^{\frac{-2r}{2r+1}}$	0	$n^{\frac{-2r}{2r+1}}$	$r > 0$

Table 2.1: Summary of assumptions and results (step-sizes, rates and conditions) for our three regions of convergence and related approaches. We focus on finite-horizon results.

We can make the following observations:

- **Dependence of the convergence rate on** α: For learning with any kernel with $\alpha > 1$ we strictly improve the asymptotic rate compared to related methods that only assume summability of eigenvalues: indeed, the function $x \mapsto x/(x+1)$ is increasing on $\mathbb{R}^+$. If we consider a given optimal prediction function and a given kernel with which we are going to learn the function, considering the decrease in eigenvalues allows to adapt the step-size and obtain an improved learning rate. Namely, we improved the previous rate $\frac{-2r}{2r+1}$ up to $\frac{-2\alpha r}{2\alpha r+1}$.
- **Worst-case result in** r: in the setting of assumptions **(a3,4)**, given δ, the optimal rate of convergence is known to be $O(n^{-1+1/\delta})$, where $\delta = 2\alpha r + 1$. We thus get the optimal rate, as soon as $\alpha < \delta < 2\alpha + 1$, while the other algorithms get the suboptimal rate $n^{\frac{\delta-1}{\delta+\alpha-1}}$ under various conditions. Note that this sub-optimal rate becomes close to the optimal rate when α is close to one, that is, in the *worst-case* situation. Thus, in the worst-case (α arbitrarily close to one), all methods behave similarly, but for any particular instance where $\alpha > 1$, our rates are better.
- **Choice of kernel**: in the setting of assumptions **(a3,4)**, given δ, in order to get the optimal rate, we may choose the kernel (*i.e.*, α) such that $\alpha < \delta < 2\alpha + 1$ (that is neither too big, nor too small), while other methods need to choose a kernel for which α is as close to one as possible, which may not be possible in practice.
- **Improved bounds**: Ying and Pontil (2008) only give asymptotic bounds, while we have exact constants for the finite horizon case. Moreover there are some logarithmic terms in Ying and Pontil (2008) which disappear in our analysis.
- **Saturation**: our method does saturate for $r > 1$, while the non-averaged framework of Ying and Pontil (2008) does not (but does not depend on the value of α). We conjecture that a proper non-uniform averaging scheme (that puts more weight on the latest iterates), we should get the best of both worlds.

2.5 Experiments on artificial data

Following Ying and Pontil (2008), we consider synthetic examples with smoothing splines on the circle, where our assumptions **(A3-4)** are easily satisfied.

2.5.1 Splines on the circle

The simplest example to match our assumptions may be found in Wahba (1990). We consider $Y = g_\rho(X) + \varepsilon$, with $X \sim \mathcal{U}[0;1]$ is a uniform random variable in $[0,1]$, and g_ρ in a particular RKHS (which is actually a Sobolev space).

Let $\mathcal{H}$ be the collection of all zero-mean periodic functions on $[0;1]$ of the form

$$f : t \mapsto \sqrt{2}\sum_{i=1}^{\infty} a_i(f)\cos(2\pi i t) + \sqrt{2}\sum_{i=1}^{\infty} b_i(f)\sin(2\pi i t),$$

with

$$\|f\|_{\mathcal{H}}^2 = \sum_{i=1}^{\infty}(a_i(f)^2 + b_i(f)^2)(2\pi i)^{2m} < \infty.$$

This means that the m-th derivative of f, $f^{(m)}$ is in $\mathcal{L}^2([0;1])$. We consider the inner product:

$$\langle f, g\rangle_{\mathcal{H}} = \sum_{i=1}^{\infty}(2\pi i)^{2m}\left(a_i(f)a_i(g) + b_i(f)b_i(g)\right).$$

It is known that $\mathcal{H}$ is an RKHS and that the reproducing kernel $R_m(s,t)$ for $\mathcal{H}$ is

$$\begin{aligned} R_m(s,t) &= \sum_{i=1}^{\infty}\frac{2}{(2\pi i)^{2m}}[\cos(2\pi i s)\cos(2\pi i t) + \sin(2\pi i s)\sin(2\pi i t)] \\ &= \sum_{i=1}^{\infty}\frac{2}{(2\pi i)^{2m}}\cos(2\pi i(s-t)). \end{aligned}$$

Moreover the study of Bernoulli polynomials gives a close formula for $R(s,t)$, that is:

$$R_m(s,t) = \frac{(-1)^{m-1}}{(2m)!}B_{2m}\left(\{s-t\}\right),$$

with B_m denoting the m-th Bernoulli polynomial and $\{s-t\}$ the fractional part of $s-t$ (Wahba, 1990).

We can derive the following proposition for the covariance operator which means that our assumption **(A3)** is satisfied for our algorithm in $\mathcal{H}$ when $X \sim \mathcal{U}[0;1]$, with $\alpha = 2m$, and $s = 2(1/2\pi)^m$.

Proposition 2.19 (Covariance operator for smoothing splines)**.** *If $X \sim \mathcal{U}[0;1]$, then in $\mathcal{H}$:*

1. *the eigenvalues of Σ are all of multiplicity 2 and are $\lambda_i = (2\pi i)^{-2m}$,*
2. *the eigenfunctions are $\phi_i^c : t \mapsto \sqrt{2}\cos(2\pi i t)$ and $\phi_i^s : t \mapsto \sqrt{2}\sin(2\pi i t)$.*

Proof. For ϕ_i^c we have (a similar argument holds for ϕ_i^s):

$$\begin{aligned} T(\phi_i^c)(s) &= \int_0^1 R(s,t)\sqrt{2}\cos(2\pi i t)dt \\ &= \left(\int_0^1 \frac{2}{(2i\pi)^{2m}}\sqrt{2}\cos(2\pi i t)^2 dt\right)\cos(2\pi i s) = \lambda_i\sqrt{2}\cos(2\pi i s) \end{aligned}$$

$$= \lambda_i \phi_i^c(s).$$

It is well known that $(\phi_i^c, \phi_i^s)_{i\geqslant 0}$ is an orthonormal system (the Fourier basis) of the functions in $L^2([0;1])$ with zero mean, and it is easy to check that $((2i\pi)^{-m}\phi_i^c, (2i\pi)^{-m}\phi_i^s)_{i\geqslant 1}$ is an orthonormal basis of our RKHS $\mathcal{H}$ (this may also be seen as a consequence of the fact that $T^{1/2}$ is an isometry). □

Finally, considering $g_\rho(x) = B_{\delta/2}(x)$ with $\delta = 2\alpha r + 1 \in 2\mathbb{N}$, our assumption **(A4)** holds. Indeed it implies **(a3-4)**, with $\alpha > 1, \delta = 2\alpha r + 1$, since for any $k \in \mathbb{N}$, $B_k(x) = -2k! \sum_{i=1}^{\infty} \frac{\cos\left(2i\pi x - \frac{k\pi}{2}\right)}{(2i\pi)^k}$ (see, e.g., Abramowitz and Stegun (1964)).

We may notice a few points:

1. Here the eigenvectors do not depend on the kernel choice, only the re-normalization constant depends on the choice of the kernel. Especially the eigenbasis of T in $L^2_{\rho_X}$ does not depend on m. That can be linked with the previous remarks made in Section 2.4.

2. Assumption **(A3)** defines here the size of the RKHS: the smaller $\alpha = 2m$ is, the bigger the space is, the harder it is to learn a function.

In the next section, we illustrate on such a toy model our main results and compare our learning algorithm to the algorithms by Ying and Pontil (2008), Tarrès and Yao (2014) and Zhang (2004) .

2.5.2 Experimental set-up

We use $g_\rho(x) = B_{\delta/2}(x)$ with $\delta = 2\alpha r + 1$, as proposed above, with $B_1(x) = x - \frac{1}{2}$, $B_2(x) = x^2 - x + \frac{1}{6}$ and $B_3(x) = x^3 - \frac{3}{2}x^2 + \frac{1}{2}x$.

We give in Figure 2.2 the functions used for simulations in a few cases that span our three regions. We also remind the choice of γ proposed for the 4 algorithms. We always use the finite horizon setting.

r	α	δ	K	g_ρ	$\frac{\log(\gamma)}{\log(n)}$ (this chapter)	$\frac{\log(\gamma)}{\log(n)}$ (previous)
0.75	2	4	R_1	B_2	$-1/2 = -0.5$	$-3/5 = -0.6$
0.375	4	4	R_2	B_2	0	$-3/7 \simeq -0.43$
1.25	2	6	R_1	B_3	$-3/7 \simeq -0.43$	$-5/7 \simeq -0.71$
0.125	4	2	R_2	B_1	0	$-1/5 = -0.2$

Table 2.2: Different choices of the parameters α, r and the corresponding convergence rates and step-sizes. The (α, δ) coordinates of the four choices of couple "(kernel, objective function)" are mapped on Figure 2.1. They are spread over the different optimality regions.

2.5.3 Optimal learning rate for our algorithm

In this section, we empirically search for the best choice of a finite horizon learning rate, in order to check if it matches our prediction. For a certain number of values for n, distributed

exponentially between 1 and $10^{3.5}$, we look for the best choice $\Gamma_{\text{best}}(n)$ of a constant learning rate for our algorithm up to horizon n. In order to do that, for a large number of constants $C_1, \cdots, C_p$, we estimate the expectation of error $\mathbb{E}[\varepsilon(\overline{g}_n(\gamma = C_i)) - \varepsilon(g_\rho)]$ by averaging over 30 independent sample of size n, then report the constant giving minimal error as a function of n in Figure 2.2. We consider here the situation $\alpha = 2, r = 0.75$. We plot results in a logarithmic scale, and evaluate the asymptotic decrease of $\Gamma_{\text{best}}(n)$ by fitting an affine approximation to the second half of the curve. We get a slope of -0.51, which matches our choice of -0.5 from Corollary 2.10. Although, our theoretical results are only upper-bounds, we conjecture that our proof technique also leads to lower-bounds in situations where assumptions (**a3-4**) hold (like in this experiment).

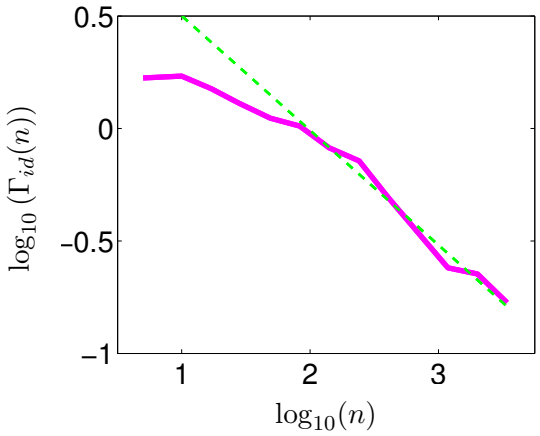
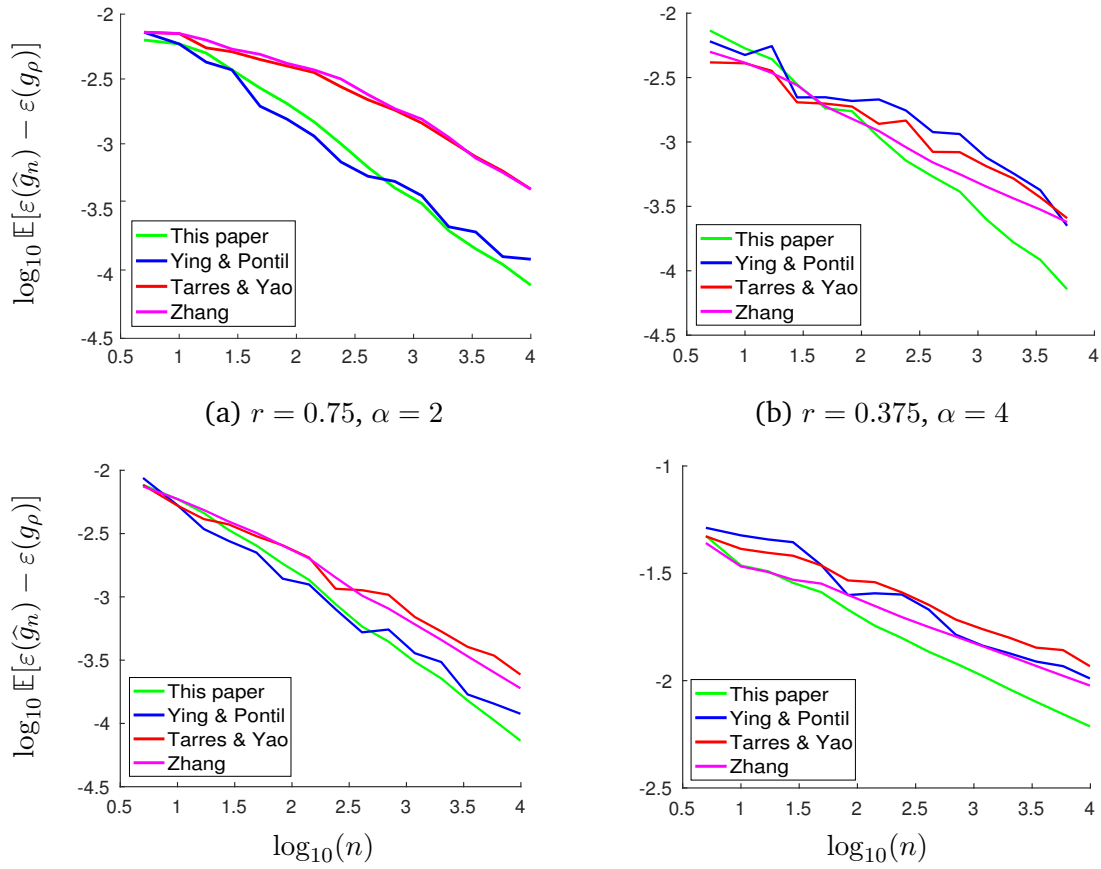


Figure 2.2: Optimal learning rate $\Gamma_{\text{best}}(n)$ for our algorithm in the finite horizon setting (plain magenta). The dashed green curve is a first order affine approximation of the second half of the magenta curve.

2.5.4 Comparison to competing algorithms

In this section, we compare the convergence rates of the four algorithms described in Section 2.4.5. We consider the different choices of (r, α) as described in Table 2.2 in order to go all over the different optimality situations. The main properties of each algorithm are described in Table 2.1. However we may note:

- For our algorithm, $\Gamma(n)$ is chosen accordingly with Corollary 2.10, with $\gamma_0 = \frac{1}{R^2}$.
- For Ying and Pontil's algorithm, accordingly to Theorem 6 in Ying and Pontil (2008), we consider $\Gamma(n) = \gamma_0 n^{-\frac{2r}{2r+1}}$. We choose $\gamma_0 = \frac{1}{R^2}$ which behaves better than the proposed $\frac{r}{64(1+R^4)(2r+1)}$.
- For Tarrès and Yao's algorithm, we refer to Theorem C in Tarrès and Yao (2014), and consider $\Gamma(n) = a\,(n_0 + n)^{-\frac{2r}{2r+1}}$ and $\Lambda(n) = \frac{1}{a}\,(n_0 + n)^{-\frac{1}{2r+1}}$. The theorem is stated for all $a \geqslant 4$: we choose $a = 4$.
- For Zhangl's algorithm, we refer to Part 2.2 in Ying and Pontil (2008), and choose $\Gamma(n) = \gamma_0 n^{-\frac{2r}{2r+1}}$ with $\gamma_0 = \frac{1}{R^2}$ which behaves better than the proposed choice $\frac{1}{4(1+R^2)}$.



(a) $r = 0.75$, $\alpha = 2$ (b) $r = 0.375$, $\alpha = 4$

(c) $r = 1.25$, $\alpha = 2$ (d) $r = 0.125$, $\alpha = 4$

Figure 2.3: Comparison between algorithms. We have chosen parameters in each algorithm accordingly with description in Section 2.4.5, especially for the choices of γ_0. The y-axis is $\log_{10}\left(\mathbb{E}[\varepsilon(\hat{g}_n) - \varepsilon(g_\rho)]\right)$, where the final output $\hat{g}_n$ may be either $\overline{g}_n$ (This chapter, Zhang) or g_n(Ying & Pontil, Yao & Tarres). This expectation is computed by averaging over 15 independent samples.

Finally, we sum up the rates that were both predicted and derived for the four algorithms in the four cases for (α, δ) in Table 2.3. It appears that (a) we approximately match the predicted rates in most cases (they would if n was larger), (b) our rates improve on existing work.

2.6 Conclusion

In this chapter, we have provided an analysis of averaged unregularized stochastic gradient methods for kernel-based least-squares regression. Our novel analysis allowed us to consider larger step-sizes, which in turn lead to optimal estimation rates for many settings of eigenvalue decay of the covariance operators and smoothness of the optimal prediction function. Moreover, we have worked on a more general setting than previous work, that includes most interesting cases of positive definite kernels.

In the finite horizon setting, the convergence rate remains sub-optimal in two situations: when the combination of the kernel and the function result in a function which is “too

	$r = 0.75$ $\alpha = 2$	$r = 0.375$ $\alpha = 4$	$r = 1.25$ $\alpha = 2$	$r = 0.125$ $\alpha = 4$
Predicted rate (our algo.)	**-0.75**	**-0.75**	**-0.8**	**-0.25**
Effective rate (our algo.)	**-0.7**	**-0.71**	**-0.69**	**-0.29**
Predicted rate (YP)	-0.6	-0.43	-0.71	-0.2
Effective rate (YP)	-0.53	-0.5	-0.63	-0.22
Predicted rate (TY)	-0.6			
Effective rate (TY)	-0.48	-0.39	-0.43	-0.2
Predicted rate (Z)		-0.43		-0.2
Effective rate (Z)	-0.53	-0.43	-0.41	-0.21

Table 2.3: Predicted and effective rates (asymptotic slope of the log-log plot) for the four different situations. We leave empty cases when the set-up does not come with existing guarantees: most algorithms seem to exhibit the expected behaviour even in such cases.

smooth" (precisely when $r > 1$), then the uniform averaging scheme is responsible from the sub-optimality: indeed, at any iteration, the averaged iterate still "strongly" depends on θ_0, which counts for one n-th of the averaged iterate. The bias cannot decrease faster than n^{-2}. This problem can be addressed using non uniform averaging schemes. For example, one can consider, for $p \in \mathbb{N}$:

$$\bar{\theta}_n^p := \frac{1}{\sum_{k=0}^{n} k^p} \sum_{k=0}^{n} k^p \theta_k \, . \tag{2.13}$$

This non-uniform averaging tends to put more weight on final iterates and thus can forget the initial condition faster in situations that were limiting before: the saturation limit changes. Typically, the bias would then decrease as $n^{2\min(p+1,r)}$ instead of $n^{2\min(1,r)}$, while the bias would be degraded by a constant factor (which depends on T).

Jain et al. (2016) later proposed an analysis of tail averaging, where one considers the uniform averaging over the last half of the iterates. While this averaging scheme cannot be compute "on-the-fly" anymore, it naturally removes the saturation effect:

$$\bar{\theta}_n^{\text{tail}} = \frac{1}{n + 1 - \lfloor n/2 \rfloor} \sum_{k=\lfloor n/2 \rfloor}^{n} \theta_k. \tag{2.14}$$

In such a situation, the regions of optimal convergence would be improved, as depicted in Figure 2.4

On the other hand, the algorithm also behaves sub-optimally on the other extreme situation, in which the optimal function is not only out of the RKHS, but really badly conditioned: then the bias term dominates as the optimal function is "too far away". We partially address this problem in Chapter 3, using acceleration to improve the speed at which initial conditions are forgotten.

The proofs of the results given in this chapter are given in the next Chapter (Ch. A): Section A.1

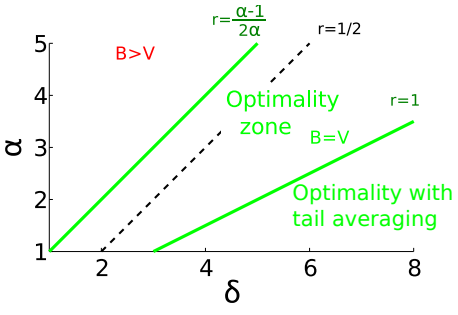


Figure 2.4: Regions of optimal convergence for tail averaging.

contains a short description of minimal assumptions,
and Section A.2 the sketch of the proofs. The following Sections A.3 and A.4 contain the details and might be skipped at first reading.

A

Appendix to Non-parametric Stochastic Approximation with Large Step-sizes

A.1 Minimal assumptions

A.1.1 Definitions

We first define the set of square ρ_X-integrable functions $\mathcal{L}^2_{\rho_X}$:

$$\mathcal{L}^2_{\rho_X} = \left\{ f \ : \mathcal{X} \to \mathbb{R} \ \Big/ \int_{\mathcal{X}} f^2(t) d\rho_X(t) < \infty \right\};$$

we will always make the assumptions that this space is separable (this is the case in most interesting situations. See Thomson et al. (2000) for more details.) $L^2_{\rho_X}$ is its quotient under the equivalence relation given by

$$f \equiv g \Leftrightarrow \int_{\mathcal{X}} (f(t) - g(t))^2 d\rho_X(t) = 0,$$

which makes it a separable Hilbert space (see, e.g., Kolmogorov and Fomin (1999)).

We denote p the canonical projection from $\mathcal{L}^2_{\rho_X}$ into $L^2_{\rho_X}$ such that $p : f \mapsto \tilde{f}$, with $\tilde{f} = \{g \in \mathcal{L}^2_{\rho_X}, \text{ s.t. } f \equiv g\}$.

Under assumptions **A1, A2** or **A1', A2'**, any function in $\mathcal{H}$ in in $\mathcal{L}^2_{\rho_X}$. Moreover, under **A1, A2** the spaces $\mathcal{H}$ and $p(\mathcal{H})$ may be identified, where $p(\mathcal{H})$ is the image of $\mathcal{H}$ via the mapping $p \circ i \ : \mathcal{H} \xrightarrow{i} \mathcal{L}^2_{\rho_X} \xrightarrow{p} L^2_{\rho_X}$, where i is the trivial injection from $\mathcal{H}$ into $\mathcal{L}^2_{\rho_X}$.

A.1.2 Isomorphism

As it has been explained in the main text, the minimization problem will appear to be an approximation problem in $\mathcal{L}^2_{\rho_X}$, for which we will build estimates in $\mathcal{H}$. However, to derive theoretical results, it is easier to consider it as an approximation problem in the Hilbert space $L^2_{\rho_X}$, building estimates in $p(\mathcal{H})$.

We thus need to define a notion of the best estimation in $p(\mathcal{H})$. We first define the closure $\overline{F}$ (with respect to $\|\cdot\|_{L^2_{\rho_X}}$) of any set $F \subset L^2_{\rho_X}$ as the set of limits of sequences in F. The space $\overline{p(\mathcal{H})}$ is a closed and convex subset in $L^2_{\rho_X}$. We can thus define $g_{\mathcal{H}} = \arg\min_{f \in \overline{p(\mathcal{H})}} \varepsilon(g)$, as the orthogonal projection of g_ρ on $\overline{p(\mathcal{H})}$, using the existence of the projection on any closed convex set in a Hilbert space. See Proposition A.1 in Section A.1 for details.

Proposition A.1 (Definition of best approximation function)**.** *Assume **(A1-2)**. The minimum of $\varepsilon(f)$ in $\overline{p(\mathcal{H})}$ is attained at a certain $g_{\mathcal{H}}$ (which is unique and well defined in $L^2_{\rho_X}$).*

Where $\overline{p(\mathcal{H})} = \left\{ f \in L^2_{\rho_X} \ / \ \exists (f_n) \subset p(\mathcal{H}), \|f_n - f\|_{L^2_{\rho_X}} \to 0 \right\}$ is the set of functions f for which we can hope for consistency, *i.e.*, having a sequence $(f_n)_n$ of estimators in $\mathcal{H}$ such that $\varepsilon(f_n) \to \varepsilon(f)$.

The properties of our estimator, especially its rate of convergence will strongly depend on some properties of both the kernel, the objective function and the distributions, which may be seen through the properties of the covariance operator which is defined in the main text. We have defined the covariance operator, $\Sigma : \mathcal{H} \to \mathcal{H}$. In the following, we extend such an operator as an endomorphism $\mathcal{T}$ from $L^2_{\rho_X}$ to $\mathcal{L}^2_{\rho_X}$ and by projection as an endomorphism $T = p \circ \mathcal{T}$ from $L^2_{\rho_X}$ to $L^2_{\rho_X}$. Note that $\mathcal{T}$ is well defined as $\int_{\mathcal{X}} g(t)\ K_t\ d\rho_{\mathcal{X}}(t)$ does not depend on the function g chosen in the class of equivalence of g.

Definition A.2 (Extended covariance operator)**.** *Assume **(A1-2)**. We define the operator $\mathcal{T}$ as follows (this expectation is formally defined as a Bochner expectation in $\mathcal{H}$.):*

$$\begin{array}{rccc} \mathcal{T} & L^2_{\rho_X} & \to & \mathcal{L}^2_{\rho_X} \\ & g & \mapsto & \displaystyle\int_{\mathcal{X}} g(t)\ K_t\ d\rho_{\mathcal{X}}(t), \end{array}$$

so that for any $z \in \mathcal{X}$, $\mathcal{T}(g)(z) = \displaystyle\int_{\mathcal{X}} g(x)\ K(x,z)\ d\rho_{\mathcal{X}}(t) = \mathbb{E}[g(X)K(X,z)]$.

A first important remark is that $\Sigma f = 0$ implies $\langle f, \Sigma f \rangle = \|f\|^2_{L^2_{\rho_X}} = 0$, that is $p(\mathrm{Ker}(\Sigma)) = \{0\}$. However, Σ may not be injective (unless $\|f\|^2_{L^2_{\rho_X}} \Rightarrow f = 0$, which is true when f is continuous and ρ_X has full support). Σ and $\mathcal{T}$ may independently be injective or not.

The operator T (which is an endomorphism of the separable Hilbert space $L^2_{\rho_X}$) can be reduced in some Hilbertian eigenbasis of $L^2_{\rho_X}$. The linear operator $\mathcal{T}$ happens to have an image included in $\mathcal{H}$, and the eigenbasis of T in $L^2_{\rho_X}$ may also be seen as eigenbasis of Σ in $\mathcal{H}$ (See proof in Section A.3.2, Proposition A.18):

Proposition A.3 (Decomposition of Σ)**.** *Assume **(A1-2)**. The image of $\mathcal{T}$ is included in $\mathcal{H}$: $Im(\mathcal{T}) \subset \mathcal{H}$, that is, for any $f \in L^2_{\rho_X}$, $\mathcal{T}f \in \mathcal{H}$. Moreover, for any $i \in I$, $\phi_i^H = \frac{1}{\mu_i}\mathcal{T}\phi_i \in \mathcal{H} \subset \mathcal{L}^2_{\rho_X}$ is a representant for the equivalence class ϕ_i, that is $p(\phi_i^H) = \phi_i$. Moreover $\mu_i^{1/2}\phi_i^H$ is an orthonormal eigen-system of the orthogonal supplement $\mathscr{S}$ of the null space Ker(Σ). That is:*

- $\forall i \in I,\ \Sigma\phi_i^H = \mu_i\phi_i^H$.
- $\mathcal{H} = \mathrm{Ker}(\Sigma) \overset{\perp}{\oplus} \mathscr{S}$.

Such decompositions allow to define $\mathcal{T}^r : L^2_{\rho_X} \to \mathcal{H}$ for $r \geqslant 1/2$. Indeed , completeness allows to define infinite sums which satisfy a Cauchy criterion. See proof in Section A.3.2, Proposition A.19. Note the different condition concerning r in the definitions. For $r \geqslant 1/2$, $T^r = p \circ \mathcal{T}^r$. We need $r \geqslant 1/2$, because $(\mu_i^{1/2}\phi^H)$ is an orthonormal system of $\mathscr{S}$.

Definition A.4 (Powers of $\mathcal{T}$)**.** *We define, for any $r \geqslant 1/2$, $\mathcal{T}^r : L^2_{\rho_X} \to \mathcal{H}$, for any $h \in \mathrm{Ker}(T)$ and $(a_i)_{i\in I}$ such that $\sum_{i\in I} a_i^2 < \infty$, through:*

$$\mathcal{T}^r\left(h + \sum_{i\in I} a_i\phi_i\right) = \sum_{i\in I} a_i\mu_i^r\phi_i^H.$$

We have two decompositions of $L^2_{\rho_X} = \mathrm{Ker}(T) \overset{\perp}{\oplus} S$ and $\mathcal{H} = \mathrm{Ker}(\Sigma) \overset{\perp}{\oplus} \mathscr{S}$. The two orthogonal supplements S and $\mathscr{S}$ happen to be related through the mapping $\mathcal{T}^{1/2}$, as stated in Proposition 2.7: $\mathcal{T}^{1/2}$ is an isomorphism from S into $\mathscr{S}$. It also has he following consequences, which generalizes Corollary 2.6:

Corollary A.5.
- *$T^{1/2}(S) = p(\mathcal{H})$, that is any element of $p(\mathcal{H})$ may be expressed as $T^{1/2}g$ for some $g \in L^2_{\rho_X}$.*
- *For any $r \geqslant 1/2$, $T^r(S) \subset \mathcal{H}$, because $T^r(S) \subset T^{1/2}(S)$, that is, with large powers r, the image of T^r is in the projection of the Hilbert space.*
- *$\forall r > 0$, $\overline{T^r(L^2_{\rho_X})} = S = \overline{T^{1/2}(L^2_{\rho_X})} = \overline{\mathcal{H}}$, because (a) $T^{1/2}(L^2_{\rho_X}) = p(\mathcal{H})$ and (b) for any $r > 0$, $\overline{T^r(L^2_{\rho_X})} = S$. In other words, elements of $\overline{p(\mathcal{H})}$ (on which our minimization problem attains its minimum), may seen as limits (in $L^2_{\rho_X}$) of elements of $T^r(L^2_{\rho_X})$, for any $r > 0$.*
- *$p(\mathcal{H})$ is dense in $L^2_{\rho_X}$ if and only if T is injective.*

A.1.3 Mercer theorem generalized

Finally, although we will not use it afterwards, we can state a generalized version of Mercer's theorem, which does not make any other assumptions than the one required for defining RKHSs.

Proposition A.6 (Kernel decomposition)**.** *Assume **(A1-2)**. We have for all $x, y \in \mathcal{X}$,*

$$K(x,y) = \sum_{i \in I} \mu_i \phi_i^H(x) \phi_i^H(y) + g(x,y),$$

and we have for all $x \in \mathcal{X}$, $\int_{\mathcal{X}} g(x,y)^2 d\rho_X(y) = 0$. Moreover, the convergence of the series is absolute.

We thus obtain a version of Mercer's theorem (see Section A.3.5) *without* any topological assumptions. Moreover, note that (a) $\mathscr{S}$ is also an RKHS, with kernel $(x,y) \mapsto \sum_{i \in I} \mu_i \phi_i^H(x) \phi_i^H(y)$ and (b) that given the decomposition above, the optimization problem in $\mathscr{S}$ and $\mathcal{H}$ have equivalent solutions. Moreover, considering the algorithm below, the estimators we consider will almost surely build equivalent functions (see Section A.3.4). Thus, we could assume without loss of generality that the kernel K is exactly equal to its expansion $\sum_{i \in I} \mu_i \phi_i^H(x) \phi_i^H(y)$.

A.1.4 Complementary (A6) assumption

Under minimal assumptions, we also have to make a complementary moment assumption:

(A6') There exists $R > 0$ and $\sigma > 0$ such that $\mathbb{E}\left[\Xi \otimes \Xi\right] \preccurlyeq \sigma^2 \Sigma$, and $\mathbb{E}(K(X,X) K_X \otimes K_X) \preccurlyeq R^2 \Sigma$ where $\preccurlyeq$ denotes the order between self-adjoint operators.

In other words, for any $f \in \mathcal{H}$, we have: $\mathbb{E}[K(X,X) f(X)^2] \leqslant R^2 \mathbb{E}[f(X)^2]$. Such an assumption is implied by **(A2)**, that is if $K(X,X)$ is almost surely bounded by R^2: this constant can then be understood as the radius of the set of our data points. However, our analysis holds in these more general set-ups where only fourth order moment of $\|K_x\|_{\mathcal{H}} = K(x,x)^{1/2}$ is finite.

A.2 Sketch of the proofs

Our main theorems are Theorem 2.9 and Theorem 2.11, respectively in the finite horizon and in the online setting. Corollaries can be easily derived by optimizing over γ the upper bound given in the theorem.

The complete proof is given in Section A.4. The proof is nearly the same for finite horizon and online setting. It relies on a refined analysis of strongly related recursions in the RKHS and on a comparison between iterates of the recursions (controlling the deviations).

We first present the sketch of the proof for the *finite-horizon setting*:
We want to analyze the error of our sequence of estimators (g_n) such that $g_0 = 0$ and

$$\begin{aligned} g_n &= g_{n-1} - \gamma_n \left[y_n - \langle g_{n-1}, K_{x_n} \rangle_{\mathcal{H}} \right] K_{x_n} \\ g_n &= (I - \gamma K_{x_n} \otimes K_{x_n}) g_{n-1} + \gamma y_n K_{x_n} \\ g_n - g_{\mathcal{H}} &= (I - \gamma \widetilde{K_{x_n} \otimes K_{x_n}})(g_{n-1} - g_{\mathcal{H}}) + \gamma \Xi_n. \end{aligned}$$

Where we have denoted $\Xi_n = (y_n - g_{\mathcal{H}}(x_n)) K_{x_n}$ the residual, which has 0 mean, and $\widetilde{K_{x_n} \otimes K_{x_n}} : L^2_{\rho_X} \to \mathcal{H}$ an a.s. defined extension of $K_{x_n} \otimes K_{x_n} : \mathcal{H} \to \mathcal{H}$, such that $\widetilde{K_{x_n} \otimes K_{x_n}}(f) = f(x_n) K_{x_n}$, that will be denoted for simplicity $K_{x_n} \otimes K_{x_n}$ in this section.

Finally, we are studying a sequence $(\eta_n)_n = (g_n - g_{\mathcal{H}})_n$ defined by:

$$\begin{aligned} \eta_0 &= g_{\mathcal{H}}, \\ \eta_n &= (I - \gamma_n K_{x_n} \otimes K_{x_n}) \eta_{n-1} + \gamma_n \Xi_n. \end{aligned}$$

We first consider splitting this recursion in two simpler recursions η_n^{init} and η_n^{noise} such that $\eta_n = \eta_n^{init} + \eta_n^{noise}$:

- $(\eta_n^{init})_n$ defined by :

  $$\eta_0^{init} = g_{\mathcal{H}} \text{ and } \eta_n^{init} = (I - \gamma K_{x_n} \otimes K_{x_n}) \eta_{n-1}^{init}.$$

  η_n^{init} is the part of $(\eta_n)_n$ which is due to the **initial conditions** (it is equivalent to assuming $\Xi_n \equiv 0$).

- Respectively, let $(\eta_n^{noise})_n$ be defined by :

  $$\eta_0^{noise} = 0 \text{ and } \eta_n^{noise} = (I - \gamma K_{x_n} \otimes K_{x_n}) \eta_{n-1}^{noise} + \gamma \Xi_n.$$

  η_n^{noise} is the part of $(\eta_n)_n$ which is due to the **noise**.

We will bound $\|\eta_n\|$ by $\|\eta_n^{init}\| + \|\eta_n^{noise}\|$ using Minkowski's inequality. *That is how the bias-variance trade-off originally appears.*

Next, we notice that $\mathbb{E}[K_{x_n} \otimes K_{x_n}] = \mathcal{T}$, and thus define "semi-stochastic" versions of the previous recursions by replacing $K_{x_n} \otimes K_{x_n}$ by its expectation:

For the initial conditions: $(\eta_n^{0,init})_{n \in \mathbb{N}}$ so that :

$$\eta_0^{0,init} = g_{\mathcal{H}}, \quad \eta_n^{0,init} = (I - \gamma \mathcal{T}) \eta_{n-1}^{0,init}.$$

which is a deterministic sequence.

An algebraic calculation gives an estimate of the norm of $\eta_n^{0,init}$, and we can also bound the residual term $\eta_n^{init} - \eta_n^{0,init}$, then conclude by Minkowski.

For the variance term: We follow the exact same idea, but have to define a sequence of "semi-stochastic recursion", to be able to bound the residual term.

This decomposition is summed up in Table A.1.

Complete recursion η_n	
variance term η_n^{noise}	bias term η_n^{init}
$\downarrow$	$\downarrow$
multiple recursion	semi stochastic variant
main terms η_n^r, $r \geqslant 1$ satisfying semi-sto recursions	residual term $\eta_n^{noise} - \sum \eta_n^r$ satisf. stochastic recursion
$\downarrow$ Lemma A.30	$\downarrow$ Lemma A.31
$\leqslant C$ Variance term	$\underset{r\to\infty}{\longrightarrow} 0$
Lemma A.26	Lemma A.25
Theorem 2.9	

main term η_n^0 satisf. semi-sto recursion	residual term $\eta_n^{init} - \eta_n^0$ satisf. stochastic recursion
$\downarrow$	$\downarrow$ Lemma A.31
$\leqslant$ Bias term	residual negligible term

Table A.1: Error decomposition in the finite horizon setting. All the references refer to Lemmas given in Section A.4.

For the *online setting*, we follow comparable ideas and end in a similar decomposition.

In Section A.3, we provide proofs of the propositions from Section 2.2 that provide the Hilbert space set-up for kernel-based learning, while in Section A.4, we prove convergence rates for the least-mean-squares algorithm.

A.3 Reproducing kernel Hilbert spaces

In this Section, we provide proofs of the results from Section 2.2 that provide the RHKS space set-up for kernel-based learning. See Aronszajn (1950); Smale and Cucker (2001); Yao (2006) for further properties of RKHSs.

We consider a reproducing kernel Hilbert space $\mathcal{H}$ with kernel K on space $\mathcal{X}$ as defined in Section 2.2.1. Unless explicitly mentioned, we do not make any topological assumption on $\mathcal{X}$.

As detailed in Section 2.2.2 we consider a set $\mathcal{X}$ and $\mathcal{Y} \subset \mathbb{R}$ and a distribution ρ on $\mathcal{X} \times \mathcal{Y}$. We denote ρ_X the marginal law on the space $\mathcal{X}$. In the following, we use the notation (X, Y) for a random variable following the law ρ. We define spaces $L^2_{\rho_X}, \mathcal{L}^2_{\rho_X}$ and the canonical projection p. In the following we further assume that $L^2_{\rho_X}$ is separable, an assumption satisfied in most cases.

We remind our assumptions:

(A1) $\mathcal{H}$ is a separable RKHS associated with kernel K on a space $\mathcal{X}$.

(A2) $\mathbb{E}\left[K(X, X)\right]$ and $\mathbb{E}[Y^2]$ are finite.

Assumption **(A2)** ensures that every function in $\mathcal{H}$ is square-integrable, that is, if $\mathbb{E}[K(X, X)] < \infty$, then $\mathcal{H} \subset \mathcal{L}^2_{\rho_X}$. Indeed, we have:

Proposition A.7. *Assume **(A1)**.*

1. *If $\mathbb{E}[K(X, X)] < \infty$, then $\mathcal{H} \subset \mathcal{L}^2_{\rho_X}$.*
2. *If $\sup_{x \in \mathcal{X}} K(x, x) < \infty$, then any function in $\mathcal{H}$ is bounded.*

Proof. Under such condition, by Cauchy-Schwartz inequality, any function $f \in \mathcal{H}$ is either bounded or integrable:

$$\begin{aligned} |f(x)|^2 &\leqslant \|f\|_K^2 K(x, x) \leqslant \|f\|_K^2 \sup_{x \in \mathcal{X}} K(x, x), \\ \int_{\mathcal{X}} |f(x)|^2 d\rho_X(x) &\leqslant \|f\|_K^2 \int_{\mathcal{X}} K(x, x) d\rho_x(x). \end{aligned}$$

□

The assumption $\mathbb{E}[K(X, X)] < \infty$ seems to be the weakest assumption to make, in order to have at least $\mathcal{H} \subset \mathcal{L}^2_{\rho_X}$. However they may exist functions $f \in \mathcal{H} \setminus \{0\}$ such that $\|f\|_{\mathcal{L}^2_{\rho_X}} = 0$. However under stronger assumptions (see Section A.3.5) we may identify $\mathcal{H}$ and $p(\mathcal{H})$.

A.3.1 Properties of the minimization problem

We are interested in minimizing the following quantity, which is the *prediction error* of a function f, which may be rewritten as follows with dot-products in $L^2_{\rho_X}$:

$$\begin{aligned}
\varepsilon(f) &= \mathbb{E}\left[(f(X)-Y)^2\right] \\
&= \|f\|^2_{L^2_{\rho_X}} - \int_{\mathcal{X}\times\mathcal{Y}} f(x) y d\rho(x,y) + c \\
&= \|f\|^2_{L^2_{\rho_X}} - \int_{\mathcal{X}} f(x) \left(\int_{\mathcal{Y}} y d\rho_{Y|X=x}(y)\right) d\rho_{|X}(x) + c \\
&= \|f\|^2_{L^2_{\rho_X}} - \left\langle f, \int_{\mathcal{Y}} y d\rho_{Y|X=.}(y)\right\rangle_{L^2_{\rho_X}} + c \qquad \text{(A.1)} \\
&= \|f\|^2_{L^2_{\rho_X}} - \langle f, \mathbb{E}\left[Y|X=\cdot\right]\rangle_{L^2_{\rho_X}} + c
\end{aligned}$$

Notice that the problem may be re-written, if f is in $\mathcal{H}$, with dot-products in $\mathcal{H}$:

$$\begin{aligned}
\varepsilon(f) &= \mathbb{E}[f(X)^2] - 2\langle f, \mathbb{E}[Y K_X]\rangle_K + \mathbb{E}[Y^2] \\
&= \langle f, \Sigma f\rangle_K - 2\langle f, \mu\rangle_K + c.
\end{aligned}$$

Interpretation: Under the form (A.1), it appears to be a minimization problem in a Hilbert space of the sum of a continuous coercive function and a linear one. Using Lax-Milgramm and Stampachia theorems (Brezis, 1983) we can conclude with the following proposition, which implies Prop. A.1 in Section 2.2:

Proposition A.8 ($g_\rho, g_\mathcal{H}$)**.** *Assume **(A1-2)**. We have the following points:*

1. *There exists a unique minimizer over the space $L^2_{\rho_X}$. This minimizer is the regression function $g_\rho : x \mapsto \int_\mathcal{Y} y d\rho_{Y|X=x}(y)$ (Lax-Milgramm).*
2. *For any non empty closed convex set, there exists a unique minimizer (Stampachia). As a consequence, there exists a unique minimizer:*

$$g_\mathcal{H} = \arg\min_{f\in\overline{p(\mathcal{H})}} \mathbb{E}\left[(f(X)-Y)^2\right]$$

over $\overline{p(\mathcal{H})}$. $g_\mathcal{H}$ is the orthogonal projection over g_ρ over $\overline{p(\mathcal{H})}$, thus satisfies the following equality: for any $\varepsilon \in \overline{H}$:

$$\mathbb{E}\left[(g_\mathcal{H}(X)-Y)\varepsilon(X)\right] = 0 \qquad \text{(A.2)}$$

A.3.2 Covariance Operator

We defined operators $\Sigma, \mathcal{T}, T$ in Section 2.2.4. We here state the main properties of these operators, then prove the two main decompositions stated in Propositions 2.2 and A.3.

Proposition A.9 (Properties of Σ)**.** *Assume **(A1-2)**.*

1. *Σ is well defined (that is for any $f \in \mathcal{H}$, $z \mapsto \mathbb{E}f(X)K(X,z)$ is in $\mathcal{H}$).*
2. *Σ is a continuous operator.*
3. *$\mathrm{Ker}(\Sigma) = \{f \in \mathcal{H} \text{ s.t. } \|f\|_{L^2_{\rho_X}} = 0\}$. Actually for any $f \in \mathcal{H}$, $\langle f, \Sigma f\rangle_K = \|f\|_{L^2_{\rho_X}}$.*
4. *Σ is a self-adjoint operator.*

Proof. 1. for any $x \in \mathcal{X}$, $f(x)K_x$ is in $\mathcal{H}$. To show that the integral $\int_{x\in\mathcal{X}} f(x)K_x$ is converging, it is sufficient to show the is is absolutely converging in $\mathcal{H}$, as absolute convergence implies convergence in any Banach space[1] (thus any Hilbert space). Moreover:

$$\begin{aligned}\int_{x\in\mathcal{X}} \|f(x)K_x\|_K &\leqslant \int_{x\in\mathcal{X}} |f(x)|\langle K_x, K_x\rangle_K^{1/2} \\ &\leqslant \int_{x\in\mathcal{X}} |f(x)|K(x,x)^{1/2} d\rho_X(x) \\ &\leqslant \left(\int_{x\in\mathcal{X}} f(x)^2 d\rho_X(x)\right)^{1/2} \left(\int_{x\in\mathcal{X}} K(x,x) d\rho_X(x)\right)^{1/2} \\ &< \infty,\end{aligned}$$

under assumption $\mathbb{E}[K(X,X)] < \infty$ (**(A2)**).

2. For any $f \in \mathcal{H}$, we have

$$\begin{aligned}\|\Sigma f\|_K &= \langle \Sigma f, \Sigma f\rangle_K = \int_{x\in\mathcal{X}} (\Sigma f)(x) f(x) d\rho_X(x) \\ &= \int_{x\in\mathcal{X}} \left(\int_{y\in\mathcal{X}} f(y)K(x,y)d\rho_X(y)\right) f(x) d\rho_X(x) \\ &= \int_{x,y\in\mathcal{X}^2} \langle f, K_x\rangle_K \langle f, K_y\rangle_K \langle K_y, K_x\rangle_K d\rho_X(x) d\rho_X(y) \\ &\leqslant \int_{x,y\in\mathcal{X}^2} \|f\|_K \|K_x\|_K \|f\|_K \|K_y\|_K \|K_x\|_K \|K_y\|_K d\rho_X(x) d\rho_X(y) \\ &\qquad\qquad \text{by Cauchy Schwartz,} \\ &\leqslant \|f\|_K^2 \left(\int_{x\in\mathcal{X}^2} \|K_x\|_K^2 d\rho_X(x)\right)^2 \\ &\leqslant \|f\|_K^2 \left(\int_{x\in\mathcal{X}^2} K(x,x) d\rho_X(x)\right)^2,\end{aligned}$$

which proves the continuity under assumption **(A2)**.

3. $\Sigma f = 0 \Rightarrow \langle f, \Sigma f\rangle = 0 \Rightarrow \mathbb{E}[f^2(X)] = 0$. Reciprocally, if $\|f\|_{L^2_{\rho_X}} = 0$, it is clear that $\|\Sigma f\|_{L^2_{\rho_X}} = 0$, then $\|\Sigma f\|_K = \mathbb{E}\left[f(X)(\Sigma f)(X)\right] = 0$, thus $f \in \mathrm{Ker}(T)$.

4. It is clear that $\langle \Sigma f, g\rangle = \langle f, \Sigma g\rangle$.

□

Proposition A.10 (Properties of $\mathcal{T}$)**.** *Assume **(A1-2)**. $\mathcal{T}$ satisfies the following properties:*

1. *$\mathcal{T}$ is a well defined, continuous operator.*
2. *For any $f \in \mathcal{H}$, $\mathcal{T}(\tilde{f}) = \Sigma f$, $\|\mathcal{T}f\|_K^2 = \int_{x,y\in\mathcal{X}^2} f(y)f(x)K(x,y)d\rho_X(y)d\rho_X(x)$.*
3. *The image of T is a subspace of $\mathcal{H}$.*

Proof. It is clear that $\mathcal{T}$ is well defined, as for any class $\tilde{f}$, $\int_{\mathcal{X}} f(t)\ K_t\ d\rho_X(t)$ does not depend on the representer f, and is converging in $\mathcal{H}$ (which is the third point), just as in the previous proof. The second point results from the definitions. Finally for continuity, we have:

$$\|\mathcal{T}f\|_K^2 = \langle \mathcal{T}f, \mathcal{T}f\rangle_K$$

[1] A Banach space is a linear normed space which is complete for the distance derived from the norm.

$$
\begin{aligned}
&= \int_{x\in\mathcal{X}}\int_{y\in\mathcal{X}} f(y)f(x)K(x,y)d\rho_X(y)d\rho_X(x)\\
&\leqslant \left(\int_{x\in\mathcal{X}^2} |f(x)K(x,x)^{1/2}|d\rho_X(x)\right)^2\\
&\leqslant \left(\int_{x\in\mathcal{X}} f(x)^2 d\rho_X(x)\right)\left(\int_{x\in\mathcal{X}} K(x,x)d\rho_X(x)\right) \leqslant C\|f\|^2_{L^2_{\rho_X}}.
\end{aligned}
$$

□

We now state here a simple lemma that will be useful later:

Lemma A.11. *Assume **(A1)**.*

1. $\mathbb{E}\left[k(X,X)\right] < \infty \Rightarrow \int_{x,y\in\mathcal{X}} k(x,y)^2 d\rho_X(x)d\rho_X(y) < \infty$.
2. $\mathbb{E}\left[|k(x,y)|\right] < \infty \Rightarrow \int_{x,y\in\mathcal{X}} k(x,y)^2 d\rho_X(x)d\rho_X(y) < \infty$.

Proposition A.12 (Properties of T)**.** *Assume **(A1-2)**. T satisfies the following properties:*

1. *T is a well defined, continuous operator.*
2. *The image of T is a subspace of $p(\mathcal{H})$.*
3. *T is a self-adjoint semi definite positive operator in the Hilbert space $L^2_{\rho_X}$.*

Proof. $T = p\circ\mathcal{T}$ is clearly well defined, using the arguments given above. Moreover:

$$
\begin{aligned}
\|Tf\|^2_{L^2_{\rho_X}} &= \int_{x\in\mathcal{X}}\left(\int_{t\in X} K(x,t)f(t)d\rho_X(t)\right)^2 d\rho_X(x)\\
&\leqslant \left(\int_{x\in\mathcal{X}}\int_{t\in X} K(x,t)^2 d\rho_X(t)d\rho_X(x)\right)\left(\int_{t\in\mathcal{X}} f^2(t)d\rho_X(t)\right) \text{ by C.S.}\\
&\leqslant C\|f\|_{\mathcal{L}^2_{\rho_X}} \quad \text{by Lemma A.11,}
\end{aligned}
$$

which is continuity[2]. Then by Proposition A.10, $\mathrm{Im}(Td) \subset p(\mathrm{Im}(\mathcal{T})) \subset p(\mathcal{H})$. Finally, for any $f, g \in \mathcal{L}^2_{\rho_X}$,

$$
\begin{aligned}
\langle f, Tg\rangle_{\mathcal{L}^2_{\rho_X}} &= \int_{\mathcal{X}} f(x)\, Tg(x)d\rho_X(x)\\
&= \int_{\mathcal{X}} f(x)\left(\int_{\mathcal{X}} g(t)K(x,t)d\rho_X(t)\right)d\rho_X(x)\\
&= \int_{\mathcal{X}\times\mathcal{X}} f(x)g(t)K(x,t)d\rho_X(t)d\rho_X(x) = \langle Tf, g\rangle_{\mathcal{L}^2_{\rho_X}}.
\end{aligned}
$$

and $\langle f, Tf\rangle_{\mathcal{L}^2_{\rho_X}} \geq 0$ as a generalization of the positive definite property of K. □

In order to show the existence of an eigenbasis for T, we now show that T is trace-class.

Proposition A.13 (Compactness of the operator)**.** *We have the following properties:*

1. *Under **(A2)**, T is a trace class operator[3]. As a consequence, it is also a Hilbert-Schmidt operator[4].*

[2] We could also use the continuity of $p : \mathcal{H} \to L^2_{\rho_X}$.

[3] Mimicking the definition for matrices, a bounded linear operator A over a separable Hilbert space H is said to be in the trace class if for some (and hence all) orthonormal bases $(e_k)_k$ of H the sum of positive terms $\mathrm{tr}|A| := \sum_k \langle (A^*A)^{1/2} e_k, e_k\rangle$ is finite.

[4] A Hilbert-Schmidt operator is a bounded operator A on a Hilbert space H with finite Hilbert–Schmidt norm: $\|A\|^2_{\mathrm{HS}} = \mathrm{tr}|(A^*A)| := \sum_{i\in I}\|Ae_i\|^2$.

2. *If $K \in L^2(\rho_X \times \rho_X)$ then T is a Hilbert-Schmidt operator.*

3. *Any Hilbert-Schmidt operator is a compact operator.*

Proof. Proofs of such facts may be found in Brezis (1983); Paulin (2009). Formally, with $(\phi_i)_i$ an Hilbertian basis in $L^2_{\rho_X}$:

$$\begin{aligned}
\mathbb{E}\left[K(X,X)\right] &= \mathbb{E}\left[\langle K_x, K_x\rangle_K\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{\infty}\langle K_x,\phi_i\rangle_K^2\right] \quad \text{by Parseval equality,} \\
&= \sum_{i=1}^{\infty}\mathbb{E}\left[\langle K_x,\phi_i\rangle_K^2\right] \\
&= \sum_{i=1}^{\infty}\langle T\phi_i,\phi_i\rangle_K = \mathrm{tr}(T).
\end{aligned}$$

□

Corollary A.14. *We have thus proved that under **(A1)** and **(A2)**, the operator T may be reduced in some Hilbertian eigenbasis: the fact that T is self-adjoint and compact implies the existence of an orthonormal eigen-system (which is an Hilbertian basis of $L^2_{\rho_X}$).*

This is a consequence of a very classical result, see for example Brezis (1983).

Definition A.15. *The null space Ker$(T) := \left\{f \in L^2_{\rho_X} \text{ s.t. } Tf = 0\right\}$ may not be $\{0\}$. We denote by S an orthogonal supplementary of Ker(T).*

Proposition 2.2 is directly derived from a slightly more complete Proposition A.16 below:

Proposition A.16 (Eigen-decomposition of T)**.** *Under **(A1)** and **(A2)**, T is a bounded self adjoint semi-definite positive operator on $L^2_{\rho_X}$, which is trace-class. There exists[5] a Hilbertian eigenbasis $(\phi_i)_{i\in I}$ of the orthogonal supplement S of the null space $\mathrm{Ker}(T)$, with summable eigenvalues $(\mu_i)_{i\in I}$. That is:*

- $\forall i \in I,\ T\phi_i = \mu_i\phi_i$, *$(\mu_i)_i$ strictly positive non increasing (or finite) sequence such that* $\sum_{i\in I}\mu_i < \infty$.
- $L^2_{\rho_X} = \mathrm{Ker}(T) \overset{\perp}{\oplus} S$.

We have[6]: $S = \overline{span\{\phi_i\}} = \left\{\sum_{i=1}^{\infty} a_i\phi_i \text{ s.t. } \sum_{i=1}^{\infty} a_i^2 < \infty\right\}$. Moreover:

$$S = \overline{p(\mathcal{H})}. \tag{A.3}$$

Proof. For any $i \in I$, $\phi_i = \frac{1}{\mu_i}L_K\phi_i \in p(\mathcal{H})$. Thus $\mathrm{span}\{\phi_i\} \subset p(\mathcal{H})$, thus $S = \overline{\mathrm{span}\{\phi_i\}} \subset \overline{p(\mathcal{H})}$. Moreover, using the following Lemma, $p(\mathcal{H}) \subset \mathrm{Ker}(T)^{\perp} = S$, which concludes the proof, by taking the closures. □

[5] S is stable by T and $T\ : S \to S$ is a self adjoint compact positive operator.

[6] We denote by $\mathrm{span}(A)$ the smallest linear space which contains A, which is in such a case the set of all finite linear combinations of $(\phi_i)_{i\in I}$.

Lemma A.17. *We have the following points:*

- *if* $T^{1/2}f = 0$ *in* $L^2_{\rho_X}$, *then* $Tf = 0$ *in* $\mathcal{H}$.
- $p(\mathcal{H}) \subset \mathrm{Ker}(T)^\perp$.

Proof. We first notice that if $T^{1/2}f = 0$ in $L^2_{\rho_X}$, then $\mathcal{T}f = 0$ in $\mathcal{H}$: indeed[7]

$$\begin{aligned} \|Tf\|^2_{\mathcal{H}} &= \left\langle \int_{\mathcal{X}} f(x)K_x d\rho_X(x), \int_{\mathcal{X}} f(y)K_y d\rho_X(y) \right\rangle_K \\ &= \int_{\mathcal{X}^2} f(x)f(y)K(x,y)d\rho_X(x)d\rho_X(y) \\ &= \langle f, Tf\rangle_{L^2_{\rho_X}} = 0 \text{ if } Tf = 0 \text{ in } L^2_{\rho_X}. \end{aligned}$$

Moreover $\mathcal{H}$ is the completed space of span $\{K_x, x \in \mathcal{X}\}$, with respect to $\|\cdot\|_K$ and for all $x \in \mathcal{X}$, for all $\psi_k \in \mathrm{Ker}(T)$:

$$\begin{aligned} \langle p(K_x), \psi_k\rangle_{L^2_{\rho_X}} &= \int_{\mathcal{X}} K_x(y)\psi_K(y)d\rho_X(y) = (T\psi_k)(x), \\ \text{however,} \quad T\psi_k &=_{L^2_{\rho_X}} 0 \quad \Rightarrow T\psi_k =_{\mathcal{H}} 0 \quad \forall x \in \mathcal{X} \Rightarrow T\psi_k(x) = 0. \end{aligned}$$

As a consequence, $\overline{\text{span}\,\{p(K_x), x \in \mathcal{X}\}} \subset \mathrm{Ker}(T)^\perp$. We just have to show that $\overline{\text{span}\,\{p(K_x), x \in \mathcal{X}\}} = p(\mathcal{H})$, as $\mathrm{Ker}(T)^\perp$ is a closed space. It is true as for any $\tilde{f} \in p(H), f \in \mathcal{H}$ there exists $f_n \subset$ span $\{K_x, x \in \mathcal{X}\}$ such that $f_n \overset{\mathcal{H}}{\to} f$, thus $p(f_n) \to \tilde{f}$ in $L^2_{\rho_X}$[8]. Finally we have proved that $p(\mathcal{H}) \subset \mathrm{Ker}(T)^\perp$. □

Similarly, Proposition A.3 is derived from Proposition A.18 below:

Proposition A.18 (Decomposition of Σ)**.** *Under* ***(A1)*** *and* ***(A2)****,* $\mathrm{Im}(\mathcal{T}) \subset \mathcal{H}$, *that is, for any* $f \in L^2_{\rho_X}$, $\mathcal{T}f \in \mathcal{H}$. *Moreover, for any* $i \in I$, $\phi_i^H = \frac{1}{\mu_i}\mathcal{T}\phi_i \in H$ *is a representant for the equivalence class* ϕ_i. *Moreover* $\left(\mu_i^{1/2}\phi_i^H\right)_{i\in I}$ *is an orthonormal eigen-system of* $\mathscr{S}$ *That is:*

- $\forall i \in I,\ \Sigma\phi_i^H = \mu_i\phi_i^H$.
- $\left(\mu_i^{1/2}\phi_i^H\right)_{i\in I}$ *is an orthonormal family in* $\mathscr{S}$.

We thus have:

$$\mathscr{S} = \left\{ \sum_{i\in I} a_i\phi_i^H \text{ s.t. } \sum_{i\in I} \frac{a_i^2}{\mu_i} < \infty \right\}.$$

Moreover $\mathscr{S}$ *is the orthogonal supplement of the null space Ker*(Σ)*:*

$$\mathcal{H} = \mathrm{Ker}(\Sigma) \overset{\perp}{\oplus} \mathscr{S}.$$

[7] In other words, we the operator defined below $T^{1/2}$

$$\begin{aligned} T^{1/2}f &=_{L^2_{\rho_X}} 0 \\ \mathcal{T}f &=_{\mathcal{H}} \Sigma^{1/2}(\mathcal{T}^{1/2}f) \\ \|\mathcal{T}f\|^2_K &= \|\Sigma^{1/2}(\mathcal{T}^{1/2}f)\|^2_K = \|(\mathcal{T}^{1/2}f)\|^2_{L^2_{\rho_X}} = 0 \\ {}^H Tf &=_{\mathcal{H}} 0. \end{aligned}$$

[8] $\|f_n - f\|_{L^2_{\rho_X}} = \|\Sigma^{1/2}(f_n - f)\|_K \to 0$ as Σ continuous.

Proof. The family $\phi_i^H = \frac{1}{\mu_i} T\phi_i$ satisfies:

- $\widetilde{\phi_i^H} = \phi_i$ (in $L^2_{\rho_X}$),
- $\phi_i^H \in \mathscr{S}$,
- $T\phi_i^H = \mu_i \phi_i$ in $L^2_{\rho_X}$,
- $\mathcal{T}\phi_i^H = \Sigma\phi_i^H = \mu_i \phi_i^H$ in $\mathcal{H}$.

All the points are clear: indeed for example $\Sigma\phi_i^H = T\phi_i = \mu_i\phi_i^H$. Moreover, we have that:

$$\begin{aligned} \|\phi_i\|^2_{L^2_{\rho_X}} = \|\phi_i^H\|^2_{L^2_{\rho_X}} &= \langle \phi_i^H, \Sigma\phi_i\rangle_K \text{ by Proposition 3} \\ &= \mu_i \|\phi_i^H\|^2_K \\ &= \|\sqrt{\mu_i}\phi_i^H\|^2_K \end{aligned}$$

That means that $(\sqrt{\mu_i}\phi_i^H)_i$ is an orthonormal family in $\mathcal{H}$.

Moreover, $\mathscr{S}$ is defined as the completion for $\|\cdot\|_K$ of this orthonormal family, which gives $\mathscr{S} = \left\{ \sum_{i\in I} a_i \phi_i^H \text{ s.t. } \sum_{i\in I} \frac{a_i^2}{\mu_i} < \infty \right\}$.

To show that $\mathcal{H} = \mathrm{Ker}(\Sigma) \overset{\perp}{\oplus} \mathscr{S}$, we use the following sequence of arguments:

- First, as Σ is a continuous operator, $\mathrm{Ker}(\Sigma)$ is a closed space in $\mathcal{H}$, thus $\mathcal{H} = \mathrm{Ker}(\Sigma) \overset{\perp}{\oplus} (\mathrm{Ker}(\Sigma))^\perp$.
- $\mathrm{Ker}(\Sigma) \subset (\mathcal{T}^{1/2}(S))^\perp$: indeed for all $f \in \mathrm{Ker}(\Sigma)$, $\langle f, \phi_i^{\mathcal{H}}\rangle = \frac{1}{\mu_i}\langle f, \Sigma\phi_i^{\mathcal{H}}\rangle = \frac{1}{\mu_i}\Sigma\langle f, \phi_i^{\mathcal{H}}\rangle = 0$, and as a consequence for any $f \in \mathrm{Ker}(\Sigma), g \in \mathcal{T}^{1/2}(S)$, there exists $(g_n) \subset \mathrm{span}(\phi_i^H)$ s.t. $g_n \overset{\mathcal{H}}{\to} g$, thus $0 = \langle g_n, f\rangle_{\mathcal{H}} \to \langle f, g\rangle$ and finally $f \in (\mathcal{T}^{1/2}(S))^\perp$. Equivalently $\mathcal{T}^{1/2}(S) \subset (\mathrm{Ker}(\Sigma))^\perp$.
- $(\mathcal{T}^{1/2}(S))^\perp \subset \mathrm{Ker}(\Sigma)$. For any i, $\phi_i^H \in \mathcal{T}^{1/2}(S)$. If $f \in (\mathcal{T}^{1/2}(S))^\perp$, then $\langle p(f), \phi_i\rangle_{L^2_{\rho_X}} = \langle f, \mathcal{T}\phi_i\rangle_{\mathcal{H}} = 0$. As a consequence $p(f) \in p(\mathcal{H}) \cap \mathrm{Ker}(T) = \{0\}$, thus $f \in \mathrm{Ker}(\Sigma)$. That is $(\mathcal{T}^{1/2}(S))^\perp \subset \mathrm{Ker}(\Sigma)$. Equivalently $\mathrm{Ker}(\Sigma)^\perp \subset (\mathcal{T}^{1/2}(S))$.
- Combining these points: $\mathcal{H} = \mathrm{Ker}(\Sigma) \overset{\perp}{\oplus} \mathscr{S}$.

□

We have two decompositions of $\mathcal{L}^2_{\rho_X} = \mathrm{Ker}(T) \overset{\perp}{\oplus} S$ and $\mathcal{H} = \mathrm{Ker}(\Sigma) \overset{\perp}{\oplus} \mathscr{S}$. They happen to be related through the mapping $\mathcal{T}^{1/2}$, which we now define.

A.3.3 Properties of T^r, $r > 0$

We defined operators T^r, $r > 0$ and $\mathcal{T}^r$, $r \geqslant 1/2$ in Section 2.2.4 in Definitions 2.4,A.4.

Proposition A.19 (Properties of T^r, $\mathcal{T}^r$)**.**

- *T^r is well defined for any $r > 0$.*
- *$\mathcal{T}^r$ is well defined for any $r \geqslant \frac{1}{2}$.*
- *$\mathcal{T}^{1/2} : S \to \mathscr{S}$ is an isometry.*

- *Moreover* $\mathrm{Im}(T^{1/2}) = p(\mathcal{H})$. *That means* $T^{1/2} : S \to p(\mathcal{H})$ *is an isomorphism.*

Proof. T^r *is well defined for any* $r > 0$.

$S = \{\sum_{i=1}^{\infty} a_i\phi_i \text{ s.t. } \sum_{i=1}^{\infty} a_i^2 < \infty\}$. For any sequence $(a_i)_{i\in I}$ such that $\sum_{i=1}^{\infty} a_i^2 < \infty$, $T^r(\sum a_i\phi_i) = \sum_i \mu_i^r a_i\phi_i$ is a converging sum in the Hilbert space $L^2_{\rho_X}$ (as $(\mu_i)_{i\in I}$ is bounded thus $\sum_i \mu_i^r a_i\phi_i$ satisfies Cauchy is criterion: $\| \sum_{i=n}^{p} \mu_i^r a_i\phi_i\|^2 \leqslant \mu_0^r \left(\sum_{i=n}^{p} a_i^2\right)^{1/2}$). And Cauchy is criterion implies convergence in Hilbert spaces.

$\mathcal{T}^r$ *is well defined for any* $r \geqslant \frac{1}{2}$.

We have shown that $(\sqrt{\mu_i}\phi_i^H)_i$ is an orthonormal family in $\mathcal{H}$. As a consequence (using the fact that (μ_i) is a bounded sequence), for any sequence $(a_i)_i$ such that $\sum a_i^2 < \infty$, $\sum_i \mu_i^r a_i\phi_i^H$ satisfies Cauchy is criterion thus is converging in $\mathcal{H}$ as $\| \sum_{i\in I'} \mu_i^r a_i\phi_i^H\|_K = \sum_{i\in I'} \mu_i^{r-1/2} a_i^2 \leqslant \mu_0^{r-1/2} \sum_{i\in I'} a_i^2 < \infty$. (We need $r \geqslant 1/2$ of course).

$\mathcal{T}^{1/2} : S \to \mathscr{S}$ *is an isometry.*

Definition has been proved. Surjectivity in $\mathscr{S}$ is by definition, as

$$\mathcal{T}^{1/2}(S) = \left\{ \sum_{i\in I} a_i\phi_i^H \text{ s.t. } \sum_{i\in I} \frac{a_i^2}{\mu_i} < \infty \right\}.$$

Moreover, the operator is clearly injective as for any $f \in S$, $Tf \neq 0$ in $L^2_{\rho_X}$ thus $Tf \neq 0$ in $\mathcal{H}$. Moreover for any $f = \sum_{i=1}^{\infty} a_i\phi_i \in S$, $\|Tf\|_K^2 = \| \sum_{i=1}^{\infty} a_i\sqrt{\mu_i}\phi_i\|_K^2 = \sum_{i=1}^{\infty} a_i^2 = \|f\|^2_{\mathcal{L}^2_{\rho_X}}$, which is the isometrical property.

It must be noticed that we cannot prove surjectivity in $\mathcal{H}$[9], that is without our "strong assumptions". However we will show that operator $T^{1/2}$ is surjective in $p(\mathcal{H})$.

$\mathrm{Im}(T^{1/2}) = p(\mathcal{H})$. *That means* $T^{1/2} : S \to p(\mathcal{H})$ *is an isomorphism.*

$\mathrm{Im}(T^{1/2}) = p(\mathrm{Im}(\mathcal{T}^{1/2})) = p(\mathscr{S})$. Moreover $p(\mathcal{H}) = p(\mathrm{Ker}(\Sigma) \oplus \mathscr{S}) = p(\mathscr{S})$. Consequently $\mathrm{Im}(T^{1/2}) = p(\mathcal{H})$. Moreover $T^{1/2} : S \to L^2_{\rho_X}$ is also injective, which give the isomorphical character.

Note that it is clear that $T^{1/2}(S) \subset p(\mathcal{H})$ and that for any $x \in \mathcal{X}$, $p(K_x) \in T^{1/2}(S)$ indeed $p(K_x) = \sum_{i=1}^{\infty} \langle K_x, \phi_i\rangle_{L^2_{\rho_X}} \phi_i = \sum_{i=1}^{\infty} \mu_i\phi_i^H(x)\phi_i$, with $\sum_{i=1}^{\infty} \frac{\left(\mu_i\phi_i^H(x)\right)^2}{\mu_i} = \sum_{i=1}^{\infty} \mu_i\phi_i^H(x)^2 < \infty$, as $K(x,x) = \sum_{i=1}^{\infty} \mu_i\phi_i^H(x)^2$ □

Finally, it has appeared that S and $\mathscr{S}$ may be identified via the isometry $\mathcal{T}^{1/2}$. We conclude by a proposition which sums up the properties of the spaces $\mathcal{T}^r(L^2_{\rho_X})$.

Proposition A.20. *The spaces* $T^r(L^2_{\rho_X}), r > 0$ *satisfy:*

$$\begin{aligned}
\forall r \geqslant r' > 0, \quad T^r\left(L^2_{\rho_X}\right) &\subset T^{r'}\left(L^2_{\rho_X}\right) \\
\forall r > 0, \quad \overline{T^r\left(L^2_{\rho_X}\right)} &= S \\
T^{1/2}\left(L^2_{\rho_X}\right) &= p(\mathcal{H}) \\
\forall r \geqslant \frac{1}{2}, \quad T^r\left(L^2_{\rho_X}\right) &\subset p(\mathcal{H})
\end{aligned}$$

[9] It is actually easy to build a counter example, f.e. with a measure of "small" support (let is say $[-1,1]$), a Hilbert space of functions on $\mathcal{X} = [-5;5]$, and a kernel like $\min(0, 1-|x-y|)$: $\mathrm{Im}\left(\mathcal{T}^{1/2}\right) \subset \{f \in \mathcal{H} \text{ s. t. } \mathrm{supp}(f) \subset [-2;2]\} \subsetneq \mathcal{H}$.

A.3.4 Kernel decomposition

We prove here Proposition A.6.

Proof. Considering our decomposition of $\mathcal{H} = \mathscr{S} \overset{\perp}{\oplus} \ker(\Sigma)$, an the fact the $(\sqrt{\mu_i}\phi_i^{\mathcal{H}})$ is a Hilbertian eigenbasis of $\mathscr{S}$, we have for any $x \in \mathcal{X}$,

$$\begin{aligned} K_x &= \sum_{i=1}^{\infty} \langle \sqrt{\mu_i}\phi_i^{\mathcal{H}}, K_x \rangle_{\mathcal{H}} \sqrt{\mu_i}\phi_i^{\mathcal{H}} + g_x \\ &= \sum_{i=1}^{\infty} \mu_i \phi_i^{\mathcal{H}}(x)\phi_i^{\mathcal{H}} + g_x \end{aligned}$$

And as it has been noticed above this sum is converging in $\mathscr{S}$ (as in $\mathcal{H}$) because $\sum_{i=1}^{\infty} \frac{(\mu_i \phi_i^{\mathcal{H}}(x))^2}{\mu_i} = \sum_{i=1}^{\infty} \mu_i (\phi_i^{\mathcal{H}}(x))^2 = K(x,x) < \infty$. However, the convergence may not be absolute in $\mathcal{H}$. Our function g_x is in $\mathrm{Ker}(\Sigma)$, which means $\int_{y\in\mathcal{X}} g_x(y)^2 d\rho_X(y) = 0$.

And as a consequence, we have for all $x, y \in \mathcal{X}$,

$$K(x,y) = \sum_{i \in I} \mu_i \phi_i^H(x)\phi_i^H(y) + g(x,y),$$

With $g(x,y) := g_x(y)$. Changing roles of x, y, it appears that $g(x,y) = g(y,x)$. And we have for all $x \in \mathcal{X}$, $\int_{\mathcal{X}} g(x,y)^2 d\rho_X(y) = 0$. Moreover, the convergence of the series is absolute

We now prove the following points

(a) $(\mathscr{S}, \|\cdot\|_{\mathcal{H}})$ is also an RKHS, with kernel $K^{\mathscr{S}} : (x,y) \mapsto \sum_{i\in I} \mu_i \phi_i^H(x)\phi_i^H(y)$

(b) given the decomposition above, almost surely the optimization problem in $\mathcal{S}$ and $\mathcal{H}$ have equivalent solutions.

(a) $(\mathscr{S}, \|\cdot\|_{\mathcal{H}})$ is a Hilbert space as a closed subspace of a Hilbert space. Then for any $x \in \mathcal{X}$: $K_x^{\mathscr{S}} := (y \mapsto K^{\mathscr{S}}(x,y)) = \sum_{i=1}^{\infty} \mu_i \phi_i^{\mathcal{H}}(x)\phi_i^{\mathcal{H}} \in \mathscr{S}$. Finally, for any $f \in \mathscr{S}$

$$\langle f, K_x^{\mathscr{S}} \rangle_{\mathcal{H}} = \langle f, K_x^{\mathscr{S}} + g_x \rangle_{\mathcal{H}} = \langle f, K_x \rangle_{\mathcal{H}} = f(x),$$

because $g_x \in \mathrm{Ker}(\Sigma) = \mathscr{S}^{\perp} \ni f$. Thus stands the reproducing property.

(b) We have that $p(\mathscr{S}) = p(\mathcal{H})$ and our best approximating function is a minimizer over this set. Moreover if $K_x^{\mathscr{S}}$ was used instead of K_x in our algorithm, both estimators are almost surely almost surely equal (*i.e.*, almost surely in the same equivalence class). Indeed, at any step n, if we denote $g_n^{\mathscr{S}}$ the sequence built in $\mathscr{S}$ with $K^{\mathscr{S}}$, if we have $g_n^{\mathscr{S}} \overset{a.s.}{=} g_n$, then almost surely $g_n^{\mathscr{S}}(x_n) = g_n(x_n)$ and moreover $K_{x_n} \overset{a.s.}{=} K_{x_n}^S$. Thus almost surely, $g_{n+1} \overset{a.s.}{=} g_{n+1}^{\mathscr{S}}$.

□

A.3.5 Alternative assumptions

As it has been noticed in this chapter, we have tried to minimize assumptions made on $\mathcal{X}$ and K. In this section, we review some of the consequences of such assumptions.

Alternative assumptions

The following have been considered previously:

1. Under the assumption that ρ *is a Borel probability measure* (with respect with some topology on $\mathbb{R}^d$) and $\mathcal{X}$ is a closed space, we may assume that $\text{supp}(\rho) = \mathcal{X}$, where $\text{supp}(\rho)$ is the smallest close space of measure one.

2. The assumption that K *is a Mercer kernel* ($\mathcal{X}$ compact, K continuous) has generally been made before (Tarrès and Yao, 2014; Smale and Zhou, 2007; Smale and Cucker, 2001; Ying and Pontil, 2008), but does not seem to be necessary here.

3. **(A2)** was replaced by the stronger assumption $\sup_{x\in\mathcal{X}} K(x,x) < \infty$ (Tarrès and Yao, 2014; Ying and Pontil, 2008; Rosasco et al., 2014) and $|Y|$ bounded (Tarrès and Yao, 2014; Rosasco et al., 2014) .

Identification $\mathcal{H}$ and $p(\mathcal{H})$

Working with mild assumptions has made it necessary to work with sub spaces of $L^2_{\rho_X}$, thus projecting $\mathcal{H}$ in $p(\mathcal{H})$. With stronger assumptions given above, the space $\mathcal{H}$ may be identified with $p(\mathcal{H})$.

Our problems are linked with the fact that a function f in $\mathcal{H}$ may satisfy both $\|f\|_{\mathcal{H}} \neq 0$ and $\|f\|_{L^2_{\rho_X}} = 0$.

- the "support" of ρ may not be $\mathcal{X}$.
- even if the support is $\mathcal{X}$, a function may be ρ-a.s. 0 but not null in $\mathcal{H}$.

Both these "problems" are solved considering the further assumptions above. We have the following Proposition:

Proposition A.21. *If we consider a Mercer kernel K (or even any continuous kernel), on a space $\mathcal{X}$ compact and a measure ρ_X on $\mathcal{X}$ such that $\text{supp}(\rho) = \mathcal{X}$ then the map:*

$$\begin{array}{rcl} p : \mathcal{H} & \to & p(\mathcal{H}) \\ f & \mapsto & \tilde{f} \end{array}$$

is injective, thus bijective.

Mercer kernel properties

We review here some of the properties of Mercer kernels, especially Mercer's theorem which may be compared to Proposition A.6.

Proposition A.22 (Mercer theorem)**.** *Let $\mathcal{X}$ be a compact domain or a manifold, ρ a Borel measure on $\mathcal{X}$, and $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a Mercer Kernel. Let λ_k be the k-th eigenvalue of T and Φ_k the corresponding eigenvectors. For all $x, t \in \mathcal{X}$, $K(x,t) = \sum_{k=1}^{\infty} \lambda_k \Phi_k(x)\Phi_k(t)$ where the convergence is absolute (for each $x, t \in \mathcal{X}^2$) and uniform on $\mathcal{X} \times \mathcal{X}$.*

The proof of this theorem is given in Hochstadt (1973).

Proposition A.23 (Mercer Kernel properties)**.** *In a Mercer kernel, we have that:*

1. $C_K := \sup_{x,t\in\mathcal{X}^2}(K(x,t)) < \infty$.

2. $\forall f \in \mathcal{H}$, *$f$ is* C^0.
3. *The sum* $\sum \lambda_k$ *is convergent and* $\sum_{k=1}^{\infty} \lambda_k = \int_X K(x,x) \leqslant \rho(\mathcal{X}) C_K$.
4. *The inclusion* $I_K : \mathcal{H} \to C(\mathcal{X})$ *is bounded with* $|||I_K||| \leqslant C_K^{1/2}$.
5. *The map*

$$\begin{aligned} \Phi : \mathcal{X} &\to \ell^2 \\ x &\mapsto (\sqrt{\lambda_k}\Phi_k(x))_{k\in\mathbb{N}} \end{aligned}$$

   is well defined, continuous, and satisfies $K(x,t) = \langle \Phi_k(x), \Phi_k(t) \rangle$.
6. *The space* $\mathcal{H}$ *is independent of the measure considered on* $\mathcal{X}$.

We can characterize $\mathcal{H}$ via the eigenvalues-eigenvectors:

$$\mathcal{H} = \left\{ f \in L^2_{\rho_X} | f = \sum_{k=1}^{\infty} a_k \Phi_k \text{ with } \sum_{k=1}^{\infty} \left(\frac{a_k}{\sqrt{\lambda_k}} \right)^2 < \infty \right\}.$$

Which is equivalent to saying that $T^{1/2}$ is an isomorphism between $L^2_{\rho_X}$ and $\mathcal{H}$. Where we have only considered $\lambda_k > 0$. It has no importance to consider the linear subspace S of $L^2_{\rho_X}$ spanned by the eigenvectors with non zero eigenvalues. However it changes the space $\overline{\mathcal{H}}$ which is in any case S, and is of some importance regarding the estimation problem.

A.4 Proofs

To get our results, we are going to derive from our recursion a new error decomposition and bound the different sources of error via algebraic calculations. We first make a few remarks on short notations that we will use in this part and difficulties that arise from the Hilbert space setting in Section A.4.1, then provide intuition via the analysis of a closely related recursion in Section A.4.2. We give in Sections A.4.3, A.4.4 the complete proof of our bound respectively in the finite horizon case (Theorem 2.9) and the online case (Theorem 2.11). We finally provide technical calculations of the main bias and variance terms in Section A.4.6.

A.4.1 Preliminary remarks

We remind that we consider a sequence of functions $(g_n)_{n\in\mathbb{N}}$ satisfying the system defined in Section 2.3.

$$\begin{aligned} g_0 &= 0 \text{ (the null function)}, \\ g_n &= \sum_{i=1}^{n} a_i K_{x_i}. \end{aligned}$$

With a sequence $(a_n)_{n\geqslant 1}$ such that for all n greater than 1 :

$$a_n := -\gamma_n(g_{n-1}(x_n) - y_n) = -\gamma_n \left(\sum_{i=1}^{n-1} a_i K(x_n, x_i) - y_i \right). \tag{A.4}$$

We output

$$\overline{g}_n = \frac{1}{n+1} \sum_{k=0}^{n} \overline{g}_k. \tag{A.5}$$

We consider a representer $g_{\mathcal{H}} \in \mathcal{L}^2_{\rho_X}$ of $g_{\mathcal{H}}$ defined by Proposition A.1. We accept to confuse notations as far as our calculations are made on $L^2_{\rho_X}$-norms, thus does not depend on our choice of the representer.

We aim to estimate :

$$\varepsilon(\overline{g}_n) - \varepsilon(g_{\mathcal{H}}) = \|\overline{g}_n - g_{\mathcal{H}}\|^2_{L^2_{\rho_X}}.$$

Notations

In order to simplify reading, we will use some shorter notations :

- For the covariance operator, we will only use Σ instead of $\Sigma, T, \mathcal{T}$,

Space :	$\mathcal{H}$
Observations :	$(x_n, y_n)_{n\in\mathbb{N}}$ i.i.d. $\sim \rho$
Best approximation function :	$g_{\mathcal{H}}$
Learning rate :	$(\gamma_i)_i$

All the functions may be split up the orthonormal eigenbasis of the operator $\mathcal{T}$. We can thus see any function as an infinite-dimensional vector, and operators as matrices. This is of course some (mild) abuse of notations if we are not in finite dimensions. For example, our operator Σ may be seen as $\mathrm{Diag}(\mu_i)_{1\leqslant i}$. Carrying on the analogy with the finite dimensional setting, a self adjoint operator, may be seen as a symmetric matrix.

We will have to deal with several "matrix products" (which are actually operator compositions). We denote :

$$\begin{aligned}
M(k,n,\gamma) &= \prod_{i=k}^{n}(I - \gamma K_{x_i}\otimes K_{x_i}) = (I - \gamma K_{x_k}\otimes K_{x_k})\cdots(I - \gamma K_{x_n}\otimes K_{x_n}) \\
M(k,n,(\gamma_i)_i) &= \prod_{i=k}^{n}(I - \gamma_i K_{x_i}\otimes K_{x_i}) \\
D(k,n,(\gamma_i)_i) &= \prod_{i=k}^{n}(I - \gamma_i \Sigma)
\end{aligned}$$

Remarks :

- As our operators may not commute, we use a somehow unusual convention by defining the products for any k, n, even with $k > n$, with $M(k,n,\gamma) = (I - \gamma K_{x_k}\otimes K_{x_k})(I - \gamma K_{x_{k-1}}\otimes K_{x_{k-1}})\cdots(I - \gamma K_{x_n}\otimes K_{x_n})$.
- We may denote $D(k,n,\gamma) = \prod_{i=k}^{n}(I - \gamma\Sigma)$ even if its clearly $(I-\gamma\Sigma)^{n-k+1}$ just in order to make the comparison between equations easier.

On norms

In the following, we will use constantly the following observation :

Lemma A.24. *Assume **A2-4** , let $\eta_n = g_n - g_{\mathcal{H}}$, $\bar{\eta}_n = \overline{g}_n - g_{\mathcal{H}}$:*

$$\begin{aligned}
\varepsilon(g_n) - \varepsilon(g_{\mathcal{H}}) &= \langle \eta_n, \Sigma\eta_n\rangle = \mathbb{E}\left[\langle x, g_n - g_{\mathcal{H}}\rangle^2\right] \left(:= \|g_n - g_{\mathcal{H}}\|^2_{L^2_{\rho_X}}\right), \\
\varepsilon(\overline{g}_n) - \varepsilon(g_{\mathcal{H}}) &= \langle \bar{\eta}_n, \Sigma\bar{\eta}_n\rangle.
\end{aligned}$$

On symmetric matrices

One has to be careful when using auto adjoint operators, especially when using the order $A \preccurlyeq B$ which means that $B - A$ is non-negative.

Some problems may arise when some self adjoint A, B do not commute, because then AB is not even in auto adjoint. It is also hopeless to compose such relations : for example $A \preccurlyeq B$ does not imply $A^2 \preccurlyeq B^2$ (while the opposite is true).

However, it is true that if $A \preccurlyeq B$, then for any C in $S_n(\mathbb{R})$, we have $C^t\,AC \preccurlyeq C^t\,BC$. We will often use this final point. Indeed for any x, $x^t\,(C^t\,BC - C^t\,AC)x = (Cx)^t(B-A)(Cx) \geqslant 0$.

Notation

In the proof, we may use, for any $x \in \mathcal{H}$:

$$\begin{aligned} \widetilde{K_x \otimes K_x} \ : L^2_{\rho_X} &\to \mathcal{H} \\ f &\mapsto f(x)\ K_x. \end{aligned}$$

We only consider functions $\mathcal{L}^2_{\rho_X}$, which are well defined at any point. The regression function is only almost surely defined but we will consider a version of the function in $\mathcal{L}^2_{\rho_X}$.

The following properties clearly hold :

- $\widetilde{K_x \otimes K_x}_{|\mathcal{H}} = K_x \otimes K_x$
- $\mathbb{E}\left(\widetilde{K_x \otimes K_x}\right) = \mathcal{T}$
- $\mathbb{E}\left(K_x \otimes K_x\right) = \Sigma$ as it has been noticed above.

For some $x \in \mathcal{X}$, we may denote $x \otimes x \ := K_x \otimes K_x$. Moreover, abusing notations, we may forget the $\sim$ in many cases.

A.4.2 Semi-stochastic recursion - intuition

We remind that :

$$g_n = (I - \gamma K_{x_n} \otimes K_{x_n}) g_{n-1} + \gamma y_n K_{x_n},$$

with $g_0 = 0$. We have denoted $\Xi_n = (y_n - g_{\mathcal{H}}(x_n))K_{x_n}$. Thus $y_n K_{x_n} = g_{\mathcal{H}}(x_n)K_{x_n} + \Xi_n \overset{\text{def}}{=} \widetilde{K_{x_n} \otimes K_{x_n}} g_{\mathcal{H}} + \Xi_n$, and our recursion may be rewritten :

$$g_n - g_{\mathcal{H}} = (I - \gamma \widetilde{K_{x_n} \otimes K_{x_n}})(g_{n-1} - g_{\mathcal{H}}) + \gamma \Xi_n, \tag{A.6}$$

Finally, we are studying a sequence $(\eta_n)_n$ defined by :

$$\begin{aligned} \eta_0 &= g_{\mathcal{H}}, \\ \eta_n &= (I - \gamma_n \widetilde{K_{x_n} \otimes K_{x_n}})\eta_{n-1} + \gamma_n \Xi_n. \end{aligned} \tag{A.7}$$

Behaviour : It appears that to understand how this will behave, we may compare it to the following recursion, which may be described as a "semi-stochastic" version of (A.7) : we keep the randomness due to the noise Ξ_n but forget the randomness due to sampling by replacing $\widetilde{K_{x_n} \otimes K_{x_n}}$ by its expectation Σ (T, more precisely) :

$$\eta_0^{ssto} = g_{\mathcal{H}},$$

$$\eta_n^{ssto} = (I - \gamma_n \Sigma)\eta_{n-1}^{ssto} + \gamma_n \Xi_n. \tag{A.8}$$

Complete proof : This comparison will give an interesting insight and the main terms of bias and variance will appear if we study (A.8). However this is not the true recursion : to get Theorem 2.9, we will have to do a bit of further work : we will first separate the error due to the noise from the error due to the initial condition, then link the true recursions to their "semi-stochastic" counterparts to make the variance and bias terms appear. That will be done in Section A.4.3.

Semi-stochastic recursion : In order to get such intuition, in both the finite horizon and on-line case, we will begin by studying the semi-stochastic equation (A.8).

First, we have, by induction:

$$\begin{aligned}
\forall j \geqslant 1 \qquad \eta_j^{ssto} &= (I - \gamma_j \Sigma)\eta_{j-1}^{ssto} + \gamma_j \Xi_j. \\
\eta_j^{ssto} &= \left[\prod_{i=1}^{j}(I - \gamma_i \Sigma)\right] \eta_0^{ssto} + \sum_{k=1}^{j} \left[\prod_{i=k+1}^{j} (I - \gamma_i \Sigma)\right] \gamma_k \Xi_k \\
\eta_j^{ssto} &= D(1, j, (\gamma_i)_i)\eta_0^{ssto} + \sum_{k=1}^{j} D(k+1, j, (\gamma_i)_i)\gamma_k \Xi_k \\
\overline{\eta}_n^{ssto} &= \frac{1}{n}\sum_{j=1}^{n} D(1, j, (\gamma_i)_i)\eta_0^{ssto} + \frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{j} D(1, j, (\gamma_i)_i)\gamma_k \Xi_k.
\end{aligned}$$

Then :

$$\begin{aligned}
\mathbb{E}\|\overline{\eta}_n^{ssto}\|_{L^2_{\rho_X}}^2 &= \frac{1}{n^2}\mathbb{E}\|\sum_{j=1}^{n} D(1, j, (\gamma_i)_i) g_{\mathcal{H}} + \sum_{j=1}^{n}\sum_{k=1}^{j} D(k+1, j, (\gamma_i)_i)\gamma_k \Xi_k\|_{L^2_{\rho_X}} \\
&= \underbrace{\frac{1}{n^2}\mathbb{E}\|\sum_{j=1}^{n} D(1, j, (\gamma_i)_i) g_{\mathcal{H}}\|_{L^2_{\rho_X}}}_{\text{Bias}(n)} \\
&+ \underbrace{2\frac{1}{n^2}\mathbb{E}\langle\sum_{j=1}^{n} D(1, j, (\gamma_i)_i) g_{\mathcal{H}}, \sum_{j=1}^{n}\sum_{k=1}^{j} D(k+1, j, (\gamma_i)_i)\gamma_k \Xi_k\rangle_{L^2_{\rho_X}}}_{=0 \text{ by (A.2)}}, \\
&+ \underbrace{\frac{1}{n^2}\mathbb{E}\|\sum_{j=1}^{n}\sum_{k=1}^{j} D(k+1, j, (\gamma_i)_i)\gamma_k \Xi_k\|_{L^2_{\rho_X}}}_{\text{var}(n)}
\end{aligned} \tag{A.9}$$

In the following, all calculations may be driven either with $\|\Sigma^{1/2} \cdot\|_K$ or in $\|\cdot\|_{L^2_{\rho_X}}$ using the isometrical character of $\Sigma^{1/2}$. In order to simplify comparison with existing work and especially (Bach and Moulines, 2013), we will mainly use the former as all calculations are only algebraic sums, we may sometimes use the notation $\langle x, \Sigma x\rangle_H$ instead of $\|\Sigma^{1/2}x\|_{\mathcal{H}}^2$. It is an abuse if $x \notin \mathcal{H}$, but however does not induce any confusion or mistake. In the following, if not explicitly specified, $\|\cdot\|$ will denote $\|\cdot\|_K$.

In the following we will thus denote:

$$\text{Bias}\left(n, (\gamma_i)_i, \Sigma, g_{\mathcal{H}}\right) = \frac{1}{n^2}\mathbb{E}\left\|\Sigma^{1/2}\sum_{j=1}^{n}\left[\prod_{i=1}^{j}(I - \gamma_i \Sigma)\right] g_{\mathcal{H}}\right\|_K^2$$

$$\mathrm{var}\left(n,(\gamma_i)_i,\Sigma,(\Xi_i)_i\right) \;=\; \frac{1}{n^2}\mathbb{E}\left\|\Sigma^{1/2}\sum_{j=1}^{n}\sum_{k=1}^{j}\left[\prod_{i=k+1}^{j}(I-\gamma_i\Sigma)\right]\gamma_k\Xi_k\right\|_K^2.$$

In section A.4.6 we will prove the following Lemmas which upper bound these bias and variance terms under different assumptions :

1. $\mathrm{Bias}\left(n,\gamma,\Sigma,g_{\mathcal{H}}\right)$ if we assume **A3,4**, γ constant,
2. $\mathrm{var}\left(n,\gamma,\Sigma,(\Xi_i)_i\right)$ if we assume **A3,6**, γ constant,
3. $\mathrm{Bias}\left(n,(\gamma_i)_i,\Sigma,g_{\mathcal{H}}\right)$ if we assume **A3,4** and $\gamma_i=\frac{1}{n^\zeta},\;\; 0\leqslant\zeta\leqslant 1$,
4. $\mathrm{var}\left(n,(\gamma_i)_i,\Sigma,(\Xi_i)_i\right)$ if we assume **A3,6** and $\gamma_i=\frac{1}{n^\zeta},\;\; 0\leqslant\zeta\leqslant 1$.

The two terms show respectively the impact :

1. of the initial setting and the hardness to forget the initial condition,
2. the noise.

Thus the first one tends to decrease when γ is increasing, whereas the second one increases when γ increases. We understand we may have to choose our step γ in order to optimize the trade-off between these two factors.

In the finite-dimensional case, it results from such a decomposition that if $C=\sigma^2\Sigma$ then $\mathbb{E}\left[\langle\overline{\alpha}_{n-1},\Sigma\overline{\alpha}_{n-1}\rangle\right]\leqslant\frac{1}{n\gamma}\|\alpha\|_0^2+\frac{\sigma^2 d}{n}$, as this upper bound is vacuous when d is either large or infinite, we can derive comparable bounds in the infinite-dimensional setting under our assumptions **A3,4,6**.

Lemma A.25 (Bias, **A3,4**, γ const.)**.** *Assume **A3-4** and let α (resp. r) be the constant in **A3** (resp. **A4**) :*

If $r\leqslant 1$ *:*

$$\mathrm{Bias}\left(n,\gamma,\Sigma,g_{\mathcal{H}}\right)\leqslant\|\Sigma^{-r}g_{\mathcal{H}}\|_{L^2_{\rho_X}}^2\left(\frac{1}{(n\gamma)^{2r}}\right)\overset{not}{=}\mathrm{bias}(n,\gamma,r).$$

If $r\geqslant 1$ *:*

$$\mathrm{Bias}\left(n,\gamma,\Sigma,g_{\mathcal{H}}\right)\leqslant\|\Sigma^{-r}g_{\mathcal{H}}\|_{L^2_{\rho_X}}^2\left(\frac{1}{n^2\gamma^r}\right)\overset{not}{=}\mathrm{bias}(n,\gamma,r).$$

Lemma A.26 (Var, **A3,4**, γ const)**.** *Assume **A3,6**, let α,s be the constants in **A3**, and σ the constant in **A6** (so that $\mathbb{E}\left[\Xi_n\otimes\Xi_n\right]\preccurlyeq\sigma^2\Sigma$).*

$$\mathrm{var}\left(n,\gamma,\Sigma,(\Xi_i)_i\right)\leqslant C(\alpha)\;s^{2/\alpha}\;\sigma^2\frac{\gamma^{\frac{1}{\alpha}}}{n^{1-\frac{1}{\alpha}}}+\frac{\sigma^2}{n}\overset{not}{=}\mathrm{var}(n,\gamma,\sigma^2,r,\alpha),$$

with $C(\alpha)=\frac{2\alpha^2}{(\alpha+1)(2\alpha-1)}$.

Lemma A.27 (Bias, **A3,4**, $(\gamma)_i$)**.** *Assume **A3-4** and let α (resp. r) be the constant in **A3** (resp. **A4**). Assume we consider a sequence $\gamma_i=\frac{\gamma_0}{i^\zeta}$ with $0<\zeta<1$ then :*

1. *if* $r(1-\zeta) < 1$*:*

$$\begin{aligned}\text{Bias}\left(n,(\gamma_i)_i,\Sigma,g_{\mathcal{H}}\right) &= O\left(\|\Sigma^{-r}g_{\mathcal{H}}\|^2_{L^2_{\rho_X}} n^{-2r(1-\zeta)}\right)\\ &= O\left(\|\Sigma^{-r}g_{\mathcal{H}}\|^2_{L^2_{\rho_X}} \frac{1}{(n\gamma_n)^{2r}}\right),\end{aligned}$$

2. *if* $r(1-\zeta) > 1$*:*

$$\text{Bias}\left(n,(\gamma_i)_i,\Sigma,g_{\mathcal{H}}\right) = O\left(\frac{1}{n^2}\right).$$

Lemma A.28 (Var, **A3,4**, $(\gamma)_i$)**.** *Assume **A3,6**, let* α, s *be the constants in **A3**, and* σ *the constant in **A6** . If we consider a sequence* $\gamma_i = \frac{\gamma_0}{i^\zeta}$ *with* $0 < \zeta < 1$ *then :*

1. *if* $0 < \zeta < \frac{1}{2}$ *then*

$$\text{var}\left(n,(\gamma_i)_i,\Sigma,(\Xi_i)_i\right) = O\left(n^{-1+\frac{1-\zeta}{\alpha}}\right) = O\left(\frac{\sigma^2(s^2\gamma_n)^{\frac{1}{\alpha}}}{n^{1-\frac{1}{\alpha}}}\right),$$

2. *and if* $\zeta > \frac{1}{2}$ *then*

$$\text{var}\left(n,(\gamma_i)_i,\Sigma,(\Xi_i)_i\right) = O\left(n^{-1+\frac{1-\zeta}{\alpha}+2\zeta-1}\right).$$

Those Lemmas are proved in section A.4.6.

Considering decomposition (A.9) and our Lemmas above, we can state a first Proposition.

Proposition A.29 (Semi-stochastic recursion)**.** *Assume **A1-6**. Let's consider the semi-stochastic recursion (that is the sequence :* $\eta_n = (I-\gamma_n\Sigma)\eta_{n-1} + \gamma_n\Xi_n$*) instead of our recursion initially defined. In the finite horizon setting, thus with* $\gamma_i = \gamma$ *for* $i \leqslant n$*, we have:*

$$\frac{1}{2}\mathbb{E}\left[\varepsilon\left(\overline{g}_n\right) - \varepsilon(g_\rho)\right] \leqslant C(\alpha)\, s^{\frac{2}{\alpha}}\, \sigma^2 \frac{\gamma^{\frac{1}{\alpha}}}{n^{1-\frac{1}{\alpha}}} + \frac{\sigma^2}{n} + \|\Sigma^{-r}g_\rho\|^2_{L^2_{\rho_X}} \left(\frac{1}{n^{2\min\{r,1\}}\gamma^{2r}}\right).$$

Theorem 2.9 must be compared to Proposition A.29 : Theorem 2.9 is just an extension but with the true stochastic recursion instead of the semi-stochastic one.

We finish this first part by a very simple Lemma which states that what we have done above is true for any semi stochastic recursion under few assumptions. Indeed, to get the complete bound, we will always come back to semi-stochastic type recursions, either without noise, or with a null initial condition.

Lemma A.30. *Let's assume:*

1. $\alpha_n = (I-\gamma\Sigma)\alpha_{n-1} + \gamma\Xi_n^\alpha$*, with* $\gamma\Sigma \preccurlyeq I$*.*
2. $(\Xi_n^\alpha) \in \mathcal{H}$ *is* $\mathcal{F}_n$ *measurable for a sequence of increasing* σ*-fields* $(\mathcal{F}_n)$*.*
3. $\mathbb{E}\left[\Xi_n^\alpha|\mathcal{F}_{n-1}\right] = 0$*,* $\mathbb{E}\left[\|\Xi_n^\alpha\|^2|\mathcal{F}_{n-1}\right]$ *is finite and* $\mathbb{E}\left[\Xi_n^\alpha \otimes \Xi_n^\alpha\right] \preccurlyeq \sigma_\alpha^2\Sigma$*.*

Then :

$$\mathbb{E}\left[\langle\overline{\alpha}_{n-1}, \Sigma\overline{\alpha}_{n-1}\rangle\right] = \text{Bias}\left(n,\gamma,\Sigma,\alpha_0\right) + \text{var}\left(n,\gamma,\Sigma,(\Xi_i^\alpha)_i\right). \tag{A.10}$$

And we may apply Lemmas A.25 and A.26 if we have good assumptions on Σ, α_0*.*

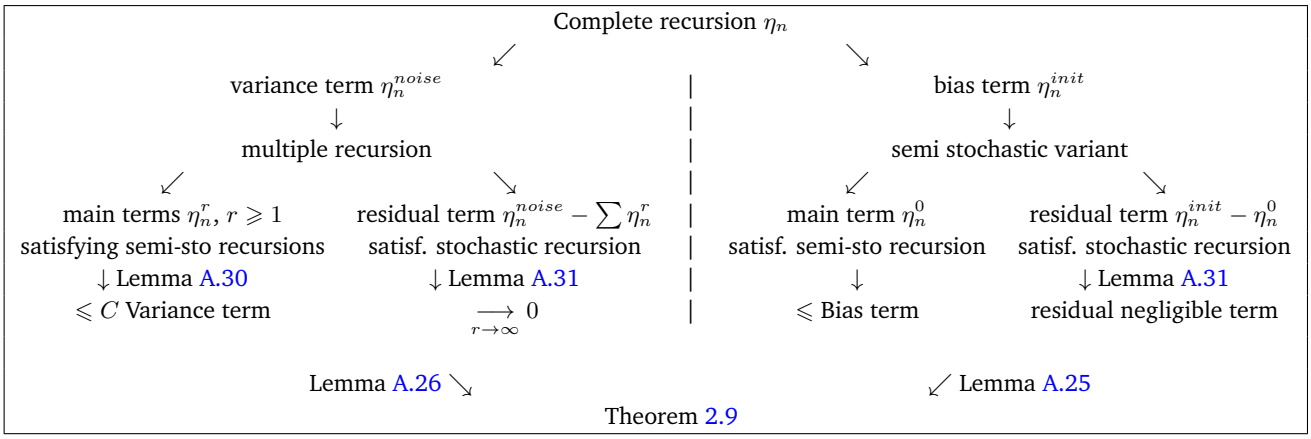


Table A.2: Error decomposition in the finite horizon setting.

A.4.3 Complete proof, Theorem 2.9 (finite horizon)

In the following, we will focus on the finite horizon setting, *i.e.*, we assume the step size is constant, but may depend on the total number of observations n: for all $1 \leqslant i \leqslant n$, $\gamma_i = \gamma = \Gamma(n)$. The main idea of the proof is to be able to :

1. separate the different sources of error (noise & initial conditions),
2. then bound the difference between the stochastic recursions and their semi-stochastic versions, a case in which we are able to compute bias and variance as it is done above.

Our main tool will be the Minkowski's inequality, which is the triangular inequality for $\mathbb{E}\left(\|\cdot\|_{L^2_{\rho_X}}\right)$. This will allow us to separate the error due to the noise from the error due to the initial conditions. The sketch of the decomposition is given in Table A.2.

We remind that $(\eta_n)_n$ is defined by :

$$\eta_0 = g_{\mathcal{H}}, \text{ and the recursion } \eta_n = (I - \gamma K_{x_n} \otimes K_{x_n})\eta_{n-1} + \gamma \Xi_n.$$

A Lemma on stochastic recursions

Before studying the main decomposition in Section A.4.3 we must give a classical Lemma on stochastic recursions which will be useful below :

Lemma A.31. *Assume $(x_n, \Xi_n) \in \mathcal{H} \times \mathcal{H}$ are $\mathcal{F}_n$ measurable for a sequence of increasing σ-fields $(\mathcal{F}_n)$. Assume that $\mathbb{E}\left[\Xi_n | \mathcal{F}_{n-1}\right] = 0$, $\mathbb{E}\left[\|\Xi_n\|^2 | \mathcal{F}_{n-1}\right]$ is finite and $\mathbb{E}\left[\|K_{x_n}\|^2 K_{x_n} \otimes K_{x_n} | \mathcal{F}_{n-1}\right] \preccurlyeq R^2\Sigma$, with $\mathbb{E}\left[K_{x_n} \otimes K_{x_n} | \mathcal{F}_{n-1}\right] = \Sigma$ for all $n \geqslant 1$, for some $R > 0$ and invertible operator Σ. Consider the recursion $\alpha_n = (I - \gamma K_{x_n} \otimes K_{x_n})\alpha_{n-1} + \gamma \Xi_n$, with $\gamma R^2 \leqslant 1$. Then :*

$$(1 - \gamma R^2)\, \mathbb{E}\left[\langle \overline{\alpha}_{n-1}, \Sigma \overline{\alpha}_{n-1} \rangle\right] + \frac{1}{2n\gamma} \mathbb{E}\|\alpha_n\|^2 \leqslant \frac{1}{2n\gamma}\|\alpha_0\|^2 + \frac{\gamma}{n} \sum_{k=1}^{n} \mathbb{E}\|\Xi_k\|^2.$$

Especially, if $\alpha_0 = 0$, we have

$$\mathbb{E}\left[\langle \overline{\alpha}_{n-1}, \Sigma \overline{\alpha}_{n-1} \rangle\right] \leqslant \frac{1}{(1 - \gamma R^2)} \frac{\gamma}{n} \sum_{k=1}^{n} \mathbb{E}\|\Xi_k\|^2.$$

Its proof may be found in Bach and Moulines (2013) : it is a direct consequence of the classical recursion to upper bound $\|\alpha_n\|^2$.

Main decomposition

We consider :

1. $(\eta_n^{init})_n$ defined by :

$$\eta_0^{init} = g_{\mathcal{H}} \text{ and } \eta_n^{init} = (I - \gamma K_{x_n} \otimes K_{x_n})\eta_{n-1}^{init}.$$

   η_n^{init} is the part of $(\eta_n)_n$ which is due to the **initial conditions** (it is equivalent to assuming $\Xi_n \equiv 0$).

2. Respectively, let $(\eta_n^{noise})_n$ be defined by :

$$\eta_0^{noise} = 0 \text{ and } \eta_n^{noise} = (I - \gamma K_{x_n} \otimes K_{x_n})\eta_{n-1}^{noise} + \gamma \Xi_n.$$

   η_n^{noise} is the part of $(\eta_n)_n$ which is due to **the noise**.

A straightforward induction shows that for any n, $\eta_n = \eta_n^{init} + \eta_n^{noise}$ and $\bar{\eta}_n = \bar{\eta}_n^{init} + \bar{\eta}_n^{noise}$. Thus Minkowski's inequality, applied to $\left(\mathbb{E}\left[\|\cdot\|^2_{L^2_{\rho_X}}\right]\right)^{1/2}$, leads to :

$$\left(\mathbb{E}\left[\|\bar{\eta}_n\|^2_{L^2_{\rho_X}}\right]\right)^{1/2} \leqslant \left(\mathbb{E}\left[\|\bar{\eta}_n^{noise}\|^2_{L^2_{\rho_X}}\right]\right)^{1/2} + \left(\mathbb{E}\left[\|\bar{\eta}_n^{init}\|^2_{L^2_{\rho_X}}\right]\right)^{1/2}$$

$$\left(\mathbb{E}\left[\langle \bar{\eta}_n, \Sigma \bar{\eta}_n \rangle\right]\right)^{1/2} \leqslant \left(\mathbb{E}\left[\langle \bar{\eta}_n^{noise}, \Sigma \bar{\eta}_n^{noise} \rangle\right]\right)^{1/2} + \left(\mathbb{E}\left[\langle \bar{\eta}_n^{init}, \Sigma \bar{\eta}_n^{init} \rangle\right]\right)^{1/2}. \tag{A.11}$$

That means we can always consider separately the effect of the noise and the effect of the initial conditions. We'll first study η_n^{noise} and then η_n^{init}.

Noise process

We remind that $(\eta_n^{noise})_n$ is defined by :

$$\eta_0^{noise} = 0 \text{ and } \eta_n^{noise} = (I - \gamma K_{x_n} \otimes K_{x_n})\eta_{n-1}^{noise} + \gamma \Xi_n. \tag{A.12}$$

We are going to define some other sequences, which are defined by the following "semi-stochastic" recursion, in which $K_{x_n} \otimes K_{x_n}$ has been replaced be its expectancy Σ : first we define $\left(\eta_n^{noise,0}\right)_n$ so that

$$\eta_0^{noise,0} = 0 \text{ and } \eta_n^{noise,0} = (I - \gamma \Sigma)\eta_{n-1}^{noise,0} + \gamma \Xi_n.$$

Triangular inequality will allow us to upper bound $\left(\mathbb{E}\left[\|\bar{\eta}_n^{noise}\|^2_{L^2_{\rho_X}}\right]\right)^{1/2}$:

$$\left(\mathbb{E}\left[\|\bar{\eta}_n^{noise}\|^2_{L^2_{\rho_X}}\right]\right)^{1/2} \leqslant \left(\mathbb{E}\left[\|\bar{\eta}_n^{noise,0}\|^2_{L^2_{\rho_X}}\right]\right)^{1/2} + \left(\mathbb{E}\left[\|\bar{\eta}_n^{noise} - \bar{\eta}_n^{noise,0}\|^2_{L^2_{\rho_X}}\right]\right)^{1/2} \tag{A.13}$$

So that we're interested in the sequence $\left(\eta_n^{noise} - \eta_n^{noise,0}\right)_n$: we have

$$\begin{aligned}
\eta_0^{noise} - \eta_0^{noise,0} &= 0, \\
\eta_n^{noise} - \eta_n^{noise,0} &= (I - \gamma K_{x_n} \otimes K_{x_n})(\eta_{n-1}^{noise} - \eta_{n-1}^{0}) + \gamma(\Sigma - K_{x_n} \otimes K_{x_n})\eta_{n-1}^{0}
\end{aligned}$$

$$= (I - \gamma K_{x_n} \otimes K_{x_n})(\eta_{n-1}^{noise} - \eta_{n-1}^{0}) + \gamma \Xi_n^1. \tag{A.14}$$

which is the same type of Equation as (A.12). We have denoted $\Xi_n^1 = (\Sigma - K_{x_n} \otimes K_{x_n})\eta_{n-1}^0$.

Thus we may consider the following sequence, satisfying the "semi-stochastic" version of recursion (A.14), changing $K_{x_n} \otimes K_{x_n}$ into its expectation Σ : we define $\left(\eta_n^{noise,1}\right)_n$ so that:

$$\eta_0^{noise,1} = 0 \text{ and } \eta_n^{noise,1} = (I - \gamma\Sigma)\eta_{n-1}^{noise,1} + \gamma\Xi_n^1. \tag{A.15}$$

Thanks to the triangular inequality, we're interested in $\left(\eta_n^{noise} - \eta_n^{noise,0} - \eta_n^{noise,1}\right)_n$, which satisfies the (A.12)-type recursion :

$$\begin{aligned}
\eta_0^{noise} - \eta_0^{noise,0} - \eta_0^{noise,1} &= 0, \\
\eta_n^{noise} - \eta_n^{noise,0} - \eta_n^{noise,1} &= (I - \gamma K_{x_n} \otimes K_{x_n})(\eta_{n-1}^{noise} - \eta_{n-1}^{noise,0} - \eta_n^{noise,1}) \\
&\quad + \gamma(\Sigma - K_{x_n} \otimes K_{x_n})\eta_{n-1}^{noise,1} \\
&= (I - \gamma K_{x_n} \otimes K_{x_n})(\eta_{n-1}^{noise} - \eta_{n-1}^{noise,0} - \eta_n^{noise,1}) + \gamma\Xi_n^{(2)}.
\end{aligned}$$

With $\Xi_n^{(2)} := (\Sigma - K_{x_n} \otimes K_{x_n})\eta_{n-1}^{noise,1}$.

And so on... For any $r \geqslant 0$ we define a sequence $(\eta_n^{noise,r})_n$ by :

$$\eta_0^{noise,r} = 0 \text{ and } \eta_n^{noise,r} = (I - \gamma\Sigma)\eta_{n-1}^{noise,r} + \gamma\Xi_n^r,$$
$$\text{with } \Xi_n^r = (\Sigma - K_{x_n} \otimes K_{x_n})\eta_{n-1}^{noise,r-1}.$$

We have, for any $r, n \in \mathbb{N}^2$:

$$\begin{aligned}
\eta_0^{noise} - \sum_{i=0}^{r} \eta_0^{noise,i} &= 0, \\
\eta_n^{noise} - \sum_{i=0}^{r} \eta_n^{noise,i} &= (I - \gamma K_{x_n} \otimes K_{x_n}) \left(\eta_{n-1}^{noise} - \sum_{i=0}^{r} \eta_{n-1}^{noise,i}\right) \\
&\quad + \gamma(\Sigma - K_{x_n} \otimes K_{x_n})\eta_{n-1}^{noise,r}. \\
&= (I - \gamma K_{x_n} \otimes K_{x_n}) \left(\eta_{n-1}^{noise} - \sum_{i=0}^{r} \eta_{n-1}^{noise,i}\right) + \gamma\Xi_n^{(r+1)}.
\end{aligned} \tag{A.16}$$

So that $(\eta_n^{noise,r+1})$ follows the "semi-stochastic" version of (A.16)...

Minkowski's inequality. Considering this decomposition, we have, for any r, using triangular inequality :

$$\left(\mathbb{E}\left[\|\bar{\eta}_n^{noise}\|_{L^2_{\rho_X}}^2\right]\right)^{1/2} \leqslant \sum_{i=0}^{r} \left(\mathbb{E}\left[\|\bar{\eta}_n^{noise,i}\|_{L^2_{\rho_X}}^2\right]\right)^{1/2} + \left(\mathbb{E}\left[\left\|\bar{\eta}_n^{noise} - \sum_{i=0}^{r} \bar{\eta}_n^{noise,i}\right\|_{L^2_{\rho_X}}^2\right]\right)^{1/2} \tag{A.17}$$

Moment Bounds. For any $i \geqslant 0$, we find that we may apply Lemma A.30 to the sequence $(\eta_n^{noise,i})$. Indeed :

1. For any $r \geqslant 0$, $(\eta_n^{noise,r})$ is defined by :

$$\eta_0^{noise,r} = 0 \text{ and } \eta_n^{noise,r} = (I - \gamma\Sigma)\eta_{n-1}^{noise,r} + \gamma\Xi_n^r,$$
$$\text{with } \Xi_n^r = \begin{cases} (\Sigma - K_{x_n} \otimes K_{x_n})\eta_{n-1}^{r-1} & \text{if } r \geqslant 1. \\ \Xi_n & \text{if } r = 0. \end{cases}$$

2. for any $r \geqslant 0$, for all $n \geqslant 0$, Ξ_n^r is $\mathcal{F}_n := \sigma\left((x_i, z_i)_{1 \leqslant i \leqslant n}\right)$ measurable. (for $r = 0$ we use the definition of Ξ_n (**H4**), and by induction, for any $r \geqslant 0$ if we have $\forall n \in \mathbb{N}$, Ξ_n^r is $\mathcal{F}_n$ measurable, then for any $n \in \mathbb{N}$, by induction on n, $\eta_n^{noise,r}$ is $\mathcal{F}_n$ measurable, thus for any $n \in \mathbb{N}$, Ξ_n^{r+1} is $\mathcal{F}_n$ measurable.)

3. for any $r, n \geqslant 0$, $\mathbb{E}\left[\Xi_n | \mathcal{F}_{n-1}\right] = 0$: as shown above, η_{n-1}^{r-1} is $\mathcal{F}_{n-1}$ measurable so $\mathbb{E}\left[\,\Xi_n\,|\mathcal{F}_{n-1}\right] = \mathbb{E}\left[\Sigma - K_{x_n} \otimes K_{x_n} | \mathcal{F}_{n-1}\right] \eta_{n-1}^{noise,r-1} = \mathbb{E}\left[\Sigma - K_{x_n} \otimes K_{x_n}\right] \eta_{n-1}^{noise,r-1} = 0$ (as x_n is independent of $\mathcal{F}_{n-1}$ by **A5** and $\mathbb{E}\left[\Sigma - K_{x_n} \otimes K_{x_n}\right] = \mathbb{E}\left[\Sigma - K_{x_n} \otimes K_{x_n}\right]$ by **H4**).

4. $\mathbb{E}\left[\|\Xi_n^r\|^2\right]$ is finite (once again, by **A2** if $r = 0$ and by a double recursion to get the result for any $r, n \geqslant 0$).

5. We have to find a bound on $\mathbb{E}\left[\Xi_n^r \otimes \Xi_n^r\right]$. To do that, we are going, once again to use induction on r.

Lemma A.32. *For any $r \geqslant 0$ we have*

$$\begin{aligned} \mathbb{E}\left[\Xi_n^r \otimes \Xi_n^r\right] &\preccurlyeq \gamma^r R^{2r} \sigma^2 \Sigma \\ \mathbb{E}\left[\eta_n^{noise,r} \otimes \eta_n^{noise,r}\right] &\preccurlyeq \gamma^{r+1} R^{2r} \sigma^2 I. \end{aligned}$$

Lemma A.32. We make an induction on n.

Initialization : for $r = 0$ we have by **A6** that $\mathbb{E}\left[\Xi_n^0 \otimes \Xi_n^0\right] \preccurlyeq \sigma^2 \Sigma$. Moreover

$$\begin{aligned} \mathbb{E}(\eta_n^0 \otimes \eta_n^0) &= \gamma^2 \sum_{k=1}^{n-1} (I - \gamma\Sigma)^{n-k} \mathbb{E}\left[\Xi_n^0 \otimes \Xi_n^0\right] (I - \gamma\Sigma)^{n-k} \\ &\preccurlyeq \gamma^2 \sigma^2 \sum_{k=1}^{n-1} (I - \gamma\Sigma)^{2(n-k)} \Sigma. \end{aligned}$$

We get

$$\forall n \geqslant 0, \quad \mathbb{E}\left[\eta_n^0 \otimes \eta_n^0\right] \preccurlyeq \gamma^2 \sigma^2 \sum_{k=1}^{n-1} (I - \gamma\Sigma)^{2n-2-k} \Sigma \preccurlyeq \gamma \sigma^2 I.$$

Recursion : If we assume that for any $n \geqslant 0, \mathbb{E}\left[\Xi_n^r \otimes \Xi_n^r\right] \preccurlyeq \gamma^r R^{2r} \sigma^2 \Sigma$ and $\mathbb{E}\left[\eta_n^r \otimes \eta_n^r\right] \preccurlyeq \gamma^{r+1} R^{2r} \sigma^2 I$ then for any $n \geqslant 0$:

$$\begin{aligned} \mathbb{E}\left[\Xi_n^{r+1} \otimes \Xi_n^{r+1}\right] &\preccurlyeq \mathbb{E}\left[(\Sigma - K_{x_n} \otimes K_{x_n}) \eta_{n-1}^r \otimes \eta_{n-1}^r (\Sigma - K_{x_n} \otimes K_{x_n})\right] \\ &= \mathbb{E}\left[(\Sigma - K_{x_n} \otimes K_{x_n}) \mathbb{E}\left[\eta_{n-1}^r \otimes \eta_{n-1}^r\right] (\Sigma - K_{x_n} \otimes K_{x_n})\right] \\ &\qquad\qquad \text{(as } \eta_{n-1} \in \mathcal{F}_{n-1}\text{)} \\ &\preccurlyeq \gamma^{r+1} R^{2r} \sigma^2 \mathbb{E}\left[(\Sigma - K_{x_n} \otimes K_{x_n})^2\right] \\ &\preccurlyeq \gamma^{r+1} R^{2r+2} \sigma^2 \Sigma. \end{aligned}$$

Once again we have $(\eta_n^{r+1}) = \gamma^2 \sum_{k=1}^{n-1} (I - \gamma\Sigma)^{n-1-k} \Xi_n^{r+1}$, for any n:

$$\begin{aligned} \mathbb{E}\left[\eta_n^{r+1} \otimes \eta_n^{r+1}\right] &\preccurlyeq \gamma^2 \mathbb{E}\left[\sum_{k=1}^{n} (I - \gamma\Sigma)^{n-1-k} \Xi_n^{r+1} \otimes \Xi_n^{r+1} (I - \gamma\Sigma)^{n-1-k}\right] \\ &\preccurlyeq \gamma^{r+3} R^{2r+2} \sigma^2 \sum_{k=1}^{n} (I - \gamma\Sigma)^{2n-2-2k} \Sigma \end{aligned}$$

$$\preccurlyeq \quad \gamma^{r+2}R^{2r+2}\sigma^2 I.$$

□

With the bound on $\mathbb{E}\left[\Xi_n^r \otimes \Xi_n^r\right]$ and as we have said, with Lemma A.30:

$$\begin{aligned}\mathbb{E}\left[\|\bar{\eta}_n^{noise,i}\|^2_{L^2_{\rho_X}}\right] = \mathbb{E}\left[\langle \bar{\eta}_n^i, \Sigma\bar{\eta}_n^i\rangle\right] &\leqslant \operatorname{var}(n,\gamma,\sigma^2\gamma^i R^{2i},s,\alpha)\\ &\leqslant \gamma^i R^{2i}\underset{\|}{\operatorname{var}}(n,\gamma,\sigma^2,s,\alpha)\,. \end{aligned} \tag{A.18}$$

Moreover, using the Lemma on stochastic recursions (Lemma A.31) for $(\bar{\eta}_n^{noise} - \sum_{i=0}^r \bar{\eta}_n^i)_n$ (all conditions are satisfied) we have :

$$\begin{aligned}(1-\gamma R^2)\,\mathbb{E}\left[\left\langle \bar{\eta}_n^{noise} - \sum_{i=0}^r \bar{\eta}_n^i, \Sigma\left(\bar{\eta}_n^{noise} - \sum_{i=0}^r \bar{\eta}_n^i\right)\right\rangle\right] &\leqslant \frac{\gamma}{n}\sum_{i=1}^n \mathbb{E}\|\Xi_k^{r+1}\|^2\\ &\leqslant \gamma\operatorname{tr}\left(\mathbb{E}\left[\Xi_k^{r+1}\otimes\Xi_k^{r+1}\right]\right)\\ &\leqslant \gamma^{r+2}R^{2r+2}\sigma^2\operatorname{tr}(\Sigma)\end{aligned}$$

$$\text{that is}\quad \mathbb{E}\left[\left\|\bar{\eta}_n^{noise} - \sum_{i=0}^r \bar{\eta}_n^{noise,i}\right\|^2_{L^2_{\rho_X}}\right] \leqslant \gamma^{r+2}R^{2r+2}\sigma^2\operatorname{tr}(\Sigma). \tag{A.19}$$

Conclusion. Thus using (A.17), (A.18) and (A.19) :

$$\begin{aligned}\left(\mathbb{E}\left[\langle \bar{\eta}_n^{noise}, \Sigma\bar{\eta}_n^{noise}\rangle\right]\right)^{1/2} &\leqslant \left(\frac{1}{1-\gamma R^2}\gamma^{r+2}\sigma^2 R^{2r+2}\operatorname{tr}(\Sigma)\right)^{1/2}\\ &\quad + \operatorname{var}(n,\gamma,\sigma^2,s,\alpha)^{1/2}\sum_{i=0}^r\left(\gamma R^2\right)^{i/2}.\end{aligned} \tag{A.20}$$

And using the fact that $\gamma R < 1$, when $r \to \infty$ we get:

$$\left(\mathbb{E}\left[\langle \bar{\eta}_n^{noise}, \Sigma\bar{\eta}_n^{noise}\rangle\right]\right)^{1/2} \leqslant \operatorname{var}(n,\gamma,\sigma^2,s,\alpha)^{1/2}\frac{1}{1-\sqrt{\gamma R^2}}. \tag{A.21}$$

Which is the main result of this part.

Initial conditions

We are now interested in getting such a bound for $\mathbb{E}\left[\langle \bar{\eta}_n^{init}, \Sigma\bar{\eta}_n^{init}\rangle\right]$. As this part stands for the initial conditions effect we may keep in mind that we would like to get an upper bound comparable to what we found for the Bias term in the proof of Proposition 1.

We remind that :

$$\eta_0^{init} = g_{\mathcal{H}} \text{ and } \eta_n^{init} = (I - \gamma K_{x_n}\otimes K_{x_n})\eta_{n-1}^{init}.$$

and define $(\eta_n^0)_{n\in\mathbb{N}}$ so that :

$$\eta_0^0 = g_{\mathcal{H}}, \quad \eta_n^0 = (I-\gamma\Sigma)\eta_{n-1}^0.$$

Minkowski's again. As above

$$\left(\mathbb{E}\left[\langle \bar{\eta}_n^{init}, \Sigma \bar{\eta}_n^{init}\rangle\right]\right)^{1/2} \leqslant \left(\mathbb{E}\left[\langle \bar{\eta}_n^{init} - \bar{\eta}_n^0, \Sigma\left(\bar{\eta}_n^{init} - \bar{\eta}_n^0\right)\rangle\right]\right)^{1/2} + \left(\mathbb{E}\left[\langle \bar{\eta}_n^0, \Sigma \bar{\eta}_n^0\rangle\right]\right)^{1/2}. \tag{A.22}$$

First for $\bar{\eta}_n^0$ we have a semi-stochastic recursion, with $\Xi_n \equiv 0$ so that we have

$$\mathbb{E}\langle \bar{\eta}_n^0, \Sigma \bar{\eta}_n^0\rangle \leqslant \text{bias}(n, \gamma, r).$$

Then , for the residual term we use Lemma A.31. Using that :

$$\eta_n^0 - \eta_n^{init} = (I - \gamma K_{x_n} \otimes K_{x_n})(\eta_n^0 - \eta_n^{init}) + \gamma(K_{x_n} \otimes K_{x_n} - \Sigma)\eta_{n-1}^0,$$

we may apply **Lemma A.31** to the recursion above with $\alpha_n = \eta_n^0 - \eta_n^{init}$ and $\Xi_n = (K_{x_n} \otimes K_{x_n} - \Sigma)\eta_{n-1}^0$. That is (as $\alpha_0 = 0$):

$$\mathbb{E}\langle \bar{\eta}_n^0 - \bar{\eta}_n^{noise}, \Sigma(\bar{\eta}_n^0 - \bar{\eta}_n^{noise})\rangle \leqslant \frac{1}{1-\gamma R^2}\frac{\gamma}{n}\mathbb{E}\left[\sum_{k=1}^n \|\Xi_k\|^2\right]. \tag{A.23}$$

Now

$$\begin{aligned}
\mathbb{E}\|\Xi_k\|^2 &= \mathbb{E}\left[\langle \eta_0, (I-\gamma\Sigma)^k(\Sigma - x_k \otimes x_k)^2(I-\gamma\Sigma)^k\eta_0\rangle\right] \\
&\leqslant \langle \eta_0, (I-\gamma\Sigma)^k R^2 \Sigma (I-\gamma\Sigma)^k \eta_0\rangle \\
&\leqslant R^2\langle \eta_0, (I-\gamma\Sigma)^{2k}\Sigma\eta_0\rangle.
\end{aligned}$$

Thus :

$$\begin{aligned}
\frac{\gamma}{n}\mathbb{E}\left[\sum_{k=1}^n \|\Xi_k\|^2\right] &\leqslant \frac{\gamma R^2}{n}\langle \eta_0, \sum_{k=1}^n (I-\gamma\Sigma)^{2k}\Sigma\eta_0\rangle \\
&\leqslant \frac{\gamma R^2}{n}\left\|\left(\sum_{k=1}^n (I-\gamma\Sigma)^{2k}\Sigma^{2r}\right)^{1/2}\Sigma^{1/2-r}\eta_0\right\|^2 \\
&\leqslant \frac{\gamma R^2}{n}\gamma^{-2r}\left|\left|\left|\sum_{k=1}^n (I-\gamma\Sigma)^{2k}(\gamma\Sigma)^{2r}\right|\right|\right| \|\Sigma^{-r}\eta_0\|_{L^2_{\rho_X}}^2.
\end{aligned}$$

$|||A^{1/2}|||^2 = |||A|||$. Moreover, as Σ is self adjoint, we have:

$$\begin{aligned}
\left|\left|\left|\sum_{k=1}^n (I-\gamma\Sigma)^{2k}(\gamma\Sigma)^{2r}\right|\right|\right| &\leqslant \sup_{0\leqslant x\leqslant 1}\sum_{k=1}^n (1-x)^{2k}(x)^{2r} \\
&\leqslant \sup_{0\leqslant x\leqslant 1}\frac{1-(1-x)^{2n}}{1-(1-x)^2}(x)^{2r} \\
&\leqslant \sup_{0\leqslant x\leqslant 1}\frac{1-(x)^{2n}}{1-x^2}(1-x)^{2r} \\
&\leqslant \sup_{0\leqslant x\leqslant 1}\frac{1-(x)^{2n}}{1+x}(1-x)^{2r-1} \\
&\leqslant \sup_{0\leqslant x\leqslant 1}(1-(1-x)^{2n})(x)^{2r-1} \\
&\leqslant n^{1-2r}
\end{aligned}$$

Where we have used inequality (A.53), if $r \leqslant 1/2$. However, this result does not stand anymore if $r \geqslant 1/2$. To deal with this particular case, we use the fact that,

$$\mathbb{E}\langle \bar{\eta}_n - \eta_*, \Sigma(\bar{\eta}_n - \eta_*)\rangle \leqslant (1 + (R^{2\alpha}\gamma^{1+\alpha}ns^2)^{\frac{2r-1}{\alpha}})\frac{\|\Sigma^{-r}\eta_0\|_{L2}^2}{(\gamma n)^{2r}}.$$

This result's proof is postponed to Lemma A.36.

So that we would get, replacing our result in (A.23) :

$$\mathbb{E}\langle \bar{\eta}_n^0 - \bar{\eta}_n^{noise}, \Sigma(\bar{\eta}_n^0 - \bar{\eta}_n^{noise})\rangle \leqslant \frac{1}{1-\gamma R^2}\frac{\gamma R^2}{(\gamma n)^{2r}}\|\Sigma^{-r}\eta_0\|^2_{L^2_{\rho_X}}. \tag{A.24}$$

Conclusion. Summing both bounds we get from (A.22) :

$$\left(\mathbb{E}\left[\langle \bar{\eta}_n^{init}, \Sigma\bar{\eta}_n^{init}\rangle\right]\right)^{1/2} \leqslant \left(\frac{1}{1-\gamma R^2}\frac{\gamma R^2}{(\gamma n)^{2r}}\|\Sigma^{-r}\eta_0\|^2_{L^2_{\rho_X}}\right)^{1/2} + (Bias(n,\gamma,g_{\mathcal{H}},\alpha))^{1/2}. \tag{A.25}$$

Conclusion

These two parts allow us to show Theorem 2.9 : using (A.25) and (A.21) in (A.11), and Lemmas A.25 and A.26 we have the final result.

Assuming **A1-6** :

1. If $r < 1$

$$\begin{aligned}\left(\mathbb{E}\left[\varepsilon\left(g_n\right) - \varepsilon(g_{\mathcal{H}})\right]\right)^{1/2} \leqslant\ & \frac{1}{1-\sqrt{\gamma R^2}}\left(C(\alpha)\, s^{\frac{2}{\alpha}}\,\sigma^2\frac{\gamma^{\frac{1}{\alpha}}}{n^{1-\frac{1}{\alpha}}} + \frac{\sigma^2}{n}\right)^{1/2} \\ & + \left(\|\Sigma^{-r}g_{\mathcal{H}}\|^2_{L^2_{\rho_X}}\left(\frac{1}{(n\gamma)^{2r}}\right)\right)^{1/2} \\ & + \left(\frac{1}{1-\gamma R^2}\frac{\gamma R^2}{(\gamma n)^{2r}}\|\Sigma^{-r}\eta_0\|^2_{L^2_{\rho_X}}\right)^{1/2}.\end{aligned}$$

2. If $r > 1$

$$\begin{aligned}\left(\mathbb{E}\left[\varepsilon\left(g_n\right) - \varepsilon(g_{\mathcal{H}})\right]\right)^{1/2} \leqslant\ & \frac{1}{1-\sqrt{\gamma R^2}}\left(C(\alpha)\, s^{\frac{2}{\alpha}}\,\sigma^2\frac{\gamma^{\frac{1}{\alpha}}}{n^{1-\frac{1}{\alpha}}} + \frac{\sigma^2}{n}\right)^{1/2} \\ & + \left(\|\Sigma^{-r}g_{\mathcal{H}}\|^2_{L^2_{\rho_X}}\left(\frac{1}{n^2\gamma^{2r}}\right)\right)^{1/2} \\ & + \left(\frac{1}{1-\gamma R^2}\frac{\gamma R^2}{(\gamma n)^{2r}}\|\Sigma^{-r}\eta_0\|^2_{L^2_{\rho_X}}\right)^{1/2}.\end{aligned}$$

Regrouping terms, we get:

Theorem A.33 (Complete bound, γ constant, finite horizon)**.** *Assume **(A1-6)** and $\gamma_i = \gamma = \Gamma(n)$, for $1 \leqslant i \leqslant n$. We have, with $C(\alpha) = \frac{2\alpha^2}{(\alpha+1)(2\alpha-1)}$:*

$$\begin{aligned}\left(\mathbb{E}\|\bar{g}_n - g_{\mathcal{H}}\|^2_{L^2_{\rho_X}}\right)^{1/2} \leqslant\ & \frac{\sigma/\sqrt{n}}{1-\sqrt{\gamma R^2}}\left(1 + C(\alpha)s^{\frac{2}{\alpha}}(\gamma n)^{\frac{1}{\alpha}}\right)^{\frac{1}{2}} \\ & + \frac{\|L_K^{-r}g_{\mathcal{H}}\|_{L^2_{\rho_X}}}{\gamma^r n^{\min\{r,1\}}}\left(1 + \frac{\sqrt{\gamma R^2}}{\sqrt{1-\gamma R^2}}\right).\end{aligned}$$

Then bounding $C(\alpha)$ by 1 and simplifying under assumption $\gamma R^2 \leqslant 1/4$, we exactly get Theorem 2.9 in the main text. In order to derive corollaries, one just has to chose $\gamma = \Gamma(n)$ in order to balance the main terms.

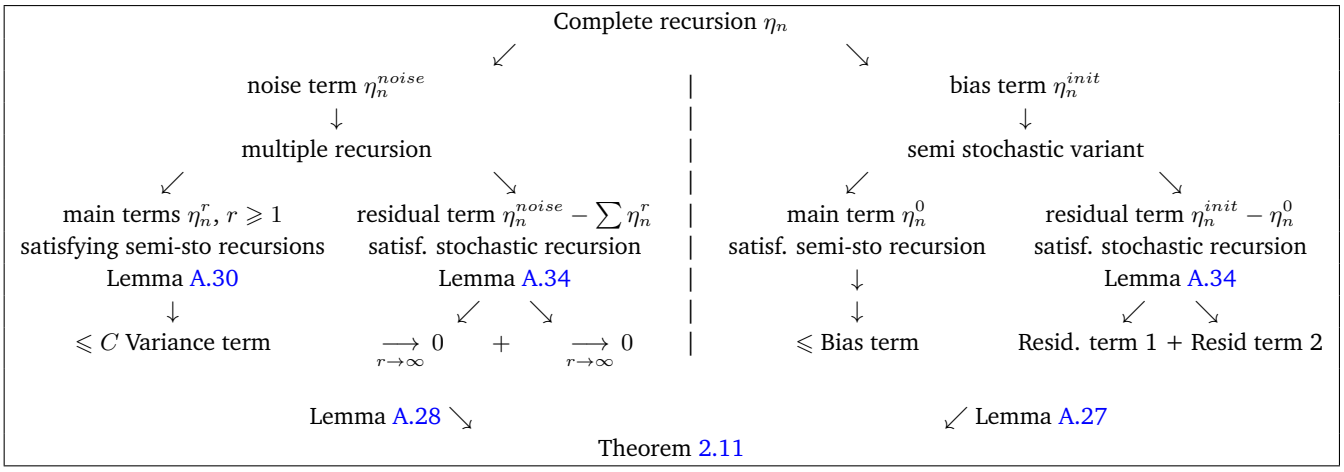


Table A.3: Sketch of the proof, on-line setting.

A.4.4 Complete proof, Theorem 2.11 (on-line setting)

The sketch of the proof is exactly the same. We just have to check that changing a constant step into a decreasing sequence of step-size does not change too much. However as most calculations make appear some weird constants, we will only look for asymptotics. The sketch of the decomposition is given in Table A.3.

A Lemma on stochastic recursions - on-line

We want to derive a Lemma comparable to Lemma A.31 in the online setting. That is considering a sequence $(\gamma_n)_n$ and the recursion $\alpha_n = (I - \gamma_n K_{x_n} \otimes K_{x_n})\alpha_{n-1} + \gamma_n \Xi_n$ we would like to have a bound on $\mathbb{E}\langle \overline{\alpha}_{n-1}, \Sigma \overline{\alpha}_{n-1} \rangle$.

Lemma A.34. *Assume $(x_n, \Xi_n) \in \mathcal{H} \times \mathcal{H}$ are $\mathcal{F}_n$ measurable for a sequence of increasing σ-fields $(\mathcal{F}_n)$. Assume that $\mathbb{E}\left[\Xi_n | \mathcal{F}_{n-1}\right] = 0$, $\mathbb{E}\left[\|\Xi_n\|^2 | \mathcal{F}_{n-1}\right]$ is finite and $\mathbb{E}\left[\|K_{x_n}\|^2 K_{x_n} \otimes K_{x_n} | \mathcal{F}_{n-1}\right] \preccurlyeq R^2\Sigma$, with $\mathbb{E}\left[K_{x_n} \otimes K_{x_n} | \mathcal{F}_{n-1}\right] = \Sigma$ for all $n \geqslant 1$, for some $R > 0$ and invertible operator Σ. Consider the recursion $\alpha_n = (I - \gamma_n K_{x_n} \otimes K_{x_n})\alpha_{n-1} + \gamma_n \Xi_n$, with $(\gamma_n)_n$ a sequence such that for any n, $\gamma_n R^2 \leqslant 1$. Then if $\alpha_0 = 0$, we have So that if $\alpha_0 = 0$:*

$$\mathbb{E}\left[\langle \overline{\alpha}_{n-1}, \Sigma \overline{\alpha}_{n-1} \rangle\right] \leqslant \frac{1}{2n(1-\gamma_0 R^2)} \left(\sum_{i=1}^{n-1} \|\alpha_i\|^2 \left(-\frac{1}{\gamma_i} + \frac{1}{\gamma_{i+1}} \right) + \sum_{k=1}^{n} \gamma_k \mathbb{E}\|\Xi_k\|^2 \right). \quad \text{(A.26)}$$

Proof.

$$2\gamma_n(1-\gamma_n R^2)\mathbb{E}\langle \Sigma\alpha_{n-1}, \alpha_{n-1} \rangle \leqslant \mathbb{E}\left(\|\alpha_{n-1}\|^2 - \|\alpha_n\|^2 + \gamma_n^2 \|\Xi_n\|^2 \right) \quad \text{(A.27)}$$

So that, if we assume that (γ_n) is non increasing:

$$\mathbb{E}\langle \Sigma\alpha_{n-1}, \alpha_{n-1} \rangle \leqslant \frac{1}{2\gamma_n(1-\gamma_0 R^2)} \mathbb{E}\left(\|\alpha_{n-1}\|^2 - \|\alpha_n\|^2 + \gamma_n^2 \|\Xi_n\|^2 \right) \quad \text{(A.28)}$$

Using convexity :

$$\mathbb{E}\left[\langle \overline{\alpha}_{n-1}, \Sigma \overline{\alpha}_{n-1} \rangle\right] \leqslant \frac{1}{2n(1-\gamma_0 R^2)} \left(\frac{\|\alpha_0\|^2}{\gamma_1} + \sum_{i=1}^{n-1} \|\alpha_i\|^2 \underbrace{\left(-\frac{1}{\gamma_i} + \frac{1}{\gamma_{i+1}} \right)}_{\geqslant 0} \right.$$

$$-\frac{\|\alpha_n\|^2}{\gamma_n}+\sum_{k=1}^{n}\gamma_k\mathbb{E}\|\Xi_k\|^2\Bigg).$$

So that if $\alpha_0=0$:

$$\mathbb{E}\left[\langle\overline{\alpha}_{n-1},\Sigma\overline{\alpha}_{n-1}\rangle\right]\leqslant\frac{1}{2n(1-\gamma_0R^2)}\left(\sum_{i=1}^{n-1}\|\alpha_i\|^2\left(-\frac{1}{\gamma_i}+\frac{1}{\gamma_{i+1}}\right)+\sum_{k=1}^{n}\gamma_k\mathbb{E}\|\Xi_k\|^2\right). \quad \text{(A.29)}$$

Note that it may be interesting to consider the weighted average $\tilde{\alpha}_n=\frac{\sum\gamma_i\alpha_i}{\sum\gamma_i}$, which would satisfy be convexity

$$\mathbb{E}\left[\langle\tilde{\alpha}_{n-1},\Sigma\tilde{\alpha}_{n-1}\rangle\right]\leqslant\frac{1}{2(\sum\gamma_i)(1-\gamma_0R^2)}\left(\frac{\|\alpha_0\|^2}{\gamma_1}-\frac{\|\alpha_n\|^2}{\gamma_n}+\sum_{k=1}^{n}\gamma_k^2\mathbb{E}\|\Xi_k\|^2\right). \quad \text{(A.30)}$$

□

Noise process

We remind that $(\eta_n^{noise})_n$ is defined by :

$$\eta_0^{noise}=0 \text{ and } \eta_n^{noise}=(I-\gamma K_{x_n}\otimes K_{x_n})\eta_{n-1}^{noise}+\gamma\Xi_n. \quad \text{(A.31)}$$

As before, for any $r\geqslant 0$ we define a sequence $(\eta_n^{noise,r})_n$ by :

$$\eta_0^{noise,r}=0 \text{ and } \eta_n^{noise,r}=(I-\gamma\Sigma)\eta_{n-1}^{noise,r}+\gamma\Xi_n^r,$$

$$\text{with } \Xi_n^r=(\Sigma-K_{x_n}\otimes K_{x_n})\eta_{n-1}^{noise,r-1}.$$

And we want to use the following upper bound

$$\left(\mathbb{E}\left[\|\bar{\eta}_n^{noise}\|_{L^2_{\rho_X}}^2\right]\right)^{1/2}\leqslant\sum_{i=0}^{r}\left(\mathbb{E}\left[\|\bar{\eta}_n^{noise,i}\|_{L^2_{\rho_X}}^2\right]\right)^{1/2}+\left(\mathbb{E}\left[\left\|\bar{\eta}_n^{noise}-\sum_{i=0}^{r}\bar{\eta}_n^{noise,i}\right\|_{L^2_{\rho_X}}^2\right]\right)^{1/2}. \quad \text{(A.32)}$$

So that we had to upper bound the noise :

Lemma A.35. *For any $r\geqslant 0$ we have $\mathbb{E}\left[\Xi_n^r\otimes\Xi_n^r\right]\preccurlyeq R^{2r}\gamma_0^r\sigma^2\Sigma$ and $\mathbb{E}\left[\eta_n^{noise,r}\otimes\eta_n^{noise,r}\right]\preccurlyeq\gamma_0^{r+1}R^{2r}\sigma^2 I$.*

Lemma A.35. We make an induction on n.
We note that :

$$\begin{aligned}\sum_{k=1}^{n}D(n,k+1,(\gamma_k)_k)\gamma_k^2\Sigma D(n,k+1,(\gamma_k)_k) &\leqslant \gamma_0\sum_{k=1}^{n}D(n,k+1,(\gamma_k)_k)\gamma_k\Sigma\\ &\leqslant \gamma_0\sum_{k=1}^{n}D(n,k+1,(\gamma_k)_k)-D(n,k,(\gamma_k)_k)\\ &\leqslant \gamma_0(I-D(n,1,(\gamma_k)_k))\\ &\leqslant \gamma_0 I \end{aligned} \quad \text{(A.33)}$$

Where we have used that : $D(n,k+1,(\gamma_k)_k)-D(n,k,(\gamma_k)_k)=D(n,k+1,(\gamma_k)_k)\gamma_k\Sigma$.
Initialization : for $r=0$ we have by **A6** that $\mathbb{E}\left[\Xi_n^0\otimes\Xi_n^0\right]\preccurlyeq\sigma^2\Sigma$. Moreover $\eta_n^0=\sum_{k=1}^{n}D(n,k+1,(\gamma_k)_k)\gamma_k\Xi_k^0$.

$$\begin{aligned}\mathbb{E}(\eta_n^0 \otimes \eta_n^0) &= \sum_{k=1}^{n} D(n,k+1,(\gamma_k)_k)\gamma_k^2 \mathbb{E}\left[\Xi_k^0 \otimes \Xi_k^0\right] D(k+1,n,(\gamma_k)_k) \\ &\preccurlyeq \sigma^2 \sum_{k=1}^{n} D(n,k+1,(\gamma_k)_k)\gamma_k^2 \Sigma D(k+1,n,(\gamma_k)_k) \\ &\preccurlyeq \sigma^2 \gamma_0 I, \quad \text{by (A.33)}\end{aligned}$$

Induction : If we assume $\forall n \geqslant 0, \quad \mathbb{E}\left[\Xi_n^r \otimes \Xi_n^r\right] \preccurlyeq \gamma_0^r R^{2r}\sigma^2\Sigma$ and $\mathbb{E}\left[\eta_n^r \otimes \eta_n^r\right] \preccurlyeq \gamma_0^{r+1} R^{2r}\sigma^2 I$ then: $\forall n \geqslant 0$,

$$\begin{aligned}\mathbb{E}\left[\Xi_n^{r+1} \otimes \Xi_n^{r+1}\right] &\preccurlyeq \mathbb{E}\left[(\Sigma - K_{x_n} \otimes K_{x_n})\eta_{n-1}^r \otimes \eta_{n-1}^r (\Sigma - K_{x_n} \otimes K_{x_n})\right] \\ &= \mathbb{E}\left[(\Sigma - K_{x_n} \otimes K_{x_n})\mathbb{E}\left[\eta_{n-1}^r \otimes \eta_{n-1}^r\right](\Sigma - K_{x_n} \otimes K_{x_n})\right] \\ &\qquad\qquad (\text{as } \eta_{n-1} \in \mathcal{F}_{n-1}) \\ &\preccurlyeq \gamma_0^{r+1} R^{2r}\sigma^2 \mathbb{E}\left[(\Sigma - K_{x_n} \otimes K_{x_n})^2\right] \\ &\preccurlyeq \gamma_0^{r+1} R^{2r+2}\sigma^2\Sigma.\end{aligned}$$

Once again we have $\eta_n^{r+1} = \sum_{k=1}^n D(n,k+1,(\gamma_k)_k)\gamma_k \Xi_k^{r+1}$, for any n:

$$\begin{aligned}\mathbb{E}\left[\eta_n^{r+1} \otimes \eta_n^{r+1}\right] &\preccurlyeq \gamma^2 \mathbb{E}\left[\sum_{k=1}^{n} (I-\gamma\Sigma)^{n-1-k}\Xi_n^{r+1} \otimes \Xi_n^{r+1}(I-\gamma\Sigma)^{n-1-k}\right] \\ &\preccurlyeq \sigma^2\gamma_0^{r+1}R^{2r}\sum_{k=1}^{n} D(n,k+1,(\gamma_k)_k)\gamma_k^2\Sigma D(k+1,n,(\gamma_k)_k) \\ &\preccurlyeq \sigma^2\gamma_0^{r+2}R^{2r} I, \quad \text{by (A.33)}\end{aligned}$$

□

With the bound on $\mathbb{E}\left[\Xi_n^r \otimes \Xi_n^r\right]$ and as we have said, with Lemma A.30:

$$\mathbb{E}\left[\|\bar{\eta}_n^{noise,i}\|_{L^2_{\rho_X}}^2\right] = \mathbb{E}\left[\langle \bar{\eta}_n^i, \Sigma\bar{\eta}_n^i\rangle\right] \leqslant \operatorname{var}(n,\gamma,\alpha,\gamma_0^i R^{2i}\sigma,s) = \gamma_0^i R^{2i}\operatorname{var}(n,\gamma,\alpha,\sigma,s). \tag{A.34}$$

Moreover, using the Lemma on stochastic recursions (Lemma A.34) for $(\alpha_n^r)_n = (\eta_n^{noise} - \sum_{i=0}^r \eta_n^i)_n$ (all conditions are satisfied) we have :

$$2(1-\gamma_0 R^2)\,\mathbb{E}\left[\left\langle \overline{\alpha}_n^r, \Sigma\overline{\alpha}_n^r\right\rangle\right] \leqslant \frac{1}{n}\left(\sum_{i=1}^{n-1}\mathbb{E}\|\alpha_i^r\|^2\left(-\frac{1}{\gamma_i}+\frac{1}{\gamma_{i+1}}\right) + \sum_{k=1}^{n}\gamma_k\mathbb{E}\|\Xi_k^{r+1}\|^2\right).$$

We are going to show that both these terms goes to 0 when r goes to infinity. Indeed :

$$\begin{aligned}\sum_{k=1}^{n}\gamma_k\mathbb{E}\|\Xi_k^{r+1}\|^2 &\leqslant \sum_{k=1}^{n}\gamma_k \operatorname{tr}\left(\mathbb{E}\left[\Xi_k^{r+1}\otimes\Xi_k^{r+1}\right]\right) \\ &\leqslant \sum_{k=1}^{n}\gamma_k\gamma_0^{r+1}R^{2r+2}\sigma^2\operatorname{tr}(\Sigma) \\ &\leqslant n\gamma_0^{r+2}R^{2r+2}\sigma^2\operatorname{tr}(\Sigma)\end{aligned}$$

Moreover, if we assume $\gamma_i = \frac{1}{i^\zeta}$:

$$\frac{1}{n}\sum_{i=1}^{n-1}\mathbb{E}\|\alpha_i^r\|^2\left(-\frac{1}{\gamma_i}+\frac{1}{\gamma_{i+1}}\right) \leqslant 2\zeta\frac{1}{n}\sum_{i=1}^{n-1}\frac{\gamma_i}{i}\mathbb{E}\|\alpha_i^r\|^2$$

And

$$\alpha_i^r = (I - \gamma_i \widetilde{K_{x_i} \otimes K_{x_i}})\alpha_{i-1}^r + \gamma_i \Xi_i$$

So that :

$$\begin{aligned}
\|\alpha_i^r\| &\leqslant |||(I - \gamma_i \widetilde{K_{x_i} \otimes K_{x_i}})||| \, \|\alpha_{i-1}^r\| + \gamma_i \, \|\Xi_i\| \\
&\leqslant \|\alpha_{i-1}^r\| + \gamma_i \, \|\Xi_i\| \\
&\leqslant \sum_{k=1}^{i} \gamma_k \, \|\Xi_k\|. \\
\text{thus : } \|\alpha_i^r\|^2 &\leqslant \sum_{k=1}^{i} \gamma_k \sum_{k=1}^{i} \gamma_k \, \|\Xi_k\|^2 \\
\mathbb{E}\|\alpha_i^r\|^2 &\leqslant \sum_{k=1}^{i} \gamma_k \sum_{k=1}^{i} \gamma_k \, \mathbb{E}\|\Xi_k\|^2 \\
\mathbb{E}\|\alpha_i^r\|^2 &\leqslant C_1 \;\; i\gamma_i \;\; i\gamma_0^{r+2} R^{2r+2} \sigma^2 \operatorname{tr}(\Sigma) \\
\frac{\gamma_i}{i} \, \mathbb{E}\|\alpha_i^r\|^2 &\leqslant C_2 \;\; i\gamma_i^2 \;\; (\gamma_0 R^2)^{r+2} \\
\frac{1}{n} \sum_{i=1}^{n-1} \mathbb{E}\|\alpha_i^r\|^2 \left(-\frac{1}{\gamma_i} + \frac{1}{\gamma_{i+1}} \right) &\leqslant C_3 \;\; n\gamma_n^2 \;\; (\gamma_0 R^2)^{r+2}.
\end{aligned}$$

That is:

$$E\left[\left\| \bar{\eta}_n^{noise} - \sum_{i=0}^{r} \bar{\eta}_n^{noise,i} \right\|_{L^2_{\rho_X}}^2 \right] \leqslant (\gamma_0 R^2)^{r+2} \left(\sigma^2 \operatorname{tr}(\Sigma) + C_3 n \gamma_n^2 \right). \tag{A.35}$$

With (A.32), (A.34),(A.35), we get :

$$\begin{aligned}
\left(\mathbb{E}\left[\|\bar{\eta}_n^{noise}\|_{L^2_{\rho_X}}^2 \right] \right)^{1/2} &\leqslant \sum_{i=0}^{r} \left(\gamma_0^i R^{2i} \operatorname{var}(n, \gamma, \alpha, \sigma, s) \right)^{1/2} \\
&\quad + \left((\gamma_0 R^2)^{r+2} \left(\sigma^2 \operatorname{tr}(\Sigma) + C_3 \, n\gamma_n^2 \right) \right)^{1/2}
\end{aligned} \tag{A.36}$$

So that, with $r \to \infty$:

$$\left(\mathbb{E}\left[\|\bar{\eta}_n^{noise}\|_{L^2_{\rho_X}}^2 \right] \right)^{1/2} \leqslant \left(C \operatorname{var}(n, \gamma, \alpha, \sigma, s) \right)^{1/2}. \tag{A.37}$$

Initial conditions

Exactly as before, we can separate the effect of initial conditions and of noise : We are interested in getting such a bound for $\mathbb{E}\left[\langle \bar{\eta}_n^{init}, \Sigma \bar{\eta}_n^{init} \rangle\right]$. We remind that :

$$\eta_0^{init} = g_{\mathcal{H}} \text{ and } \eta_n^{init} = (I - \gamma_n K_{x_n} \otimes K_{x_n}) \eta_{n-1}^{init}.$$

and define $(\eta_n^0)_{n \in \mathbb{N}}$ so that :

$$\eta_0^0 = g_{\mathcal{H}}, \quad \eta_n^0 = (I - \gamma_n \Sigma) \eta_{n-1}^0.$$

Minkowski's again : As above

$$\left(\mathbb{E}\left[\langle \bar{\eta}_n^{init}, \Sigma \bar{\eta}_n^{init} \rangle \right] \right)^{1/2} \leqslant \left(\mathbb{E}\left[\langle \bar{\eta}_n^{init} - \bar{\eta}_n^0, \Sigma \left(\bar{\eta}_n^{init} - \bar{\eta}_n^0 \right) \rangle \right] \right)^{1/2} + \left(\mathbb{E}\left[\langle \bar{\eta}_n^0, \Sigma \bar{\eta}_n^0 \rangle \right] \right)^{1/2}. \tag{A.38}$$

First for $\overline{\eta}_n^0$ we have a semi-stochastic recursion, with $\Xi_n \equiv 0$ so that we have

$$\langle \overline{\eta}_n^0, \Sigma \overline{\eta}_n^0 \rangle \leqslant \mathrm{Bias}(n, (\gamma_n)_n, g_{\mathcal{H}}, r). \tag{A.39}$$

Then, for the residual term we use **Lemma** A.34 for the recursion above with $\alpha_n = \eta_n^0 - \eta_n^{init}$. Using that :

$$\eta_n^0 - \eta_n^{init} = (I - \gamma K_{x_n} \otimes K_{x_n})(\eta_n^0 - \eta_n^{init}) + \gamma_n (K_{x_n} \otimes K_{x_n} - \Sigma)\eta_{n-1}^0,$$

That is (as $\alpha_0 = 0$):

$$\begin{aligned} \mathbb{E}\langle \bar{\eta}_n^0 - \bar{\eta}_n^{noise}, \Sigma(\bar{\eta}_n^0 - \bar{\eta}_n^{noise})\rangle &\leqslant \frac{1}{2n(1-\gamma_0 R^2)} \left(\sum_{i=1}^{n-1} \mathbb{E}\|\alpha_i\|^2 \left(-\frac{1}{\gamma_i} + \frac{1}{\gamma_{i+1}} \right) \right. \\ &\qquad \left. + \sum_{k=1}^{n} \gamma_k \mathbb{E}\|\Xi_k\|^2 \right). \end{aligned} \tag{A.40}$$

Now

$$\begin{aligned} \mathbb{E}\|\Xi_k\|^2 &= \mathbb{E}\left[\langle \eta_0, D(n,1,(\gamma_i)_i)(\Sigma - x_k \otimes x_k)^2 D(1,n,(\gamma_i)_i)\eta_0 \rangle \right] \\ &\leqslant R^2 \langle \eta_0, D(1,n,(\gamma_i)_i)^2 \Sigma \eta_0 \rangle. \end{aligned}$$

Thus :

$$\begin{aligned} \mathbb{E}\left[\sum_{k=1}^{n} \gamma_k \|\Xi_k\|^2 \right] &\leqslant R^2 \langle \eta_0, \sum_{k=1}^{n} \gamma_k D(1,n,(\gamma_i)_i)^2 \Sigma \eta_0 \rangle \\ &\leqslant R^2 \left\| \left(\sum_{k=1}^{n} \gamma_k D(1,n,(\gamma_i)_i)^2 \Sigma^{2r} \right)^{1/2} \Sigma^{1/2-r} \eta_0 \right\|^2 \\ &\leqslant R^2 \left\| \left\| \sum_{k=1}^{n} D(1,n,(\gamma_i)_i)^2 \gamma_k \Sigma^{2r} \right\| \right\| \|\Sigma^{-r}\eta_0\|_{L^2_{\rho_X}}^2. \end{aligned} \tag{A.41}$$

Now :

$$\begin{aligned} \left\| \left\| \sum_{k=1}^{n} D(1,n,(\gamma_i)_i)^2 \gamma_k \Sigma^{2r} \right\| \right\| &\leqslant \sup_{0 \leqslant x \leqslant 1/\gamma_0} \sum_{k=1}^{n} \prod_{i=1}^{n} (1-\gamma_i x)^2 \gamma_k x^{2r} \\ &\leqslant \sup_{0 \leqslant x \leqslant 1/\gamma_0} \sum_{k=1}^{n} \exp\left(-\sum_{i=1}^{k} \gamma_i x \right) \gamma_k x^{2r} \\ &\leqslant \sup_{0 \leqslant x \leqslant 1/\gamma_0} \sum_{k=1}^{n} \exp\left(-k\gamma_k x\right) \gamma_k x^{2r} \quad \text{if } (\gamma_k)_k \text{ is decreasing} \\ &\leqslant \gamma_0 \sup_{0 \leqslant x \leqslant 1/\gamma_0} \sum_{k=1}^{n} \exp\left(-k\gamma_k x\right) x^{2r} \\ &\leqslant \gamma_0 \sup_{0 \leqslant x \leqslant 1/\gamma_0} \sum_{k=1}^{n} \exp\left(-k^{1-\rho}\gamma_0 x\right) x^{2r} \quad \text{if } (\gamma_k)_i = \frac{\gamma_0}{k^\rho} \\ &\leqslant \gamma_0 \sup_{0 \leqslant x \leqslant 1/\gamma_0} x^{2r} \int_{u=0}^{n} \exp\left(-u^{1-\rho}\gamma_0 x\right) du \end{aligned}$$

$$\int_{u=0}^{n-1} \exp\left(-u^{1-\rho}\gamma_0 x\right) du \leqslant n \quad \text{clearly, but also}$$

$$\int_{u=0}^{n-1} \exp\left(-u^{1-\rho}\gamma_0 x\right) du \leqslant \int_{t=0}^{\infty} \exp\left(-t^{1-\rho}\right) (x\gamma_0)^{-\frac{1}{1-\rho}} dt \quad \text{changing variables. So that :}$$

$$\begin{aligned}\left|\left|\left|\sum_{k=1}^{n} D(1,n,(\gamma_i)_i)^2\gamma_k\Sigma^{2r}\right|\right|\right| &\leqslant \gamma_0 \sup_{0\leqslant x\leqslant 1/\gamma_0} x^{2r}\left(n\wedge I(x\gamma_0)^{-\frac{1}{1-\rho}}\right)\\ &\leqslant \gamma_0 C_1 \sup_{0\leqslant x\leqslant 1/\gamma_0}\left(nx^{2r}\wedge x^{2r-\frac{1}{1-\rho}}\right) \text{ and if } 2r-\frac{1}{1-\rho}<0\\ &\leqslant \gamma_0 C_1 n^{1-2r(1-\rho)}. \end{aligned} \tag{A.42}$$

And finally, using (A.41), (A.42) :

$$\begin{aligned}\frac{1}{2n(1-\gamma_0 R^2)}\sum_{k=1}^{n}\gamma_k\mathbb{E}\|\Xi_k\|^2 &\leqslant \frac{\gamma_0 C_1\ \|\Sigma^{-r}\eta_0\|^2_{L^2_{\rho_X}} R^2}{2(1-\gamma_0 R^2)}(n\gamma_n)^{-2r}\\ &\leqslant K(n\gamma_n)^{-2r}. \end{aligned} \tag{A.43}$$

To conclude, we have to upper bound :

$$\frac{1}{2n(1-\gamma_0 R^2)}\sum_{i=1}^{n-1}\mathbb{E}\|\alpha_i\|^2\left(-\frac{1}{\gamma_i}+\frac{1}{\gamma_{i+1}}\right).$$

By the induction we make to get **Lemma** A.34, we have :

$$\begin{aligned}\|\alpha_i\|^2 &\leqslant \|\alpha_{i-1}\|^2+\gamma_i^2\|\Xi_i\|^2\\ &\leqslant \sum_{k=1}^{i}\gamma_k^2\|\Xi_k\|^2\\ &\leqslant \sum_{k=1}^{i}\gamma_k\|\Xi_k\|^2\\ &\leqslant Ci\ (i\gamma_i)^{-2r}. \end{aligned}$$

So that (C changes during calculation) :

$$\begin{aligned}\frac{1}{2n(1-\gamma_0 R^2)}\sum_{i=1}^{n-1}\mathbb{E}\|\alpha_i\|^2\left(-\frac{1}{\gamma_i}+\frac{1}{\gamma_{i+1}}\right) &\leqslant C\frac{1}{n}\sum_{i=1}^{n-1}\mathbb{E}\|\alpha_i\|^2\frac{\gamma_i}{i}\\ &\leqslant C\frac{1}{n}\sum_{i=1}^{n-1} i\ (i\gamma_i)^{-2r}\frac{\gamma_i}{i}\\ &\leqslant C\frac{1}{n}\sum_{i=1}^{n-1}(i\gamma_i)^{-2r}\gamma_i\\ &\leqslant C\frac{\gamma_n}{(n\gamma_n)^{2r}}. \end{aligned}$$

So that we would get, replacing our result in (A.40) :

$$\mathbb{E}\langle\bar{\eta}_n^0-\bar{\eta}_n^{noise},\Sigma(\bar{\eta}_n^0-\bar{\eta}_n^{noise})\rangle = O\left(\frac{1}{n\gamma_n}\right)^{2r}+O\left(\frac{\gamma_n}{n\gamma_n}\right)^{2r}=O\left(\frac{1}{n\gamma_n}\right)^{2r}. \tag{A.44}$$

And finally, with (A.39) and (A.44) in (A.38),

$$\begin{aligned}\left(\mathbb{E}\left[\langle\bar{\eta}_n^{init},\Sigma\bar{\eta}_n^{init}\rangle\right]\right)^{1/2} &\leqslant \left(\mathbb{E}\left[\langle\bar{\eta}_n^{init}-\bar{\eta}_n^0,\Sigma\left(\bar{\eta}_n^{init}-\bar{\eta}_n^0\right)\rangle\right]\right)^{1/2}+\left(\mathbb{E}\left[\langle\bar{\eta}_n^0,\Sigma\bar{\eta}_n^0\rangle\right]\right)^{1/2}\\ &\leqslant \left(O\left(\frac{1}{n\gamma_n}\right)^{2r}\right)^{1/2}+\text{bias}(n,(\gamma_n)_n,g_{\mathcal{H}},r)^{1/2}. \end{aligned} \tag{A.45}$$

Conclusion

We conclude with both (A.37) and (A.45) in (A.11) :

$$\left(\mathbb{E}\left[\|\bar{\eta}_n\|^2_{L^2_{\rho_X}}\right]\right)^{1/2} \leqslant (C\,\mathrm{var}(n,\gamma,\alpha,\sigma,s))^{1/2} + \left(O\left(\frac{1}{n\gamma_n}\right)^{2r}\right)^{1/2} + \mathrm{bias}(n,(\gamma_n)_n,g_{\mathcal{H}},r)^{1/2}. \tag{A.46}$$

Which gives Theorem 2.11 using Lemmas A.27 and A.28. Once again, deriving corollaries is simple.

A.4.5 A lemma on stochastic recursion, $r \geqslant 1/2$

Lemma A.36. *If we consider the recursion $\eta_{n+1} = (I - \gamma x_n \otimes x_n)\eta_n$, with $\eta_0 = g_{\mathcal{H}}$, we have*

$$\mathbb{E}\langle \bar{\eta}_n, \Sigma\bar{\eta}_n\rangle \quad \leqslant \quad (1 + (R^{2\alpha}\gamma^{1+\alpha}ns^2)^{\frac{2r-1}{\alpha}})\frac{\|\Sigma^{-r}\eta_0\|^2_{L2}}{(\gamma n)^{2r}}.$$

Let us first state a few properties of symmetric matrices that are useful here. First we recall that the order $\preccurlyeq$ is defined by $M \preccurlyeq N$ if $N - M$ is sdp. This is an order on S_n, which is not a total order. We say that a function f is matrix increasing if $M \preccurlyeq N$ implies $f(M) \preccurlyeq f(N)$. We have the following special cases:

- The function $M \mapsto M^2$ is not matrix increasing on S_n^+.
- For any $q \in [0;1]$, the function $M \mapsto M^q$ is matrix increasing on S_n^+.
- For any N, the function $M \mapsto N^\top MN$ is matrix increasing.
- For some N, the function $M \mapsto N^\top M + MN$ is not matrix increasing.
- For any N, the function $M \mapsto (N \otimes \mathrm{I} + \mathrm{I} \otimes N)^{-1}$ is matrix increasing.
- $\exp$ is not matrix increasing, $\log$ is matrix increasing.
- if $A \preccurlyeq B$, $A \preccurlyeq C$ and $BC = CB$, then for any $q \in [0;1]$, $A \preccurlyeq B^q C^{1-q}$.

It is important to notice that it often occurs that f is matrix increasing and its inverse f^{-1} is not (square/square root, exp/log, left and right multiplication).

Notation $\wedge$ is somehow ill defined as $A \wedge B$ may be neither A nor B. However, we use this notation as a shortcut for a matrix C such that $C \preccurlyeq A$ and $C \preccurlyeq B$.

To prove Lemma A.36, we consider the stochastic recursion, in the case $r \geqslant 1/2$.

We consider a full expansion of the function value $\|\Sigma^{1/2}(\bar{\eta}_n)\|^2$. This corresponds to

$$\eta_n = \theta_n - \theta_0 = M(n,1)(\eta_0) = M(n,1)(\theta_0 - g_{\mathcal{H}})$$

Where the matrix $M(n,k)$ is $(I - \gamma x_n \otimes x_n)\cdots(I - \gamma x_k \otimes x_k)$, for $k \leqslant n$ We consider the recursion without noise and rely on an explicit decomposition.

$$\begin{aligned} n^2\mathbb{E}\langle \bar{\eta}_n, \Sigma(\bar{\eta}_n)\rangle &= \mathbb{E}\sum_{i=0}^{n}\sum_{j=0}^{n}\langle \eta_i, \Sigma(\eta_j)\rangle \\ &= \mathbb{E}\sum_{i=0}^{n}\langle \eta_i, \Sigma(\eta_i)\rangle + 2\mathbb{E}\sum_{i=0}^{n-1}\sum_{j=i+1}^{n}\langle \eta_i, \Sigma(\eta_j)\rangle. \end{aligned}$$

Moreover,

$$\begin{aligned}
&\mathbb{E}\sum_{i=0}^{n-1}\sum_{j=i+1}^{n}\langle\eta_i,\Sigma(\eta_j)\rangle\\
=\;&\mathbb{E}\sum_{i=0}^{n-1}\sum_{j=i+1}^{n}\left\langle\eta_i,\Sigma\Big[M(j,i+1)(\eta_i)\Big]\right\rangle\\
=\;&\mathbb{E}\sum_{i=0}^{n-1}\sum_{j=i+1}^{n}\langle\eta_i,\Sigma(\mathrm{I}-\gamma\Sigma)^{j-i}(\eta_i)\rangle \text{ because } M(j,i+1) \text{ and } \eta_i \text{ are independent,}\\
=\;&\mathbb{E}\sum_{i=0}^{n-1}\Big\langle\eta_i,(\gamma^{-1}\big[(\mathrm{I}-\gamma\Sigma)-(\mathrm{I}-\gamma\Sigma)^{n-i+1}\big]\wedge n\Sigma(\mathrm{I}-\gamma\Sigma))(\eta_i)\Big\rangle\\
\leqslant\;&\mathbb{E}\sum_{i=0}^{n}\langle\eta_i,A_{i,n}(\eta_i)\rangle-\mathbb{E}\sum_{i=0}^{n}\langle\eta_i,\Sigma(\eta_i)\rangle.
\end{aligned}$$

with $A_{i,n}\preccurlyeq(\gamma^{-1}I\wedge n\Sigma)$ (meaning $A_{i,n}\preccurlyeq\gamma^{-1}I$ and $A_{i,n}\preccurlyeq n\Sigma$). As for $i\in[0;n]$, $A_{i,n}\preccurlyeq A_{0,n}=:A$, we only need to get an upper bound on: $\mathbb{E}\sum_{i=0}^{n}\langle\eta_i,A(\eta_i)\rangle$, to get a bound on $n^2\mathbb{E}\|\Sigma^{1/2}(\bar{\eta}_n)\|^2$.

However,

$$\begin{aligned}
\mathbb{E}\sum_{i=0}^{n}\langle\eta_i,A(\eta_i)\rangle&=\mathbb{E}\sum_{i=0}^{n}\langle(\eta_0),(M(i,1))^*AM(i,1)(\eta_0)\rangle\\
&=\langle\langle\mathbb{E}\sum_{i=0}^{n}(M(i,1))^\top AM(i,1),E_0\rangle\rangle
\end{aligned}$$

as $\eta_i=M(i,1)(\eta_0)$, with $E_0=(\eta_0)(\eta_0)^*$. and $\langle\langle\cdot,\cdot\rangle\rangle$ denoting the Froebenius scalar product between matrices.

We consider the operator T from symmetric matrices to symmetric matrices defined as

$$TA=\Sigma A+A\Sigma-\gamma E\big[x_n\otimes x_nAx_n\otimes x_n\big].$$

of the form $TA=\Sigma A+A\Sigma-\gamma SA$.

We can make the following remarks:

- Operator $(I-\gamma T)$ is matrix increasing, as it is by definition $M\mapsto\mathbb{E}[x_n\otimes x_n\Sigma x_n\otimes x_n]$.
- The operator S is self-adjoint and positive.

We have for any symmetric matrix A:

$$\mathbb{E}M(i,1)^*AM(i,1)=(\mathrm{I}-\gamma T)^iA.$$

Thus,

$$\mathbb{E}\sum_{i=0}^{n}M(i,1)^*AM(i,1)=\sum_{i=0}^{n}(\mathrm{I}-\gamma T)^iA$$

We have from previous calculations (case $r=1/2$ previously) that the main quantity of interest $\sum_{i=0}^{n}(\mathrm{I}-\gamma T)^iA$ satisfies:

$$\gamma\sum_{i=0}^{n}(\mathrm{I}-\gamma T)^iA\preccurlyeq nI. \tag{A.47}$$

We now show the following lemma:

Lemma A.37.

$$\gamma \sum_{i=0}^{n} (\mathrm{I} - \gamma T)^i A \leqslant \gamma^{-1}\Sigma^{-1} + (n\gamma)^{1/\alpha} \operatorname{tr}(\Sigma^{\alpha})\Sigma^{-1}. \tag{A.48}$$

Proof. Let us denote P the quantity of interest $P = \sum_{i=0}^{n} (\mathrm{I} - \gamma T)^i A$. We clearly have that $P \preccurlyeq nA$, and $P \preccurlyeq \gamma^{-1}T^{-1}A$. As a consequence, we first consider an upper bound on $M_A := T^{-1}A$.

We have:

$$A \quad = \quad \Sigma M_A + M_A \Sigma - \gamma S M_A. \tag{A.49}$$

Thus:

$$M_A \quad = \quad \left[\Sigma \otimes \mathrm{I} + \mathrm{I} \otimes \Sigma\right]^{-1} A + \gamma \left[\Sigma \otimes \mathrm{I} + \mathrm{I} \otimes \Sigma\right]^{-1} S M_A$$

As $\left[\Sigma \otimes \mathrm{I} + \mathrm{I} \otimes \Sigma\right]^{-1}$ is a matrix increasing operator, we have that: $\left[\Sigma \otimes \mathrm{I} + \mathrm{I} \otimes \Sigma\right]^{-1} A \preccurlyeq \left[\Sigma \otimes \mathrm{I} + \mathrm{I} \otimes \Sigma\right]^{-1} n\Sigma = \frac{n}{2}\mathrm{I}$ and $\left[\Sigma \otimes \mathrm{I} + \mathrm{I} \otimes \Sigma\right]^{-1} A \preccurlyeq \left[\Sigma \otimes \mathrm{I} + \mathrm{I} \otimes \Sigma\right]^{-1} \gamma^{-1}\mathrm{I} = \frac{1}{2\gamma}\Sigma^{-1}$.

Moreover,

$$S M_A \quad \preccurlyeq \quad \operatorname{tr}(S M_A)\mathrm{I} \tag{A.50}$$

Moreover we can upper bound $\operatorname{tr}(SM)$: as

$$\operatorname{tr}(A) = 2 \operatorname{tr}(\Sigma M_A) - \gamma \operatorname{tr} \mathbb{E}(x_n \otimes x_n M_A x_n \otimes x_n)$$

And

$$\operatorname{tr} \mathbb{E}(x_n \otimes x_n M_A x_n \otimes x_n) \leqslant R^2 \operatorname{tr} M_A \Sigma.$$

This implies

$$\operatorname{tr} A \quad \geqslant \quad \frac{1}{R^2} \operatorname{tr} S M_A.$$

And as $A \preccurlyeq n^{1/\alpha}\gamma^{-1/\alpha}\Sigma^{1/\alpha}$, we finally have:

$$S M_A \quad \preccurlyeq \quad R^2 n^{1/\alpha}\gamma^{-1+1/\alpha} \operatorname{tr}(\Sigma^{1/\alpha})\mathrm{I}$$

Thus:

$$\begin{aligned} P \quad &\preccurlyeq \quad \gamma^{-2}\Sigma^{-1} + R^2 n^{1/\alpha}\gamma^{-1+1/\alpha} \operatorname{tr}(\Sigma^{1/\alpha})\Sigma^{-1} \\ \gamma P \quad &\preccurlyeq \quad (\gamma^{-1} + R^2 n^{1/\alpha}\gamma^{1/\alpha} \operatorname{tr}(\Sigma^{1/\alpha}))\Sigma^{-1}. \end{aligned}$$

□

Combining equation (A.47) and (A.48) we get, for any $q \in [0; 1]$:

$$\gamma P \quad \preccurlyeq \quad (\gamma^{-1} + R^2 (n^{1/\alpha}\gamma^{1/\alpha} \operatorname{tr}(\Sigma^{1/\alpha}))^q n^{1-q}\Sigma^{-q} \tag{A.51}$$

Thus with $q = -1 + 2r$, for $r \in [1/2; 1]$, we get

$$\gamma P \quad \preccurlyeq \quad (\gamma^{-1} + R^2 (n^{1/\alpha}\gamma^{1/\alpha} \operatorname{tr}(\Sigma^{1/\alpha}))^{2r-1} n^{2-2r}\Sigma^{1-2r} \tag{A.52}$$

Thus

$$\begin{aligned}\mathbb{E}\langle\langle P, E_0\rangle\rangle &\leqslant \frac{1}{\gamma}(\gamma^{-1} + R^2\gamma^{1/\alpha}n^{1/\alpha}\operatorname{tr}(\Sigma^{1/\alpha}))^{2r-1}n^{2-2r}\|\Sigma^{-r}\eta_0\|_{L2}^2 \\ &= \frac{1}{\gamma^{2r}}(1 + R^2\gamma^{1+1/\alpha}n^{1/\alpha}\operatorname{tr}(\Sigma^{1/\alpha}))^{2r-1}n^{2-2r}\|\Sigma^{-r}\eta_0\|_{L2}^2\end{aligned}$$

And the quantity $R^2\gamma^{1+1/\alpha}n^{1/\alpha}$ is always going to 0 for the optimal choice of γ. Moreover, the exponent $2r-1$ is logically vanishing when $r \to 1/2$.

And we have

$$\begin{aligned}\mathbb{E}\langle\bar{\eta}_n, \Sigma(\bar{\eta}_n)\rangle &\leqslant (1 + R^2\gamma^{1+1/\alpha}n^{1/\alpha}\operatorname{tr}(\Sigma^{1/\alpha}))^{2r-1}\frac{\|\Sigma^{-r}\eta_0\|_{L2}^2}{(\gamma n)^{2r}} \\ &\leqslant (1 + (R^{2\alpha}\gamma^{1+\alpha}ns^2)^{\frac{2r-1}{\alpha}})\frac{\|\Sigma^{-r}\eta_0\|_{L2}^2}{(\gamma n)^{2r}}\end{aligned}$$

which concludes the proof of the Lemma.

A.4.6 Some quantities

In this section, we bound the main quantities which are involved above.

Lemma A.25

Lemma A.25.

If $0 \leqslant r \leqslant 1$:

$$\begin{aligned}\operatorname{Bias}(n,\gamma,g_{\mathcal{H}},r) &= \frac{1}{n^2}\langle\sum_{k=0}^{n-1}(I-\gamma\Sigma)^k g_{\mathcal{H}}, \sum_{k=0}^{n-1}(I-\gamma\Sigma)^k\ \Sigma g_{\mathcal{H}}\rangle \\ &= \frac{1}{n^2}\langle\sum_{k=0}^{n-1}(I-\gamma\Sigma)^k\Sigma^{2r}\Sigma^{-r+1/2}g_{\mathcal{H}}, \sum_{k=0}^{n-1}(I-\gamma\Sigma)^k\ \Sigma^{-r+1/2}g_{\mathcal{H}}\rangle \\ &= \frac{1}{n^2}\left\|\sum_{k=0}^{n-1}(I-\gamma\Sigma)^k\Sigma^r(\Sigma^{-r+1/2}g_{\mathcal{H}})\right\|^2 \\ &\leqslant \frac{1}{n^2}\left|\left|\left|\sum_{k=0}^{n-1}(I-\gamma\Sigma)^k\Sigma^r\right|\right|\right|^2\ \left\|\Sigma^{-r+1/2}g_{\mathcal{H}}\right\|^2 \\ &= \frac{1}{n^2}\gamma^{-2r}\left|\left|\left|\sum_{k=0}^{n-1}(I-\gamma\Sigma)^k\gamma^r\Sigma^r\right|\right|\right|^2\ \left\|\Sigma^{-r}g_{\mathcal{H}}\right\|_{\mathcal{L}^2_\rho}^2 \\ &\leqslant \frac{1}{n^2}\gamma^{-2r}\sup_{0\leqslant x\leqslant 1}\left(\sum_{k=0}^{n-1}(1-x)^k x^r\right)^2\left\|\Sigma^{-r}g_{\mathcal{H}}\right\|_{\mathcal{L}^2_\rho}^2 \\ &\leqslant \left(\frac{1}{(n\gamma)^{2r}}\right)\left\|\Sigma^{-r}g_{\mathcal{H}}\right\|_{L^2_{\rho_X}}^2.\end{aligned}$$

Using the inequality:

$$\sup_{0\leqslant x\leqslant 1}\left(\sum_{k=0}^{n-1}(1-x)^k x^r\right) \leqslant n^{1-r}. \tag{A.53}$$

Indeed:

$$\left(\sum_{k=0}^{n-1}(1-x)^k x^r\right) = \frac{1-(1-x)^n}{x}x^r$$

$$= \quad (1-(1-x)^n)x^{r-1}.$$

And we have, for any $n \in \mathbb{N}, r \in [\,0;1], x \in [\,0;1]$: $(1-(1-x)^n) \leqslant (nx)^{1-r}$:

1. if $nx \leqslant 1$ then $(1-(1-x)^n) \leqslant nx \leqslant (nx)^{1-r}$ (the first inequality can be proved by deriving the difference).
2. if $nx \geqslant 1$ then $(1-(1-x)^n) \leqslant 1 \leqslant (nx)^{1-r}$.

If $r \geqslant 1$, $x \mapsto (1-(1-x)^n)$ is increasing on $[\,0;1]$ so $\sup_{0\leqslant x\leqslant 1}\left(\sum_{k=0}^{n-1}(1-x)^k x^r\right) = 1$: there is no improvement in comparison to $r = 1$:

$$\mathrm{Bias}(n,\gamma,g_{\mathcal{H}},r) \quad \leqslant \quad \left(\frac{1}{n^2\gamma^{2r}}\right)\left|\left|\Sigma^{-r}g_{\mathcal{H}}\right|\right|^2_{L^2_{\rho_X}}.$$

□

Lemma A.26

Lemma A.26 .
In the following proof, we consider $s = 1$. It's easy to get the complete result replacing in the proof below " γ" by " $s^2\gamma$". We have, for $j \in \mathbb{N}$, still assuming $\gamma\Sigma \preccurlyeq I$, and by a comparison to the integral:

$$\begin{aligned}
\mathrm{tr}\left(I-(I-\gamma\Sigma)^j\right)^2\Sigma^{-1}C &= \sigma^2\,\mathrm{tr}\left(I-(I-\gamma\Sigma)^j\right)^2 \\
&\leqslant 1+\sigma^2\int_{u=1}^{\infty}\left(1-\left(1-\frac{\gamma}{u^\alpha}\right)^j\right)^2 du \\
&\quad \text{(1 stands for the first term in the sum)} \\
&= 1+\sigma^2\int_{u=1}^{(\gamma j)^{\frac{1}{\alpha}}}\left(1-\left(1-\frac{\gamma}{u^\alpha}\right)^j\right)^2 du \\
&\qquad +\sigma^2\int_{u=(\gamma j)^{\frac{1}{\alpha}}}^{\infty}\left(1-\left(1-\frac{\gamma}{u^\alpha}\right)^j\right)^2 du.
\end{aligned}$$

Note that the first integral may be empty if $\gamma j \leqslant 1$. We also have:

$$\mathrm{tr}\left(I-(I-\gamma\Sigma)^j\right)^2\Sigma^{-1}C \quad \geqslant \quad \sigma^2\int_{u=1}^{\infty}\left(1-\left(1-\frac{\gamma}{u^\alpha}\right)^j\right)^2 du.$$

Considering that $g_j : u \mapsto \left(1-(1-\frac{\gamma}{u^\alpha})^j\right)^2$ is a decreasing function of u we get:

$$\forall u \in [1;(\gamma j)^{\frac{1}{\alpha}}], \quad (1-e^{-1})^2 \leqslant g_j(u) \leqslant 1.$$

Where we have used the fact that $\left(1-\frac{1}{j}\right)^j \leqslant e^{-1}$ for the left hand side inequality. Thus we have proved:

$$(1-e^{-1})^2(\gamma j)^{\frac{1}{\alpha}} \leqslant \int_{u=1}^{(\gamma j)^{\frac{1}{\alpha}}}\left(1-\left(1-\frac{\gamma}{u^\alpha}\right)^j\right)^2 du \leqslant (\gamma j)^{\frac{1}{\alpha}}.$$

For the other part of the sum, we consider $h_j : u \mapsto \left(\frac{1-\left(1-\frac{\gamma}{u^\alpha}\right)^j}{\frac{\gamma}{u^\alpha}}\right)^2$ which is an increasing function of u. So:

$$\forall u \in [(\gamma j)^{\frac{1}{\alpha}}; +\infty], \quad (1-e^{-1})^2 j^2 \leqslant h_j(u) \leqslant j^2,$$

using the same trick as above. Thus:

$$\begin{aligned}
\int_{u=(\gamma j)^{\frac{1}{\alpha}}}^{\infty} \left(1-\left(1-\frac{\gamma}{u^\alpha}\right)^j\right)^2 du &= \int_{u=(\gamma j)^{\frac{1}{\alpha}}}^{\infty} h_j(u) \left(\frac{\gamma}{u^\alpha}\right)^2 du \\
&\leqslant j^2 \int_{u=(\gamma j)^{\frac{1}{\alpha}}}^{\infty} \left(\frac{\gamma}{u^\alpha}\right)^2 du \\
&\leqslant j^2\gamma^2 \int_{u=(\gamma j)^{\frac{1}{\alpha}}}^{\infty} \left(\frac{1}{u^\alpha}\right)^2 du \\
&= j^2\gamma^2 \left[\frac{1}{(1-2\alpha)u^{2\alpha-1}}\right]_{u=(\gamma j)^{\frac{1}{\alpha}}}^{\infty} \\
&= j^2\gamma^2 \frac{1}{(2\alpha-1)((\gamma j)^{\frac{1}{\alpha}})^{2\alpha-1}} \\
&= \frac{1}{(2\alpha-1)}(j\gamma)^{\frac{1}{\alpha}}.
\end{aligned}$$

And we could get, by a similar calculation:

$$\int_{u=(\gamma j)^{\frac{1}{\alpha}}+1}^{\infty} \left(1-\left(1-\frac{\gamma}{u^\alpha}\right)^j\right)^2 du \geqslant (1-e^{-1})^2 \frac{1}{(2\alpha-1)}(j\gamma)^{\frac{1}{\alpha}}.$$

Finally, we have shown that:

$$C_1(j\gamma)^{\frac{1}{\alpha}} \leqslant \operatorname{tr}\left(I-(I-\gamma\Sigma)^j\right)^2 \leqslant C_2(j\gamma)^{\frac{1}{\alpha}} + 1.$$

Where $C_1 = (1-e^{-1})^2\,(1+\frac{1}{(2\alpha-1)})$ and $C_2 = (1+\frac{1}{(2\alpha-1)})$ are real constants. To get the complete variance term we have to calculate: $\frac{\sigma^2}{n^2}\sum_{j=1}^{n-1} \operatorname{tr}\left(I-(I-\gamma\Sigma)\right)^j$. We have:

$$\begin{aligned}
\frac{\sigma^2}{n^2}\sum_{j=1}^{n-1} \operatorname{tr}\left(I-(I-\gamma\Sigma)^j\right)^2 &\leqslant \frac{\sigma^2}{n^2}\sum_{j=1}^{n-1}\left(C_2(j\gamma)^{\frac{1}{\alpha}}+1\right) \\
&\leqslant \frac{\sigma^2}{n^2} C_2\gamma^{\frac{1}{\alpha}} \int_{u=2}^{n} u^{\frac{1}{\alpha}} du + \frac{\sigma^2}{n} \\
&\leqslant \frac{\sigma^2}{n^2} C_2\gamma^{\frac{1}{\alpha}} \frac{\alpha}{\alpha+1} n^{\frac{\alpha+1}{\alpha}} + \frac{\sigma^2}{n} \\
&\leqslant \frac{\alpha\,\sigma^2\,C_2}{\alpha+1}\,\frac{\gamma^{\frac{1}{\alpha}}}{n^{1-\frac{1}{\alpha}}} + \frac{\sigma^2}{n}.
\end{aligned}$$

That is:

$$(1-e^{-1})^2\,C(\alpha)\,\sigma^2 \frac{\gamma^{\frac{1}{\alpha}}}{(n-1)^{1-\frac{1}{\alpha}}} \leqslant \operatorname{var}(n,\gamma,\alpha,\sigma^2) \leqslant C(\alpha)\,\sigma^2 \frac{\gamma^{\frac{1}{\alpha}}}{n^{1-\frac{1}{\alpha}}} + \frac{\sigma^2}{n},$$

with $C(\alpha) = \frac{2\alpha^2}{(\alpha+1)(2\alpha-1)}$.

□

Lemma A.27

Proof.

$$\begin{aligned}\frac{1}{n^2}\left\|\Sigma^{1/2}\sum_{k=1}^{n}\prod_{i=1}^{k}\left(I-\gamma_i\Sigma\right)g_{\mathcal{H}}\right\|_K^2 &\leqslant \frac{1}{n^2}\left|\left|\left|\sum_{k=1}^{n}\prod_{i=1}^{k}\left(I-\gamma_i\Sigma\right)\Sigma^r\right|\right|\right|^2||\Sigma^{1/2-r}g_{\mathcal{H}}||_K^2\\ &\leqslant \frac{1}{n^2}\left|\left|\left|\sum_{k=1}^{n}\prod_{i=1}^{k}\left(I-\gamma_i\Sigma\right)\Sigma^r\right|\right|\right|^2||\Sigma^{-r}g_{\mathcal{H}}||_{L^2_{\rho_X}}^2.\end{aligned}$$

Moreover:

$$\begin{aligned}\left|\left|\left|\sum_{k=1}^{n}\prod_{i=1}^{k}\left(I-\gamma_i\Sigma\right)\Sigma^r\right|\right|\right| &\leqslant \sup_{0\leqslant x\leqslant 1}\sum_{k=1}^{n}\prod_{i=1}^{k}\left(I-\gamma_i x\right)x^r\\ &\leqslant \sup_{0\leqslant x\leqslant 1}\sum_{k=1}^{n}\exp\left(-\sum_{i=1}^{k}\gamma_i x\right)x^r\\ &\leqslant \sup_{0\leqslant x\leqslant 1}\sum_{k=1}^{n}\exp\left(-k\gamma_k x\right)x^r \quad \text{if } (\gamma_k)_k \text{ is decreasing}\\ &\leqslant \sup_{0\leqslant x\leqslant 1}\sum_{k=1}^{n}\exp\left(-k^{1-\zeta}x\right)x^r \quad \text{if } \gamma_k=\frac{1}{k^\zeta}\\ &\leqslant \sup_{0\leqslant x\leqslant 1}x^r\int_{u=0}^{n}\exp\left(-u^{1-\zeta}x\right)du \quad \text{by comparison to the integral}\\ \int_{u=0}^{n}\exp\left(-u^{1-\zeta}x\right)du &\leqslant n \quad \text{clearly, but also}\\ \int_{u=0}^{n}\exp\left(-u^{1-\zeta}x\right)du &\leqslant \int_{t=0}^{\infty}\exp\left(-t^{1-\zeta}\right)(x)^{-\frac{1}{1-\zeta}}dt \quad \text{changing variables. So that:}\\ \left|\left|\left|\sum_{k=1}^{n}\prod_{i=1}^{k}\left(I-\gamma_i\Sigma\right)\Sigma^r\right|\right|\right| &\leqslant K\sup_{0\leqslant x\leqslant 1}x^r\left(n\wedge x^{-\frac{1}{1-\zeta}}\right)\\ &\leqslant K\sup_{0\leqslant x\leqslant 1}\left(nx^r\wedge x^{r-\frac{1}{1-\zeta}}\right) \text{ and if } r-\frac{1}{1-\zeta}<0\\ &\leqslant Kn^{1-r(1-\zeta)}.\end{aligned}$$

So that:

$$\begin{aligned}\frac{1}{n^2}\left\langle\sum_{k=1}^{n}\prod_{i=1}^{k}\left(I-\gamma_i\Sigma\right)g_{\mathcal{H}},\sum_{k=1}^{n}\prod_{i=1}^{k}\left(I-\gamma_i\Sigma\right)\Sigma g_{\mathcal{H}}\right\rangle &\leqslant \frac{1}{n^2}\left(Kn^{1-r(1-\zeta)}\right)^2||\Sigma^{-r}g_{\mathcal{H}}||_{L^2_{\rho_X}}^2\\ &\leqslant K^2||\Sigma^{-r}g_{\mathcal{H}}||_{L^2_{\rho_X}}^2 n^{-2r(1-\zeta)}.\end{aligned}$$

Else if $r-\frac{1}{1-\zeta}>0$, then $\sup_{0\leqslant x\leqslant 1}\left(nx^r\wedge x^{r-\frac{1}{1-\zeta}}\right)=1$, so that

$$\frac{1}{n^2}\left\langle\sum_{k=1}^{n}\prod_{i=1}^{k}\left(I-\gamma_i\Sigma\right)g_{\mathcal{H}},\sum_{k=1}^{n}\prod_{i=1}^{k}\left(I-\gamma_i\Sigma\right)\Sigma g_{\mathcal{H}}\right\rangle = O\left(\frac{||\Sigma^{-r}g_{\mathcal{H}}||_{L^2_{\rho_X}}^2}{n^2}\right).$$

□

Lemma A.28

Proof. To get corollary A.28, we will just replace in the following calculations γ by $s^2\gamma$ We remind that:

$$\operatorname{var}\left(n,(\gamma_i)_i,\Sigma,(\xi_i)_i\right)=\frac{1}{n^2}\mathbb{E}\left\langle\sum_{j=1}^{n}\sum_{k=1}^{j}\left[\prod_{i=k+1}^{j}(I-\gamma_i\Sigma)\right]\gamma_k\xi_k,\Sigma\sum_{j=1}^{n}\sum_{k=1}^{j}\left[\prod_{i=k+1}^{j}(I-\gamma_i\Sigma)\right]\gamma_k\xi_k\right\rangle. \tag{A.54}$$

For shorter notation, in the following proof, we note $\operatorname{var}(n)=\operatorname{var}\left(n,(\gamma_i)_i,\Sigma,(\xi_i)_i\right)$.

$$\begin{aligned}
\operatorname{var}(n) &= \frac{1}{n^2}\mathbb{E}\left\langle\sum_{j=1}^{n}\sum_{k=1}^{j}\left[\prod_{i=k+1}^{j}(I-\gamma_i\Sigma)\right]\gamma_k\xi_k,\Sigma\sum_{j=1}^{n}\sum_{k=1}^{j}\left[\prod_{i=k+1}^{j}(I-\gamma_i\Sigma)\right]\gamma_k\xi_k\right\rangle\\
&= \frac{1}{n^2}\mathbb{E}\left\langle\sum_{k=1}^{n}\left(\sum_{j=k}^{n}\left[\prod_{i=k+1}^{j}(I-\gamma_i\Sigma)\right]\right)\gamma_k\xi_k,\Sigma\sum_{k=1}^{n}\left(\sum_{j=k}^{n}\left[\prod_{i=k+1}^{j}(I-\gamma_i\Sigma)\right]\right)\gamma_k\xi_k\right\rangle\\
&= \frac{1}{n^2}\sum_{k=1}^{n}\mathbb{E}\left\langle\left(\sum_{j=k}^{n}\left[\prod_{i=k+1}^{j}(I-\gamma_i\Sigma)\right]\right)\gamma_k\xi_k,\Sigma\left(\sum_{j=k}^{n}\left[\prod_{i=k+1}^{j}(I-\gamma_i\Sigma)\right]\right)\gamma_k\xi_k\right\rangle\\
&= \frac{1}{n^2}\sum_{k=1}^{n}\mathbb{E}\left\langle M_{n,k}\gamma_k\xi_k,\Sigma M_{n,k}\gamma_k\xi_k\right\rangle \qquad \text{with } M_{n,k}:=\left(\sum_{j=k}^{n}\left[\prod_{i=k+1}^{j}(I-\gamma_i\Sigma)\right]\right)\\
&= \frac{1}{n^2}\sum_{k=1}^{n}\gamma_k^2\,\mathbb{E}\left\langle M_{n,k}\xi_k,\Sigma M_{n,k}\xi_k\right\rangle=\frac{1}{n^2}\sum_{k=1}^{n}\gamma_k^2\,\mathbb{E}\operatorname{tr}\left(M_{n,k}\Sigma M_{n,k}\xi_k\otimes\xi_k\right)\\
&\leqslant \frac{1}{n^2}\sum_{k=1}^{n}\gamma_k^2\sigma^2\operatorname{tr}\left(M_{n,k}^2\Sigma\Sigma\right)\\
&\leqslant \frac{1}{n^2}\sum_{k=1}^{n}\gamma_k^2\sigma^2\operatorname{tr}\left(\left(\sum_{j=k}^{n}\left[\prod_{i=k+1}^{j}(I-\gamma_i\Sigma)\right]\right)\Sigma\right)^2\\
&\leqslant \frac{1}{n^2}\sum_{k=1}^{n}\gamma_k^2\sigma^2\sum_{t=1}^{\infty}\left(\left(\sum_{j=k}^{n}\left[\prod_{i=k+1}^{j}\left(1-\gamma_i\frac{1}{t^\alpha}\right)\right]\right)\frac{1}{t^\alpha}\right)^2.
\end{aligned}$$

Let's first upper bound:

$$\begin{aligned}
\left[\prod_{i=k+1}^{j}\left(1-\gamma_i\frac{1}{t^\alpha}\right)\right] &\leqslant \exp\sum_{i=k+1}^{j}\left(\gamma_i\frac{1}{t^\alpha}\right)\\
&= \exp-\sum_{i=k+1}^{j}\left(\frac{1}{i^\zeta}\frac{1}{t^\alpha}\right)\ \text{if}\ \gamma_i=\frac{1}{i^\zeta}\\
&\leqslant \exp-\frac{1}{t^\alpha}\int_{u=k+1}^{j+1}\left(\frac{1}{u^\zeta}du\right)\\
&\leqslant \exp-\frac{1}{t^\alpha}\frac{(j+1)^{1-\zeta}-(k+1)^{1-\zeta}}{1-\zeta}.
\end{aligned}$$

Then

$$\sum_{j=k}^{n}\prod_{i=k+1}^{j}\left(1-\gamma_i\frac{1}{t^\alpha}\right)\leqslant\sum_{j=k}^{n}\exp-\frac{1}{t^\alpha}\frac{(j+1)^{1-\zeta}-(k+1)^{1-\zeta}}{1-\zeta}$$

$$\begin{aligned} &\leqslant \int_{u=k}^{n} \exp -\frac{1}{t^\alpha}\frac{(u+1)^{1-\zeta}-(k+1)^{1-\zeta}}{1-\zeta}du \\ &\leqslant (n-k) \quad \text{clearly} \end{aligned}$$

(this upper bound is good when $t >> n^{1-\zeta}$), but we also have:

$$\int_{u=k}^{n} \exp -\frac{1}{t^\alpha}\frac{(u+1)^{1-\zeta}-(k+1)^{1-\zeta}}{1-\zeta}du = \int_{u=k+1}^{n+1} \exp -\frac{1}{t^\alpha}\frac{u^{1-\zeta}-(k+1)^{1-\zeta}}{1-\zeta}du.$$

With $\rho = 1-\zeta, K_\zeta := \frac{1}{(1-\zeta)^{1/\rho}t^{\alpha/\rho}}$ and

$$\begin{aligned} v^\rho &= \frac{1}{t^\alpha}\frac{(u)^\rho-(k+1)^\rho}{(1-\zeta)} \\ v &= \frac{1}{(1-\zeta)^{1/\rho}t^{\alpha/\rho}}\left((u)^\rho-(k+1)^\rho\right)^{1/\rho} \\ dv &= K_\zeta\frac{1}{\rho}\left(u^\rho-(k+1)^\rho\right)^{1/\rho-1}\rho u^{\rho-1}du \\ dv &= K_\zeta\left(1-\left(\frac{k+1}{u}\right)^\rho\right)^{1/\rho-1}du \\ dv\frac{1}{K_\zeta\left(1-\left(\frac{(k+1)^\rho}{t^\alpha Cv^\rho+(k+1)^\rho}\right)\right)^{1/\rho-1}} &= du \\ dv\frac{1}{K_\zeta}\left(\frac{t^\alpha Cv^\rho+(k+1)^\rho}{t^\alpha Cv^\rho+(k+1)^\rho-(k+1)^\rho}\right)^{1/\rho-1} &= du \\ dv\frac{1}{K_\zeta}\left(\frac{t^\alpha Cv^\rho+(k+1)^\rho}{t^\alpha Cv^\rho}\right)^{1/\rho-1} &= du \\ dv\frac{1}{K_\zeta}\left(1+\frac{(k+1)^\rho}{t^\alpha Cv^\rho}\right)^{1/\rho-1} &= du \end{aligned}$$

$$\begin{aligned} \int_{u=k}^{n} \exp -\frac{1}{t^\alpha}\frac{(u+1)^{\frac{\alpha}{\alpha+\beta}}-(k+1)^{\frac{\alpha}{\alpha+\beta}}}{(1-\zeta)}du &\leqslant \int_0^\infty \frac{1}{K_\zeta}\left(1+\frac{(k+1)^\rho}{t^\alpha Cv^\rho}\right)^{1/\rho-1}\exp\left(-v^\rho\right)dv \\ &\leqslant \frac{2^{1/\rho-1}}{K_\zeta}\int_0^\infty\left(1\vee\frac{(k+1)^\rho}{t^\alpha Cv^\rho}\right)^{1/\rho-1}\exp\left(-v^\rho\right)dv \\ &\leqslant 2^{1/\rho-1}(1-\zeta)^{1/\rho}t^{\alpha/\rho}\int_0^\infty\left(1\vee\frac{(k+1)^{1-\rho}}{(t^\alpha C)^{1/\rho-1}v^{1-\rho}}\right)\exp\left(-v^\rho\right)dv. \\ &\leqslant Kt^{\alpha/\rho}\left(I_1\vee I_2\frac{(k+1)^{1-\rho}}{(t^\alpha)^{1/\rho-1}}\right) \\ &\leqslant K\left(t^{\frac{\alpha}{1-\zeta}}\vee t^\alpha(k+1)^\zeta\right). \end{aligned}$$

Finally:

$$\begin{aligned} \mathrm{var}(n) &\leqslant \frac{1}{n^2}\sum_{k=1}^{n}\gamma_k^2\sigma^2\sum_{t=1}^{\infty}\frac{1}{t^{2\alpha}}\left((n-k)\wedge K\left(t^{\frac{\alpha}{1-\zeta}}\vee t^\alpha(k+1)^\zeta\right)\right)^2 \\ \mathrm{var}(n) &\leqslant \frac{1}{n^2}\sum_{k=1}^{n}\gamma_k^2\sigma^2\sum_{t=1}^{\infty}\frac{1}{t^{2\alpha}}\left((n-k)^2\wedge K\left(t^{2\frac{\alpha}{1-\zeta}}+t^{2\alpha}k^{2\zeta}\right)\right) \end{aligned}$$

$$\leqslant \underbrace{\frac{1}{n^2}\sum_{k=1}^{n}\gamma_k^2\sigma^2\sum_{t=1}^{\infty}\frac{1}{t^{2\alpha}}\left((n-k)^2 \wedge K\left(t^{2\frac{\alpha}{1-\zeta}}\right)\right)}_{S_1}$$
$$+ \underbrace{\frac{1}{n^2}\sum_{k=1}^{n}\gamma_k^2\sigma^2\sum_{t=1}^{\infty}\frac{1}{t^{2\alpha}}\left((n-k)^2 \wedge t^{2\alpha}k^{2\zeta}\right)}_{S_2}$$

$$\begin{aligned}
S_1 &\leqslant K\frac{1}{n^2}\sum_{k=1}^{n}\gamma_k^2\sigma^2\left(\sum_{t=1}^{(n-k)^{\frac{1-\zeta}{\alpha}}}\frac{1}{t^{2\alpha}}t^{2\frac{\alpha}{1-\zeta}} + \sum_{t=(n-k)^{\frac{1-\zeta}{\alpha}}}^{\infty}\frac{1}{t^{2\alpha}}(n-k)^2\right)\\
&\leqslant K\frac{1}{n^2}\sum_{k=1}^{n}\gamma_k^2\sigma^2\left(\sum_{t=1}^{(n-k)^{\frac{1-\zeta}{\alpha}}}t^{\frac{2\alpha\zeta}{1-\zeta}} + (n-k)^2\sum_{t=(n-k)^{\frac{1-\zeta}{\alpha}}}^{\infty}\frac{1}{t^{2\alpha}}\right)\\
&\leqslant G\frac{1}{n^2}\sum_{k=1}^{n}\gamma_k^2\sigma^2\left((n-k)^{\frac{1-\zeta}{\alpha}(\frac{2\alpha\zeta}{1-\zeta}+1)} + (n-k)^2\frac{1}{(n-k)^{\frac{1-\zeta}{\alpha}(2\alpha-1)}}\right)\\
&\leqslant G\frac{1}{n^2}\sum_{k=1}^{n}\gamma_k^2\sigma^2\left((n-k)^{\frac{(2\alpha-1)\zeta+1}{\alpha}} + (n-k)^{2-\frac{1-\zeta}{\alpha}(2\alpha-1)}\right)\\
&= 2G\sigma^2\frac{1}{n^2}\sum_{k=1}^{n}\frac{1}{k^{2\zeta}}(n-k)^{\frac{(2\alpha-1)\zeta+1}{\alpha}}\\
&\leqslant 2G\sigma^2\frac{1}{n^2}\sum_{k=1}^{n}\left(\frac{n}{k}-1\right)^{\frac{(2\alpha-1)\zeta+1}{\alpha}}k^{\frac{1-\zeta}{\alpha}}\\
&= 2G\sigma^2 n^{-1+\frac{1-\zeta}{\alpha}}\frac{1}{n}\sum_{k=1}^{n}\left(\frac{1}{k/n}-1\right)^{\frac{(2\alpha-1)\zeta+1}{\alpha}}\left(\frac{k}{n}\right)^{\frac{1-\zeta}{\alpha}}\\
&= 2G\sigma^2 n^{-1+\frac{1-\zeta}{\alpha}}\left(\frac{1}{n}\sum_{k=1}^{n}\left(\frac{1}{k/n}-1\right)^{\frac{(2\alpha-1)\zeta+1}{\alpha}}\left(\frac{k}{n}\right)^{\frac{1-\zeta}{\alpha}}\right)\\
&= 2G\sigma^2 n^{-1+\frac{1-\zeta}{\alpha}}\left(\frac{1}{n}\sum_{k=1}^{n}\left(\frac{1}{k/n}-1\right)^{2\zeta}\left(1-\frac{k}{n}\right)^{\frac{1-\zeta}{\alpha}}\right).
\end{aligned}$$

If $\zeta < \frac{1}{2}$ then

$$\int_0^1\left(\frac{1}{x}-1\right)^{2\zeta}(1-x)^{\frac{1-\zeta}{\alpha}}dx < \infty$$

and

$$\begin{aligned}
S_1 &\leqslant Hn^{-1+\frac{1-\zeta}{\alpha}}\left(\frac{1}{n}\sum_{k=1}^{n}\left(\frac{1}{k/n}-1\right)^{2\zeta}\left(1-\frac{k}{n}\right)^{\frac{1-\zeta}{\alpha}}\right)\\
&\leqslant H'n^{-1+\frac{1-\zeta}{\alpha}}.
\end{aligned}$$

If $\zeta > \frac{1}{2}$ then

$$\int_0^1\left(\frac{1}{x}-1\right)^{2\zeta}(1-x)^{\frac{1-\zeta}{\alpha}} - \left(\frac{1}{x}\right)^{2\zeta}dx < \infty.$$

and

$$S_1 \leqslant Hn^{-1+\frac{1-\zeta}{\alpha}}\left(\frac{1}{n}\sum_{k=1}^{n}\left(\frac{1}{k/n}-1\right)^{2\zeta}\left(1-\frac{k}{n}\right)^{\frac{1-\zeta}{\alpha}} - \left(\frac{n}{k}\right)^{2\zeta} + \frac{1}{n}\sum_{k=1}^{n}\left(\frac{n}{k}\right)^{2\zeta}\right)$$

$$\begin{aligned} &\leqslant Hn^{-1+\frac{1-\zeta}{\alpha}}\left(C+n^{2\zeta-1}\right) \\ &\leqslant Cn^{-1+\frac{1-\zeta+\alpha(2\zeta-1)}{\alpha}}. \end{aligned}$$

$$\begin{aligned} S_2 &= \frac{1}{n^2}\sum_{k=1}^{n}\gamma_k^2\sigma^2\sum_{t=1}^{\infty}\frac{1}{t^{2\alpha}}\left((n-k)^2\wedge t^{2\alpha}k^{2\zeta}\right) \\ &\leqslant \frac{1}{n^2}\sum_{k=1}^{n}\gamma_k^2\sigma^2\left(\sum_{t=1}^{t_\ell}\frac{1}{t^{2\alpha}}t^{2\alpha}k^{2\zeta}+\sum_{t=t_\ell}^{\infty}\frac{1}{t^{2\alpha}}(n-k)^2\right) \quad \text{with} \quad t_\ell=\frac{(n-k)^{\frac{1}{\alpha}}}{k^{\frac{\zeta}{\alpha}}} \\ &\leqslant \frac{1}{n^2}\sum_{k=1}^{n}\gamma_k^2\sigma^2\left(k^{2\zeta}\sum_{t=1}^{t_\ell}1+(n-k)^2\sum_{t=t_\ell}^{\infty}\frac{1}{t^{2\alpha}}\right) \\ &\leqslant \frac{1}{n^2}\sum_{k=1}^{n}\gamma_k^2\sigma^2\left(k^{2\zeta}\frac{(n-k)^{\frac{1}{\alpha}}}{k^{\frac{\zeta}{\alpha}}}+(n-k)^2\left(\frac{(n-k)^{\frac{1}{\alpha}}}{k^{\frac{\zeta}{\alpha}}}\right)^{1-2\alpha}\right) \\ &= \frac{1}{n^2}\sum_{k=1}^{n}\gamma_k^2\sigma^2\left(k^{2\zeta-\frac{\zeta}{\alpha}}(n-k)^{\frac{1}{\alpha}}+(n-k)^{\frac{1}{\alpha}}k^{\frac{\zeta}{\alpha}(2\alpha-1)}\right) \\ &= \frac{2\sigma^2}{n^2}\sum_{k=1}^{n}\frac{1}{k^{2\zeta}}(n-k)^{\frac{1}{\alpha}}k^{\frac{\zeta}{\alpha}(2\alpha-1)} \\ &= \frac{2\sigma^2}{n^2}\sum_{k=1}^{n}k^{-\frac{\zeta}{\alpha}}(n-k)^{\frac{1}{\alpha}} \\ &= 2\sigma^2 n^{\left(-1+-\frac{\zeta}{\alpha}+\frac{1}{\alpha}\right)}\frac{1}{n}\sum_{k=1}^{n}\left(\frac{k}{n}\right)^{-\frac{\zeta}{\alpha}}\left(1-\frac{k}{n}\right)^{\frac{1}{\alpha}} \\ &\leqslant Kn^{\left(-1+\frac{1-\zeta}{\alpha}\right)}. \end{aligned}$$

As we have a Riemann sum which converges.

Finally we get: if $0<\zeta<\frac{1}{2}$ then

$$\begin{aligned} \operatorname{var}(n) &= O\left(\sigma^2 n^{-1+\frac{1-\zeta}{\alpha}}\right) \\ &= O\left(\sigma^2\frac{\sigma^2(s^2\gamma_n)1/\alpha}{n^{1-1/\alpha}}n^{-1+\frac{1-\zeta}{\alpha}}\right) \end{aligned}$$

where we have re-used the constants s by formally replacing in the proof γ by γs^2.

and if $\zeta>\frac{1}{2}$ then

$$\operatorname{var}(n)=O\left(\sigma^2 n^{-1+\frac{1-\zeta}{\alpha}+2\zeta-1}\right).$$

Which is substantially Lemma A.28.

□

3

Faster Convergence Rates for Least-Squares Regression

We consider the optimization of a quadratic objective function whose gradients are only accessible through a stochastic oracle that returns the gradient at any given point plus a zero-mean finite variance random error. We present the first algorithm that achieves jointly the optimal prediction error rates for least-squares regression, both in terms of forgetting the initial conditions in $O(1/n^2)$, and in terms of dependence on the noise and dimension d of the problem, as $O(d/n)$. Our new algorithm is based on averaged accelerated regularized gradient descent, and may also be analyzed through finer assumptions on initial conditions and the Hessian matrix, leading to dimension-free quantities that may still be small in some distances while the "optimal" terms above are large. In order to characterize the tightness of these new bounds, we consider an application to non-parametric regression and use the known lower bounds on the statistical performance (without computational limits), which happen to match our bounds obtained from a single pass on the data and thus show optimality of our algorithm in a wide variety of particular trade-offs between bias and variance.

This chapter is based on our work *Harder, Better, Faster, Stronger Convergence Rates for Least-Squares Regression*, A. Dieuleveut, N. Flammarion and F.Bach, accepted for publication in Journal of Machine Learning Research (JMLR), 2017.

Contents

3.1 Introduction 115
3.2 Least-Squares Regression 117
3.2.1 Statistical Assumptions 117
3.2.2 Averaged Gradient Methods and Acceleration 119
3.2.3 Additive versus Multiplicative Stochastic Oracles on the Gradient . . 119
3.3 Averaged Stochastic Gradient Descent 121
3.3.1 Additive Noise 121
3.3.2 Multiplicative/Additive Noise 122
3.4 Accelerated Stochastic Averaged Gradient Descent 123
3.5 Tighter Dimension-independent Convergence Rates 125
3.6 Rates of Convergence for Kernel Regression 126
3.6.1 Averaged SGD 128
3.6.2 Averaged-accelerated SGD 129
3.7 Conclusion 131

3.1 Introduction

Many supervised machine learning problems are naturally cast as the minimization of a smooth function defined on a Euclidean space. This includes least-squares regression, logistic regression (see, e.g., Hastie et al., 2001) or generalized linear models (McCullagh and Nelder, 1989). While small problems with few or low-dimensional input features may be solved precisely by many potential optimization algorithms (e.g., Newton method), large-scale problems with many high-dimensional features are typically solved with simple gradient-based iterative techniques whose per-iteration cost is small.

In this chapter, we consider a quadratic objective function f whose gradients are only accessible through a stochastic oracle that returns the gradient at any given point plus a zero-mean finite variance random error. In this stochastic approximation framework (Robbins and Monro, 1951), it is known that two quantities dictate the behavior of various algorithms, namely the covariance matrix V of the noise in the gradients, and the deviation $\theta_0 - \theta_*$ between the initial point of the algorithm θ_0 and any of the global minimizer θ_* of f. This leads to a "bias/variance" decomposition (Bach and Moulines, 2013; Hsu et al., 2014) of the performance of most algorithms as the sum of two terms: (a) the bias term characterizes how fast initial conditions are forgotten and thus is increasing in a well-chosen norm of $\theta_0 - \theta_*$; while (b) the variance term characterizes the effect of the noise in the gradients, independently of the starting point, and with a term that is increasing in the covariance of the noise.

For quadratic functions with (a) a noise covariance matrix V which is proportional (with constant σ^2) to the Hessian of f (a situation which corresponds to least-squares regression) and (b) an initial point characterized by the norm $\|\theta_0 - \theta_*\|^2$, the optimal bias and variance terms are known *separately* from the optimization and statistical theories. On the one hand, the optimal bias dependency after n iterations is proportional to $\frac{L\|\theta_0-\theta_*\|^2}{n^2}$, where L is the largest eigenvalue of the Hessian of f. This rate is achieved by accelerated gradient descent (Nesterov, 1983, 2004), and is known to be optimal if the number of iterations n is less than the dimension d of the underlying predictors, but the algorithm is not robust to random or deterministic noise in the gradients (d'Aspremont, 2008; Schmidt et al., 2011; Devolder et al., 2014). On the other hand, the optimal variance term is proportional to $\frac{\sigma^2 d}{n}$ (Tsybakov, 2003); it is known to be achieved by averaged gradient descent (Bach and Moulines, 2013), for which the bias term only achieves $\frac{L\|\theta_0-\theta_*\|^2}{n}$ instead of $\frac{L\|\theta_0-\theta_*\|^2}{n^2}$.

Our first contribution in this chapter is to present a novel algorithm which attains optimal rates for *both the variance and the bias terms*. This algorithm analyzed in Section 3.4 is averaged accelerated gradient descent; beyond obtaining jointly optimal rates, our result shows that averaging is beneficial for accelerated techniques and provides a provable robustness to noise.

While optimal when measuring performance in terms of the dimension d and the initial distance to optimum $\|\theta_0 - \theta_*\|^2$, these rates are not adapted in many situations where either d is larger than the number of iterations n (*i.e.*, the number of observations for regular stochastic gradient descent) or $L\|\theta_0 - \theta_*\|^2$ is much larger than n^2. Our second contribution is to provide in Section 3.5 an analysis of a new algorithm (based on some additional regularization) that can adapt our bounds to finer assumptions on $\theta_0 - \theta_*$ and the Hessian of the problem, leading in particular to dimension-free quantities that can thus be extended to the Hilbert space setting (in particular for non-parametric estimation).

In order to characterize the optimality of these new bounds, our third contribution is

to consider an application to non-parametric regression in Section 3.6 and use the known lower bounds on the statistical performance (without computational limits), which happen to match our bounds obtained from a single pass on the data and thus show optimality of our algorithm in a wide variety of particular trade-offs between bias and variance.

This chapter is organized as follows: in Section 3.2, we present the main problem we tackle, namely least-squares regression, then introduce the two algorithms that we consider in Section 3.2.2, as well as the two types of oracles on the gradient in Section 3.2.3. In Section 3.3, we present new results for averaged stochastic gradient descent that set the stage for Section 3.4, where we present our main novel result leading to an accelerated algorithm which is robust to noise. This tighter analysis of convergence rates based on finer dimension-free quantities is presented in Section 3.5, and their optimality for kernel-based non-parametric regression is studied in Section 3.6. Organization of the main results is summarized in the Table 3.1 bellow.

Proofs are given in Chapter B.

	Averaged Algo.	**Averaged Accelerated Algo.**
Dimension dependent rates	*Section 3.3*	*Section 3.4*
Additive Noise	Lemma 3.1$^{\diamondsuit}$	Theorem 3.3
Multiplicative Noise	Theorem 3.2$^{\diamondsuit}$	$\natural$
Dimension independent rates	*Section 3.5*	*Section 3.5*
Additive Noise	$\sharp$	Theorem 3.5
Multiplicative Noise	4^{th} remark after Cor. 3.6$^{\flat}$	$\natural$
Kernel regression setting	*Section 3.6*	*Section 3.6*
Additive Noise	$\sharp$	Theorem 3.8
Multiplicative Noise	Theorem 3.7$^{\flat}$	$\natural$

Table 3.1: Organization of the chapter. $\diamondsuit$: We extend results from (Bach and Moulines, 2013) to the setting in which extra regularization is added; $\sharp$: apart from Lemma 3.1 which is useful to develop intuition of the different terms in the upper bound, we do not state result for the averaged algorithm with additive noise, as the most powerful result is for the multiplicative noise; $\flat$: these results recover results from Chapter 2 (with the use of an extra regularization); $\natural$: it is still an open problem to get results in the accelerated setting for a multiplicative noise oracle.

Collaboration with Nicolas Flammarion: this work was done in collaboration with another PhD student, Nicolas Flammarion, and we both equally contributed to the entire paper. We both include parts of the paper in our thesis but with a different focus: while the core acceleration result is present in the two thesis, Nicolas mainly focused on the finite-dimensional part (and does not cover the non-parametric setting), I focused more on the non-parametric setting (e.g., the experimental part which is done in finite dimension is not covered).

3.2 Least-Squares Regression

In this section, we present the least-squares regression framework, which is risk minimization with the square loss, together with the main assumptions regarding the model and the algorithms. These algorithms will rely on stochastic gradient oracles, which will come in two kinds, an additive noise which does not depend on the current iterate, which will correspond in practice to the full knowledge of the covariance matrix, and a "multiplicative/additive" noise, which corresponds to the regular stochastic gradient obtained from a single pair of observations. This second oracle is much harder to analyze.

3.2.1 Statistical Assumptions

We consider the following general setting:

- $\mathcal{H}$ is a d-dimensional Euclidean space with $d \geqslant 1$. The (temporary) restriction to finite dimension will be relaxed in Section 3.6.
- The observations $(x_n, y_n) \in \mathcal{H} \times \mathbb{R}$, $n \geqslant 1$, are independent and identically distributed (i.i.d.), and such that $\mathbb{E}\|x_n\|^2$ and $\mathbb{E}y_n^2$ are finite.
- We consider the *least-squares regression* problem, namely the minimization of the expected loss $f(\theta) = \frac{1}{2}\mathbb{E}(\langle x_n, \theta\rangle - y_n)^2$ which is a quadratic function.

We first introduce an assumption on the distribution of x_n.

Covariance matrix. We denote by $\Sigma = \mathbb{E}(x_n \otimes x_n) \in \mathbb{R}^{d\times d}$ the population covariance matrix, which is the Hessian of f at all points. Without loss of generality, we can assume Σ is invertible by reducing $\mathcal{H}$ to the minimal subspace where all x_n, $n \geqslant 1$, lie almost surely. This implies that all eigenvalues of Σ are strictly positive (but they may be arbitrarily small). Following Bach and Moulines (2013), we assume there exists $R > 0$ such that

$$\mathbb{E}\|x_n\|^2 x_n \otimes x_n \preccurlyeq R^2\Sigma, \tag{$\mathcal{A}_1$}$$

where $A \preccurlyeq B$ means that $B - A$ is positive semi-definite. This assumption implies in particular that (a) $\mathbb{E}\|x_n\|^4$ is finite and (b) $\operatorname{tr}\Sigma = \mathbb{E}\|x_n\|^2 \leqslant R^2$ since taking the trace of the previous inequality we get $\mathbb{E}\|x_n\|^4 \leqslant R^2\mathbb{E}\|x_n\|^2$ and using Cauchy-Schwarz inequality we get $\mathbb{E}\|x_n\|^2 \leqslant \sqrt{\mathbb{E}\|x_n\|^4} \leqslant R\sqrt{\mathbb{E}\|x_n\|^2}$.

Assumption ($\mathcal{A}_1$) is satisfied, for example, for least-square regression with almost surely bounded data, since $\|x_n\|^2 \leqslant R^2$ almost surely implies $\mathbb{E}\|x_n\|^2 x_n \otimes x_n \preccurlyeq \mathbb{E}\big[R^2 x_n \otimes x_n\big] = R^2\Sigma$. This assumption is also true for data with infinite support and a bounded *kurtosis* for the projection of the covariates x_n on any direction $z \in \mathcal{H}$, e.g, for which there exists $\kappa > 0$, such that:

$$\forall z \in \mathcal{H}, \quad \mathbb{E}\langle z, x_n\rangle^4 \leqslant \kappa\langle z, \Sigma z\rangle^2. \tag{3.1}$$

Indeed, by Cauchy-Schwarz inequality, Equation (3.1) implies for all $(z, t) \in \mathcal{H}^2$, the following bound $\mathbb{E}\langle z, x_n\rangle^2\langle t, x_n\rangle^2 \leqslant \kappa\langle z, \Sigma z\rangle\langle t, \Sigma t\rangle$, which in turn implies that for all positive semi-definite symmetric matrices M, N, we have $\mathbb{E}\langle x_n, Mx_n\rangle\langle x_n, Nx_n\rangle \leqslant \kappa \operatorname{tr}(M\Sigma)\operatorname{tr}(N\Sigma)$. Equation (3.1), which is true for Gaussian vectors with $\kappa = 3$, thus implies ($\mathcal{A}_1$) for $R^2 = \kappa \operatorname{tr}\Sigma = \kappa\mathbb{E}\|x_n\|^2$.

In the next two paragraphs, we introduce some quantities that will be important in the analysis, in order to get tighter bounds.

Eigenvalue decay. Most convergence bounds depend on the dimension d of $\mathcal{H}$. However it is possible to derive dimension-free and often tighter convergence rates by considering bounds depending on the value $\operatorname{tr}\Sigma^b$ for $b \in [0,1]$. Given b, if we consider the eigenvalues of Σ ordered in decreasing order, which we denote by s_i, then $\operatorname{tr}\Sigma^b = \sum_i s_i^b$, and the eigenvalues decay[1] at least as $\frac{(\operatorname{tr}\Sigma^b)^{1/b}}{i^{1/b}}$. Moreover, it is known that $(\operatorname{tr}\Sigma^b)^{1/b}$ is decreasing in b and thus, the smaller the b, the stronger the assumption. For b going to 0 then $\operatorname{tr}\Sigma^b$ tends to d and we are back in the classical low-dimensional case. When $b = 1$, we simply get $\operatorname{tr}\Sigma = \mathbb{E}\|x_n\|^2$, which will correspond to the weakest assumption in our context.

Optimal predictor. In finite dimension the regression function $f(\theta) = \frac{1}{2}\mathbb{E}(\langle x_n, \theta\rangle - y_n)^2$ always admits a global minimum $\theta_* = \Sigma^\dagger \mathbb{E}(y_n x_n)$. When initializing algorithms at $\theta_0 = 0$ or regularizing by the squared norm, rates of convergence generally depend on $\|\theta_*\|$, a quantity which could be arbitrarily large.

However there exists a systematic upper-bound[2] $\|\Sigma^{\frac{1}{2}}\theta_*\| \leqslant 2\sqrt{\mathbb{E}y_n^2}$. This leads naturally to the consideration of convergence bounds depending on $\|\Sigma^{r/2}\theta_*\|$ for $r \leqslant 1$. In infinite dimension this will correspond to assuming $\|\Sigma^{r/2}\theta_*\| < \infty$. This new assumption relates the optimal predictor with sources of ill-conditioning (since Σ is the Hessian of the objective function f), the smaller r, the stronger our assumption, with $r = 1$ corresponding to no assumption at all, $r = 0$ to θ_* in $\mathcal{H}$ and $r = -1$ to a convergence of the bias of least-squares regression with averaged stochastic gradient descent in $O(\frac{\|\Sigma^{-1/2}\theta_*\|^2}{n^2})$ (as in Chapter 2, or in Défossez and Bach, 2015). In this chapter, we will use arbitrary initial points θ_0 and thus our bounds will depend on $\|\Sigma^{r/2}(\theta_0 - \theta_*)\|$.

Finally , we make an assumption on the joint distribution of (x_n, y_n).

Noise. We denote by $\varepsilon_n = y_n - \langle \theta_*, x_n\rangle$ the residual for which we have $\mathbb{E}[\varepsilon_n x_n] = 0$. Although we do not have $\mathbb{E}[\varepsilon_n | x_n] = 0$ in general unless the model is well-specified, we assume the noise to be a structured process such that there exists $\sigma > 0$ with

$$\mathbb{E}[\varepsilon_n^2 x_n \otimes x_n] \preccurlyeq \sigma^2 \Sigma. \tag{$\mathcal{A}_2$}$$

Assumption ($\mathcal{A}_2$) is satisfied for example for data almost surely bounded or when the model is well-specified, (e.g., $y_n = \langle \theta_*, x_n\rangle + \varepsilon_n$, with $(\varepsilon_n)_{n\in\mathbb{N}}$ i.i.d. of variance σ^2 and independent of x_n).

⚠ *In order to slightly simplify notations, notations change between Chapter 2 and Chapter 3. Especially, we use b instead of $\frac{1}{\alpha}$ for the constant characterizing the eigenvalue decay (in order to write $\operatorname{tr}(\Sigma^b)$ and not $\operatorname{tr}(\Sigma^{1/\alpha})$). Similarly, the r in this chapter corresponds to $1-2r$ with the r from Chapter 2: we summarize connections in the following tabular:*

	Chapter 2	Chapter 3	Connections
Capacity condition parameter	α	b	$\alpha = 1/b$
Source condition parameter	$r_{\text{Ch.2}}$	$r_{\text{Ch.3}}$	$r_{\text{Ch.3}} = 1 - 2r_{\text{Ch.2}}$

As a consequence, e.g., the exponent on n in the optimal rate is $\frac{-2\alpha r_{\text{Ch.2}}}{2\alpha r_{\text{Ch.2}}+1}$ in Chapter 2 and $-\frac{1-r_{\text{Ch.3}}}{b+1-r_{\text{Ch.3}}}$ in Chapter 3.

[1] Indeed for any $i \geqslant 1$, we have $is_i^b \leqslant \sum_{t=1}^i s_t^b \leqslant \operatorname{tr}(\Sigma^b)$.

[2] Indeed for all $\theta \in \mathbb{R}^d$ and in particular $\theta = 0$, by Minkowski's inequality, $\|\Sigma^{\frac{1}{2}}\theta_*\| - \sqrt{\mathbb{E}y_n^2} = \sqrt{\mathbb{E}\langle\theta_*, x_n\rangle^2} - \sqrt{\mathbb{E}y_n^2} \leqslant \sqrt{\mathbb{E}(\langle\theta_*, x_n\rangle - y_n)^2} \leqslant \sqrt{\mathbb{E}(\langle\theta, x_n\rangle - y_n)^2} \leqslant \sqrt{\mathbb{E}(y_n)^2}$.

3.2.2 Averaged Gradient Methods and Acceleration

We focus in this chapter on stochastic gradient methods with and without acceleration for the least-squares function regularized by $\frac{\lambda}{2}\|\theta - \theta_0\|^2$ for $\lambda \in \mathbb{R}^+$. The regularization will be useful when deriving tighter convergence rates in Section 3.5, and it has the additional benefit of making the problem λ-strongly-convex. Stochastic gradient descent (referred to from now on as "SGD"), applied to the regularized problem, can be described for $n \geqslant 1$ as

$$\theta_n = \theta_{n-1} - \gamma f_n'(\theta_{n-1}) - \gamma\lambda(\theta_{n-1} - \theta_0), \tag{3.2}$$

starting from $\theta_0 \in \mathcal{H}$, where $\gamma > 0$ is either called the step-size in optimization or the learning rate in machine learning, and $f_n'(\theta_{n-1})$ is an unbiased estimate of the gradient of f at θ_{n-1}, that is, its conditional expectation given all other sources of randomness is equal to $f'(\theta_{n-1})$.

Accelerated stochastic gradient descent is defined, for the regularized problem, by an iterative system with two parameters (θ_n, ν_n) satisfying for $n \geqslant 1$

$$\begin{aligned} \theta_n &= \nu_{n-1} - \gamma f_n'(\nu_{n-1}) - \gamma\lambda(\nu_{n-1} - \theta_0) \\ \nu_n &= \theta_n + \delta(\theta_n - \theta_{n-1}), \end{aligned} \tag{3.3}$$

starting from $\theta_0 = \nu_0 \in \mathcal{H}$, with $\gamma, \delta \in \mathbb{R}^2$ and $f_n'(\theta_{n-1})$ described as before. It may be reformulated as the following second-order recursion

$$\theta_n = (1 - \gamma\lambda)\big(\theta_{n-1} + \delta(\theta_{n-1} - \theta_{n-2})\big) - \gamma f_n'\big(\theta_{n-1} + \delta(\theta_{n-1} - \theta_{n-2})\big) + \gamma\lambda\theta_0.$$

The *momentum* coefficient $\delta \in \mathbb{R}$ is chosen to accelerate the convergence rate (Nesterov, 1983; Beck and Teboulle, 2009) and has its roots in the heavy-ball algorithm from Polyak (1964). We especially concentrate here, following Polyak and Juditsky (1992), on the average of the sequence

$$\bar{\theta}_n = \frac{1}{n+1}\sum_{i=0}^{n} \theta_n, \tag{3.4}$$

and we note that it can be computed online as $\bar{\theta}_n = \frac{n}{n+1}\bar{\theta}_{n-1} + \frac{1}{n+1}\theta_n$.

The key ingredient in the algorithms presented above is the unbiased estimate on the gradient $f_n'(\theta)$, which can take two forms that we now describe in our setting.

3.2.3 Additive versus Multiplicative Stochastic Oracles on the Gradient

We consider the standard stochastic approximation framework (Kushner and Yin, 2003). That is, we let $(\mathcal{F}_n)_{n\geqslant 0}$ be the increasing family of σ-fields that are generated by all variables (x_i, y_i) for $i \leqslant n$, and such that for each $\theta \in \mathcal{H}$ the random variable $f_n'(\theta)$ is square-integrable and $\mathcal{F}_n$-measurable with $\mathbb{E}[f_n'(\theta)|\mathcal{F}_{n-1}] = f'(\theta)$, for all $n \geqslant 0$. Consequently it is of the form

$$f_n'(\theta) = f'(\theta) - \xi_n, \tag{$\mathcal{A}_3$}$$

where the noise process ξ_n is $\mathcal{F}_n$-measurable with $\mathbb{E}[\xi_n|\mathcal{F}_{n-1}] = 0$ and $\mathbb{E}[\|\xi_n\|^2]$ is finite. We will consider two different gradient oracles.

Additive noise. The first oracle is the sum of the true gradient $f'(\theta)$ and an independent zero-mean noise that does not depend on θ. This oracle is equal to

$$f'_n(\theta) = \Sigma\theta - y_n x_n. \tag{3.5}$$

Since $f'(\theta) = \Sigma\theta - \mathbb{E}y_n x_n$, the oracle above has a noise vector $\xi_n = y_n x_n - \mathbb{E}y_n x_n$ independent of θ and therefore satisfies Assumption ($\mathcal{A}_3$). Furthermore we also assume that there exists $\tau \in \mathbb{R}$ such that

$$\mathbb{E}[\xi_n \otimes \xi_n] \preccurlyeq \tau^2 \Sigma, \tag{$\mathcal{A}_4$}$$

that is, the noise has a particular structure adapted to least-squares regression. For optimal results for unstructured noise, with convergence rate for the noise part in $O(1/\sqrt{n})$, see Lan (2012). The oracle above with an additive noise which is independent of the current iterate corresponds to the first setting studied in stochastic approximation (Robbins and Monro, 1951; Duflo, 1997; Polyak and Juditsky, 1992). While used by Bach and Moulines (2013) as an artifact of proof, for least-squares regression, such an additive noise corresponds to the situation where the distribution of x is known so that the population covariance matrix is computable, but the distribution of the outputs $(y_n)_{n\in\mathbb{N}}$ remains unknown. Thus it may be seen as an intermediate set-up between regression estimation with fixed and random design (see, e.g., Györfi et al., 2002, Section 1.9).

Assumption ($\mathcal{A}_4$) will be satisfied, for example if the outputs are almost surely bounded because $\mathbb{E}[\xi_n \otimes \xi_n] \preccurlyeq \mathbb{E}[y_n^2 x_n \otimes x_n] \preccurlyeq \tau^2\Sigma$ if $y_n^2 \leqslant \tau^2$ almost surely. But it will also be for data satisfying Equation (3.1) since we will have

$$\begin{aligned}
\mathbb{E}[\xi_n \otimes \xi_n] &\preccurlyeq \mathbb{E}[y_n^2 x_n \otimes x_n] = \mathbb{E}[(\langle \theta_*, x_n\rangle + \varepsilon_n)^2 x_n \otimes x_n] \\
&\preccurlyeq 2\mathbb{E}[\langle \theta_*, x_n\rangle^2 x_n \otimes x_n] + 2\sigma^2\Sigma \preccurlyeq 2(\kappa\|\Sigma^{1/2}\theta_*\|^2 + \sigma^2)\Sigma \\
&\preccurlyeq 2(4\kappa\mathbb{E}[y_n^2] + \sigma^2)\Sigma,
\end{aligned}$$

and thus Assumption ($\mathcal{A}_4$) is satisfied with $\tau^2 = 2(4\kappa\mathbb{E}[y_n^2] + \sigma^2)$.

Stochastic noise ("multiplicative/additive"). This corresponds to:

$$f'_n(\theta) = (\langle x_n, \theta\rangle - y_n)x_n = (\Sigma + \zeta_n)(\theta - \theta_*) - \Xi_n, \tag{3.6}$$

with $\zeta_n = x_n \otimes x_n - \Sigma$ and $\Xi_n = (y_n - \langle x_n, \theta_*\rangle)x_n = \varepsilon_n x_n$. This oracle corresponds to regular SGD, which is often referred to as the least-mean-square (LMS) algorithm for least-squares regression, where the noise comes from sampling a single pair of observations. While still satisfying Assumption ($\mathcal{A}_3$), it combines an additive noise Ξ_n independent of θ as in Equation (3.5) and a multiplicative noise ζ_n. This multiplicative noise makes this stochastic oracle harder to analyze which explains why it is often approximated by an additive noise oracle. However it is the most widely used and most practical one. Note that for the oracle in Equation (3.6), from Equation ($\mathcal{A}_2$), we have $\mathbb{E}[\Xi_n \otimes \Xi_n] \preccurlyeq \sigma^2\Sigma$. It has a similar form to Assumption ($\mathcal{A}_4$) which is valid for the additive noise oracle in Equation (3.5): we use different constants σ^2 and τ^2 to highlight the difference between these two oracles.

3.3 Averaged Stochastic Gradient Descent

In this section, we provide convergence bounds for regularized averaged stochastic gradient descent. The main novelty compared to the work of Bach and Moulines (2013) is (a) the presence of regularization, which will be useful when deriving tighter convergence rates in Section 3.5 and (b) a much simpler proof. We first consider the additive noise in Section 3.3.1 before considering the multiplicative/additive noise in Section 3.3.2.

3.3.1 Additive Noise

We study here the convergence of the averaged SGD recursion defined by Equation (3.2) under the simple oracle defined in Equation (3.5). For least-squares regression, it takes the form:

$$\theta_n = \big[\mathrm{I} - \gamma\Sigma - \gamma\lambda\mathrm{I}\big]\theta_{n-1} + \gamma y_n x_n + \lambda\gamma\theta_0. \tag{3.7}$$

This is an easy adaptation of the work of Bach and Moulines (2013, Lemma 2) for the regularized case.

Lemma 3.1. *Assume* $(\mathcal{A}_4)$. *Consider the recursion in Equation* (3.7) *with any regularization parameter* $\lambda \in \mathbb{R}_+$ *and any constant step-size* γ *such that* $\gamma(\Sigma + \lambda\mathrm{I}) \preccurlyeq \mathrm{I}$. *Then*

$$\mathbb{E}f(\bar{\theta}_n) - f(\theta_*) \leqslant \Big(\lambda + \frac{1}{\gamma n}\Big)^2 \|\Sigma^{1/2}(\Sigma + \lambda\mathrm{I})^{-1}(\theta_0 - \theta_*)\|^2 + \frac{\tau^2 \operatorname{tr}\big[\Sigma^2(\Sigma + \lambda\mathrm{I})^{-2}\big]}{n}. \tag{3.8}$$

We can make the following observations:

- The proof (see Section B.1) relies on the fact that $\theta_n - \theta_*$ is obtainable in closed form since the cost function is quadratic and thus the recursions are linear, and follows from Polyak and Juditsky (1992).
- The constraint on the step-size γ is equivalent to $\gamma(L + \lambda) \leqslant 1$ where L is the largest eigenvalue of Σ and we thus recover the usual step-size from deterministic gradient descent (Nesterov, 2004).
- When n tends to infinity, the algorithm converges to the minimum of $f(\theta) + \frac{\lambda}{2}\|\theta - \theta_0\|^2$ and our performance guarantee becomes $\lambda^2\|\Sigma^{1/2}(\Sigma + \lambda\mathrm{I})^{-1}(\theta_0 - \theta_*)\|^2$. This is the standard "bias term" from regularized ridge regression (Hsu et al., 2014) which we naturally recover here. The term $\frac{\tau^2}{n}\operatorname{tr}\big[\Sigma^2(\Sigma + \lambda\mathrm{I})^{-2}\big]$ is usually referred to as the "variance term" (Hsu et al., 2014), and is equal to $\frac{\tau^2}{n}$ times the quantity $\operatorname{tr}\big[\Sigma^2(\Sigma + \lambda\mathrm{I})^{-2}\big]$, which is often called the degrees of freedom of the ridge regression problem (Gu, 2002).
- For finite n, the first term in Equation (3.8) is the usual bias term which depends on the distance from the initial point θ_0 to the objective point θ_* with an appropriate norm. It includes a regularization-based component which is proportional to λ^2 and optimization-based component which depends on $(\gamma n)^{-2}$. The regularization-based bias appears because the algorithm tends to minimize the regularized function instead of the true function f.
- Given Equation (3.8), it is natural to set $\lambda\gamma = \frac{1}{n}$, and the two components of the bias term are exactly of the same order leading to $\frac{4}{\gamma^2 n^2}\|\Sigma^{1/2}(\Sigma + \lambda\mathrm{I})^{-1}(\theta_0 - \theta_*)\|^2$. It corresponds up to a constant factor to the bias term of regularized least-squares (Hsu

et al., 2014), but it is achieved by an algorithm accessing only n stochastic gradients. Note that when λ or γ depend on n, this term is not necessarily of order $O(n^{-2})$, as the numerator might be arbitrarily large. Note also that here as in the rest of the chapter, we only prove results in the finite horizon setting, meaning that the number of samples is known in advance and the parameters γ, λ may be chosen as functions of n, but remain constant along the iterations (when λ or γ depend on n, our bounds only hold for the last iterate).

- Note that the bias term can also be bounded by $\frac{1}{\gamma n}\|\Sigma^{1/2}(\Sigma+\lambda \mathrm{I})^{-1/2}(\theta_0-\theta_*)\|^2$ only when $\|\theta_0-\theta_*\|$ is finite (note the difference in the powers of n and $(\Sigma+\lambda \mathrm{I})^{-1}$). See the proof in Section B.1.2 for details.

- The second term in Equation (3.8) is the variance term. It depends on the noise in the gradient. When this one is not structured the variance turns to be also bounded by $\gamma \operatorname{tr}\left(\Sigma(\Sigma+\lambda \mathrm{I})^{-1}\mathbb{E}[\xi_n \otimes \xi_n]\right)$ (see Section B.1.3) and we recover for $\gamma = O(1/\sqrt{n})$, the usual rate of $\frac{1}{\sqrt{n}}$ for SGD in the smooth case (Shalev-Shwartz et al., 2009).

- Overall we get the same performance as the empirical risk minimizer with fixed design, but with an algorithm that performs a single pass over the data.

- When $\lambda = 0$ we recover Lemma 2 of Bach and Moulines (2013). In this case the variance term $\frac{\tau^2 d}{n}$ is optimal over all estimators in $\mathcal{H}$ (Tsybakov, 2003) even without computational limits, in the sense that no estimator that uses the same information can improve upon this rate.

3.3.2 Multiplicative/Additive Noise

When the general stochastic oracle in Equation (3.6) is considered, the regularized LMS algorithm defined by Equation (3.2) takes the form:

$$\theta_n = \left[\mathrm{I} - \gamma x_n \otimes x_n - \gamma\lambda \mathrm{I}\right]\theta_{n-1} + \gamma y_n x_n + \lambda\gamma\theta_0. \tag{3.9}$$

We have a very similar result with an additional corrective term (second line below) compared to Lemma 3.1.

Theorem 3.2. *Assume $(\mathcal{A}_{1,2})$. Consider the recursion in Equation (3.9). For any regularization parameter $\lambda \in \mathbb{R}^+$ and for any constant step-size γ such that $2\gamma(R^2+2\lambda) \leqslant 1$ we have:*

$$\begin{aligned}\mathbb{E} f(\bar{\theta}_n) - f(\theta_*) \quad \leqslant \quad & 3\Big(2\lambda + \frac{1}{\gamma n}\Big)^2 \|\Sigma^{1/2}(\Sigma+\lambda \mathrm{I})^{-1}(\theta_0-\theta_*)\|^2 + \frac{6\sigma^2}{n+1}\operatorname{tr}\left[\Sigma^2(\Sigma+\lambda \mathrm{I})^{-2}\right] \\ & + 3\frac{\left\|(\Sigma+\lambda \mathrm{I})^{-1/2}(\theta_0-\theta_*)\right\|^2 \operatorname{tr}(\Sigma(\Sigma+\lambda \mathrm{I})^{-1})}{\gamma^2(n+1)^2}.\end{aligned}$$

We can make the following remarks:

- The proof (see Section B.2) relies on a bias-variance decomposition, each term being treated separately. We adapt a proof technique from Bach and Moulines (2013) which considers the difference between the recursions in Equation (3.9) and in Equation (3.7).

- As in Lemma 3.1, the bias term can also be bounded by $\frac{1}{\gamma n}\|\Sigma^{1/2}(\Sigma+\lambda \mathrm{I})^{-1/2}(\theta_0-\theta_*)\|^2$ and the variance term by $\gamma \operatorname{tr}[\Sigma(\Sigma+\lambda \mathrm{I})^{-1}\xi_n \otimes \xi_n]$ (see proof in Sections B.2.4 and B.2.5). This is useful in particular when considering unstructured noise.
- The variance term is the same as in the previous case. However there is a residual term that now appears when we go to the fully stochastic oracle (second line). This term will go to zero when γ tends to zero and can be compared to the corrective term which also appears when Hsu et al. (2014) go from fixed to random design. Nevertheless our bounds are more concise than theirs, making significantly fewer assumptions and relying on an efficient single-pass algorithm.
- In this setting, the step-size may not exceed $1/(2(R^2+2\lambda))$, whereas with an additive noise in Lemma 3.1 the condition is $\gamma \leqslant 1/(L+\lambda)$, a quantity which can be much bigger than $1/(2(R^2+2\lambda))$, as L is the spectral radius of Σ whereas R^2 is of the order of $\operatorname{tr}(\Sigma)$. Note that in practice, computing L is as hard as computing θ_* so that the step-size $\gamma \propto 1/R^2$ is a good practical choice. See Défossez and Bach (2015) for larger allowed step-sizes that require more information.
- For $\lambda = 0$ the error is bounded by $\frac{3(1+d)}{(\gamma n)^2}\|\Sigma^{-1/2}(\theta_0-\theta_*)\|^2 + \frac{6\sigma^2 d}{n+1}$. We recover results from Défossez and Bach (2015) with a non-asymptotic bound but we lose the advantage of having an asymptotic equivalent (*i.e.*, a limit rather than an upper-bound). We note that the assumption $(\mathcal{A}_{1,2})$ are close to the minimal assumptions required to obtain the optimal rate of convergence of $\sigma^2 d/n$ (Lecué and Mendelson, 2016; Oliveira, 2016)

3.4 Accelerated Stochastic Averaged Gradient Descent

We study the convergence under the stochastic oracle from Equation (3.5) of averaged *accelerated* stochastic gradient descent defined by Equation (3.3) which can be rewritten for the least-squares function f as a second-order iterative system with constant coefficients:

$$\theta_n = \big[\mathrm{I} - \gamma\Sigma - \gamma\lambda \mathrm{I}\big]\big[\theta_{n-1} + \delta(\theta_{n-1}-\theta_{n-2})\big] + \gamma y_n x_n + \gamma\lambda\theta_0. \tag{3.10}$$

When using averaging, we refer to this algorithm as "averaged-accelerated-SGD".

Theorem 3.3. *Assume $(\mathcal{A}_4)$. For any regularization parameter $\lambda \in \mathbb{R}_+$ and for any constant step-size $\gamma(\Sigma+\lambda \mathrm{I}) \preccurlyeq \mathrm{I}$, we have for any $\delta \in [\frac{1-\sqrt{\gamma\lambda}}{1+\sqrt{\gamma\lambda}}, 1]$, for the recursion in Equation (3.10):*

$$\mathbb{E}f(\bar{\theta}_n) - f(\theta_*) \leqslant 2\Big(\lambda + \frac{36}{\gamma(n+1)^2}\Big)\|\Sigma^{1/2}(\Sigma+\lambda I)^{-1/2}(\theta_0-\theta_*)\|^2 + 8\tau^2\frac{\operatorname{tr}\big[\Sigma^2(\Sigma+\lambda I)^{-2}\big]}{n+1}.$$

The numerical constants are partially artifacts of the proof (see Sections B.3 and B.5). Thanks to a wise use of tight inequalities, the bound is independent of δ and valid for all $\lambda \in \mathbb{R}_+$. This results in the simple following corollary for $\lambda = 0$, which corresponds to the particularly simple recursion (with averaging to obtain $\bar{\theta}_n$):

$$\theta_n = \big[\mathrm{I} - \gamma\Sigma\big](2\theta_{n-1} - \theta_{n-2}) + \gamma y_n x_n. \tag{3.11}$$

Corollary 3.4. *Assume $(\mathcal{A}_4)$. For any constant step-size $\gamma\Sigma \preccurlyeq \mathrm{I}$, we have for $\delta = 1$,*

$$\mathbb{E}f(\bar{\theta}_n) - f(\theta_*) \leqslant 36\frac{\|\theta_0-\theta_*\|^2}{\gamma(n+1)^2} + 8\frac{\tau^2 d}{n+1}. \tag{3.12}$$

We can make the following observations:

- The proof technique relies on direct moment computations in each eigensubspace obtained by O'Donoghue and Candès (2013) in the deterministic case. Indeed as Σ is a symmetric matrix, the space can be decomposed on an orthonormal eigenbasis of Σ, and the iterations are decoupled in such an eigenbasis. Although we only provide an upper-bound, this is in fact an equality plus other exponentially small terms as shown in the proof which relies on linear algebra, with difficulties arising from the fact that this second-order system can be expressed as a linear stochastic dynamical system with non-symmetric matrices. We only provide a result for additive noise.

- The first bound $\frac{1}{\gamma n^2}\|\theta_0 - \theta_*\|^2$ in Equation (3.12) corresponds to the usual accelerated rate. It has been shown by Nesterov (2004) to be the optimal rate of convergence for optimizing a quadratic function with a first-order method that can access only to sequences of gradients when $n \leqslant d$. We recover by averaging an algorithm dedicated to strongly-convex function the traditional convergence rate for non-strongly convex functions. Even if it seems surprising, the algorithm works also for $\lambda = 0$ and $\delta = 1$.

- The second bound in Equation (3.12) also matches the optimal statistical performance $\frac{\tau^2 d}{n}$ described in the observations following Lemma 3.1. Accordingly this algorithm achieves joint bias/variance optimality (when measured in terms of τ^2 and $\|\theta_0 - \theta_*\|^2$).

- We have the same rate of convergence for the bias when compared to the regular Nesterov acceleration without averaging studied by Flammarion and Bach (2015), which corresponds to choosing $\delta_n = 1 - 2/n$ for all n. However if the problem is μ-strongly convex, this latter was shown to also converge at the linear rate $O((1-\gamma\mu)^n)$ and thus is adaptive to hidden strong-convexity (since the algorithm does not need to know μ to run). This explains that it ends up converging faster for quadratic function since for large n the convergence at rate $1/n^2$ becomes slower than the one at rate $(1-\gamma\mu)^n$ even for very small μ. Thanks to this adaptivity, we can also show using the same tools and considering its weighted average $\tilde{\theta}_n = \frac{2}{n(n+1)}\sum_{k=0}^n k\theta_k$ that the bias term of $\mathbb{E}f(\tilde{\theta}_n) - f(\theta_*)$ has a convergence rate of order $(\lambda^2 + \frac{1}{\gamma^2(n+1)^4})\|\Sigma^{1/2}(\Sigma + \lambda I)^{-1}(\theta_0 - \theta_*)\|^2$ without any change in the variance term. This has to be compared to the bias of averaged SGD $(\lambda + \frac{1}{\gamma(n+1)^2})\|\Sigma^{1/2}(\Sigma + \lambda I)^{-1}(\theta_0 - \theta_*)\|^2$ in Section 3.3 and may lead to faster convergence for the bias in presence of hidden strong-convexity.

- Overall, the bias term is improved whereas the variance term is not degraded and acceleration is thus robust to noise in the gradients. Thereby, while second-order finite difference methods for optimizing quadratic functions in the singular case, such as conjugate gradient (Polyak, 1987, Section 6.1) are notoriously highly sensitive to noise, we are able to propose a version which is robust to stochastic noise.

- Note that when there is no assumption on the covariance of the noise we still have the variance bounded by $\frac{\gamma n}{2}\operatorname{tr}\left[\Sigma(\Sigma + \lambda I)^{-1}V\right]$; setting $\gamma = 1/n^{3/2}$ and $\lambda = 0$ leads to the bound $\frac{\|\theta_0 - \theta_*\|^2}{\sqrt{n}} + \frac{\operatorname{tr} V}{\sqrt{n}}$. We recover the usual rate for accelerated stochastic gradient in the non-strongly-convex case (Xiao, 2010). When the values of the bias and the variance are known, we can achieve the optimal trade-off of Lan (2012) $\frac{R^2\|\theta_0 - \theta_*\|^2}{n^2} + \frac{\|\theta_0 - \theta_*\|\sqrt{\operatorname{tr} V}}{\sqrt{n}}$ for $\gamma = \min\left\{1/R^2, \frac{\|\theta_0 - \theta_*\|}{\sqrt{\operatorname{tr} V}n^{3/2}}\right\}$.

3.5 Tighter Dimension-independent Convergence Rates

We have seen in Corollary 3.4 above that the averaged accelerated gradient algorithm matches the lower bounds $\tau^2 d/n$ and $\frac{L}{n^2}\|\theta_0 - \theta_*\|^2$ for the prediction error. However the algorithm performs better in almost all cases except the worst-case scenarios corresponding to the lower bounds. For example the algorithm may still predict well when the dimension d is much bigger than n. Similarly the norm of the optimal predictor $\|\theta_*\|^2$ may be huge and the prediction still good, as gradients algorithms happen to be adaptive to the difficulty of the problem: indeed, if the problem is simpler, the convergence rate of the gradient algorithm will be improved. In this section, we provide such a theoretical guarantee.

The following bound stands for the averaged *accelerated* algorithm. It extends bounds from Chapter 2 in the kernel least-mean-squares setting.

Theorem 3.5. *Assume ($\mathcal{A}_4$); for any regularization parameter $\lambda \in \mathbb{R}_+$ and for any constant step-size such that $\gamma(\Sigma + \lambda \mathrm{I}) \preccurlyeq \mathrm{I}$ we have for $\delta \in [\frac{1-\sqrt{\gamma\lambda}}{1+\sqrt{\gamma\lambda}}, 1]$, for the recursion in Equation (3.10):*

$$\mathbb{E} f(\bar{\theta}_n) - f(\theta_*) \leqslant \min_{r\in[0,1],\ b\in[0,1]} \left[2\|\Sigma^{r/2}(\theta_0 - \theta_*)\|^2 \, \lambda^{-r} \left(\frac{36}{\gamma(n+1)^2} + \lambda \right) + 8 \frac{\tau^2 \operatorname{tr}(\Sigma^b) \lambda^{-b}}{n+1} \right].$$

The proof is straightforward by upper bounding the terms coming from regularization, depending on $\Sigma(\Sigma + \lambda \mathrm{I})^{-1}$, by a power of λ times the considered quantities. More precisely, the quantity $\operatorname{tr}(\Sigma(\Sigma + \lambda \mathrm{I})^{-1})$ can be seen as an effective dimension of the problem (Gu, 2002), and is upper bounded by $\lambda^{-b} \operatorname{tr}(\Sigma^b)$ for any $b \in [0; 1]$. Similarly, $\|\Sigma^{1/2}(\Sigma + \lambda I)^{-1/2}\theta_*\|^2$ can be upper bounded by $\lambda^{-r}\|\Sigma^{r/2}(\theta_0 - \theta_*)\|^2$. A detailed proof of these results is given in Section B.4.

In order to benefit from the acceleration, we choose $\lambda = (\gamma n^2)^{-1}$. With such a choice we have the following corollary:

Corollary 3.6. *Assume ($\mathcal{A}_4$), for any constant step-size $\gamma(\Sigma + \lambda \mathrm{I}) \preccurlyeq \mathrm{I}$, we have for $\lambda = \frac{1}{\gamma(n+1)^2}$ and $\delta \in \left[1 - \frac{2}{n+2}, 1\right]$, for the recursion in Equation (3.10):*

$$\mathbb{E} f(\bar{\theta}_n) - f(\theta_*) \leqslant \min_{r\in[0,1],\ b\in[0,1]} \left[74 \, \frac{\|\Sigma^{r/2}(\theta_0 - \theta_*)\|^2}{\gamma^{1-r}(n+1)^{2(1-r)}} + 8 \frac{\tau^2 \gamma^b \operatorname{tr}(\Sigma^b)}{(n+1)^{1-2b}} \right].$$

We can make the following observations:

- The algorithm is independent of r and b, thus all the bounds for different values of (r, b) are valid. This is a strong property of the algorithm, which is indeed adaptive to the regularity and the effective dimension of the problem (once γ is chosen). In situations in which either d is larger than n or $L\|\theta_0 - \theta_*\|^2$ is larger than n^2, the algorithm can still enjoy good convergence properties, by adapting to the best values of b and r.

- For $b = 0$ we recover the variance term of Corollary 3.4, but for $b > 0$ and fast decays of eigenvalues of Σ, the bound may be much smaller; note that we lose in the dependency in n, but typically, for large d, this can be advantageous.

- For $r = 0$ we recover the bias term of Corollary 3.4 and for $r = 1$ (no assumption at all) the bias is bounded by $\|\Sigma^{1/2}\theta_*\|^2 \leqslant 4R^2$, which is not going to zero. The smaller r is, the stronger the decrease of the bias with respect to n is (which is coherent with

the fact that we have a stronger assumption). Moreover, r is only considered between 0 and 1: indeed, if $r < 0$, the constant$\|(\gamma\Sigma)^{r/2}(\theta_0 - \theta_*)\|$ is bigger than $\|\theta_0 - \theta_*\|$, but the dependence on n cannot improve beyond $(\gamma n^2)^{-1}$. This is a classical phenomenon called "saturation" (Engl et al., 1996). It is linked with the uniform averaging scheme: here, the bias term cannot forget the initial condition faster than n^{-2}.

- A similar result happens to hold, for averaged gradient descent, with $\lambda = (\gamma n)^{-1}$:

$$\mathbb{E} f(\bar{\theta}_n) - f(\theta_*) \leqslant \min_{\substack{r\in[-1,1]\\ b\in[0,1]}} \left[(18 + \mathrm{Res}(b,r,n,\gamma)) \frac{\|\Sigma^{r/2}(\theta_0 - \theta_*)\|^2}{\gamma^{1-r}(n+1)^{(1-r)}} + 6\frac{\sigma^2\gamma^b \operatorname{tr}(\Sigma^b)}{(n+1)^{1-b}}\right] \quad (3.13)$$

where $\mathrm{Res}(b,r,n,\gamma))$ corresponds to a residual term, which is smaller than $\operatorname{tr}(\Sigma^b)n^b\gamma^{1+b}$ if $r \geqslant 0$ and does not exist otherwise. The bias term's dependence on n is degraded, thus the "saturation" limit is logically pushed down to $r = -1$, which explains the $[-1;1]$ interval for r. The choice $\lambda = (\gamma n)^{-1}$ arises from Th. 3.2, in order to balance both components of the bias term $\lambda + (\gamma n)^{-1}$. This result is proved in Section B.4. This recovers the result of Chapter 2.

- Considering a non-uniform averaging, as proposed after Theorem 3.2 the $\min_{0\leqslant r\leqslant 1}$ in Th. 3.5 and Corollary 3.6 can be extended to $\min_{-1\leqslant r\leqslant 1}$. Indeed, considering a non-uniform averaging allows to have a faster decreasing bias, pushing the saturation limit observed below.

In finite dimension these bounds for the bias and the variance cannot be said to be optimal independently in any sense we are aware of. Indeed, in finite dimension, the asymptotic rate of convergence for the bias (respectively the variance), when n goes to ∞ is governed by $L\|\theta_0 - \theta_*\|^2/n^2$ (resp. $\tau^2 d/n$). However, we show in the next section that in the setting of non parametric learning in kernel spaces, these bounds lead to the optimal statistical rate of convergence among all estimators (independently of their computational cost). Moving to the infinite-dimensional setting allows to characterize the optimality of the bounds by showing that they achieve the statistical rate when optimizing the bias/variance tradeoff in Corollary 3.6.

3.6 Rates of Convergence for Kernel Regression

Computational convergence rates give the speed at which an objective function can decrease depending on the amount of computation which is allowed. Typically, they show how the error decreases with respect to the number of iterations, as in Theorem 3.2. Statistical rates, however, show how close one can get to some objective given some amount of information which is provided. Statistical rates do not depend on some chosen algorithm: these bounds do not involve computation, on the contrary, they state the best performance that no algorithm can beat, given the information, and without computational limits. In particular, any lower bound on the statistical rate implies a lower bound on the computational rates, if each iteration corresponds to access to some new information, here pairs of observations. Interestingly, many algorithms for the past few years have proved to match, with minimal computations (in general one pass through the data), the statistical rate, emphasizing the importance of carrying together optimization and approximation in large scale learning, as described by Bottou and Bousquet (2008). In a similar flavor, it also appears that

regularization can be accomplished through early stopping (Yao et al., 2007; Rudi et al., 2015), highlighting this interplay between computation and statistics.

To characterize the optimality of the bounds, we will show that averaged-accelerated-SGD matches the statistical lower bound in the context of non-parametric estimation. Even if it may be computationally hard or impossible to implement averaged-accelerated-SGD with additive noise in the kernel-based framework below (see remarks following Theorem 3.8), it leads to the optimal statistical rate for a broader class of problems than averaged-SGD, showing that for a wider set of trade-offs, acceleration is optimal.

A natural extension of the finite-dimensional analysis is the non-parametric setting, especially with reproducing kernel Hilbert spaces. In the setting of non-parametric regression, we consider a probability space $\mathcal{X} \times \mathbb{R}$ with probability distribution ρ, and assume that we are given an i.i.d. sample $(x_i, y_i)_{i=1,\dots,n} \sim \rho^{\otimes n}$, and denote by ρ_X the marginal distribution of x_n in $\mathcal{X}$; the aim of non-parametric least-squares regression is to find a function $g : \mathcal{X} \to \mathbb{R}$, which minimizes the expected risk:

$$f(g) = \frac{1}{2}\mathbb{E}_\rho[(g(x_n) - y_n)^2]. \tag{3.14}$$

The optimal function g is the conditional expectation $g(x) = \mathbb{E}_\rho(y_n|x)$. In the kernel regression setting, we consider as hypothesis space a reproducing kernel Hilbert space (Aronszajn, 1950; Steinwart and Christmann, 2008; Schölkopf and Smola, 2002) associated with a kernel function K. The space $\mathcal{H}$ is a subspace of the space of squared integrable functions $L^2_{\rho_X}$. We look for a function $g_\mathcal{H}$ which satisfies: $f(g_\mathcal{H}) = \inf_{g\in\mathcal{H}} f(g)$, and $g_\mathcal{H}$ belongs to the closure $\bar{\mathcal{H}}$ of $\mathcal{H}$ (meaning that there exists a sequence of function $g_n \in \mathcal{H}$ such that $\|g_n - g_H\|_{L^2_{\rho_X}} \to 0$). When $\mathcal{H}$ is dense, the minimum is attained for the regression function defined above. This function however *is not* in $\mathcal{H}$ in general. Moreover there exists an operator $\Sigma : \mathcal{H} \to \mathcal{H}$, which extends the finite-dimensional population covariance matrix, that will allow the characterization of the smoothness of $g_\mathcal{H}$. This operator is known to be trace class when $\mathbb{E}_{\rho_X}[K(x_n, x_n)] < \infty$.

Data points x_i are mapped into the RKHS, via the feature map: $x \mapsto K_x$, where $K_x : \mathcal{H} \to \mathbb{R}$ is a function in the RKHS, such that $K_x : y \mapsto K(x, y)$. The reproducing property[3] allows to express the minimization problem (3.14) as a least-squares linear regression problem: for any $g \in \mathcal{H}$, $f(g) = \frac{1}{2}\mathbb{E}_\rho[(\langle g, K_{x_n}\rangle_\mathcal{H} - y_n)^2]$, and can thus be seen as an extension to the infinite-dimensional setting of linear least-squares regression.

However, in such a setting, both quantities $\|\Sigma^{r/2}\theta_*\|_\mathcal{H}$ (where $\|\cdot\|_\mathcal{H}$ stands for the norm associated with the inner product in the Hilbert space $\mathcal{H}$) and $\mathrm{tr}(\Sigma^b)$ *may exist or not*. It thus arises as a natural *assumption* to consider the smaller $r \in [-1; 1]$ and the smaller $b \in [0; 1]$ such that

- $\|\Sigma^{r/2}\theta_*\|_\mathcal{H} < \infty$ (meaning that $\Sigma^{r/2}\theta_* \in \mathcal{H}$), ($\mathcal{A}_5$)
- $\mathrm{tr}(\Sigma^b) < \infty$. ($\mathcal{A}_6$)

The quantities considered in Sections 3.2 and 3.5 are the natural finite-dimensional twins of these assumptions. However in infinite dimension a quantity may exist or not and it is thus an assumption to consider its existence, whereas it can only be characterized by its value, big or small, in finite dimension. The first assumption is generally called the "source condition", the second one the "capacity condition".

[3] It states that for any function $g \in \mathcal{H}$, $\langle g, K_x\rangle_\mathcal{H} = g(x)$, where $\langle\cdot,\cdot\rangle_\mathcal{H}$ denotes the scalar product in the Hilbert space.

In the last decade, De Vito et al. (2005); Smale and Cucker (2002) studied non-parametric least-squares regression in the RKHS framework. These works were extended to derive rates of convergence depending on assumption $(\mathcal{A}_5)$: Ying and Pontil (2008) studied un-regularized stochastic gradient descent and derived asymptotic rate of convergence $O(n^{-\frac{1-r}{2-r}})$, for $r \leqslant 1$ and proved that one could derive similar rates of convergence for $0 \leqslant r \leqslant 1$ from Zhang (2004), who studies stochastic gradient descent with averaging; whereas Tarrès and Yao (2014) give similar performance for $-1 \leqslant r \leqslant 0$. Interestingly, Ying and Pontil (2008) do not have saturation, meaning that the rate still improves for r smaller than -1. As it will appear, any algorithm based on a uniform averaging scheme faces a saturation issue: one cannot forget initial conditions faster than n^{-2}, which makes the algorithm sub-optimal in situations in which the optimal predictor is very smooth ($\mathcal{A}_5$ holds with $r \leqslant -1$). However, these papers only prove rates in the capacity-independent setting, meaning without assumption on the spectrum of the covariance matrix. Although the rate $O(n^{-\frac{1-r}{2-r}})$ is optimal in this setting, it comes from a worst-case analysis. Considering the capacity-dependent setting is more challenging, but allows to derive tighter and more realistic rates (a capacity condition always stands under the trace class assumption that is made). Moreover, the capacity-independent setting also does not allow to recover finite-dimensional rates. Up to our knowledge, there is no one pass stochastic gradient algorithm which does not have saturation while getting the minimax rate under both the capacity condition and source condition. In a recent work, Lin and Rosasco (2016) achieves optimality without saturation with multiple passes. We show in the next paragraphs that we can derive a tighter and optimal rate for both averaged-SGD (recovering results from Chapter 2) and averaged-accelerated-SGD, for a larger class of kernels for the latter. Note that the averaging scheme for the RKHS setting was originally considered by Yao (2006).

We will first describe results for averaged-SGD, then increase the validity region of these rates (which depends on r, b) using averaged accelerated SGD. We show that the derived rates match statistical rates for our setting and thus our algorithms reach the optimal prediction performance for certain b and r.

3.6.1 Averaged SGD

We have the following result, proved in Section B.4 and following from Theorem 3.2: for some fixed b, r, we choose the best step-size γ, that optimizes the bias-variance trade-off, while still satisfying the constraint $\gamma \leqslant 1/(2R^2)$. We get a result for the stochastic oracle (multiplicative/additive noise).

Theorem 3.7. *With $\lambda = \frac{1}{\gamma n}$, we have, if $r \leqslant b$, under Assumptions $(\mathcal{A}_{1,2,5,6})$ and the stochastic oracle Equation (3.6), for any constant step-size γ such that $2\gamma(R^2+2\lambda) \leqslant 1$, with $\gamma \propto n^{\frac{-b+r}{b+1-r}}$, for the recursion in Equation (3.9):*

$$\mathbb{E}f(\bar{\theta}_n) - f(\theta_*) \quad \leqslant \quad \left((27 + o(1)) \|\Sigma^{r/2}(\theta_0 - \theta_*)\|^2 + 6\sigma^2 \operatorname{tr}(\Sigma^b) \right) n^{-\frac{1-r}{b+1-r}}.$$

We can make the following remarks:

- The term $o(1)$ stands for a quantity which is decreasing to 0 when $n \to \infty$. More specifically, this constant is smaller than $3\operatorname{tr}(\Sigma^b)$ divided by n^χ, where χ is bigger than 0 (see Section B.4). The result comes from Equation (3.13) (which follows from Theorem 3.5), with the choice of the optimal step-size.

- We recover the same errors bounds as in Chapter 2, but with a simpler analysis resulting from the consideration of the regularized version of the problem associated with a choice of λ. However, we only recover rates in the finite horizon setting.
- This result shows that we get the optimal rate of convergence under Assumptions $(\mathcal{A}_{5,6})$, for $r \leqslant b$. This point will be discussed in more details after Theorem 3.8.

We now turn to the averaged accelerated SGD algorithm. We prove that it enjoys the optimal rate of convergence for a larger class of problems, but only for the additive noise which corresponds to knowing the distribution of x_n.

3.6.2 Averaged-accelerated SGD

Similarly, choosing the best step-size γ, it comes from Theorem 3.5, that in the RKHS setting, under additional Assumptions $(\mathcal{A}_{5,6})$, we have for the the averaged accelerated algorithm the following result:

Theorem 3.8. *With $\lambda = \frac{1}{\gamma n^2}$, we have, if $r \leqslant b + 1/2$, under Assumptions $(\mathcal{A}_{4,5,6})$, for any constant step-size $\gamma \leqslant \frac{1}{L+\lambda}$, with $\gamma \propto n^{\frac{-2b+2r-1}{b+1-r}}$, for the recursion in Equation (3.10):*

$$\mathbb{E} f(\bar{\theta}_n) - f(\theta_*) \quad \leqslant \quad \left(74 \|\Sigma^{r/2}(\theta_0 - \theta_*)\|^2 + 8\tau^2 \operatorname{tr}(\Sigma^b) \right) n^{-\frac{1-r}{b+1-r}}.$$

We can make the following remarks:

- The rate $\frac{1-r}{b+1-r}$ is always between 0 and 1, and improves when our assumptions gets stronger (r getting smaller, b getting smaller). Ultimately, with $b \to 0$, and $r \to -1$, we recover the finite-dimensional n^{-1} rate.
- We can achieve this optimal rate when $r \leqslant b + 1/2$. Beyond, if $r > b + 1/2$, the rate is only $n^{-2(1-r)}$. Indeed, the bias term cannot decrease faster than $n^{-2(1-r)}$, as γ is compelled to be upper bounded.
- The same phenomenon appears in the un-accelerated averaged situation, as shown by Theorem 3.7, but the critical value was then $r \leqslant b$. There is thus a region (precisely $b < r \leqslant b + 1/2$) in which only the accelerated algorithm gets the optimal rate of convergence. Note that we increase the optimality region towards optimization problems which are more ill-conditioned, naturally benefiting from acceleration. This is represented in Figure 3.1.
- This algorithm cannot be computed in practice (at least with computational limits). Indeed, without any further assumption on the kernel K, it is not possible to compute images of vectors by the covariance operator Σ in the RKHS. However, as explained in the following remark, this is enough to show some form of optimality of our algorithm.

  Note that the easy computability is a great advantage of the multiplicative/additive noise variant of the algorithms, for which the current point θ_n can always be expressed as a finite sum of features $\theta_n = \sum_{i=1}^n \alpha_i K_{x_i}$, with $\alpha_i \in \mathbb{R}$, leading to a tractable algorithm. An accelerated variant of SGD naturally arises from our algorithm, when considering this stochastic oracle from Equation (3.6). Such a variant can be implemented but does not behave similarly for large step sizes, say, $\gamma \simeq 1/(2R^2)$. It is an

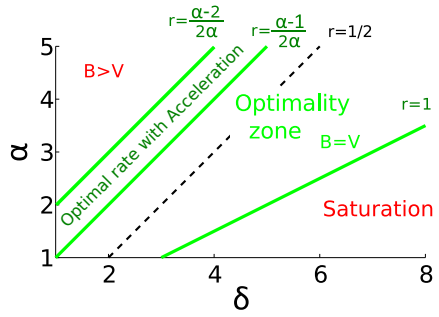


Figure 3.1: Regions of theoretical optimal convergence with acceleration. ⚠: in this figure only, for the sake of comparison, we use the notations from Chapter 2.

open problem to prove convergence results for averaged accelerated gradient under this multiplicative/additive noise.

- These rates happen to be optimal from a statistical perspective, meaning that no algorithm which is given access to the sample points and the distribution of x_n can perform better for all functions that satisfy assumption $(\mathcal{A}_6)$, for a kernel satisfying $(\mathcal{A}_5)$. Indeed it is equivalent to assuming that the function lives in some ellipsoid in the space of squared integrable functions. Note that the statistical minimization problem (and thus the lower bound) does not depend on the kernel, and is valid without computational limits. The case of learning with kernels is studied by Caponnetto and De Vito (2007) which shows these minimax convergence rates under $(\mathcal{A}_{5,6})$, under assumption that $-1 \leqslant r \leqslant 0$ (but state that it can be easily extended to $0 \leqslant r \leqslant 1$). They do not assume knowledge of the distribution of the inputs; however, Massart (2007) and Tsybakov (2008) discuss optimal rates on ellipsoids, and Györfi et al. (2002) proves similar results for certain class of functions under a known distribution for the input data, showing that the knowledge of the distribution does not make any difference. This minimax statistical rate stands without computational limits and is thus valid for both algorithms (additive noise that corresponds to knowing Σ, and multiplicative/additive noise). The optimal tradeoff is derived for an extended region of b, r (namely $r \leqslant b + 1/2$ instead of $r \leqslant b$) in the accelerated case which shows the improvement upon non-accelerated averaged SGD.

- The choice of the optimal γ is difficult in practice, as the parameters b, r are unknown, and this remains an open problem in general (see, e.g., Birgé, 2001, for some methods for non-parametric regression), even if in the capacity-independent setting, Orabona (2014) has proposed an algorithm that adapts to the unknown parameter r.

- Note that we do not give rates in terms of norm in the RHKS (*i.e.*, an upper bound on $\|\bar{\theta}_n - \theta_*\|_{\mathcal{H}}$), because we mainly aim at extending optimality of prediction error rate to ill-conditioned cases (*i.e.*, situations for which $r \geqslant b \geqslant 0$). In such a situation, Hilbert spaces norm bounds would not be relevant as the optimal estimator does not even live in the RKHS.

3.7 Conclusion

In this chapter, we showed that stochastic *averaged* accelerated gradient descent was robust to structured noise in the gradients present in least-squares regression. Beyond being the first algorithm which is jointly optimal in terms of both bias and finite-dimensional variance, it is also adapted to finer assumptions such as fast decays of the covariance matrices or optimal predictors with large norms.

In Chapters 2 and 3, we have focused on least squares regression. While it is a very classical setting in practice, many methods, for example in Classification, rely on non quadratic risk functions. In Chapter 4, we consider a more general strongly convex and smooth function and analyze SGD with constant step size.

The proofs of the results given in this chapter are given in the next chapter (Ch. C): it might be skipped at first reading.

B

Appendix to Faster Convergence Rates for Least-Squares Regression

B.1 Proofs of Section 3.3

B.1.1 Proof of Lemma 3.1

We proof here Lemma 3.1 which is the extension of Lemma 2 of Bach and Moulines (2013) for the regularized case. The proof technique relies on the fact that recursions in Equation (3.7) are linear since the cost function is quadratic which allows us to obtain $\theta_n - \theta_*$ in closed form.

For any regularization parameter $\lambda \in \mathbb{R}_+$ and any constant step-size $\gamma(\Sigma + \lambda \mathrm{I}) \preccurlyeq \mathrm{I}$ we may rewrite the regularized stochastic gradient recursion in Equation (3.7) as:

$$\theta_n - \theta_* = \big[\mathrm{I} - \gamma\Sigma - \gamma\lambda\mathrm{I}\big](\theta_{n-1} - \theta_*) + \gamma\xi_n + \lambda\gamma(\theta_0 - \theta_*).$$

We thus get for $n \geqslant 1$ the expansion

$$\begin{aligned}
\theta_n - \theta_* &= (\mathrm{I} - \gamma\Sigma - \gamma\lambda\mathrm{I})^n(\theta_0 - \theta_*) + \gamma\sum_{k=1}^{n}(\mathrm{I} - \gamma\Sigma - \gamma\lambda\mathrm{I})^{n-k}\xi_k \\
&\quad + \gamma\lambda\sum_{k=1}^{n}(\mathrm{I} - \gamma\Sigma - \gamma\lambda\mathrm{I})^{n-k}(\theta_0 - \theta_*) \\
&= (\mathrm{I} - \gamma\Sigma - \gamma\lambda\mathrm{I})^n(\theta_0 - \theta_*) + \gamma\sum_{k=1}^{n}(\mathrm{I} - \gamma\Sigma - \gamma\lambda\mathrm{I})^{n-k}\xi_k \\
&\quad + \lambda\big[\mathrm{I} - (\mathrm{I} - \gamma\Sigma - \gamma\lambda\mathrm{I})^n\big](\Sigma + \lambda\mathrm{I})^{-1}(\theta_0 - \theta_*) \\
&= (\mathrm{I} - \gamma\Sigma - \gamma\lambda\mathrm{I})^n[\mathrm{I} - \lambda(\Sigma + \lambda\mathrm{I})^{-1}](\theta_0 - \theta_*) + \gamma\sum_{k=1}^{n}(\mathrm{I} - \gamma\Sigma - \gamma\lambda\mathrm{I})^{n-k}\xi_k \\
&\quad + \lambda(\Sigma + \lambda\mathrm{I})^{-1}(\theta_0 - \theta_*).
\end{aligned}$$

We then have using the definition of the average

$$\begin{aligned}
n(\bar{\theta}_{n-1} - \theta_*) &= \sum_{j=0}^{n-1}(\theta_j - \theta_*) \\
&= \sum_{j=0}^{n-1}(\mathrm{I} - \gamma\Sigma - \gamma\lambda\mathrm{I})^j[\mathrm{I} - \lambda(\Sigma + \lambda\mathrm{I})^{-1}](\theta_0 - \theta_*)
\end{aligned}$$

$$+\gamma\sum_{j=0}^{n-1}\sum_{k=1}^{j}(\mathrm{I}-\gamma\Sigma-\gamma\lambda\mathrm{I})^{n-k}\xi_k+n\lambda(\Sigma+\lambda\mathrm{I})^{-1}(\theta_0-\theta_*).$$

For which we will compute the two sums separately

$$\sum_{j=0}^{n-1}(\mathrm{I}-\gamma\Sigma-\gamma\lambda\mathrm{I})^{j}[\mathrm{I}-\lambda(\Sigma+\lambda\mathrm{I})^{-1}](\theta_0-\theta_*)$$
$$=\frac{1}{\gamma}[\mathrm{I}-(\mathrm{I}-\gamma\Sigma-\gamma\lambda\mathrm{I})^{n}](\Sigma+\lambda\mathrm{I})^{-1}[\mathrm{I}-\lambda(\Sigma+\lambda\mathrm{I})^{-1}](\theta_0-\theta_*),$$

and

$$\begin{aligned}\gamma\sum_{j=0}^{n-1}\sum_{k=1}^{j}(\mathrm{I}-\gamma\Sigma-\gamma\lambda\mathrm{I})^{j-k}\xi_k &= \gamma\sum_{k=1}^{n-1}\Big(\sum_{j=k}^{n-1}(\mathrm{I}-\gamma\Sigma-\gamma\lambda\mathrm{I})^{j-k}\Big)\xi_k\\ &= \gamma\sum_{k=1}^{n-1}\Big(\sum_{j=0}^{n-1-k}(\mathrm{I}-\gamma\Sigma-\gamma\lambda\mathrm{I})^{j}\Big)\xi_k\\ &= \sum_{k=1}^{n-1}[\mathrm{I}-(\mathrm{I}-\gamma\Sigma-\gamma\lambda\mathrm{I})^{n-k}](\Sigma+\lambda\mathrm{I})^{-1}\xi_k.\end{aligned}$$

Gathering the three terms together, we thus have

$$\begin{aligned}n(\bar\theta_{n-1}-\theta_*) &= \frac{1}{\gamma}[\mathrm{I}-(\mathrm{I}-\gamma\Sigma-\gamma\lambda\mathrm{I})^{n}](\Sigma+\lambda\mathrm{I})^{-1}[\mathrm{I}-\lambda(\Sigma+\lambda\mathrm{I})^{-1}](\theta_0-\theta_*)\\ &\quad+\sum_{k=1}^{n-1}[\mathrm{I}-(\mathrm{I}-\gamma\Sigma-\gamma\lambda\mathrm{I})^{n-k}](\Sigma+\lambda\mathrm{I})^{-1}\xi_k+n\lambda(\Sigma+\lambda\mathrm{I})^{-1}(\theta_0-\theta_*)\\ &= \Big[\frac{1}{\gamma}[\mathrm{I}-(\mathrm{I}-\gamma\Sigma-\gamma\lambda\mathrm{I})^{n}][\mathrm{I}-\lambda(\Sigma+\lambda\mathrm{I})^{-1}]+n\lambda\mathrm{I}\Big](\Sigma+\lambda\mathrm{I})^{-1}(\theta_0-\theta_*)\\ &\quad+\sum_{k=1}^{n-1}[\mathrm{I}-(\mathrm{I}-\gamma\Sigma-\gamma\lambda\mathrm{I})^{n-k}](\Sigma+\lambda\mathrm{I})^{-1}\xi_k.\end{aligned}$$

Using standard martingale square moment inequalities which amount to consider ξ_i, $i=1,\dots,n$ independent, the variance of the sum is the sum of variances and we have for $V=\mathbb{E}\xi_n\otimes\xi_n$

$$\begin{aligned}&n^2\mathbb{E}\|\Sigma^{1/2}(\bar\theta_{n-1}-\theta_*)\|^2=\textstyle\sum_{k=1}^{n-1}\operatorname{tr}\big[\mathrm{I}-(\mathrm{I}-\gamma\Sigma-\gamma\lambda\mathrm{I})^{n-k}\big]^2\Sigma(\Sigma+\lambda\mathrm{I})^{-2}V\\ &+\Big\|\Big[\frac{1}{\gamma}[\mathrm{I}-(\mathrm{I}-\gamma\Sigma-\gamma\lambda\mathrm{I})^{n}][\mathrm{I}-\lambda(\Sigma+\lambda\mathrm{I})^{-1}]+n\lambda\mathrm{I}\Big]\Sigma^{1/2}(\Sigma+\lambda\mathrm{I})^{-1}(\theta_0-\theta_*)\Big\|^2.\end{aligned}\tag{B.1}$$

Since all the matrices in this equality are symmetric positive-definite we are allowed to bound

$$\begin{aligned}\Big[\frac{1}{\gamma}[\mathrm{I}-(\mathrm{I}-\gamma\Sigma-\gamma\lambda\mathrm{I})^{n}][\mathrm{I}-\lambda(\Sigma+\lambda\mathrm{I})^{-1}]+n\lambda\mathrm{I}\Big] &\preccurlyeq \Big(\frac{1}{\gamma}+n\lambda\Big)\mathrm{I}\\ [\mathrm{I}-(\mathrm{I}-\gamma\Sigma-\gamma\lambda\mathrm{I})^{n-k}]^2 &\preccurlyeq \mathrm{I}.\end{aligned}\tag{B.2}$$

This concludes proof of the Lemma 3.1

$$\mathbb{E}\|\Sigma^{1/2}(\bar\theta_{n-1}-\theta_*)\|^2\leqslant\Big(\frac{1}{n\gamma}+\lambda\Big)^2\|\Sigma^{1/2}(\Sigma+\lambda\mathrm{I})^{-1}(\theta_0-\theta_*)\|^2$$
$$+\frac{1}{n}\operatorname{tr}\Sigma(\Sigma+\lambda\mathrm{I})^{-2}V.\tag{B.3}$$

B.1.2 Proof when only $\|\theta_0 - \theta_*\|$ is finite

Unfortunately $\|\Sigma^{-1}(\theta_0 - \theta_*)\|$ may not be finite. However we can use that for all $u \in [0,1]$ we have $\frac{1-(1-u)^n}{nu} \leqslant 1$[1] and have therefore the bound

$$\begin{aligned}
&\Big[\frac{1}{\gamma}\big[\mathrm{I} - (\mathrm{I} - \gamma\Sigma - \gamma\lambda\mathrm{I})^n\big]\big[\mathrm{I} - \lambda(\Sigma + \lambda\mathrm{I})^{-1}\big] + n\lambda\mathrm{I}\Big][\Sigma + \lambda\mathrm{I}]^{-1} \\
&\preccurlyeq \Big[\frac{1}{\gamma}\big[\mathrm{I} - (\mathrm{I} - \gamma\Sigma - \gamma\lambda\mathrm{I})^n\big] + n\lambda\mathrm{I}\Big][\Sigma + \lambda\mathrm{I}]^{-1} \\
&\preccurlyeq \Big[\frac{1}{\gamma}\big[\mathrm{I} - (\mathrm{I} - \gamma\Sigma - \gamma\lambda\mathrm{I})^n\big][\Sigma + \lambda\mathrm{I}]^{-1} + n\lambda[\Sigma + \lambda\mathrm{I}]^{-1}\Big] \\
&\preccurlyeq \mathrm{I} + n\mathrm{I}.
\end{aligned}$$

Combining with Equation (B.2) we have

$$\begin{aligned}
\Big\|\Big[\frac{1}{\gamma}\big[\mathrm{I} - (\mathrm{I} - \gamma\Sigma - \gamma\lambda\mathrm{I})^n\big]\big[\mathrm{I} - \lambda(\Sigma + \lambda\mathrm{I})^{-1}\big] + n\lambda\mathrm{I}\Big]\Sigma^{1/2}(\Sigma + \lambda\mathrm{I})^{-1}(\theta_0 - \theta_*)\Big\|^2 \\
\leqslant (n+1)\Big(\frac{1}{\gamma} + n\lambda\Big)\|\Sigma^{1/2}(\Sigma + \lambda\mathrm{I})^{-1/2}(\theta_0 - \theta_*)\|^2,
\end{aligned}$$

which implies that

$$\begin{aligned}
\mathbb{E}\|\Sigma^{1/2}(\bar{\theta}_{n-1} - \theta_*)\|^2 \leqslant 2\Big(\frac{1}{n\gamma} + \lambda\Big)\|\Sigma^{1/2}(\Sigma + \lambda\mathrm{I})^{-1/2}(\theta_0 - \theta_*)\|^2 \\
+ \frac{1}{n}\operatorname{tr}\Sigma(\Sigma + \lambda\mathrm{I})^{-2}V, \quad \text{(B.4)}
\end{aligned}$$

which is interesting when only $\|\theta_0 - \theta_*\|$ is finite.

B.1.3 Proof when the noise is not structured

The bound in Equation (B.3) becomes less interesting when the noise is not structured. However using the same technique we have that $\big[\mathrm{I} - (\mathrm{I} - \gamma\Sigma - \gamma\lambda\mathrm{I})^{n-k}\big]^2(\Sigma + \lambda\mathrm{I})^{-1} \preccurlyeq (n-k)\gamma\mathrm{I}$ and we get the following upper-bound on the variance

$$\begin{aligned}
\sum_{k=1}^{n}\operatorname{tr}\big[\mathrm{I} - (\mathrm{I} - \gamma\Sigma - \gamma\lambda\mathrm{I})^{n-k}\big]^2\Sigma(\Sigma + \lambda\mathrm{I})^{-2}V &\leqslant \gamma\sum_{k=1}^{n}(n-k)\operatorname{tr}\Sigma(\Sigma + \lambda\mathrm{I})^{-1}V \\
&\leqslant \gamma\frac{n(n+1)}{2}\operatorname{tr}\Sigma(\Sigma + \lambda\mathrm{I})^{-1}V.
\end{aligned}$$

Therefore we get

$$\begin{aligned}
\mathbb{E}\|\Sigma^{1/2}(\bar{\theta}_{n-1} - \theta_*)\|^2 \leqslant \Big(\frac{1}{n\gamma} + \lambda\Big)^2\|\Sigma^{1/2}(\Sigma + \lambda\mathrm{I})^{-1}(\theta_0 - \theta_*)\|^2 \\
+ \gamma\operatorname{tr}\Sigma(\Sigma + \lambda\mathrm{I})^{-1}V, \quad \text{(B.5)}
\end{aligned}$$

which is meaningful when the noise is not structured.

B.2 Proof of Theorem 3.2

In this section, we will prove Theorem 3.2. The proof relies on a decomposition of the error as the sum of three main terms which will be studied separately. We state decomposition in Section B.2.1 then prove upper bounds for the different terms in Sections B.2.2 and B.2.3.

[1] since $\frac{1-(1-u)^n}{u} = \sum_{k=0}^{n}(1-u)^k \leqslant n$

B.2.1 Expansion of the recursion

We may rewrite the regularized stochastic gradient recursion as:

$$\begin{aligned}\theta_n &= [\mathrm{I} - \gamma x_n \otimes x_n - \gamma\lambda \mathrm{I}]\theta_{n-1} + \gamma\varepsilon_n x_n + \gamma\langle x_n, \theta_*\rangle x_n + \lambda\gamma\theta_0 \\ \theta_n - \theta_* &= [\mathrm{I} - \gamma x_n \otimes x_n - \gamma\lambda \mathrm{I}](\theta_{n-1} - \theta_*) + \gamma\varepsilon_n x_n + \lambda\gamma(\theta_0 - \theta_*).\end{aligned}$$

For $i \geqslant k$, let

$$M(i,k) = [\mathrm{I} - \gamma x_i \otimes x_i - \gamma\lambda \mathrm{I}] \cdots [\mathrm{I} - \gamma x_k \otimes x_k - \gamma\lambda \mathrm{I}]$$

be an operator from $\mathcal{H}$ to $\mathcal{H}$. We have the expansion

$$\theta_n - \theta_* = M(n,1)(\theta_0 - \theta_*) + \gamma \sum_{k=1}^n M(n,k+1)\varepsilon_k x_k + \gamma \sum_{k=1}^n M(n,k+1)\lambda(\theta_0 - \theta_*).$$

Our goal is to study these three terms separately and bound $\|\Sigma^{1/2}(\bar{\theta}_n - \theta_*)\|$ for each of them.

B.2.2 Regularization-based bias term

This is the term: $\theta_n - \theta_* = \gamma \sum_{k=1}^n M(n,k+1)\lambda(\theta_0 - \theta_*)$, which corresponds to the recursion

$$\theta_n - \theta_* = (\mathrm{I} - \gamma x_n \otimes x_n - \gamma\lambda \mathrm{I})(\theta_{n-1} - \theta_*) + \lambda\gamma(\theta_0 - \theta_*), \tag{B.6}$$

initialized with $\theta_0 = \theta_*$, and no noise.

Following the proof technique of Bach and Moulines (2013), we are going to consider a related recursion by replacing in Equation (B.6) the operator $x_n \otimes x_n$ by its expectation Σ. Thus, we consider η_n defined as

$$\eta_n - \theta_* = \gamma \sum_{k=1}^n (\mathrm{I} - \gamma\Sigma - \lambda\gamma \mathrm{I})^{n-k}\lambda(\theta_0 - \theta_*),$$

which satisfies the recursion (with initialization $\eta_0 = \theta_*$) and

$$\eta_n - \theta_* = [\mathrm{I} - \gamma\Sigma - \lambda\gamma \mathrm{I}](\eta_{n-1} - \theta_*) + \lambda\gamma(\theta_0 - \theta_*).$$

In order to bound $\|\Sigma^{1/2}(\theta_n - \theta_*)\|$, we will independently bound $\|\Sigma^{1/2}(\eta_n - \theta_*)\|$ and $\|\Sigma^{1/2}(\theta_n - \eta_n)\|$ using Minkowski's inequality.

Bounding $\|\Sigma^{1/2}(\theta_n - \eta_n)\|$. We have $\theta_0 - \eta_0 = 0$, and

$$\theta_n - \eta_n = [\mathrm{I} - \gamma x_n \otimes x_n - \lambda\gamma \mathrm{I}](\theta_{n-1} - \eta_{n-1}) + \gamma[\Sigma - x_n \otimes x_n](\eta_{n-1} - \theta_*).$$

We can now bound the recursion for $\theta_n - \eta_n$ as follows, using standard online learning proofs (Nemirovski et al., 2009):

$$\begin{aligned}\|\theta_n - \eta_n\|^2 \leqslant\ & \|\theta_{n-1} - \eta_{n-1}\|^2 - 2\gamma\langle \theta_{n-1} - \eta_{n-1}, (x_n \otimes x_n + \lambda \mathrm{I})(\theta_{n-1} - \eta_{n-1})\rangle \\ & + 2\gamma\langle \theta_{n-1} - \eta_{n-1}, [\Sigma - x_n \otimes x_n](\eta_{n-1} - \theta_*)\rangle \\ & + \gamma^2 \big\| [x_n \otimes x_n + \lambda \mathrm{I}](\theta_{n-1} - \eta_{n-1}) - [\Sigma - x_n \otimes x_n](\eta_{n-1} - \theta_*)\big\|^2.\end{aligned}$$

By taking conditional expectations given $\mathcal{F}_{n-1}$, we get, using first the fact that $\mathbb{E}(\Sigma - x_n \otimes x_n | \mathcal{F}_{n-1}) = 0$ and the inequality $(a+b)^2 \leqslant 2(a^2 + b^2)$, then developing and using $\mathbb{E}[(x_n \otimes x_n)^2] \leqslant R^2\Sigma$, which is assumption $\mathcal{A}_1$.

$$\mathbb{E}(\|\theta_n - \eta_n\|^2 | \mathcal{F}_{n-1}) \leqslant \|\theta_{n-1} - \eta_{n-1}\|^2 - 2\gamma\langle \theta_{n-1} - \eta_{n-1}, (\Sigma + \lambda \mathrm{I})(\theta_{n-1} - \eta_{n-1})\rangle$$

$$
\begin{aligned}
&+ 2\gamma^2 \mathbb{E}(\| [x_n \otimes x_n + \lambda \mathrm{I}](\theta_{n-1} - \eta_{n-1})\|^2 | \mathcal{F}_{n-1}) \\
&+ 2\gamma^2 \mathbb{E}(\| [\Sigma - x_n \otimes x_n](\eta_{n-1} - \theta_*)\|^2 | \mathcal{F}_{n-1}) \\
&\leqslant \|\theta_{n-1} - \eta_{n-1}\|^2 - 2\gamma \langle \theta_{n-1} - \eta_{n-1}, (\Sigma + \lambda \mathrm{I})(\theta_{n-1} - \eta_{n-1}) \rangle \\
&+ 2\gamma^2 \langle \theta_{n-1} - \eta_{n-1}, (R^2 \Sigma + \lambda^2 \mathrm{I} + 2\lambda \Sigma)(\theta_{n-1} - \eta_{n-1}) \rangle \\
&+ 2\gamma^2 R^2 \langle \eta_{n-1} - \theta_*, \Sigma \rangle \\
&\leqslant \|\theta_{n-1} - \eta_{n-1}\|^2 - 2\gamma [1 - \gamma(R^2 + 2\lambda)] \langle \theta_{n-1} - \eta_{n-1}, \Sigma(\theta_{n-1} - \eta_{n-1}) \rangle \\
&+ 2\gamma^2 R^2 \langle \eta_{n-1} - \theta_*, \Sigma(\eta_{n-1} - \theta_*) \rangle.
\end{aligned}
$$

This leads by taking full expectations and moving terms to

$$
\begin{aligned}
\mathbb{E}\langle \theta_{n-1} - \eta_{n-1}, \Sigma(\theta_{n-1} - \eta_{n-1}) \rangle \leqslant & \frac{1}{2\gamma[1 - \gamma(R^2 + 2\lambda)]} [\mathbb{E}\|\theta_{n-1} - \eta_{n-1}\|^2 - \mathbb{E}\|\theta_n - \eta_n\|^2] \\
& + \frac{\gamma R^2}{1 - \gamma(R^2 + 2\lambda)} \langle \eta_{n-1} - \theta_*, \Sigma(\eta_{n-1} - \theta_*) \rangle.
\end{aligned}
$$

Thus, if $\gamma(R^2 + 2\lambda) \leqslant \frac{1}{2}$

$$
\begin{aligned}
\mathbb{E}\langle \theta_{n-1} - \eta_{n-1}, \Sigma(\theta_{n-1} - \eta_{n-1}) \rangle \leqslant & \frac{1}{\gamma} [\mathbb{E}\|\theta_{n-1} - \eta_{n-1}\|^2 - \mathbb{E}\|\theta_n - \eta_n\|^2] \\
& + 2\gamma R^2 \mathbb{E}\langle \eta_{n-1} - \theta_*, \Sigma(\eta_{n-1} - \theta_*) \rangle.
\end{aligned}
$$

This leads to, summing and using initial conditions $\theta_0 - \eta_0 = 0$, then using convexity to upper bound $\langle \bar{\theta}_n - \bar{\eta}_n, \Sigma(\bar{\theta}_n - \bar{\eta}_n) \rangle \leqslant \frac{1}{n+1} \sum_{k=0}^n \langle \theta_k - \eta_k, \Sigma(\theta_k - \eta_k) \rangle$,

$$
\mathbb{E}\langle \bar{\theta}_n - \bar{\eta}_n, \Sigma(\bar{\theta}_n - \bar{\eta}_n) \rangle \leqslant \frac{2\gamma R^2}{n+1} \sum_{k=0}^n \langle \eta_k - \theta_*, \Sigma(\eta_k - \theta_*) \rangle.
$$

Bounding $\|\Sigma^{1/2}(\eta_n - \theta_*)\|$. Moreover we have:

$$
\begin{aligned}
\eta_n - \theta_* =& !\lambda(\Sigma + \lambda \mathrm{I})^{-1}(\theta_0 - \theta_*) - (\mathrm{I} - \gamma\Sigma - \lambda\gamma \mathrm{I})^n [\lambda(\Sigma + \lambda \mathrm{I})^{-1}(\theta_0 - \theta_*)] \\
\bar{\eta}_n - \theta_* =& \lambda(\Sigma + \lambda \mathrm{I})^{-1}(\theta_0 - \theta_*) - \frac{1}{n+1} \sum_{k=0}^n (\mathrm{I} - \gamma\Sigma - \lambda\gamma \mathrm{I})^k [\lambda(\Sigma + \lambda \mathrm{I})^{-1}(\theta_0 - \theta_*)] \\
=& \lambda(\Sigma + \lambda \mathrm{I})^{-1}(\theta_0 - \theta_*) \\
& - \frac{1}{n+1} \gamma^{-1} (\Sigma + \lambda \mathrm{I})^{-1} [\mathrm{I} - (\mathrm{I} - \gamma\Sigma - \lambda\gamma \mathrm{I})^{n+1}] [\lambda(\Sigma + \lambda \mathrm{I})^{-1}(\theta_0 - \theta_*)] \ .
\end{aligned}
$$

This leads using Minkowski inequality to

$$
\begin{aligned}
(\mathbb{E}\|\Sigma^{1/2}(\eta_n - \theta_*)\|^2)^{1/2} \leqslant & \|\lambda \Sigma^{1/2}(\Sigma + \lambda \mathrm{I})^{-1}(\theta_0 - \theta_*)\| \\
(\mathbb{E}\|\Sigma^{1/2}(\bar{\eta}_n - \theta_*)\|^2)^{1/2} \leqslant & \|\lambda \Sigma^{1/2}(\Sigma + \lambda \mathrm{I})^{-1}(\theta_0 - \theta_*)\|.
\end{aligned}
$$

Thus this part is such that

$$
\begin{aligned}
(\mathbb{E}\|\Sigma^{1/2}(\bar{\theta}_n - \theta_*)\|^2)^{1/2} &\leqslant \|\lambda \Sigma^{1/2}(\Sigma + \lambda \mathrm{I})^{-1}(\theta_0 - \theta_*)\| + \sqrt{2\gamma R^2} \|\lambda \Sigma^{1/2}(\Sigma + \lambda \mathrm{I})^{-1}(\theta_0 - \theta_*)\| \\
&\leqslant \|\lambda \Sigma^{1/2}(\Sigma + \lambda \mathrm{I})^{-1}(\theta_0 - \theta_*)\| (1 + \sqrt{2\gamma R^2}),
\end{aligned}
$$

that gives the first bound on the regularization-based bias

$$
\mathbb{E}\|\Sigma^{1/2}(\bar{\theta}_n - \theta_*)\|^2 \leqslant \|\lambda \Sigma^{1/2}(\Sigma + \lambda \mathrm{I})^{-1}(\theta_0 - \theta_*)\|^2 (1 + \sqrt{2\gamma R^2})^2. \tag{B.7}
$$

B.2.3 Expansion without the regularization term

We will follow here the outline of the proof of Györfi and Walk (1996) which considers a full expansion of the function value $\|\Sigma^{1/2}(\bar{\theta}_n - \theta_*)\|^2$. This corresponds to

$$\theta_n - \theta_* = M(n,1)(\theta_0 - \theta_*) + \gamma \sum_{k=1}^{n} M(n,k+1)\varepsilon_k x_k.$$

We have

$$\mathbb{E}\sum_{i=0}^{n}\sum_{j=0}^{n}\langle \theta_i - \theta_*, \Sigma(\theta_j - \theta_*)\rangle = \mathbb{E}\sum_{i=0}^{n}\langle \theta_i - \theta_*, \Sigma(\theta_i - \theta_*)\rangle + 2\mathbb{E}\sum_{i=0}^{n-1}\sum_{j=i+1}^{n}\langle \theta_i - \theta_*, \Sigma(\theta_j - \theta_*)\rangle.$$

Moreover,

$$\begin{aligned}
&\mathbb{E}\sum_{i=0}^{n-1}\sum_{j=i+1}^{n}\langle \theta_i - \theta_*, \Sigma(\theta_j - \theta_*)\rangle \\
&= \mathbb{E}\sum_{i=0}^{n-1}\sum_{j=i+1}^{n}\left\langle \theta_i - \theta_*, \Sigma\Big[M(j,i+1)(\theta_i - \theta_*) + \sum_{k=i+1}^{j} M(j,k+1)\gamma\varepsilon_k x_k\Big]\right\rangle \\
&= \mathbb{E}\sum_{i=0}^{n-1}\sum_{j=i+1}^{n}\langle \theta_i - \theta_*, \Sigma M(j,i+1)(\theta_i - \theta_*)\rangle \text{ because } \varepsilon_k x_k \text{ and } \theta_i \text{ are independent,} \\
&= \mathbb{E}\sum_{i=0}^{n-1}\sum_{j=i+1}^{n}\langle \theta_i - \theta_*, \Sigma(\mathrm{I} - \gamma\Sigma - \gamma\lambda\mathrm{I})^{j-i}(\theta_i - \theta_*)\rangle \text{ as } M(j,i+1) \text{ and } \theta_i \text{ are independent,} \\
&= \mathbb{E}\sum_{i=0}^{n-1}\left\langle \theta_i - \theta_*, \gamma^{-1}\Sigma(\Sigma + \lambda\mathrm{I})^{-1}\big[(\mathrm{I} - \gamma\Sigma - \gamma\lambda\mathrm{I}) - (\mathrm{I} - \gamma\Sigma - \gamma\lambda\mathrm{I})^{n-i+1}\big](\theta_i - \theta_*)\right\rangle \\
&\leqslant \mathbb{E}\sum_{i=0}^{n}\left\langle \theta_i - \theta_*, \gamma^{-1}\Sigma(\Sigma + \lambda\mathrm{I})^{-1}(\mathrm{I} - \gamma\Sigma - \gamma\lambda\mathrm{I})(\theta_i - \theta_*)\right\rangle \text{ using } (\Sigma + \lambda\mathrm{I}) \preccurlyeq \mathrm{I}, \\
&= \gamma^{-1}\mathbb{E}\sum_{i=0}^{n}\langle \theta_i - \theta_*, \Sigma(\Sigma + \lambda\mathrm{I})^{-1}(\theta_i - \theta_*)\rangle - \mathbb{E}\sum_{i=0}^{n}\langle \theta_i - \theta_*, \Sigma(\theta_i - \theta_*)\rangle.
\end{aligned}$$

We thus simply need to bound $\gamma^{-1}\mathbb{E}\sum_{i=0}^{n}\langle \theta_i - \theta_*, \Sigma(\Sigma + \lambda\mathrm{I})^{-1}(\theta_i - \theta_*)\rangle$, to get a bound on $n^2\mathbb{E}\|\Sigma^{1/2}(\bar{\theta}_n - \theta_*)\|^2$.

Recursion on operators. We have:

$$\begin{aligned}
\mathbb{E}\big[M(i,k)\Sigma(\Sigma + \lambda\mathrm{I})^{-1}M(i,k)^*\big] &= \mathbb{E}\Big[M(i,k+1)\big[\mathrm{I} - \gamma x_k \otimes x_k - \gamma\lambda\mathrm{I}\big]\Sigma(\Sigma + \lambda\mathrm{I})^{-1} \\
&\quad \big[\mathrm{I} - \gamma x_k \otimes x_k - \gamma\lambda\mathrm{I}\big]M(i,k+1)^*\Big] \\
&= \mathbb{E}\Big[M(i,k+1)\Big(\Sigma(\Sigma + \lambda\mathrm{I})^{-1} - 2\gamma\Sigma + \gamma^2\big[x_k \otimes x_k \\
&\quad + \lambda\mathrm{I}\big]\Sigma(\Sigma + \lambda\mathrm{I})^{-1}\big[x_k \otimes x_k + \lambda\mathrm{I}\big]\Big)M(i,k+1)^*\Big] \\
&\preccurlyeq \mathbb{E}\Big[M(i,k+1)\big[\Sigma(\Sigma + \lambda\mathrm{I})^{-1} - 2\gamma\Sigma \\
&\quad + \gamma^2(R^2 + 2\lambda)\Sigma\big]M(i,k+1)^*\Big] \\
&= \mathbb{E}\Big[M(i,k+1)\Sigma(\Sigma + \lambda\mathrm{I})^{-1}M(i,k+1)^*\Big] \\
&\quad - \gamma(2 - \gamma(R^2 + 2\lambda))\mathbb{E}\Big[M(i,k+1)\Sigma M(i,k+1)^*\Big],
\end{aligned}$$

which leads to

$$\mathbb{E}\Big[M(i,k+1)\Sigma M(i,k+1)^*\Big] \preccurlyeq \frac{1}{\gamma(2-\gamma(R^2+2\lambda))}\Big(E\Big[M(i,k+1)\Sigma(\Sigma+\lambda\mathrm{I})^{-1}M(i,k+1)^*\Big] - E\Big[M(i,k)\Sigma(\Sigma+\lambda\mathrm{I})^{-1}M(i,k)^*\Big]\Big). \tag{B.8}$$

Using the operator T on matrices defined below, this corresponds to showing

$$(\mathrm{I}-\gamma T)\big[\Sigma(\Sigma+\lambda\mathrm{I})\big] \preccurlyeq \Sigma(\Sigma+\lambda\mathrm{I}) - \gamma\Sigma.$$

Noise term. For $\theta_0 - \theta_* = 0$, we have:

$$\begin{aligned}
& \mathbb{E}\langle \theta_i - \theta_*, \Sigma(\Sigma+\lambda\mathrm{I})^{-1}(\theta_i-\theta_*)\rangle \\
= \;& \gamma^2 \mathbb{E}\sum_{k=1}^{i}\sum_{j=1}^{i} \varepsilon_j x_j^* M(i,j+1)^*\Sigma(\Sigma+\lambda\mathrm{I})^{-1}M(i,k+1)\varepsilon_k x_k \text{ by expanding all terms,} \\
= \;& \gamma^2 \mathbb{E}\sum_{k=1}^{i} \varepsilon_k x_k^* M(i,k+1)^*\Sigma(\Sigma+\lambda\mathrm{I})^{-1}M(i,k+1)\varepsilon_k x_k \text{ using independence,} \\
= \;& \gamma^2 \operatorname{tr}\Big(\sum_{k=1}^{i} \mathbb{E}\varepsilon_k^2 x_k x_k^* \mathbb{E}M(i,k+1)^*\Sigma(\Sigma+\lambda\mathrm{I})^{-1}M(i,k+1)\Big) \\
\leqslant \;& \gamma^2\sigma^2 \operatorname{tr}\Big(\sum_{k=1}^{i} \mathbb{E}M(i,k+1)\Sigma M(i,k+1)^*\Sigma(\Sigma+\lambda\mathrm{I})^{-1}\Big),
\end{aligned}$$

using our assumption regarding the noise. Then using the recurrence between operators

$$\begin{aligned}
& \mathbb{E}\langle \theta_i - \theta_*, \Sigma(\Sigma+\lambda\mathrm{I})^{-1}(\theta_i-\theta_*)\rangle \\
\leqslant \;& \frac{\gamma\sigma^2}{2-\gamma(R^2+2\lambda)} \operatorname{tr}\sum_{k=1}^{i}\Big(E\Big[M(i,k+1)\Sigma(\Sigma+\lambda\mathrm{I})^{-1}M(i,k+1)^*\Sigma(\Sigma+\lambda\mathrm{I})^{-1}\Big] \\
& -E\Big[M(i,k)\Sigma(\Sigma+\lambda\mathrm{I})^{-1}M(i,k)^*\Sigma(\Sigma+\lambda\mathrm{I})^{-1}\Big]\Big) \\
\leqslant \;& \frac{\gamma\sigma^2}{2-\gamma(R^2+2\lambda)} \operatorname{tr}\Big(E\Big[M(i,i+1)\Sigma(\Sigma+\lambda\mathrm{I})^{-1}M(i,i+1)^*\Sigma(\Sigma+\lambda\mathrm{I})^{-1}\Big] \\
& -E\Big[M(i,1)\Sigma(\Sigma+\lambda\mathrm{I})^{-1}M(i,1)^*\Sigma(\Sigma+\lambda\mathrm{I})^{-1}\Big]\Big) \text{ by summing,} \\
\leqslant \;& \frac{\gamma\sigma^2}{2-\gamma(R^2+2\lambda)} \operatorname{tr}\Sigma^2(\Sigma+\lambda\mathrm{I})^{-2}.
\end{aligned}$$

This implies that for the noise process

$$\mathbb{E}\|\Sigma^{1/2}(\bar{\theta}_n-\theta_*)\|^2 \leqslant \Big(\frac{\sigma^2}{n+1}\operatorname{tr}\big[\Sigma^2(\Sigma+\lambda\mathrm{I})^{-2}\big]\Big)\frac{1}{1-\gamma(R^2/2+\lambda)}.$$

Note that when γ tends to zero, we recover the optimal variance term.

Noiseless term. Without noise, we then need to bound:

$$\gamma^{-1}\mathbb{E}\sum_{i=0}^{n}\langle \theta_i-\theta_*, \Sigma(\Sigma+\lambda\mathrm{I})^{-1}(\theta_i-\theta_*)\rangle,$$

with $\theta_i - \theta_* = M(i,1)(\theta_0 - \theta_*)$, that is

$$\gamma^{-1}\mathbb{E}\sum_{i=0}^{n} \operatorname{tr}\left[M(i,1)^*\Sigma(\Sigma+\lambda \mathrm{I})^{-1}M(i,1)(\theta_0-\theta_*)(\theta_0-\theta_*)^*\right].$$

We follow here the proof of Défossez and Bach (2015) and consider the operator T from symmetric matrices to symmetric matrices defined as

$$TA = (\Sigma+\lambda\mathrm{I})A + A(\Sigma+\lambda\mathrm{I}) - \gamma E\big[(x_n\otimes x_n+\lambda\mathrm{I})A(x_n\otimes x_n+\lambda\mathrm{I})\big].$$

of the form $TA = (\Sigma+\lambda\mathrm{I})A + (\Sigma+\lambda\mathrm{I})A - \gamma SA$.

The operator S is self-adjoint and positive. Moreover:

$$\begin{aligned}
\langle A, SA\rangle &= \mathbb{E}\operatorname{tr}\left[A(x_n\otimes x_n+\lambda\mathrm{I})A(x_n\otimes x_n+\lambda\mathrm{I})\right] \\
&= \operatorname{tr}\left[2A^2\lambda\Sigma+\lambda^2A^2\right] + \mathbb{E}\operatorname{tr}\left[\langle x_n, Ax_n\rangle^2\right] \\
&\leqslant \operatorname{tr}\left[2A^2\lambda\Sigma+\lambda^2A^2\right] + \mathbb{E}\operatorname{tr}\left[\|x_n\|^2 x_n\otimes x_n, A^2\right] \text{ with Cauchy-Schwarz ineq.,} \\
&\leqslant \operatorname{tr}\left[2A^2\lambda\Sigma+\lambda^2A^2\right] + R^2\operatorname{tr}\Sigma A^2 \\
&\leqslant (R^2+2\lambda)\operatorname{tr}\left[\Sigma+\lambda\mathrm{I}\right]A^2.
\end{aligned}$$

We have for any symmetric matrix A:

$$\mathbb{E}M(i,1)^*AM(i,1) = (\mathrm{I}-\gamma T)^i A.$$

Thus,

$$\gamma^{-1}\mathbb{E}\sum_{i=0}^{n}\operatorname{tr}\left[M(i,1)^*\Sigma(\Sigma+\lambda\mathrm{I})^{-1}M(i,1)(\theta_0-\theta_*)(\theta_0-\theta_*)^*\right] = \gamma^{-1}\mathbb{E}\sum_{i=0}^{n}\langle\langle(\mathrm{I}-\gamma T)^iA, E_0\rangle\rangle$$

with $E_0 = (\theta_0-\theta_*)(\theta_0-\theta_*)^*$ and $A = \Sigma(\Sigma+\lambda\mathrm{I})^{-1}$. This leads to

$$\gamma^{-1}\mathbb{E}\langle\langle\gamma^{-1}T^{-1}(\mathrm{I}-(\mathrm{I}-\gamma T)^{n+1})A, E_0\rangle\rangle,$$

where $\langle\langle\cdot,\cdot\rangle\rangle$ denote the dot-product between self-adjoint operators.

The sum is less than its limit for $n\to\infty$, and thus, we can get rid of the term $(\mathrm{I}-\gamma T)^{n+1}$, and we need to bound

$$\gamma^{-2}\langle\langle M, E_0\rangle\rangle = \gamma^{-2}\langle\langle T^{-1}(\Sigma(\Sigma+\lambda\mathrm{I})^{-1}), E_0\rangle\rangle,$$

with $M := T^{-1}\left[\Sigma(\Sigma+\lambda\mathrm{I})^{-1}\right]$, *i.e.*, such that

$$\begin{aligned}
\Sigma(\Sigma+\lambda\mathrm{I})^{-1} &= (\Sigma+\lambda\mathrm{I})M + M(\Sigma+\lambda\mathrm{I}) - \gamma\mathbb{E}(x_n\otimes x_n+\lambda\mathrm{I})M(x_n\otimes x_n+\lambda\mathrm{I}) \\
&= (\Sigma+\lambda\mathrm{I})M + M(\Sigma+\lambda\mathrm{I}) - \gamma SM. \qquad \text{(B.9)}
\end{aligned}$$

So that:

$$\begin{aligned}
M &= \left[(\Sigma+\lambda\mathrm{I})\otimes\mathrm{I} + \mathrm{I}\otimes(\Sigma+\lambda\mathrm{I})\right]^{-1}\left[\Sigma(\Sigma+\lambda\mathrm{I})^{-1}\right] + \gamma\left[(\Sigma+\lambda\mathrm{I})\otimes\mathrm{I}+\mathrm{I}\otimes(\Sigma+\lambda\mathrm{I})\right]^{-1}SM \\
&= \frac{1}{2}\Sigma(\Sigma+\lambda\mathrm{I})^{-2} + \gamma\left[(\Sigma+\lambda\mathrm{I})\otimes\mathrm{I}+\mathrm{I}\otimes(\Sigma+\lambda\mathrm{I})\right]^{-1}SM.
\end{aligned}$$

The operator $(\Sigma+\lambda\mathrm{I})\otimes\mathrm{I}+\mathrm{I}\otimes(\Sigma+\lambda\mathrm{I})$ is self adjoint, and so is its inverse, thus:

$$\gamma^{-2}\langle\langle M, E_0\rangle\rangle = \gamma^{-2}\langle\langle\frac{1}{2}\Sigma(\Sigma+\lambda\mathrm{I})^{-2} + \gamma\left[(\Sigma+\lambda\mathrm{I})\otimes\mathrm{I}+\mathrm{I}\otimes(\Sigma+\lambda\mathrm{I})\right]^{-1}SM, E_0\rangle\rangle$$

$$= \frac{1}{2\gamma^2}\langle\langle \Sigma(\Sigma+\lambda \mathrm{I})^{-2}, E_0\rangle\rangle + \gamma^{-1}\langle\langle SM, \left[(\Sigma+\lambda \mathrm{I})\otimes \mathrm{I} + \mathrm{I}\otimes(\Sigma+\lambda \mathrm{I})\right]^{-1} E_0\rangle\rangle$$
$$= \frac{1}{2\gamma^2}\operatorname{tr}(\Sigma(\Sigma+\lambda \mathrm{I})^{-2}E_0) + \gamma^{-1}\langle\langle SM, \left[(\Sigma+\lambda \mathrm{I})\otimes \mathrm{I} + \mathrm{I}\otimes(\Sigma+\lambda \mathrm{I})\right]^{-1} E_0\rangle\rangle.$$

Moreover,

$$\begin{aligned} E_0 &= (\theta_0-\theta_*)(\theta_0-\theta_*)^* \\ &= (\Sigma+\lambda \mathrm{I})^{1/2}(\Sigma+\lambda \mathrm{I})^{-1/2}(\theta_0-\theta_*)(\theta_0-\theta_*)^*(\Sigma+\lambda \mathrm{I})^{-1/2}(\Sigma+\lambda \mathrm{I})^{+1/2} \\ &\preccurlyeq [(\theta_0-\theta_*)^*(\Sigma+\lambda \mathrm{I})^{-1}(\theta_0-\theta_*)]\ (\Sigma+\lambda \mathrm{I}), \\ &\quad \text{as } (\Sigma+\lambda \mathrm{I})^{-1/2}(\theta_0-\theta_*)(\theta_0-\theta_*)^*(\Sigma+\lambda \mathrm{I})^{-1/2} \preccurlyeq (\theta_0-\theta_*)^*(\Sigma+\lambda \mathrm{I})^{-1}(\theta_0-\theta_*)I. \end{aligned}$$

Thus, as $\left[(\Sigma+\lambda \mathrm{I})\otimes \mathrm{I} + \mathrm{I}\otimes(\Sigma+\lambda \mathrm{I})\right]^{-1}$ is an non-decreasing operator on $(S_n(\mathbb{R}), \preccurlyeq)$ (see technical Lemma B.7 in Section B.5):

$$\begin{aligned} &\left[(\Sigma+\lambda \mathrm{I})\otimes \mathrm{I} + \mathrm{I}\otimes(\Sigma+\lambda \mathrm{I})\right]^{-1} E_0 \\ \preccurlyeq\ &\left[(\Sigma+\lambda \mathrm{I})\otimes \mathrm{I} + \mathrm{I}\otimes(\Sigma+\lambda \mathrm{I})\right]^{-1} \Big([(\theta_0-\theta_*)^*(\Sigma+\lambda \mathrm{I})^{-1}(\theta_0-\theta_*)](\Sigma+\lambda \mathrm{I})\Big) \\ =\ &\frac{(\theta_0-\theta_*)^*(\Sigma+\lambda \mathrm{I})^{-1}(\theta_0-\theta_*)}{2} I. \end{aligned}$$

Thus as SM is positive:

$$\gamma^{-2}\langle\langle M, E_0\rangle\rangle \leqslant \frac{1}{2\gamma^2}\operatorname{tr}(\Sigma(\Sigma+\lambda \mathrm{I})^{-2}E_0) + \frac{(\theta_0-\theta_*)^*(\Sigma+\lambda \mathrm{I})^{-1}(\theta_0-\theta_*)}{2\gamma}\operatorname{tr}(SM).$$

Moreover we can upper bound $\operatorname{tr}(SM)$: using Equation (B.9) we have

$$\operatorname{tr}(\Sigma(\Sigma+\lambda \mathrm{I})^{-1}) = 2\operatorname{tr}(\Sigma+\lambda \mathrm{I})M - \gamma \operatorname{tr}\mathbb{E}(x_n\otimes x_n+\lambda \mathrm{I})M(x_n\otimes x_n+\lambda \mathrm{I})$$

then, using Assumption ($\mathcal{A}_1$):

$$\operatorname{tr}\mathbb{E}(x_n\otimes x_n+\lambda \mathrm{I})M(x_n\otimes x_n+\lambda \mathrm{I}) \leqslant R^2 \operatorname{tr} M\Sigma + 2\operatorname{tr} M\Sigma\lambda + \lambda^2 \operatorname{tr} M \leqslant (R^2+2\lambda)\operatorname{tr} M(\Sigma+\lambda \mathrm{I}).$$

This implies

$$\begin{aligned} \operatorname{tr}\left[\Sigma(\Sigma+\lambda \mathrm{I})^{-1}\right] &\geqslant \Big(\frac{2}{R^2+2\lambda} - \gamma\Big)\operatorname{tr}\mathbb{E}(x_n\otimes x_n+\lambda \mathrm{I})M(x_n\otimes x_n+\lambda \mathrm{I}), \\ &\geqslant \frac{1}{R^2+2\lambda}\operatorname{tr}\mathbb{E}(x_n\otimes x_n+\lambda \mathrm{I})M(x_n\otimes x_n+\lambda \mathrm{I}) \text{ since } \gamma(R^2+2\lambda)\leqslant 1, \\ &\geqslant \frac{1}{R^2+2\lambda}\operatorname{tr} SM. \end{aligned}$$

Thus finally:

$$\begin{aligned} \gamma^{-2}\langle\langle M, E_0\rangle\rangle \leqslant\ &\frac{1}{2\gamma^2}\operatorname{tr} E_0\Sigma(\Sigma+\lambda \mathrm{I})^{-2} \\ &+ \frac{(\theta_0-\theta_*)^*(\Sigma+\lambda \mathrm{I})^{-1}(\theta_0-\theta_*)}{2\gamma}(R^2+2\lambda)\operatorname{tr}(\Sigma(\Sigma+\lambda \mathrm{I})^{-1}), \end{aligned}$$

which leads to the desired error term.

B.2.4 Proof when only $\|\theta_0 - \theta_*\|$ is finite

When $\lambda = 0$, without noise, we then need to bound:

$$\gamma^{-1}\mathbb{E}\sum_{i=0}^{n}\langle\theta_i - \theta_*, \theta_i - \theta_*\rangle,$$

with $\theta_i - \theta_* = M(i,1)(\theta_0 - \theta_*)$, that is

$$\gamma^{-1}\mathbb{E}\sum_{i=0}^{n}\operatorname{tr}\Big[M(i,1)^*M(i,1)(\theta_0 - \theta_*)(\theta_0 - \theta_*)^*\Big].$$

By definition of $M(i,1)$ we have that $\mathbb{E}M(i,1)^*M(i,1) \preccurlyeq \mathrm{I}$ leading to

$$\gamma^{-1}\mathbb{E}\sum_{i=0}^{n}\langle\theta_i - \theta_*, \theta_i - \theta_*\rangle \leqslant \frac{(n+1)\|\theta_0 - \theta_*\|^2}{\gamma}.$$

For the regularization-based bias we also have

$$\|\lambda\Sigma^{1/2}(\Sigma + \lambda\mathrm{I})^{-1}(\theta_0 - \theta_*)\|^2 \leqslant \lambda\|\Sigma^{1/2}(\Sigma + \lambda\mathrm{I})^{-1/2}(\theta_0 - \theta_*)\|^2.$$

B.2.5 Proof when the noise is not structured

For $\|\theta_0 - \theta_*\| = 0$ we have $\theta_n - \theta_* = \gamma\sum_{k=1}^{n} M(n,k+1)\varepsilon_k x_k$ which leads to

$$\mathbb{E}\|\Sigma^{1/2}(\theta_n - \theta_*)\|^2 = \gamma^2\sum_{k=1}^{n}\operatorname{tr}\mathbb{E}M(n,k+1)^*\Sigma M(n,k+1)V,$$

where $V = \mathbb{E}\varepsilon_k^2 x_k x_k^*$. And using the recursion on operators in Equation (B.8) by changing order of elements we have

$$\begin{aligned}\mathbb{E}\Big[M(n,k+1)^*\Sigma M(n,k+1)\Big] \preccurlyeq \frac{1}{\gamma(2-\gamma(R^2+2\lambda))}\Big(&E\Big[M(n,k+1)^*\Sigma(\Sigma+\lambda\mathrm{I})^{-1}M(n,k+1)\Big]\\ &- E\Big[M(n,k)^*\Sigma(\Sigma+\lambda\mathrm{I})^{-1}M(n,k)\Big]\Big).\end{aligned}$$

And by adding the terms

$$\mathbb{E}\|\Sigma^{1/2}(\theta_n - \theta_*)\|^2 \preccurlyeq \frac{\gamma^2}{\gamma(2-\gamma(R^2+2\lambda))}\operatorname{tr}\Sigma(\Sigma+\lambda\mathrm{I})^{-1}V,$$

We conclude by convexity

$$\mathbb{E}\|\Sigma^{1/2}(\bar{\theta}_n - \theta_*)\|^2 \preccurlyeq \frac{\gamma^2}{\gamma(2-\gamma(R^2+2\lambda))}\operatorname{tr}\Sigma(\Sigma+\lambda\mathrm{I})^{-1}V.$$

B.3 Convergence of Accelerated Averaged Stochastic Gradient Descent

We now prove Theorem 3.3. We thus consider iterates satisfying Equation (3.10), under Assumptions ($\mathcal{A}_3$), ($\mathcal{A}_4$). We consider a fixed step size γ such that $\gamma(\Sigma + \lambda I) \preccurlyeq I$. Seeing Equation (3.10) as a linear second order for θ_n, we will derive from exact calculations a decomposition of the errors a sum of three terms that will be studied independently. The proof is organized as follows: in Section B.3.1, we state the formulation as a second order linear system and derive the three main terms that have to be studied (see

Lemma B.1). Section B.3.2 studies asymptotic behaviors of the three terms, ignoring some exponentially decreasing terms, in order to give insight of how they behave. This section is not necessary for the proof, indeed a direct and exact calculation in the eigenbasis of Σ, following O'Donoghue and Candès (2013), is provided in Section B.3.3. Results are summed up in Section B.3.4.

B.3.1 General expansion

We study the regularized stochastic accelerated gradient descent recursion defined for $n \geqslant 1$ by

$$\begin{aligned} \theta_n &= \nu_{n-1} - \gamma f'(\nu_{n-1}) - \gamma\lambda(\nu_n - \theta_0) + \gamma\xi_n \\ \nu_n &= \theta_n + \delta(\theta_n - \theta_{n-1}), \end{aligned}$$

starting from $\theta_0 = \nu_0 \in \mathcal{H}$. We may rewrite it for a quadratic function $f : \theta \mapsto \frac{1}{2}\langle \theta - \theta_*, \Sigma(\theta - \theta_*)\rangle$ for $n \geqslant 2$ as

$$\theta_n = \big[\mathrm{I} - \gamma\Sigma - \gamma\lambda\mathrm{I}\big]\big[\theta_{n-1} + \delta(\theta_{n-1} - \theta_{n-2})\big] + \gamma\xi_n + \gamma\lambda\theta_0 + \gamma\Sigma\theta_*,$$

with $\theta_0 \in \mathcal{H}$ and $\theta_1 = \big[\mathrm{I} - \gamma\Sigma - \gamma\lambda\mathrm{I}\big]\theta_0 + \gamma\xi_1 + \gamma\lambda\theta_0 + \gamma\Sigma\theta_*$.

And by centering around the optimum, we get:

$$\theta_n - \theta_* = \big[\mathrm{I} - \gamma\Sigma - \gamma\lambda\mathrm{I}\big]\big[\theta_{n-1} - \theta_* + \delta(\theta_{n-1} - \theta_* - \theta_{n-2} + \theta_*)\big] + \gamma\xi_n + \lambda\gamma(\theta_0 - \theta_*).$$

Thus this is a second order iterative system which is standard to cast in a linear form

$$\Theta_n = F\Theta_{n-1} + \gamma\Xi_n + \gamma\lambda\Theta_\lambda, \tag{B.10}$$

with $T = \mathrm{I} - \gamma\Sigma - \gamma\lambda\mathrm{I}$, $F = \begin{pmatrix} (1+\delta)T & -\delta T \\ I & 0 \end{pmatrix}$, $\Theta_n = \begin{pmatrix} \theta_n - \theta_* \\ \theta_{n-1} - \theta_* \end{pmatrix}$, $\Theta_0 = \begin{pmatrix} \theta_0 - \theta_* \\ \theta_0 - \theta_* \end{pmatrix}$, $\Xi_n = \begin{pmatrix} \xi_n \\ 0 \end{pmatrix}$ and $\Theta_\lambda = \begin{pmatrix} \theta_0 - \theta_* \\ 0 \end{pmatrix}$.

We are interested in the behavior of the average $\bar{\Theta}_n = \frac{1}{n+1}\sum_{k=0}^n \Theta_k$ for which we have the following general convergence result:

Lemma B.1. *For all $\lambda \in \mathbb{R}_+$ and γ such that $\gamma(\Sigma + \lambda\mathrm{I}) \preccurlyeq \mathrm{I}$ and any matrix C the average of the iterates Θ_n defined by Equation (B.10) satisfy for $P_k \overset{(def)}{=} C^{1/2}(I - F^k)(I - F)^{-1}$, with $\tilde{\Theta}_0 = \Theta_0 - \gamma\lambda(I - F)^{-1}\Theta_\lambda$,*

$$\begin{aligned} \mathbb{E}\langle \bar{\Theta}_n, C\bar{\Theta}_n\rangle \leqslant\ & 2\,(\gamma\lambda)^2\,\|C^{1/2}(I - F)^{-1}\Theta_\lambda\|^2 + \frac{2}{(n+1)^2}\|P_{n+1}\tilde{\Theta}_0\|^2 \\ & + \frac{\gamma^2}{(n+1)^2}\sum_{j=1}^n \operatorname{tr} P_j V P_j^\top. \end{aligned}$$

The error thus decomposes as the sum of three main terms:

- the two first ones are bias terms, one arising from the regularization (the first one), and one arising computation (the second one),
- a variance term. which is the last one.

We remark that as we have assumed that Σ is invertible, the matrix $I - F$ can be shown to be invertible for all the considered δ.

The regularization-based term will be studied directly whereas the two others will be studied in two stages. First a heuristic will lead to an asymptotic bound then an exact computation will give a non-asymptotic bound. Then using $C = H = \begin{pmatrix} \Sigma & 0 \\ 0 & 0 \end{pmatrix}$ would give a convergence result on the function value and $C = \begin{pmatrix} \mathrm{I} & 0 \\ 0 & 0 \end{pmatrix}$ a result on the iterate. The end of the section is devoted to the proof of this lemma.

Proof. The sequence Θ_n satisfies a linear recursion, from which we get, for all $n \geqslant 1$:

$$\begin{aligned} \Theta_n &= F^n\Theta_0 + \gamma\sum_{k=1}^{n} F^{n-k}\Xi_k + \gamma\lambda\sum_{k=1}^{n} F^{n-k}\Theta_\lambda \\ &= F^n\Theta_0 + \gamma\sum_{k=1}^{n} F^{n-k}\Xi_k + \gamma\lambda(I-F^n)(I-F)^{-1}\Theta_\lambda. \end{aligned}$$

We study the averaged sequence: $\bar{\Theta}_n = \frac{1}{n+1}\sum_{k=0}^{n}\Theta_k$. Using the identity $\sum_{k=0}^{n-1} F^k = (I-F^n)(I-F)^{-1}$, we get

$$\bar{\Theta}_n = \frac{1}{n+1}\sum_{k=0}^{n} F^k\Theta_0 + \frac{\gamma}{n+1}\sum_{k=1}^{n}\sum_{j=1}^{k} F^{k-j}\Xi_j + \frac{\gamma\lambda}{n+1}\sum_{k=1}^{n}(I-F^k)(I-F)^{-1}\Theta_\lambda.$$

With

$$\tilde{\Theta}_0 = \Theta_0 - \gamma\lambda(I-F)^{-1}\Theta_\lambda,$$

and $\sum_{k=1}^{n}(I-F^k) = \sum_{k=0}^{n}(I-F^k) = [n+1-(I-F^{n+1})(I-F)^{-1}]$.

Using summation formulas for geometric series, we derive:

$$\begin{aligned} \bar{\Theta}_n &= \frac{1}{n+1}(I-F^{n+1})(I-F)^{-1}\tilde{\Theta}_0 + \frac{\gamma}{n+1}\sum_{k=1}^{n}\sum_{j=1}^{k} F^{k-j}\Xi_j + \gamma\lambda(I-F)^{-1}\Theta_\lambda \\ &= \frac{1}{n+1}(I-F^{n+1})(I-F)^{-1}\tilde{\Theta}_0 + \frac{\gamma}{n+1}\sum_{j=1}^{n}(\sum_{k=j}^{n} F^{k-j})\Xi_j + \gamma\lambda(I-F)^{-1}\Theta_\lambda \\ &= \frac{1}{n+1}(I-F^{n+1})(I-F)^{-1}\tilde{\Theta}_0 + \frac{\gamma}{n+1}\sum_{j=1}^{n}(\sum_{k=0}^{n-j} F^{k})\Xi_j + \gamma\lambda(I-F)^{-1}\Theta_\lambda \\ &= \frac{1}{n+1}(I-F^{n+1})(I-F)^{-1}\tilde{\Theta}_0 + \frac{\gamma}{n+1}\sum_{j=1}^{n}(I-F^{n+1-j})(I-F)^{-1}\Xi_j + \gamma\lambda(I-F)^{-1}\Theta_\lambda \\ &= \frac{1}{n+1}(I-F^{n+1})(I-F)^{-1}\tilde{\Theta}_0 + \frac{\gamma}{n+1}\sum_{j=1}^{n}(I-F^{j})(I-F)^{-1}\Xi_{n+1-j} + \gamma\lambda(I-F)^{-1}\Theta_\lambda. \end{aligned}$$

Using martingale square moment inequalities which amount to consider $\Xi_i, i = 1, ..., n$ independent, so that the variance of the sum is the sum of variances, and denoting by $V = \mathbb{E}[\Xi_n \otimes \Xi_n]$ we have for any positive semi-definite C,

$$\mathbb{E}\langle\bar{\Theta}_n, C\bar{\Theta}_n\rangle = \left\| C^{1/2}\left(\frac{1}{n+1}(I-F^{n+1})(I-F)^{-1}\tilde{\Theta}_0 + \gamma\lambda(I-F)^{-1}\Theta_\lambda\right)\right\|^2$$

$$+\frac{\gamma^2}{(n+1)^2}\sum_{j=1}^{n}\operatorname{tr}(I-F^j)(I-F)^{-1}V(I-F^\top)^{-1}(I-F^j)^\top C,$$

where $C^{1/2}$ denotes a symmetric square root of C. Define $P_k \overset{(def)}{=} C^{1/2}(I-F^k)(I-F)^{-1}$, we have, Using Minkowski's inequality and inequality $(a+b)^2 \leqslant 2(a^2+b^2)$ for any $a, b \in \mathbb{R}$,

$$\begin{aligned}\mathbb{E}\langle\bar{\Theta}_n, C\bar{\Theta}_n\rangle &= \left\|\frac{1}{n+1}P_{n+1}\tilde{\Theta}_0 + \gamma\lambda C^{1/2}(I-F)^{-1}\Theta_\lambda\right\|^2 + \frac{\gamma^2}{(n+1)^2}\sum_{j=1}^{n}\operatorname{tr}P_jVP_j^\top \\ &\leqslant 2\left(\gamma\lambda\right)^2\|C^{1/2}(I-F)^{-1}\Theta_\lambda\|^2 + \frac{2\|P_{n+1}\tilde{\Theta}_0\|^2}{(n+1)^2} + \frac{\gamma^2}{(n+1)^2}\sum_{j=1}^{n}\operatorname{tr}P_jVP_j^\top.\end{aligned}$$

This concludes proof of Lemma B.1. □

B.3.2 Asymptotic expansion

To give the main terms that we expect, we first provide an asymptotic analysis, which shall only be understood as an insight and is not necessary for the proof. Operator F will have only eigenvalues smaller than 1, thus $|||F^j|||$ will decrease exponentially to 0 as $j \to \infty$ (even if $|||F|||$[2] might be bigger than 1). The asymptotic analysis relies on ignoring all terms in which F^j appears. We thus approximately have:

$$\begin{aligned}\mathbb{E}\langle\bar{\Theta}_n, C\bar{\Theta}_n\rangle &\leqslant 2\left(\gamma\lambda\right)^2\|C^{1/2}(I-F)^{-1}\Theta_\lambda\|^2 + 2\left\|C^{1/2}\frac{1}{n+1}(I-F^{n+1})(I-F)^{-1}\tilde{\Theta}_0\right\|^2 \\ &\quad+\frac{\gamma^2}{(n+1)^2}\sum_{j=1}^{n}\operatorname{tr}(I-F^j)(I-F)^{-1}V(I-F^\top)^{-1}(I-F^j)^\top C \\ &\approx 2\left(\gamma\lambda\right)^2\|C^{1/2}(I-F)^{-1}\Theta_\lambda\|^2 + 2\left\|C^{1/2}\frac{1}{n+1}(I-F)^{-1}\tilde{\Theta}_0\right\|^2 \\ &\quad+\frac{\gamma^2}{(n+1)^2}\sum_{j=1}^{n}\operatorname{tr}(I-F)^{-1}V(I-F^\top)^{-1}C,\end{aligned}$$

where, as it has been explained $\approx$ stands for an equality up to terms that will decay exponentially. However, these terms have to be studied very carefully, what will be done in the Section B.3.3.

Using the matrix inversion lemma we have for $C = \begin{pmatrix} c & 0 \\ 0 & 0 \end{pmatrix}$,

$$\begin{aligned}\mathrm{I}-F &= \begin{pmatrix}(1+\delta)(\gamma\Sigma+\gamma\lambda\mathrm{I})-\delta\mathrm{I} & \delta(\mathrm{I}-(\gamma\Sigma+\gamma\lambda\mathrm{I})) \\ -\mathrm{I} & \mathrm{I}\end{pmatrix} \\ (\mathrm{I}-F)^{-1} &= \begin{pmatrix}(\gamma\Sigma+\gamma\lambda\mathrm{I})^{-1} & \delta(\mathrm{I}-(\gamma\Sigma+\gamma\lambda\mathrm{I})^{-1}) \\ (\gamma\Sigma+\gamma\lambda\mathrm{I})^{-1} & (1+\delta)\mathrm{I}-\delta(\gamma\Sigma+\gamma\lambda\mathrm{I})^{-1}\end{pmatrix} \\ C^{1/2}(\mathrm{I}-F)^{-1} &= \begin{pmatrix}c^{1/2}(\gamma\Sigma+\gamma\lambda\mathrm{I})^{-1} & \delta c^{1/2}(\mathrm{I}-(\gamma\Sigma+\gamma\lambda\mathrm{I})^{-1}) \\ 0 & 0\end{pmatrix}.\end{aligned} \tag{B.11}$$

Regularization based term. This gives for the regularization based term

$$\left\|C^{1/2}(\mathrm{I}-F)^{-1}\Theta_\lambda\right\|^2 = \left\|\begin{pmatrix}c^{1/2}(\gamma\Sigma+\gamma\lambda\mathrm{I})^{-1} & \delta c^{1/2}(\mathrm{I}-(\gamma\Sigma+\gamma\lambda\mathrm{I})^{-1}) \\ 0 & 0\end{pmatrix}\begin{pmatrix}\theta_0-\theta_* \\ 0\end{pmatrix}\right\|^2$$

[2] $|||F|||$ denotes the operator norm of F, *i.e.*, $\sup_{\|x\|\leqslant 1}\|Fx\|$.

$$= \left(\frac{1}{\gamma}\right)^2 \|(c^{1/2}(\Sigma+\lambda I)^{-1}(\theta_0-\theta_*))\|^2. \tag{B.12}$$

The computation of this term is exact (not asymptotic).

Bias term. For the bias term we have

$$\begin{aligned}
\tilde{\Theta}_0 &= \Theta_0 - \gamma\lambda(I-F)^{-1}\Theta_\lambda \\
&= \begin{pmatrix}\theta_0-\theta_*\\ \theta_0-\theta_*\end{pmatrix} - \gamma\lambda \begin{pmatrix}(\gamma\Sigma+\gamma\lambda\mathrm{I})^{-1} & \delta(\mathrm{I}-(\gamma\Sigma+\gamma\lambda\mathrm{I})^{-1})\\ (\gamma\Sigma+\gamma\lambda\mathrm{I})^{-1} & (1+\delta)\mathrm{I}-\delta(\gamma\Sigma+\gamma\lambda\mathrm{I})^{-1}\end{pmatrix}\begin{pmatrix}\theta_0-\theta_*\\ 0\end{pmatrix} \\
&= \begin{pmatrix}\theta_0-\theta_*\\ \theta_0-\theta_*\end{pmatrix} - \gamma\lambda \begin{pmatrix}(\gamma\Sigma+\gamma\lambda\mathrm{I})^{-1}(\theta_0-\theta_*)\\ (\gamma\Sigma+\gamma\lambda\mathrm{I})^{-1}(\theta_0-\theta_*)\end{pmatrix} \\
&= \begin{pmatrix}[\mathrm{I}-\lambda(\Sigma+\lambda\mathrm{I})^{-1}](\theta_0-\theta_*)\\ [\mathrm{I}-\lambda(\Sigma+\lambda\mathrm{I})^{-1}](\theta_0-\theta_*)\end{pmatrix}.
\end{aligned}$$

Thus this gives for the dominant term

$$\begin{aligned}
\left\|C^{1/2}(\mathrm{I}-F)^{-1}\tilde{\Theta}_0\right\|^2 &= \left\| \begin{pmatrix} c^{1/2}(\gamma\Sigma+\gamma\lambda\mathrm{I})^{-1} & \delta c^{1/2}(\mathrm{I}-(\gamma\Sigma+\gamma\lambda\mathrm{I})^{-1})\\ 0 & 0\end{pmatrix}\tilde{\Theta}_0\right\|^2 \\
&= \|(c^{1/2}[(1-\delta)(\gamma\Sigma+\gamma\lambda\mathrm{I})^{-1}+\delta\mathrm{I}][\mathrm{I}-\lambda(\Sigma+\lambda\mathrm{I})^{-1}](\theta_0-\theta_*)\|^2.
\end{aligned}$$

And if c commutes with Σ we have the bound for $\delta \in [\frac{1-\sqrt{\gamma\lambda}}{1+\sqrt{\gamma\lambda}}, 1]$

$$\begin{aligned}
\left\|C^{1/2}(\mathrm{I}-F)^{-1}\tilde{\Theta}_0\right\|^2 &\leqslant (\frac{(1-\delta)}{\gamma\lambda}+\delta)\|(c^{1/2}[\mathrm{I}-\lambda(\Sigma+\lambda\mathrm{I})^{-1}](\theta_0-\theta_*)\|^2 \\
&\leqslant (\frac{2}{\sqrt{\gamma\lambda}}+1)\|(c^{1/2}[\mathrm{I}-\lambda(\Sigma+\lambda\mathrm{I})^{-1}](\theta_0-\theta_*)\|^2.
\end{aligned}$$

Variance term. And for the variance term with $V = \begin{pmatrix} v & 0\\ 0 & 0\end{pmatrix}$, we have $C^{1/2}(\mathrm{I}-F)^{-1}V^{1/2} = \begin{pmatrix} c^{1/2}(\gamma\Sigma+\gamma\lambda\mathrm{I})^{-1}v^{1/2} & 0\\ 0 & 0\end{pmatrix}$, and

$$\operatorname{tr} C^{1/2}(I-F)^{-1}V(I-F^\top)^{-1}C^{1/2} = \operatorname{tr} c(\gamma\Sigma+\gamma\lambda\mathrm{I})^{-1}v(\gamma\Sigma+\gamma\lambda\mathrm{I})^{-1}.$$

This gives the three dominant terms. However in order to control the remainders we have to compute the eigenvalues more carefully, as done in the next section.

B.3.3 Direct computation without the regularization based term

We derive now direct computation both the bias and variance terms. This is not required for the regularization based term whose previous expression in Equation (B.12) is already non-asymptotic. Following O'Donoghue and Candès (2013) we consider an eigen-decomposition of the matrix F, in order to study independently the recursion on eigenspaces. We assume Σ has eigenvalues (s_i) and we decompose vectors in an eigenvector basis of Σ we denote by (p_i), with $\theta_n^i = p_i^\top \theta_n$ and $\xi_n^i = p_i^\top \xi_n$ and we have the reduced equation:

$$\Theta_{n+1}^i = F_i\Theta_n^i + \gamma\Xi_{n+1}^i.$$

with $\Theta_0^i = \tilde{\Theta}_0^i$, $F_i = \begin{pmatrix}(1+\delta)T_i & -\delta T_i\\ 1 & 0\end{pmatrix}$, with $T_i = 1-\gamma s_i - \gamma\lambda$.

Computing initial point $\tilde{\Theta}_0^i$. $\tilde{\Theta}_0^i = \Theta_0^i - \gamma\lambda(I - F_i)^{-1}\Theta_\lambda^i$, with $\Theta_0^i = \begin{pmatrix}\theta_0^i - \theta_*^i \\ \theta_0^i - \theta_*^i\end{pmatrix}$, $\Theta_\lambda^i = \begin{pmatrix}\theta_0^i - \theta_*^i \\ 0\end{pmatrix}$ and $(I - F_i)^{-1}$ given in Equation (B.11). Thus

$$\begin{aligned}\tilde{\Theta}_0^i &= \begin{pmatrix}\theta_0^i - \theta_*^i \\ \theta_0^i - \theta_*^i\end{pmatrix} - \frac{\gamma\lambda}{(\gamma s_i + \gamma\lambda)}\begin{pmatrix}1 & \delta((\gamma s_i + \gamma\lambda) - 1) \\ 1 & (1+\delta)(\gamma s_i + \gamma\lambda) - \delta\end{pmatrix}\begin{pmatrix}\theta_0^i - \theta_*^i \\ 0\end{pmatrix} \\ &= \begin{pmatrix}(1 - \frac{\lambda}{\lambda + s_i})(\theta_0^i - \theta_*^i) \\ (1 - \frac{\lambda}{\lambda + s_i})(\theta_0^i - \theta_*^i)\end{pmatrix}. \end{aligned} \tag{B.13}$$

Study of spectrum of F_i. Depending on δ, F_i may have two distinct complex eigenvalues of same modulus, only one (double) eigenvalue, or two real eigenvalues. We only consider the two former cases, which we detail below.

Indeed, the characteristic polynomial

$$\chi_{F_i}(X) \overset{def}{=} \det(X\mathrm{I} - F_i) = X^2 - (1+\delta)(1 - \gamma(s_i + \lambda))X + \delta(1 - \gamma(s_i + \lambda))$$

has discriminant $\Delta_i = (1 - \gamma(s_i + \lambda))((1+\delta)^2(1 - \gamma(s_i + \lambda)) - 4\delta)$ which is non positive as far as $\delta \in [\delta_-; \delta_+]$, with $\delta_- = \frac{1 - \sqrt{\gamma(s_i + \lambda)}}{1 + \sqrt{\gamma(s_i + \lambda)}}$, $\delta_+ = \frac{1 + \sqrt{\gamma(s_i + \lambda)}}{1 - \sqrt{\gamma(s_i + \lambda)}}$.

Two distinct eigenvalues

We first assume that F_i has two distinct complex eigenvalues $r_\pm = \frac{(1+\delta)(1-\gamma(s_i+\lambda)) \pm \sqrt{-1}\sqrt{-\Delta_i}}{2}$ which are conjugate. Thus the roots are of the form $\rho_i e^{\pm i\omega_i}$ with $\rho_i = \sqrt{\delta(1 - \gamma(s_i + \lambda))}$, $\cos(\omega_i) = \frac{(1+\delta)(1-\gamma(s_i+\lambda))}{2\rho_i}$, $\omega_i \in [-\pi/2; \pi/2]$ and $\sin(\omega_i) = \frac{\sqrt{-\Delta_i}}{2\rho_i}$.

Let $Q_i = \begin{pmatrix}r_i^- & r_i^+ \\ 1 & 1\end{pmatrix}$ be the transfer matrix into an eigenbasis of F_i, *i.e.*, $F_i = Q_i D_i Q_i^{-1}$ with $D_i = \begin{pmatrix}r_i^- & 0 \\ 0 & r_i^+\end{pmatrix}$ and $Q_i^{-1} = \frac{1}{r_i^- - r_i^+}\begin{pmatrix}1 & -r_i^+ \\ -1 & r_i^-\end{pmatrix}$.

Computing $P_{i,k}$. We first compute the matrix $P_{i,k}$: With

$$C_i^{1/2} = \begin{pmatrix}\sqrt{c_i} & 0 \\ 0 & 0\end{pmatrix}, C_i^{1/2}Q_i = \begin{pmatrix}r_i^-\sqrt{c_i} & r_i^+\sqrt{c_i} \\ 0 & 0\end{pmatrix}$$

we have

$$C_i^{1/2}Q_i(I - D_i^k)(I - D_i)^{-1} = \sqrt{c_i}\begin{pmatrix}\frac{1-(r_i^-)^k}{1-r_i^-}r_i^- & \frac{1-(r_i^+)^k}{1-r_i^+}r_i^+ \\ 0 & 0\end{pmatrix},$$

and, when developing and regrouping terms which depend on k, we get:

$$\begin{aligned}P_{i,k} &= C_i^{1/2}Q_i(I - D_i^k)(I - D_i)^{-1}Q_i^{-1} \\ &= \frac{\sqrt{c_i}}{r_i^- - r_i^+}\begin{pmatrix}\frac{1-(r_i^-)^k}{1-r_i^-}r_i^- - \frac{1-(r_i^+)^k}{1-r_i^+}r_i^+ & \frac{1-(r_i^+)^k}{1-r_i^+}r_i^-r_i^+ - \frac{1-(r_i^-)^k}{1-r_i^-}r_i^+r_i^- \\ 0 & 0\end{pmatrix} \\ &= \sqrt{c_i}\begin{pmatrix}\frac{1}{(1-r_i^-)(1-r_i^+)} & \frac{-r_i^+r_i^-}{(1-r_i^-)(1-r_i^+)} \\ 0 & 0\end{pmatrix} \\ &\quad -\frac{\sqrt{c_i}}{r_i^- - r_i^+}\begin{pmatrix}\frac{(r_i^-)^{k+1}}{1-r_i^-} - \frac{(r_i^+)^{k+1}}{1-r_i^+} & \frac{(r_i^+)^{k+1}}{1-r_i^+}r_i^- - \frac{(r_i^-)^{k+1}}{1-r_i^-}r_i^+ \\ 0 & 0\end{pmatrix}.\end{aligned}$$

We also have $P_{i,k} = C_i^{1/2} Q_i (I - D_i^k)(I - D_i)^{-1} Q_i^{-1} = \sum_{j=0}^{k-1} R_{i,j}$ with

$$\begin{aligned} R_{i,j} &= C_i^{1/2} Q_i D_i^j Q_i^{-1} \\ &= \sqrt{c_i} \begin{pmatrix} (r_i^-)^{j+1} & (r_i^+)^{j+1} \\ 0 & 0 \end{pmatrix} Q_i^{-1} \\ &= \frac{\sqrt{s_i}}{r_i^- - r_i^+} \begin{pmatrix} (r_i^-)^{j+1} - (r_i^+)^{j+1} & -r_i^+(r_i^-)^{j+1} + r_i^-(r_i^+)^{j+1} \\ 0 & 0 \end{pmatrix}, \end{aligned}$$

but computing error terms based in $R_{i,j}$ before summing these errors gives a looser error bound than a tight calculation using $P_{i,k}$. More precisely, if we use $P_{i,k}\Theta_0^i = \sum_{j=0}^{k-1} R_{i,j}\Theta_0^i$ to upper bound $\|P_{i,k}\Theta_0^i\| \leqslant \sum_{j=0}^{k-1} \|R_{i,j}\Theta_0^i\|$, we end up with a worse bound.

Bias term. Thus, for the bias term:

$$\begin{aligned} P_{i,k}\Theta_0^i &= \sqrt{c_i}\theta_0^i \frac{1 - r_i^+ r_i^-}{(1 - r_i^-)(1 - r_i^+)} - \frac{\sqrt{c_i}\theta_0^i}{r_i^- - r_i^+} \begin{pmatrix} \left[(r_i^-)^{k+1}\frac{1-r_i^+}{1-r_i^-} - (r_i^+)^{k+1}\frac{1-r_i^-}{1-r_i^+}\right] \\ 0 \end{pmatrix} \\ &= \frac{\sqrt{c_i}\theta_0^i}{\sqrt{(1 - r_i^-)(1 - r_i^+)}} \begin{pmatrix} \frac{\left[(1-r_i^+r_i^-) - \rho_i^k A_1\right]}{\sqrt{(1-r_i^-)(1-r_i^+)}} \\ 0 \end{pmatrix}, \end{aligned}$$

where

$$\rho_i^k A_1 = \frac{(r_i^-)^{k+1}(1 - r_i^+)^2 - (r_i^+)^{k+1}(1 - r_i^-)^2}{r_i^- - r_i^+}.$$

This can be bound with the following lemma

Lemma B.2. *For all $\rho \in (0, 1)$ and $\omega \in [-\pi/2; \pi/2]$ and $r^\pm = \rho(\cos(\omega) \pm \sqrt{-1}\sin(\omega))$ we have:*

$$\left|\frac{1 - r^+r^- - \rho^k|A_1|}{|1 - r^+|}\right| \leqslant 3 + 3\rho^k \leqslant 6 \tag{B.14}$$

We note that the exact constant seems empirically to be 2. This lemma is proved as Lemma B.8 in Section B.5. This gives for the bias term

$$\begin{aligned} \|P_{i,k}\Theta_0^i\| &= \frac{\sqrt{c_i}(\theta_0^i)}{\sqrt{(1 - r_i^-)(1 - r_i^+)}} \left[\frac{1}{\sqrt{(1 - r_i^-)(1 - r_i^+)}} \left((1 - r_i^+r_i^-) - \rho_i^k A_1\right)\right] \\ &\leqslant 6\frac{\sqrt{c_i}(\theta_0^i)}{\sqrt{\gamma(s_i + \lambda)}}, \end{aligned}$$

since:

$$\begin{aligned} (1 - r_i^-)(1 - r_i^+) &= 1 - 2\,\mathfrak{Re}\,(r_i^+) + |r_i^+|^2 \\ &= 1 - (1 + \delta)(1 - \gamma(s_i + \lambda)) + \delta(1 - \gamma(s_i + \lambda)) \\ &= \gamma(s_i + \lambda). \end{aligned}$$

We also have a looser bound using $P_{i,k}\Theta_0^i = \sum_{j=0}^{k-1} R_{i,j}\Theta_0^i$.

$$\begin{aligned} R_{i,j}\Theta_0^i &= \frac{\sqrt{c_i}\theta_0^i}{r_i^- - r_i^+}\left((1 - r_i^+)(r_i^-)^{j+1} - (1 - r_i^-)(r_i^+)^{j+1}\right) \\ &= \sqrt{c_i}\theta_0^i \left(\frac{(r_i^-)^{j+1} - (r_i^+)^{j+1}}{r_i^- - r_i^+} - \frac{r_i^+(r_i^-)^{j+1} - r_i^-(r_i^+)^{j+1}}{r_i^- - r_i^+}\right) \end{aligned}$$

$$
\begin{aligned}
& & & \text{using De Moivre's formula,} \\
&= \sqrt{c_i}\theta_0^i\left(\frac{\rho_i^{j+1}\sin(\omega_i(j+1))}{\rho_i\sin(\omega_i)} - \frac{\rho_i e^{i\omega_i}\rho_i^{j+1}e^{-i\omega_i(j+1)} - \rho_i e^{-i\omega_i}\rho_i^{j+1}e^{+i\omega_i(j+1)}}{\rho_i e^{-i\omega_i} - \rho_i e^{i\omega_i}}\right) \\
&= \sqrt{c_i}\theta_0^i\left(\frac{\rho_i^{j+1}\sin(\omega_i(j+1))}{\rho_i\sin(\omega_i)} - \rho_i^{j+1}\frac{e^{-i\omega_i j} - e^{+i\omega_i j}}{e^{-i\omega_i} - e^{i\omega_i}}\right) \\
&= \sqrt{c_i}\theta_0^i\left(\frac{\rho_i^{j}\sin(\omega_i(j+1))}{\sin(\omega_i)} - \rho_i^{j+1}\frac{\sin(\omega_i j)}{\sin(\omega_i)}\right) \\
&\leqslant (1+e^{-1})\sqrt{c_i}\theta_0^i \quad \text{using Lemma B.9 (see proof in Section B.5),}
\end{aligned}
$$

which also gives for the bias term

$$\|P_{i,k}\Theta_0^i\| \leqslant (1+e^{-1})\sqrt{c_i}\theta_0^i k.$$

Thus we have the final bound:

$$\|P_{i,k}\Theta_0^i\|^2 \leqslant \min\left\{36\frac{c_i(\theta_0^i)^2}{\gamma(s_i+\lambda)}, 6n(1+e^{-1})\frac{c_i(\theta_0^i)^2}{\sqrt{\gamma(s_i+\lambda)}}, n^2(1+e^{-1})^2 c_i(\theta_0^i)^2\right\}. \quad \text{(B.15)}$$

Variance term. As for the variance term, with $V_i = \begin{pmatrix} v_i & 0 \\ 0 & 0 \end{pmatrix}$, we have $\operatorname{tr} P_{i,k}V_iP_{i,k} = \left\|P_{i,k}\begin{pmatrix}\sqrt{v_i}\\0\end{pmatrix}\right\|^2$.

$$
\begin{aligned}
\left\|P_{i,k}\begin{pmatrix}\sqrt{v_i}\\0\end{pmatrix}\right\| &= \frac{\sqrt{v_i c_i}}{(1-r_i^-)(1-r_i^+)}\left[1 + \frac{(r_i^-)^{k+1}(1-r_i^+) - (r_i^+)^{k+1}(1-r_i^-)}{r_i^+ - r_i^-}\right] \\
&= \frac{\sqrt{v_i c_i}}{\gamma(s_i+\lambda)}\left[1 - \rho_i^k B_{i,k}\right],
\end{aligned}
$$

where

$$\rho_i^k B_{i,k} = -\frac{(r_i^-)^{k+1}(1-r_i^+) - (r_i^+)^{k+1}(1-r_i^-)}{r_i^+ - r_i^-},$$

which we can bound using the following Lemma:

Lemma B.3. *For all $\rho \in (0,1)$ and $\omega \in [-\pi/2;\pi/2]$ and $r^\pm = \rho(\cos(\omega) \pm \sqrt{-1}\sin(\omega))$ we have:*

$$\left|\rho^k B_k\right| \leqslant 1.75.$$

Where we note that the exact upper bound seems to be 1.3. This Lemma is proved as Lemma B.10 in Section B.5.

We can also have a looser bound using $P_{i,k}\begin{pmatrix}v_i^{1/2}\\0\end{pmatrix} = \sum_{j=0}^{k-1} R_{i,j}\begin{pmatrix}v_i^{1/2}\\0\end{pmatrix}$ and

$$
\begin{aligned}
R_{i,j}\begin{pmatrix}v_i^{1/2}\\0\end{pmatrix} &= \frac{\sqrt{c_i v_i}}{r_i^- - r_i^+}\left((r_i^-)^{j+1} - (r_i^+)^{j+1}\right) \\
&= \sqrt{c_i v_i}\,\frac{\rho_i^{j+1}\sin(\omega_i(j+1))}{\rho_i\sin(\omega_i)} \\
&\leqslant (j+1)\sqrt{c_i v_i}, \text{ using the inequality } |\sin(k\omega_i)| \leqslant k|\sin(\omega_i)|
\end{aligned}
$$

and $\|P_{i,k}\begin{pmatrix} v_i^{1/2} \\ 0 \end{pmatrix}\| \leqslant \frac{\sqrt{c_i v_i}(k+1)k}{2}$.
This gives for the Variance term

$$\begin{aligned}\sum_{k=1}^{n} \operatorname{tr} P_{i,k} V_i P_{i,k} &\leqslant v_i c_i \sum_{k=1}^{n} \min\left\{ \frac{\left[1-\rho_i^k B_{1,k}\right]^2}{\gamma^2(s_i+\lambda)^2}, \frac{[1-\rho_i^k B_{1,k}]k(k+1)}{2\gamma(s_i+\lambda)}, \frac{k^2(k+1)^2}{4} \right\} \\ &\leqslant v_i c_i \min\left\{ \frac{8n}{\gamma^2(s_i+\lambda)^2}, \frac{(n+1)^3}{2\gamma(s_i+\lambda)}, \frac{(n+1)^5}{20} \right\}. \end{aligned} \tag{B.16}$$

One coalescent eigenvalue

We now turn to the case where F has two coalescent eigenvalues, which happens when the discriminant $\Delta = 0$. We assume that F_i has one coalescent eigenvalue $r_i = \frac{(1+\delta)(1-\gamma(s_i+\lambda))}{2}$. Then, with $\delta = \frac{1-\sqrt{\gamma(s_i+\lambda)}}{1+\sqrt{\gamma(s_i+\lambda)}}$, $r_i = \frac{(1+\delta)(1-\gamma(s_i+\lambda))}{2} = 1 - \sqrt{\gamma(s_i+\lambda)}$. Then F_i can be trigonalized as $F_i = Q_i D_i Q_i^{-1}$ with $Q_i = \begin{pmatrix} r_i & 1 \\ 1 & 0 \end{pmatrix}$, $D_i = \begin{pmatrix} r_i & 1 \\ 0 & r_i \end{pmatrix}$ and $Q_i^{-1} = \begin{pmatrix} 0 & 1 \\ 1 & -r_i \end{pmatrix}$.
We note that for all $k \geqslant 0$, then $D_i^k = r_i^{k-1}\begin{pmatrix} r_i & k \\ 0 & r_i \end{pmatrix}$.

Computing $P_{i,k}$. We first compute $P_{i,k}$:

$$(I_2 - D_i)^{-1} = \begin{pmatrix} \frac{1}{1-r_i} & \frac{1}{(1-r_i)^2} \\ 0 & \frac{1}{1-r_i} \end{pmatrix}$$

and

$$(I_2 - D_i^k)(I_2 - D_i)^{-1} = \begin{pmatrix} \frac{1-r_i^k}{1-r_i} & \frac{1-r_i^k}{(1-r_i)^2} - \frac{kr_i^{k-1}}{1-r_i} \\ 0 & \frac{1-r_i^k}{1-r_i} \end{pmatrix}.$$

Thus with $C_i^{1/2}Q_i = \begin{pmatrix} \sqrt{c_i}r_i & \sqrt{c_i} \\ 0 & 0 \end{pmatrix}$ we have

$$C_i^{1/2}Q_i(I_2 - D_i^k)(I_2 - D_i)^{-1} = \sqrt{c_i}\begin{pmatrix} \frac{1-r_i^k}{1-r_i}r_i & \frac{1-r_i^k}{(1-r_i)^2} - \frac{kr_i^k}{1-r_i} \\ 0 & 0 \end{pmatrix}.$$

And, computing as previously the matrices products, we derive:

$$\begin{aligned} P_{i,k} &= C_i^{1/2}Q_i(I_2 - D_i^k)(I_2 - D_i)^{-1}Q_i^{-1} \\ &= \sqrt{c_i}\begin{pmatrix} \frac{1-r_i^k}{(1-r_i)^2} - \frac{kr_i^k}{1-r_i} & \frac{1-r_i^k}{1-r_i}r_i - (\frac{1-r_i^k}{(1-r_i)^2} - \frac{kr_i^k}{1-r_i})r_i \\ 0 & 0 \end{pmatrix} \\ &= \sqrt{c_i}\begin{pmatrix} \frac{1-r_i^k}{(1-r_i)^2} - \frac{kr_i^k}{1-r_i} & \frac{1-r_i^k}{(1-r_i)^2}(r_i)^2 + \frac{kr_i^{k+1}}{1-r_i} \\ 0 & 0 \end{pmatrix} \\ &= \frac{\sqrt{c_i}}{1-r_i}\begin{pmatrix} \frac{1-r_i^k}{1-r_i} - kr_i^k & -\frac{1-r_i^k}{1-r_i}(r_i)^2 + kr_i^{k+1} \\ 0 & 0 \end{pmatrix}. \end{aligned}$$

Bias term. We thus have:

$$\begin{aligned} P_{i,k}\Theta_0^i &= \frac{\sqrt{c_i}}{1-r_i}\begin{pmatrix} \frac{1-r_i^k}{1-r_i} - kr_i^k & -\frac{1-r_i^k}{1-r_i}(r_i)^2 + kr_i^{k+1} \\ 0 & 0 \end{pmatrix}\begin{pmatrix} \theta_0^i \\ \theta_0^i \end{pmatrix} \\ &= \theta_0^i\sqrt{c_i}\begin{pmatrix} (1-r_i^k)\frac{1+r_i}{1-r_i} - kr_i^k \\ 0 \end{pmatrix}, \end{aligned}$$

and this gives for the bias term:

$$\begin{aligned} \|P_{i,k}\Theta_0^i\|^2 &= (\theta_0^i)^2 c_i\Big[(1-r_i^k)\frac{1+r_i}{1-r_i} - kr_i^k\Big]^2 \\ &= (\theta_0^i)^2 c_i\Big[\frac{1+r_i}{1-r_i} - \Big(k+\frac{1+r_i}{1-r_i}\Big)r_i^k\Big]^2 \\ &\qquad \text{developing the product, then using formulas for } r_i, \\ &= (\theta_0^i)^2 c_i\Big[\frac{2-\sqrt{\gamma(s_i+\lambda)}}{\sqrt{\gamma(s_i+\lambda)}} - \Big(k+\frac{2-\sqrt{\gamma(s_i+\lambda)}}{\sqrt{\gamma(c_i+\lambda)}}\Big)(1-\sqrt{\gamma(s_i+\lambda)})^k\Big]^2 \\ &= \frac{(\theta_0^i)^2 c_i}{\gamma(s_i+\lambda)}\Big[2-\sqrt{\gamma(s_i+\lambda)} \\ &\qquad\qquad - \Big(k\sqrt{\gamma(s_i+\lambda)} + 2 - \sqrt{\gamma(s_i+\lambda)}\Big)\Big(1-\sqrt{\gamma(s_i+\lambda)}\Big)^k\Big]^2 \\ &= \frac{(\theta_0^i)^2 c_i}{\gamma(s_i+\lambda)}\Big[2-\sqrt{\gamma(s_i+\lambda)} - \Big((2+(k-1)\sqrt{\gamma(s_i+\lambda)}\Big)\Big(1-\sqrt{\gamma(s_i+\lambda)}\Big)^k\Big]^2 \\ &\leqslant 4\frac{(\theta_0^i)^2 c_i}{\gamma(s_i+\lambda)}, \text{ using Lemma B.11 in Section B.5.} \end{aligned} \tag{B.17}$$

Variance term. With $V = \begin{pmatrix} v_i & 0 \\ 0 & 0 \end{pmatrix}$,

$$\begin{aligned} &\operatorname{tr} P_{i,k}VP_{i,k} \\ =\ & \frac{s_i}{(1-r_i)^2}\begin{pmatrix} \frac{1-r_i^k}{1-r_i} - kr_i^k & -\frac{1-r_i^k}{1-r_i}(r_i)^2 + kr_i^{k+1} \\ 0 & 0 \end{pmatrix}\begin{pmatrix} v_i & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} \frac{1-r_i^k}{1-r_i} - kr_i^k & 0 \\ -\frac{1-r_i^k}{1-r_i}(r_i)^2 + kr_i^{k+1} & 0 \end{pmatrix} \\ =\ & \frac{s_i v_i}{(1-r_i)^2}\Big[\frac{1-r_i^k}{1-r_i} - kr_i^k\Big]^2 \\ =\ & \frac{v_i h_i}{\gamma(s_i+\lambda)}\Big[\frac{1-r_i^k}{1-r_i} - kr_i^k\Big]^2 \\ =\ & \frac{v_i h_i}{\gamma(s_i+\lambda)(1-r_i)^2}\Big[1-r_i^k-(1-r_i)kr_i^k\Big]^2 \\ =\ & \frac{v_i h_i}{\gamma^2(s_i+\lambda)^2}\Big[1-(1+k\sqrt{\gamma(s_i+\lambda)})(1-\sqrt{\gamma(s_i+\lambda)})^k\Big]^2, \end{aligned}$$

and

$$\begin{aligned} \sum_{k=1}^n \operatorname{tr} P_{i,k}VP_{i,k} &= \frac{v_i s_i}{\gamma^2(s_i+\lambda)^2}\sum_{k=1}^n\Big[1-(1+k\sqrt{\gamma(s_i+\lambda)})(1-\sqrt{\gamma(s_i+\lambda)})^k\Big]^2 \\ &\leqslant n\frac{v_i s_i}{\gamma^2(s_i+\lambda)^2} \text{ using Lemma B.11 in Section B.5.} \end{aligned} \tag{B.18}$$

Alternative bounds for the bias and the variance term, as in Equations(B.12), (B.15) may be derived as well. Combining all these results, we are now able to state Theorem 3.3.

B.3.4 Conclusion

Combining results from Lemma B.1, and Equations (B.12), (B.15), (B.16), with $c = \Sigma$, and using the following simple facts:

- For the least squares regression function, with $c = \Sigma$, $\mathbb{E}\langle \bar{\Theta}_n, C\bar{\Theta}_n \rangle = \mathbb{E}f(\bar{\theta}_n) - f(\theta_*)$.
- Under assumption $\mathcal{A}_3$, $\mathcal{A}_4$, we have $V \preccurlyeq \tau^2\Sigma$.
- The squared norm of a vector is the sum of its squared components on the orthonormal eigenbasis. For example $\|P_{n+1}\Theta_0\|^2 = \sum_{i=1}^d \|P_{i,n+1}\Theta_0^i\|^2$.
- For any regularization parameter $\lambda \in \mathbb{R}_+$ and for any constant step-size $\gamma(\Sigma + \lambda \mathrm{I}) \preccurlyeq \mathrm{I}$, for any $\delta \in [\frac{1-\sqrt{\gamma\lambda}}{1+\sqrt{\gamma\lambda}}, 1]$, matrix F will have only two distinct complex eigenvalues or two coalescent eigenvalues.

Proposition B.4. *Under ($\mathcal{A}_{4,5}$), for any regularization parameter $\lambda \in \mathbb{R}_+$ and for any constant step-size $\gamma(\Sigma + \lambda \mathrm{I}) \preccurlyeq \mathrm{I}$ we have for any $\delta \in [\frac{1-\sqrt{\gamma\lambda}}{1+\sqrt{\gamma\lambda}}, 1]$, for the recursion in Equation (3.10):*

$$\begin{aligned}
\mathbb{E}f(\bar{\theta}_n) - f(\theta_*) \leqslant\ & 2\lambda \|\lambda^{1/2}\Sigma^{1/2}(\Sigma + \lambda I)^{-1}(\theta_0 - \theta_*)\|^2 \\
& + \sum_{i=1}^d \frac{2}{(n+1)^2} \min\left\{ 36\frac{c_i(\tilde{\theta}_0^i)^2}{\gamma(s_i+\lambda)}, 6n(1+e^{-1})\frac{c_i(\tilde{\theta}_0^i)^2}{\sqrt{\gamma(s_i+\lambda)}}, n^2(1+e^{-1})^2 c_i(\tilde{\theta}_0^i)^2 \right\} \\
& + \sum_{i=1}^d \frac{\gamma^2}{(n+1)^2} v_i c_i \min\left\{ \frac{8n}{\gamma^2(s_i+\lambda)^2}, \frac{(n+1)^3}{2\gamma(s_i+\lambda)}, \frac{(n+1)^5}{20} \right\}.
\end{aligned}$$

This implies, using the Equation (B.13) for the initial point, using $c_i = \sigma_i$ and regrouping sums as traces or norms:

$$\begin{aligned}
\mathbb{E}f(\bar{\theta}_n) - f(\theta_*) \leqslant\ & 2\lambda \|\lambda^{1/2}\Sigma^{1/2}(\Sigma + \lambda I)^{-1}(\theta_0 - \theta_*)\|^2 \\
& + 2\min\left\{ \frac{36\|\Sigma^{1/2}(\Sigma+\lambda I)^{-1/2}(\theta_0-\theta_*)\|^2}{\gamma(n+1)^2}, (1+e^{-1})^2\|\Sigma^{1/2}(\theta_0-\theta_*)\|^2 \right\} \\
& + \min\left\{ \frac{8\operatorname{tr}(V\Sigma(\Sigma+\lambda I)^{-2})}{n+1}, n\gamma \operatorname{tr}(V\Sigma(\Sigma+\lambda I)^{-1}) \right\},
\end{aligned}$$

which gives exactly Theorem 3.3 using $V \preccurlyeq \tau^2\Sigma$ in the Variance term, and $\lambda^{1/2}(\Sigma+\lambda I)^{-1/2} \preccurlyeq I$ in the first term.

B.4 Tighter bounds

B.4.1 Simple upper-bounds

In this section, we chow how tighter bounds naturally appear from the regularized quantities appearing in Theorems. It only relies on simple algebraic majorations, even if one has to be careful with the allowed intervals for r, b.

Lemma B.5. *For any $\lambda \geqslant 0$, for any $b \in [0; 1]$, if $\operatorname{tr}(\Sigma^b)$ exists, we have:*

$$\begin{aligned}
\operatorname{tr}(\Sigma(\Sigma+\lambda I)^{-1}) &\leqslant \frac{\operatorname{tr}(\Sigma^b)}{\lambda^b} \\
\operatorname{tr}(\Sigma^{-2}(\Sigma+\lambda I)^{-2}) &\leqslant \frac{\operatorname{tr}(\Sigma^b)}{\lambda^b}.
\end{aligned}$$

Proof. As all operators can be diagonalized in a same eigenbasis with positive eigenvalues, we have,

$$\begin{aligned}
\operatorname{tr}(\Sigma(\Sigma+\lambda I)^{-1}) &\leqslant \left|\!\left|\!\left|\Sigma^{1-b}(\Sigma+\lambda I)^{-1}\right|\!\right|\!\right| \operatorname{tr}(\Sigma^b) \\
|||\Sigma^{1-b}(\Sigma+\lambda I)^{-1}||| &\leqslant \sup_{0\leqslant x} \frac{x^{1-b}}{(x+\lambda)} \\
&\leqslant \sup_{0\leqslant x} x^{1-b}\left(\frac{1}{\lambda}\wedge\frac{1}{x}\right) \\
&\leqslant \sup_{0\leqslant x} x^{1-b}\left(\frac{1}{\lambda}\right)^b\left(\frac{1}{x}\right)^{1-b} = \lambda^{-b}.
\end{aligned}$$

The calculations are exactly the same for $\operatorname{tr}(\Sigma^{-2}(\Sigma+\lambda I)^{-2}) \leqslant \frac{\operatorname{tr}(\Sigma^b)}{\lambda^b}$. □

As for the bias term, we need to bound the following quantities:

Lemma B.6. *For any $\lambda \geqslant 0$, for any $r \in [-1;1]$, we have:*

$$\left\|\Sigma^{1/2}(\Sigma+\lambda\mathrm{I})^{-1}(\theta_0-\theta_*)\right\|^2 \leqslant \lambda^{-(1+r)}\left\|\Sigma^{r/2}(\theta_0-\theta_*)\right\|^2.$$

For any $\lambda \geqslant 0$, for any $r \in [-1;0]$, we have:

$$\left\|(\Sigma+\lambda\mathrm{I})^{-1/2}(\theta_0-\theta_*)\right\|^2 \leqslant \lambda^{-(1+r)}\left\|\Sigma^{r/2}(\theta_0-\theta_*)\right\|^2.$$

For any $\lambda \geqslant 0$, for any $r \in [0;1]$, we have:

$$\|\Sigma^{1/2}(\Sigma+\lambda I)^{-1/2}(\theta_0-\theta_*)\|^2 \leqslant \lambda^{-r}\left\|\Sigma^{r/2}(\theta_0-\theta_*)\right\|^2$$

(No result when $r \leqslant 0$ because of saturation effect).

Proof. Proof relies of simple following calculations:

$$\begin{aligned}
\|\Sigma^{1/2}(\Sigma+\lambda\mathrm{I})^{-1}(\theta_0-\theta_*)\| &\leqslant \left|\!\left|\!\left|\Sigma^{1/2-r/2}(\Sigma+\lambda\mathrm{I})^{-1}\right|\!\right|\!\right| \|\Sigma^{r/2}(\theta_0-\theta_*)\| \\
&\leqslant \left(\frac{1}{\lambda}\right)^{1-(1/2-r/2)} \|\Sigma^{r/2}(\theta_0-\theta_*)\| \\
&\leqslant \lambda^{-\frac{1+r}{2}}\|\Sigma^{r/2}(\theta_0-\theta_*)\|
\end{aligned}$$

$$\begin{aligned}
\|(\Sigma+\lambda\mathrm{I})^{-1/2}(\theta_0-\theta_*)\| &\leqslant \left|\!\left|\!\left|\Sigma^{-r/2}(\Sigma+\lambda\mathrm{I})^{-1/2}\right|\!\right|\!\right| \|\Sigma^{r/2}(\theta_0-\theta_*)\| \\
&\leqslant \left(\frac{1}{\lambda}\right)^{\frac{1+r}{2}} \|\Sigma^{r/2}(\theta_0-\theta_*)\| \\
&\leqslant \lambda^{-\frac{1+r}{2}}\|\Sigma^{r/2}(\theta_0-\theta_*)\|
\end{aligned}$$

$$\begin{aligned}
\|\Sigma^{1/2}(\Sigma+\lambda I)^{-1/2}(\theta_0-\theta_*)\| &\leqslant \left|\!\left|\!\left|\Sigma^{1/2-r/2}(\Sigma+\lambda\mathrm{I})^{-1/2}\right|\!\right|\!\right| \|\Sigma^{r/2}(\theta_0-\theta_*)\| \\
&\leqslant \left(\frac{1}{\lambda}\right)^{\frac{1-(1-r)}{2}} \|\Sigma^{r/2}(\theta_0-\theta_*)\| \\
&\leqslant \lambda^{-\frac{r}{2}}\|\Sigma^{r/2}(\theta_0-\theta_*)\|.
\end{aligned}$$

□

B.4.2 Theorem 3.5 and Equation (3.13)

Theorem 3.5 and Equation (3.13) are directly derived from Theorem 3.2 and Theorem 3.3, using Lemmas B.5 and B.6.

To derive corollaries for the optimal γ, one has to find the γ that balances the bias and variance term and to compute the products for such a step size.

Equation (3.13)

We derive from Theorem 3.2, when choosing $\gamma = (\lambda n)^{-1}$, and using Lemmas B.5 and B.6, the following bound, under assumptions of Theorem 3.2:

$$\mathbb{E} f(\bar{\theta}_n) - f(\theta_*) \leqslant \frac{(18 + \mathrm{Res}(n, b, r, \gamma))\|\Sigma^{r/2}(\theta_0 - \theta_*)\|^2}{(\gamma n)^{\frac{1-r}{2}}} + \frac{6\sigma^2 \operatorname{tr}(\Sigma^b)\gamma^b}{n^{1-b}}.$$

Where $\mathrm{Res}(n, b, r, \gamma) := 3\gamma^{1+b} n^b \operatorname{tr}(\Sigma^b)$ if $-1 \leqslant r \leqslant 0$ and $\mathrm{Res}(n, b, r, \gamma) := 0$ if $0 \leqslant r \leqslant 1$. When choosing the optimal $\gamma \propto n^{\frac{-b+r}{b+1-r}}$, we have that $\gamma^{1+b} n^b = n^{-1+\frac{1+b}{1+b-r}} = n^{\chi}$, with $\chi = \frac{-r}{1+b-r} \geqslant 0$ if $r \leqslant 0$. Thus the residual term is always vanishing for $r \leqslant 0$ and does not exist for $r \geqslant 0$.

Theorem 3.5

Theorem 3.5 directly follows from Lemmas B.5 and B.6 and the choice of $\gamma \propto n^{\frac{-2b+2r-1}{b+1-r}}$.

B.5 Technical Lemmas

The following sequence of Lemmas appear in the proof. They are mostly independent and rely on simple calculations.

Lemma B.7. *The operator* $\left[(\Sigma + \lambda \mathrm{I}) \otimes \mathrm{I} + \mathrm{I} \otimes (\Sigma + \lambda \mathrm{I})\right]^{-1}$ *is a non-decreasing operator on* $(S_n, \preccurlyeq)$.

Proof. Lemma means that for two matrices $M, N \in S_n(\mathbb{R})$ such that $M \preccurlyeq N$, then

$$\left[(\Sigma + \lambda \mathrm{I}) \otimes \mathrm{I} + \mathrm{I} \otimes (\Sigma + \lambda \mathrm{I})\right]^{-1} M \preccurlyeq \left[(\Sigma + \lambda \mathrm{I}) \otimes \mathrm{I} + \mathrm{I} \otimes (\Sigma + \lambda \mathrm{I})\right]^{-1} N.$$

It is equivalent to show that for any symmetric positive matrix $A \in S_n^+$,

$$\left[(\Sigma + \lambda \mathrm{I}) \otimes \mathrm{I} + \mathrm{I} \otimes (\Sigma + \lambda \mathrm{I})\right]^{-1} A \in S_n^+(\mathbb{R}).$$

We consider a matrix $A \in S_n^+(\mathbb{R})$. A can be decomposed as a sum of (at most) n rank one matrices $A = \sum_{i=1}^n \omega_i \omega_i^\top$, with $\omega_i \in \mathbb{R}^n$. We thus just have to prove that for some $\omega \in \mathbb{R}^n$, $\left[(\Sigma + \lambda \mathrm{I}) \otimes \mathrm{I} + \mathrm{I} \otimes (\Sigma + \lambda \mathrm{I})\right]^{-1} \omega\omega^\top \in S_n^+(\mathbb{R})$.

Let $\Sigma = \sum_{i \geqslant 0} \mu_i e_i \otimes e_i$ is the eigenvalue decomposition of Σ, then

$$\left[(\Sigma + \lambda \mathrm{I}) \otimes \mathrm{I} + \mathrm{I} \otimes (\Sigma + \lambda \mathrm{I})\right]^{-1} \omega\omega^\top = \sum_{i,j \geqslant 0} \frac{\langle \omega, e_i \rangle \langle \omega, e_j \rangle}{\mu_i + \mu_j + 2\lambda} e_i \otimes e_j.$$

Thus, in the orthonormal basis of eigenvectors, this is thus Hadamard product between

$$\sum_{i,j \geqslant 0} \langle \omega, e_i \rangle \langle \omega, e_j \rangle e_i \otimes e_j = \omega\omega^\top$$

and the matrix $C = \left((\frac{1}{\mu_i + \mu_j + 2\lambda})_{i,j \geqslant 0} \right)$. Matrix C is a Cauchy matrix and is thus positive. Moreover the Hadamard product of two positive matrices is positive, which concludes the proof. □

Remark: surprisingly, the inverse operator $(\Sigma + \lambda \mathrm{I}) \otimes \mathrm{I} + \mathrm{I} \otimes (\Sigma + \lambda \mathrm{I})$ is not non-decreasing. Indeed, $\preccurlyeq$ is not a total order on S_n so we may have that an operator is non-decreasing and its inverse is not.

Lemma B.8. *For all $\rho \in (0,1)$ and $\omega \in [-\pi/2; \pi/2]$ and $r^{\pm} = \rho(\cos(\omega) \pm \sqrt{-1}\sin(\omega))$ we have:*

$$\left| \frac{1 - r^+ r^- - \rho^k |A_1|}{|1 - r^+|} \right| \leqslant \min\{1 + \rho + e^{-1} + 4\rho^k, 2 + \rho + \sqrt{5}\rho^{k+1}\} \leqslant 6. \tag{B.19}$$

Proof. We note that $\rho_i^k A_1$ is a real number as is is a quotient of pure complex numbers, which come from the difference between a complex and its conjugate. We first write A_1 as a combination of sine and cosine functions:

$$\begin{aligned}
\rho_i^k A_1 &= \frac{(r_i^-)^{k+1}(1 - r_i^+)^2 - (r_i^+)^{k+1}(1 - r_i^-)^2}{r_i^- - r_i^+} \\
&= -\frac{(r_i^-)^{k+1} - (r_i^+)^{k+1} - 2r_i^- r_i^+((r_i^-)^k - (r_i^+)^k) + (r_i^- r_i^+)(r_i^-)^{k-1} - (r_i^+)^{k-1})}{\rho_i \sin \omega_i} \\
&= -\frac{\rho_i^{k+1}\sin((k+1)\omega_i) - 2\rho_i^{k+2}\sin(k\omega_i) + \rho_i^{k+3}\sin((k-1)\omega_i)}{\rho_i \sin \omega_i}.
\end{aligned}$$

This quantity can be simplified when $\rho \to 1$ or $\omega \to 0$. We thus modify the expression of A_1 to make these dependencies clearer:

$$\begin{aligned}
-A_1 &= \frac{\sin((k+1)\omega_i) - 2\rho_i \sin(k\omega_i) + \rho_i^2 \sin((k-1)\omega_i)}{\sin \omega_i} \\
&= \frac{(\cos(\omega) - \rho)(\sin(k\omega) - \rho\sin((k-1)\omega)) + \cos(k\omega)\sin(\omega) - \rho\cos((k-1)\omega)\sin(\omega)}{\sin \omega_i} \\
&\quad \text{developing } \sin(a+b) = \sin(a)\cos(b) + \cos(a)\sin(b) \text{ and regrouping terms,} \\
&= \frac{(\cos(\omega) - \rho)^2 \sin((k-1)\omega) + (\cos(\omega) - \rho)\sin(\omega)\cos((k-1)\omega) + \cos(k\omega)\sin(\omega)}{\sin \omega_i} \\
&\quad - \frac{\rho\cos((k-1)\omega)\sin(\omega)}{\sin \omega_i} \\
&= \frac{(\cos(\omega) - \rho)^2 \sin((k-1)\omega)}{\sin \omega_i} + (\cos(\omega) - \rho)\cos((k-1)\omega) + \cos(k\omega) \\
&\quad - \rho\cos((k-1)\omega) \\
&\quad \text{simplifying expression, then developing the cosine,} \\
&= \frac{(\cos(\omega) - \rho)^2 \sin((k-1)\omega)}{\sin \omega_i} + 2(\cos(\omega) - \rho)\cos((k-1)\omega) + \sin(\omega)\sin((k-1)\omega) \\
&\quad .
\end{aligned} \tag{B.20}$$

So that in that final expression all the terms behave relatively simply when $\rho \to 1$ or $\omega \to 0$. We want to upper bound:

$$\left| \frac{1 - r^+ r^- - \rho^k |A_1|}{|1 - r^+|} \right|.$$

We thus consider separately the first and second term.

$$\frac{1 - r_i^+ r_i^-}{|1 - r_i^+|} = \frac{1 - \rho^2}{|1 - r_i^+|} \leqslant 1 + \rho \qquad \text{(exact if } \omega = 0\text{)}.$$

Then, using Equation (B.20):

$$\frac{-\rho_i^k|A_1|}{|1-r_i^+|} = \rho^k \frac{\frac{(\cos(\omega)-\rho)^2\sin((k-1)\omega)}{\sin\omega_i} + 2(\cos(\omega)-\rho)\cos((k-1)\omega) + \sin(\omega)\sin((k-1)\omega)}{\sqrt{(1-\rho\cos\omega)^2+\rho^2\sin^2(\omega)}}.$$

Considering separately the three terms in the numerator, using numerous times that for any $a, b \in [0;1]$, $|a-b| \leqslant 1-ab$:

$$\begin{aligned}
\Diamond &= \left|\rho^k \frac{\frac{(\cos(\omega)-\rho)^2\sin((k-1)\omega)}{\sin\omega_i}}{\sqrt{(1-\rho\cos\omega)^2+\rho^2\sin^2(\omega)}}\right| \\
&\leqslant \rho^k \frac{(\cos(\omega)-\rho)\sin((k-1)\omega)}{\sin\omega_i} \\
&\qquad \text{as } |(\cos(\omega)-\rho)| \leqslant 1-\rho\cos(\omega), \\
&\leqslant \rho^k \frac{(\cos(\omega)-1)\sin((k-1)\omega)}{\sin\omega_i} + \rho^k \frac{(1-\rho)\sin((k-1)\omega)}{\sin\omega_i} \\
&\qquad \text{writing } \cos(\omega)-\rho = \cos(\omega)-1+1-\rho \\
&\leqslant \rho^k(1-\rho)(k-1) + \rho^k \frac{(\cos(\omega)-1)\sin((k-1)\omega)}{\sin\omega_i} \\
&\qquad \text{as } |\sin((k-1)\omega)| \leqslant |(k-1)\sin(\omega)|, \\
&\leqslant \rho^k(1-\rho)k - (1-\rho)\rho^k + \rho^k \frac{(\cos(\omega)-1)\sin((k-1)\omega)}{\sin\omega_i} \\
&\qquad \text{writing } \cos(\omega)-1 = 2\sin^2(\omega/2), \\
&\leqslant \rho^k(1+(1-\rho))^k - \rho^k - (1-\rho)\rho^k + \rho^k \frac{2\sin^2(\omega/2)}{\sin\omega_i} \\
&\qquad \text{using } 1+(1-\rho)k \leqslant (1+(1-\rho))^k, \\
&\leqslant \rho^k(1+(1-\rho))^k - \rho^k - (1-\rho)\rho^k + \rho^k\tan(\omega/2) \\
&\qquad \text{and as } \tan(\omega/2) \leqslant 1 \text{ for } |\omega| \leqslant \pi/2, \\
&\leqslant 1-(1-\rho)\rho^k \\
&\qquad \text{using } \rho^k(1+(1-\rho))^k = (1-(1-\rho)^2)^k \leqslant 1,
\end{aligned}$$

And for the second and third term:

$$\begin{aligned}
2\left|\rho^k \frac{(\cos(\omega)-\rho)\cos((k-1)\omega)}{\sqrt{(1-\rho\cos\omega)^2+\rho^2\sin^2(\omega)}}\right| &\leqslant 2\rho^k, \\
\left|\rho^k \frac{+\sin(\omega)\sin((k-1)\omega)}{\sqrt{(1-\rho\cos\omega)^2+\rho^2\sin^2(\omega)}}\right| &\leqslant \rho^k.
\end{aligned}$$

Thus:

$$\left|\frac{1-r_i^+r_i^- - \rho_i^k|A_1|}{|1-r_i^+|}\right| \leqslant 1+\rho+1+3\rho^k.$$

We also have

$$\Diamond = |\rho^k \frac{\frac{(\cos(\omega)-\rho)^2\sin((k-1)\omega)}{\sin\omega_i}}{\sqrt{(1-\rho\cos\omega)^2+\rho^2\sin^2(\omega)}}|$$

$$\begin{aligned}
&\leqslant \rho^k \frac{(\cos(\omega)-\rho)\sin((k-1)\omega)}{\sin\omega_i} \\
&\leqslant \rho^k(1-\rho)(k-1) + \rho^k \frac{(\cos(\omega)-1)\sin((k-1)\omega)}{\sin\omega_i} \\
&\leqslant (1-\frac{1}{k+1})^{k+1} - (1-\rho)\rho^k + \rho^k \frac{(\cos(\omega)-1)\sin((k-1)\omega)}{\sin\omega_i} \\
&\leqslant e^{-1} - (1-\rho)\rho^k + \rho^k \frac{\sin^2(\omega/2)}{\sin\omega_i}.
\end{aligned}$$

Using that

$$\begin{aligned}
k \sup_{x\in[0;1]} x^k(1-x) &= k\frac{1}{k+1}(1-\frac{1}{k+1})^k = (1-\frac{1}{k+1})^{k+1} &\text{(B.21)}\\
&= \exp((k+1)\ln((1-\frac{1}{k+1})) \leqslant e^{-1}, &\text{(B.22)}
\end{aligned}$$

we get

$$\left|\frac{1-r_i^+r_i^- - \rho_i^k|A_1|}{|1-r_i^+|}\right| \leqslant 1+\rho+e^{-1}+4\rho^k$$

We can also change $3\rho^k$ into $\sqrt{5}\rho^k$ We have used that $|(\rho-\cos(\omega))| \leqslant (1-\rho\cos(\omega))$. □

Lemma B.9. *For any $\rho_i \in (0;1)$, for any $\omega_i \in [-\pi/2;\pi/2]$*

$$\frac{\rho_i^j \sin(\omega_i(j+1))}{\sin(\omega_i)} - \rho_i^{j+1}\frac{\sin(\omega_i j)}{\sin(\omega_i)} \leqslant 1+e^{-1}.$$

Proof.

$$\begin{aligned}
\frac{\rho_i^j \sin(\omega_i(j+1))}{\sin(\omega_i)} - \rho_i^{j+1}\frac{\sin(\omega_i j)}{\sin(\omega_i)} &= \rho_i^j\left(\frac{\sin(\omega_i(j+1)) - \rho_i\sin(\omega_i j)}{\sin(\omega_i)}\right) \\
&= \rho_i^j\left(\frac{(\cos(\omega_i)-\rho_i)\sin(\omega_i j)}{\sin(\omega_i)} + \cos(j\omega_i)\right) \\
&\leqslant \rho_i^j\left((1-\rho_i)j+1\right) \\
&\leqslant 1+e^{-1} \text{ using (B.22).}
\end{aligned}$$

□

Lemma B.10. *For all $\rho \in (0,1)$ and $\omega \in [-\pi/2;\pi/2]$ and $r^\pm = \rho(\cos(\omega) \pm \sqrt{-1}\sin(\omega))$ we have:*

$$\left|\rho_i^k B_{1,k}\right| \leqslant 1.75 \tag{B.23}$$

Proof. Once again, as the considered quantity is real, we first express it as a combination of sine and cosine functions. We then use some simple trigonometric tricks to upper bound the quantity.

$$\begin{aligned}
\rho_i^k B_{1,k} &= -\frac{(r_i^-)^{k+1}(1-r_i^+) - (r_i^+)^{k+1}(1-r_i^-)}{r_i^+ - r_i^-} \\
&= -\frac{2\mathfrak{Im}[(r_i^-)^{k+1}(1-r_i^+)]}{\sqrt{-\Delta_i}}
\end{aligned}$$

as it is the difference between a complex and its conjugate,

$$= -\frac{\mathfrak{Im}\left[\rho_i^k e^{-(k+1)i\omega_i}(1-\rho_i\cos(\omega_i)-i\rho_i\sin(\omega_i))\right]}{\sin\omega_i\rho_i} \text{ developing the product,}$$

$$= \rho_i^k\frac{\cos((k+1)\omega_i)\sin(\omega_i)\rho_i+\sin((k+1)\omega_i)(1-\rho_i\cos(\omega_i))}{\sin\omega_i\rho_i}$$

$$= \rho_i^k\Big[\rho_i\cos((k+1)\omega_i)+(1-\rho_i\cos(\omega_i))\frac{\sin((k+1)\omega_i)}{\sin\omega_i}\Big] \text{ and simplifying.}$$

Let's turn our interest to the second part of the quantity:

$$\begin{aligned}
\Diamond &= \left|\rho_i^k(1-\rho_i\cos(\omega_i))\frac{\sin((k+1)\omega_i)}{\sin\omega_i}\right| \\
&= \left|\rho_i^k(1-\rho_i+\rho_i(1-\cos(\omega_i)))\frac{\sin((k+1)\omega_i)}{\sin\omega_i}\right| \\
&\quad \text{introducing an artificial } +\rho_i-\rho_i, \\
&\leqslant \rho_i^k\left|(1-\rho_i)\frac{\sin((k+1)\omega_i)}{\sin\omega_i}\right|+\rho_i^k\left|\rho_i(1-\cos(\omega_i))\frac{\sin((k+1)\omega_i)}{\sin\omega_i}\right| \\
&\quad \text{by triangular inequality,} \\
&\leqslant \rho_i^k\left|(1-\rho_i)(k+1)\right|+\rho_i^k\left|\rho_i\sin^2(\frac{\omega}{2})\frac{1}{2\cos(\frac{\omega}{2})\sin(\frac{\omega}{2})}\right| \\
&\quad \text{using } 1-\cos(\omega_i)=2\sin^2(\frac{\omega}{2}) \\
&\leqslant \rho_i^k(1-\rho_i)k+\rho_i^k(1-\rho)+\rho_i^k\left|\rho_i\sin^2(\frac{\omega}{2})\frac{1}{2\cos(\frac{\omega}{2})\sin(\frac{\omega}{2})}\right| \\
&\leqslant (1-(1-\rho_i))^k(1+(1-\rho_i))^k-\rho_i^k+\frac{1}{2(k+1)}+\rho_i^k\left|\frac{\rho_i}{2}\tan(\frac{\omega}{2})\right| \\
&\leqslant (1-(1-\rho_i)^2)^k+\frac{1}{4}+\frac{1}{2}\leqslant 1+\frac{1}{4}+\frac{1}{2}-\rho_i^k.
\end{aligned}$$

Thus

$$\left|\rho_i^k B_{1,k}\right| = \rho_i^k+1+\frac{1}{4}+\frac{1}{2}-\rho_i^k\leqslant 1+\frac{1}{4}+\frac{1}{2}=1.75.$$

□

Lemma B.11. *For any $s_i,\gamma,\lambda\in\mathbb{R}_+^3$ such that $\gamma(s_i+\lambda)\leqslant 1$, for any $k\in\mathbb{N}$, we have the two following highly related identities:*

$$\begin{aligned}
0\leqslant 2-\sqrt{\gamma(s_i+\lambda)}-(2+(k-1)\sqrt{\gamma(s_i+\lambda)})(1-\sqrt{\gamma(s_i+\lambda)})^k &\leqslant 2 \\
0\leqslant 1-(1+k\sqrt{\gamma(s_i+\lambda)})(1-\sqrt{\gamma(s_i+\lambda)})^k &\leqslant 1.
\end{aligned}$$

Proof. Proof relies on the trick, for any $\alpha\in\mathbb{R},n\in\mathbb{N}$: $1+n\alpha\leqslant(1+\alpha)^n$. For the first one:

$$\begin{aligned}
&\sqrt{\gamma(s_i+\lambda)}+(2+(k-1)\sqrt{\gamma(s_i+\lambda)})(1-\sqrt{\gamma(s_i+\lambda)})^k= \\
=\ &\sqrt{\gamma(s_i+\lambda)}+(1-\sqrt{\gamma(s_i+\lambda)})^k+(1+(k-1)\sqrt{\gamma(s_i+\lambda)})(1-\sqrt{\gamma(s_i+\lambda)})^k \\
\leqslant\ &\sqrt{\gamma(s_i+\lambda)}+(1-\sqrt{\gamma(s_i+\lambda)})+(1+(k-1)\sqrt{\gamma(s_i+\lambda)})(1-\sqrt{\gamma(s_i+\lambda)})^{k-1} \\
\leqslant\ &1+(1-\gamma(s_i+\lambda))^{k-1}\leqslant 2.
\end{aligned}$$

For the second one:

$$0\leqslant(1+k\sqrt{\gamma(s_i+\lambda)})(1-\sqrt{\gamma(s_i+\lambda)})^k \leqslant (1-\gamma(s_i+\lambda))^k\leqslant 1.$$

□

4

Bridging the Gap between Constant Step Size Stochastic Gradient Descent and Markov Chains

We consider the minimization of an objective function given access to unbiased estimates of its gradient through stochastic gradient descent (SGD) with constant step-size. While the detailed analysis was only performed for quadratic functions, we provide an explicit asymptotic expansion of the moments of the averaged SGD iterates that outlines the dependence on initial conditions, the effect of noise and the step-size, as well as the lack of convergence in the general (non-quadratic) case. For this analysis, we bring tools from Markov chain theory into the analysis of stochastic gradient and create new ones (similar but different from stochastic MCMC methods). We then show that Richardson-Romberg extrapolation may be used to get closer to the global optimum and we show empirical improvements of the new extrapolation scheme.

This chapter is based on our work *Bridging the Gap between Constant Step Size Stochastic Gradient Descent and Markov Chains*, A. Dieuleveut, A. Durmus and F.Bach.

Contents

4.1 Introduction 160
4.2 Main results 162
4.2.1 Setting 162
4.2.2 Related work 163
4.2.3 Summary and discussion of main results 164
4.3 Detailed analysis 165
4.3.1 Expansion of moments under π_γ when γ is in a neighborhood of 0 . 165
4.3.2 Expansion for a given $\gamma > 0$ when k tends to $+\infty$ 167
4.3.3 Continuous interpretation of SGD and weak error expansion 168
4.4 Experiments 171
4.5 Conclusion 171

4.1 Introduction

We consider the minimization of an objective function given access to unbiased estimates of the function values or its gradients. This key methodological problem has raised interest in different communities: in large-scale machine learning (Bottou and Bousquet, 2008; Shalev-Shwartz et al., 2009, 2007), optimization (Nemirovski et al., 2009; Nesterov and Vial, 2008), and stochastic approximation (Kushner and Yin, 2003; Polyak and Juditsky, 1992; Ruppert, 1988). The most widely used algorithms are stochastic gradient descent (SGD), a.k.a. Robbins-Monro algorithm (Robbins and Monro, 1951), and some of its modifications based on averaging of the iterates (Polyak and Juditsky, 1992; Rakhlin et al., 2011; Shamir and Zhang, 2013).

While the choice of the step-size may be done robustly in the deterministic case (see, e.g., Bertsekas, 1995), this remains a traditional theoretical and practical issue in the stochastic case. Indeed, early work suggested to use step-size decaying with the number k of iterations as $O(1/k)$ (Robbins and Monro, 1951), but it appeared to be non-robust to ill-conditioning and slower decays such as $O(1/\sqrt{k})$ together with averaging lead to both good practical and theoretical performance (Bach, 2014).

We consider in this chapter constant step-size SGD, which is often used in practice. Although the algorithm is not converging in general to the global optimum of the objective function, constant step-sizes come with benefits: (a) there is single parameter value to set as opposed to the several choices of parameters to deal with decaying step-sizes, e.g., as $1/(\square k + \triangle)^\circ$; the initial conditions are forgotten exponentially fast for well-conditioned (e.g., strongly convex) problems (Nedić and Bertsekas, 2001; Needell et al., 2014), and the performance, although not optimal, is sufficient in practice (in a machine learning set-up, being only 0.1% away from the optimal prediction often does not matter).

The main goals of this chapter are (a) to gain a complete understanding of the properties of constant-step-size SGD in the strongly convex case, and (b) to propose provable improvements to get closer to the optimum when precision matters or in high-dimensional settings. We consider the iterates of the SGD recursion on $\mathbb{R}^d$ defined starting from $\theta_0 \in \mathbb{R}^d$, for $k \geqslant 0$, and a step-size $\gamma > 0$ by

$$\theta_{k+1}^{(\gamma)} = \theta_k^{(\gamma)} - \gamma\big[f'(\theta_k^{(\gamma)}) + \varepsilon_{k+1}(\theta_k^{(\gamma)})\big] \, , \tag{4.1}$$

where f is the objective function to minimize (in machine learning the generalization performance), $\varepsilon_{k+1}(\theta_k^{(\gamma)})$ the zero-mean statistically independent noise (in machine learning, obtained from a single i.i.d. observation of a data point). Following Bach and Moulines (2013), we leverage the property that the sequence of iterates $(\theta_k^{(\gamma)})_{k\geqslant 0}$ is an *homogeneous Markov chain*.

This interpretation allows us to capture the general behavior of the algorithm. In the strongly convex case, this Markov chain converges exponentially fast to its unique stationary distribution π_γ (see Section 4.3.1) highlighting the facts that (a) initial conditions of the algorithms are forgotten quickly and (b) the algorithm does not converge to a point but oscillates around the mean of π_γ. See an illustration in Figure 4.1 (left). It is known that the oscillations of the non-averaged iterates have an average magnitude of $\gamma^{1/2}$ (Pflug, 1986).

Consider the average process $(\bar{\theta}_k^{(\gamma)})_{k\geqslant 0}$ given for all $k \geqslant 0$ by

$$\bar{\theta}_k^{(\gamma)} = \frac{1}{k+1}\sum_{j=0}^{k} \theta_j^{(\gamma)} \,. \tag{4.2}$$

Then under appropriate conditions on the Markov chain $(\theta_k^{(\gamma)})_{k\geqslant 0}$, a central limit theorem on $(\bar{\theta}_k^{(\gamma)})_{k\geqslant 0}$ holds which implies that $\bar{\theta}_k^{(\gamma)}$ converges at rate $O(1/\sqrt{k})$ to

$$\bar{\theta}_\gamma = \int_{\mathbb{R}^d} \vartheta \, \mathrm{d}\pi_\gamma(\vartheta) \,. \tag{4.3}$$

The deviation between $\bar{\theta}_k^{(\gamma)}$ and the global optimum θ_* is thus composed of a stochastic part $\bar{\theta}_k^{(\gamma)} - \bar{\theta}_\gamma$ and a deterministic part $\bar{\theta}^{(\gamma)} - \theta_*$.

For quadratic functions, it turns out that the deterministic part vanishes (Bach and Moulines, 2013), that is, $\bar{\theta}_{(\gamma)} = \theta_*$ and thus averaged SGD with a constant step-size does converge. However, it is not true for general objective functions where we can only show that $\bar{\theta}_\gamma - \theta_* = O(\gamma)$, and this deviation is the reason why constant step-size SGD is not convergent.

The first main contribution of the chapter is to provide an explicit asymptotic expansion that highlights all dependencies on initial conditions and noise variance, as achieved for least-squares by Défossez and Bach (2015), with an explicit decomposition into "bias" and "variance" terms: the bias term characterizes how fast initial conditions are forgotten and thus is increasing in a well-chosen norm of $\theta_0 - \theta_*$; while the variance term characterizes the effect of the noise in the gradient, independently of the starting point, and increases with the covariance of the noise.

Moreover, akin to weak error results for ergodic diffusions, we achieve a non-asymptotic weak error expansion in the step-size between π_γ and the Dirac at θ_*. Namely, we prove that for all functions $g : \mathbb{R}^d \to \mathbb{R}$, regular enough, $\int_{\mathbb{R}^d} g(\theta)\mathrm{d}\pi_\gamma(\theta) = g(\theta_*) + \gamma C + O(\gamma^2)$ for some $C \in \mathbb{R}$ independent of γ. Given this expansion, we can now use a very simple trick from numerical analysis, namely Richardson-Romberg extrapolation (Stoer and Bulirsch, 2013): if we run two SGD recursions $(\theta_k^{(\gamma)})_{k\geqslant 0}$ and $(\theta_k^{(2\gamma)})_{k\geqslant 0}$ with the two different step-sizes γ and 2γ, then the averaged iterates $(\bar{\theta}_k^{(\gamma)})_{k\geqslant 0}$ and $(\bar{\theta}_k^{(2\gamma)})_{k\geqslant 0}$ will converge to $\bar{\theta}_\gamma$ and $\bar{\theta}_{2\gamma}$ respectively. Since $\bar{\theta}_\gamma = \theta_* + \Delta\gamma + O(\gamma^2)$ and $\bar{\theta}_{2\gamma} = \theta_* + 2\Delta\gamma + O(\gamma^2)$, for $\Delta \in \mathbb{R}^d$ independent of γ, the combined iterate $2\bar{\theta}_k^{(\gamma)} - \bar{\theta}_k^{(2\gamma)}$ will converge to a point which is $\theta_* + O(\gamma^2)$ and we have thus gained one order in the convergence rate. See illustration in Figure 4.1(right).

In summary, we make the following contributions:

- We provide in Section 4.2 an asymptotic expansion of the mean of the averaged SGD iterate that outlines the dependence on initial conditions, the effect of noise and the step-size.
- We show in Section 4.2 that Richardson-Romberg extrapolation may be used to get closer to the global optimum.
- We bring and adapt in Section 4.3 tools from analysis of discretization of diffusion processes into the one of SGD and create new ones. We believe that this analogy and the associated ideas are interesting in their own right.
- We show in Section 4.4 empirical improvements of the extrapolation schemes.

Proofs are given in Chapter C.

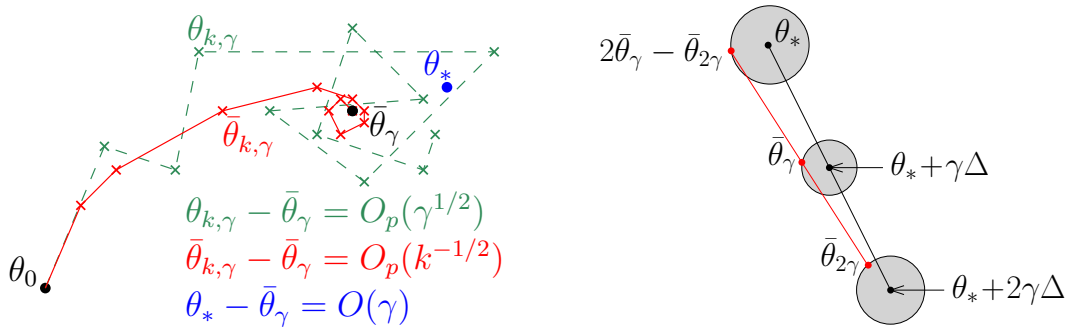


Figure 4.1: (Left) Convergence of iterates $\theta_k^{(\gamma)}$ and averaged iterates $\bar{\theta}_k^{(\gamma)}$ to the mean $\bar{\theta}^{(\gamma)}$ under the stationary distribution π_γ. (Right) Richardson-Romberg extrapolation, the disks are of radius $O(\gamma^2)$.

4.2 Main results

In this section, we describe the assumptions underlying our analysis and give our main results.

4.2.1 Setting

Let $f : \mathbb{R}^d \to \mathbb{R}$ be an objective function, satisfying the following assumptions:

A1. *The function f is strongly convex with convexity constant μ, i.e., $f - \frac{\mu}{2}\|\cdot\|^2$ is convex.*

A2. *The function f is four times continuously differentiable with uniformly second to fourth bounded derivatives. Especially f is L-smooth: $\forall \theta \in \mathbb{R}^d$, the largest eigenvalue of $f''(\theta)$ is less than L.*

If there exists a positive definite matrix $\Sigma \in \mathbb{R}^{d\times d}$, such that the function f is a quadratic function $f_\Sigma : \theta \mapsto \|\Sigma^{1/2}(\theta - \theta_*)\|^2$, then Assumptions **A1**, **A2** are satisfied.

In the definition of SGD given by (4.1), $(\varepsilon_k)_{k\geqslant 1}$ is a sequence of random functions from $\mathbb{R}^d$ to $\mathbb{R}$ satisfying the following properties.

A3. *There exists a filtration $(\mathcal{F}_k)_{k\geqslant 0}$ (i.e., for all $k \in \mathbb{N}$, $\mathcal{F}_k \subset \mathcal{F}_{k+1}$) on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that for any $k \in \mathbb{N}$, for any $\theta \in \mathbb{R}^d$, $\varepsilon_{k+1}(\theta)$ is an $\mathcal{F}_{k+1}$-measurable random variable and $\mathbb{E}[\varepsilon_{k+1}(\theta)|\mathcal{F}_k] = 0$. In addition, $(\varepsilon_k)_{k\in\mathbb{N}^*}$ are independent and identically distributed (i.i.d.) random fields.*

A3 expresses that we observe a noisy gradient $f'_{k+1}(\theta_k^{(\gamma)}) = f'(\theta_k^{(\gamma)}) - \varepsilon_{k+1}(\theta_k^{(\gamma)})$ which are unbiased estimator of f'. Note that we do not assume that the random vectors $(\varepsilon_{k+1}(\theta_k^{(\gamma)}))_{k\in\mathbb{N}}$ are i.i.d., a stronger assumption generally referred to as the semi-stochastic setting. We also consider the following conditions on the noise, for $p \geqslant 2$:

A4 (p). *For any $k \in \mathbb{N}^*$, f'_k is almost surely L-co-coercive (with the same constant as in **A2**): for any $\eta, \theta \in \mathbb{R}^d$, $L\langle f'_k(\theta) - f'_k(\eta), \theta - \eta\rangle \geqslant \|f'_k(\theta) - f'_k(\eta)\|^2$. Moreover, for any $k \in \mathbb{N}^*$, there exists $\tau_p \geqslant 0$, such that $\varepsilon_k(\theta_*)$ admits bounded moments up to the order p: $\mathbb{E}^{1/p}[\|\varepsilon_k(\theta_*)\|^p] \leqslant \tau_p$.*

Almost sure L-co-coercivity (Zhu and Marcotte, 1996) is for example satisfied if for any $k \in \mathbb{N}^*$, there exist a random function f_k (such that $f'_k = (f_k)'$) which is a.s. convex and

L-smooth. Note that weaker assumptions could be made on the noise (see Section C.1.3 for a discussion). Also note that we could remove the "for any $k \in \mathbb{N}^*$" quantification in Assumption **A**4, and only make the assumption for $k = 1$: as the functions are already assumed to be i.i.d., the assumption for $k = 1$ is equivalent to the assumption for all $k \in \mathbb{N}^*$.

Learning from i.i.d. observations. Our main motivation comes from machine learning; namely, we consider sets $\mathcal{X}, \mathcal{Y}$, a convex loss function $\ell : \mathcal{X} \times \mathcal{Y} \times \mathbb{R}^d \to \mathbb{R}$. The objective function is the generalization error $f_\ell(\theta) = \mathbb{E}_{X,Y}[\ell(X, Y, \theta)]$. For any $k \geqslant 1$, we define $\varepsilon_k(\theta) = \ell(x_k, y_k, \theta) - f_\ell(\theta)$ which corresponds to following the negative gradient of a single i.i.d. observation $(x_k, y_k)_{k \geqslant 1}$; Assumption **A**3 is then satisfied with $\mathcal{F}_k := \sigma((x_j, y_j)_{1 \leqslant j \leqslant k})$.

Two classical situations are worth mentioning: in *least-squares regression*, $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \mathbb{R}$, and the loss function is $\ell(X, Y, \theta) = (\langle X, \theta \rangle - Y)^2$. Then f_ℓ is a quadratic function f_Σ, with $\Sigma = \mathbb{E}[XX^\top]$, thus satisfies Assumption **A**2. For any $p \geqslant 2$, Assumption **A**4(p) is satisfied as soon as the iterates are a.s. bounded, while **A**1 is satisfied if the second moment matrix is invertible or additional regularization is added. In this setting, ε_k can be decomposed as $\varepsilon_k = \varrho_k + \xi_k$ where ϱ_k is the multiplicative part, ξ_k the additive part, given for $\theta \in \mathbb{R}^d$ by $\varrho_k(\theta) = (x_k x_k^\top - \Sigma)(\theta - \theta_*)$ and

$$\xi_k = (x_k^\top \theta_* - y_k) x_k \, . \tag{4.4}$$

Note that for all $k \geqslant 1$, ξ_k does not depend on θ. This two parts in the noise will appear in Corollary 4.5. In *logistic regression*, where $\ell(X, Y, \theta) = \log(1 + \exp(-Y \langle X, \theta \rangle))$. Assumptions **A**4 or **A**2 are similarly satisfied, while **A**1 needs an additional restriction to a compact set. Using self-concordance assumptions (Bach, 2014) would allow a direct unconstrained application.

4.2.2 Related work

Constant step-size SGD. Several attempts have been made to improve convergence of SGD. Bach and Moulines (2013) propose an online Newton algorithm which converges to the optimal point with constant steps. While it behaves very well in practice, this algorithm has no convergence guarantees.

The quadratic case was studied by Bach and Moulines (2013), for the (uniform) average iterate: the variance term is upper bounded by $\sigma^2 d/n$ and the squared bias term by $\|\theta_*\|^2/(\gamma n)$. This last term was improved to $\|\Sigma^{-1/2}\theta_*\|^2/(\gamma n)^2$ in Chapter 2 and by Défossez and Bach (2015). See also (Lan, 2012). Analysis has been extended to "tail averaging" (Jain et al., 2016), to improve the dependence on the initial conditions. Note that this procedure can be seen as a Richardson-Romberg trick with respect to k. Other strategies were proposed to improve the speed at which initial conditions were forgotten, for example using acceleration when the noise is additive (as in Chapter 3; or in Jain et al., 2017).

Link between discretization of ergodic diffusions and SGD. In the context of discretization of ergodic diffusions, weak error estimates between the stationary distribution of the discretization and the invariant distribution of the associated diffusion have been first shown by Talay and Tubaro (1990) and Mattingly et al. (2002) in the case of the Euler-Maruyama discretization. Then Talay and Tubaro (1990) suggested the use of Richardson-Romberg interpolation to improve the accuracy of estimates of integrals with

respect to the invariant distribution of the diffusion. Extension of these results have been obtained for other types of discretization by Abdulle et al. (2014) and Chen et al. (2015). We show in Section 4.3.3 that a weak error expansion in the step size γ also holds for SGD between π_γ and δ_{θ_*}. Interestingly similarly to the Euler-Maruyama discretization, SGD has a weak error of order γ. Finally, Durmus et al. (2016) proposed and analyzed the use of Richardson-Romberg extrapolation applied to the stochastic gradient Langevin dynamics (SGLD) algorithm. This method introduced by Welling and Teh (2011) combines SGD and the Euler-Maruyama discretization of the Langevin diffusion associated to a target probability measure (see also Dalalyan (2014)). Note that this method is however completely different from SGD, in part because Gaussian noise of order $\gamma^{1/2}$ (instead of γ) is injected in SGD which changes the overall dynamics.

4.2.3 Summary and discussion of main results

Under the stated assumptions, the Markov chain $(\theta_k^{(\gamma)})_{k\geqslant 0}$ admits a unique invariant/stationary distribution π_γ which has a moment of order 2, see Theorem 4.4 in Section 4.3. Recall that π_γ is a stationary distribution of this Markov chain if, when $\theta_0^{(\gamma)}$ is distributed according to π_γ, then $\theta_1^{(\gamma)}$ is distributed according to π_γ as well. In the next section, by two different methods (Theorem 4.3 and Theorem 4.6), we show that under suitable conditions on f and the noise $(\varepsilon_k)_{k\geqslant 1}$, there exists $C \geqslant 0$ such that for all $\gamma \geqslant 0$, small enough

$$\bar{\theta}_\gamma = \int_{\mathbb{R}^d} \vartheta \pi_\gamma(\mathrm{d}\vartheta) = \theta_* + C\gamma + O(\gamma^2) \; .$$

Using Theorem 4.3, we get that for γ small enough and all $k \geqslant 1$,

$$\mathbb{E}(\bar{\theta}_k^{(\gamma)} - \theta_*) = \frac{A(\theta_0, \gamma)}{k} + C\gamma + O(\gamma^2) + O(e^{-k\mu\gamma}) \; . \tag{4.5}$$

This expansion in the step size γ shows that a Richardson-Romberg extrapolation can be used to have better estimates of θ_*. Consider the average iterates $(\bar{\theta}_{2\gamma}^{(k)})_{k\geqslant 0}$ and $(\bar{\theta}_k^{(\gamma)})_{k\geqslant 0}$ associated with SGD with step size 2γ and γ respectively. Then (4.5) shows that $(2\bar{\theta}_k^{(\gamma)} - \bar{\theta}_k^{(2\gamma)})_{k\geqslant 0}$ satisfies

$$\mathbb{E}(2\bar{\theta}_k^{(\gamma)} - \bar{\theta}_k^{(2\gamma)} - \theta_*) = \frac{A(\theta_0, \gamma) - A(\theta_0, 2\gamma)}{k} + O(\gamma^2) + O(e^{-k\mu\gamma}) \; ,$$

and therefore is closer to the optimum θ_*. This very simple trick improves the convergence by a factor of γ (at the expense of a slight increase of the variance). In practice, while the un-averaged gradient iterate $\theta_k^{(\gamma)}$ saturates rapidly, $\bar{\theta}_k^{(\gamma)}$ may already perform well enough to avoid saturation on real data-sets (Bach and Moulines, 2013). The Richardson-Romberg extrapolated iterate $2\bar{\theta}_k^{(\gamma)} - \bar{\theta}_k^{(2\gamma)}$ very rarely reaches saturation in practice. This appears in synthetic experiments presented in Section 4.4. Moreover, this procedure only requires to compute two parallel SGD recursions, either with the same inputs, or with different ones, and is naturally parallelizable.

In Section 4.3.2, we give a quantitative version of the central limit theorem for a fixed $\gamma > 0$ and k goes to $+\infty$ for $(\bar{\theta}_k^{(\gamma)})_{k\geqslant 0}$, *i.e.*, under appropriate conditions, there exist $B_1(\gamma)$ and $B_2(\gamma)$ such that

$$\mathbb{E}\left[\left\|\bar{\theta}_k^{(\gamma)} - \bar{\theta}_\gamma\right\|^2\right] = \frac{B_1(\gamma)}{k} + \frac{B_2(\gamma)}{k^2} \; . \tag{4.6}$$

Combining (4.5) and (4.6) characterizes the bias/variance trade-off of SGD used to estimate θ_*.

4.3 Detailed analysis

In this Section, we describe in detail our approach. A first step is to describe the existence of a unique stationary distribution π_γ for the Markov chain $(\theta_k^{(\gamma)})_{k\geqslant 0}$ and the convergence of this Markov chain to π_γ. The convergence is quantified with the Wasserstein distance (see e.g., Chapter 6 in Villani, 2009).

Limit distribution. A fundamental tool in Markov chain theory is the *Markov kernel*, which is the equivalent for continuous spaces of the *transition matrix* in finite state spaces. Let R_γ be the Markov kernel on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ associated with the SGD iterates $(\theta_k^{(\gamma)})_{k\geqslant 0}$, where $\mathcal{B}(\mathbb{R}^d)$ is the Borel σ-field of $\mathbb{R}^d$. We refer to (Meyn and Tweedie, 2009) for an introduction to Markov chain theory. For all initial distributions ν_0 on $\mathcal{B}(\mathbb{R}^d)$ and $k \in \mathbb{N}$, $\nu_0 R_\gamma^k$ denotes the law of $\theta_k^{(\gamma)}$ starting at θ_0 distributed according to ν_0. For any measure π on $\mathcal{B}(\mathbb{R}^d)$ and any measurable function $h : \mathbb{R}^d \to \mathbb{R}$, $\pi(h)$ denotes $\int h(\theta) d\pi(\theta)$ when it exists. Finally, for all $\theta \in \mathbb{R}^d$ and measurable function $h : \mathbb{R}^d \to \mathbb{R}$, $k \geqslant 1$, set $R_\gamma^k(\theta, \cdot) = \delta_\theta R_\gamma^k$ the distribution of $\theta_k^{(\gamma)}$ starting at θ and $R_\gamma^k h(\theta) = \int_{\mathbb{R}^d} h(\vartheta) \left\{\delta_\theta R_\gamma^k\right\}(\mathrm{d}\vartheta)$.

To show that $(\theta_k^{(\gamma)})_{k\geqslant 0}$ admits a unique stationary distribution π_γ and quantify the convergence of $(\nu_0 R_\gamma^k)_{k\geqslant 0}$ to π_γ, we introduce the Wasserstein distance. For all probability measures ν and λ on $\mathcal{B}(\mathbb{R}^d)$, such that $\int_{\mathbb{R}^d} \|\theta\|^2 \,\mathrm{d}\nu(\theta) < +\infty$ and $\int_{\mathbb{R}^d} \|\theta\|^2 \,\mathrm{d}\lambda(\theta) \leqslant +\infty$, define the Wasserstein distance of order 2 between λ and ν by $W_2(\lambda, \nu) := \inf_{\xi \in \Pi(\lambda,\nu)} \left(\int \|x - y\|^2 \xi(dx, dy) \right)^{1/2}$, where $\Pi(\mu, \nu)$ is the set of probability measure ξ on $\mathcal{B}(\mathbb{R}^d \times \mathbb{R}^d)$ satisfying for all $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$, $\xi(\mathsf{A} \times \mathbb{R}^d) = \nu(\mathsf{A})$, $\xi(\mathbb{R}^d \times \mathsf{A}) = \lambda(\mathsf{A})$.

Proposition 4.1. *Assume A1-A2-A3-A4(2), for any step size $\gamma < L^{-1}$, the Markov chain $(\theta_k^{(\gamma)})_{k\geqslant 0}$ defined by the recursion (4.1), admits a unique stationary distribution π_γ such that $\int_{\mathbb{R}^d} \|\vartheta\|^2 \,\mathrm{d}\pi_\gamma(\vartheta) < +\infty$. In addition for all $\theta \in \mathbb{R}^d$, $k \in \mathbb{N}$:*

$$W_2^2(R_\gamma^k(\theta, \cdot), \pi_\gamma) \leqslant (1 - 2\mu\gamma(1 - \gamma L))^k \int_{\mathbb{R}^d} \|\theta - \vartheta\|^2 \,\mathrm{d}\pi_\gamma(\vartheta) \,.$$

Proof. The proof is postponed to Section C.2.1. □

To prove the existence of the limit, one shows that for any x, $(R_\gamma^k(x, \cdot))_{k\geqslant 0}$ is a Cauchy sequence in a particular Polish space. We can thus define a point-wise limit, and show that it is unique. This uses the strong convexity, smoothness and the Lipschitzness of the noise.

As a consequence of Proposition 4.1, the expectation of $\bar{\theta}_k^{(\gamma)} = \frac{1}{k+1} \sum_{i=0}^k \theta_i^{(\gamma)}$ converges $\int_{\mathbb{R}^d} \vartheta \mathrm{d}\pi_\gamma(\vartheta)$ as k goes to infinity at a rate of order $O(k^{-1})$, see Theorem C.8 in Section C.3.

4.3.1 Expansion of moments under π_γ when γ is in a neighborhood of 0

In this paragraph, we analyze the properties of the chain starting at θ_0 distributed according to π_γ. As a result, we prove that the mean of the stationary distribution $\bar{\theta}_\gamma = \int_{\mathbb{R}^d} \vartheta \pi_\gamma(\mathrm{d}\vartheta)$ is such that $\bar{\theta}_\gamma = \theta_* + O(\gamma)$. By simple developments of Equation (4.1) at the equilibrium, we propose expansions of the first two moments of the chain. It extends (Pflug, 1986; Ljung et al., 1992) which showed that $(\gamma^{-1/2}(\pi_\gamma - \delta_{\theta_*}))_{\gamma>0}$ converges in distribution to a normal law as $\gamma \to 0$.

Quadratic case. When f is a quadratic function, *i.e.*, f' is affine, we have the following result.

Lemma 4.2 (Properties under stationarity, Quadratic case)**.** *Let $\gamma < 1/L$ and assume A1-A2-A3-A4(4). Then for a quadratic function $f_\Sigma : \theta \mapsto \left\|\Sigma^{1/2}(\theta - \theta_*)\right\|^2$,*

$$\begin{aligned} \bar{\theta}_\gamma &= \mathbb{E}_{\pi_\gamma}[\theta] = \theta_* \\ \int_{\mathbb{R}^d} (\theta - \theta_*)^{\otimes 2} \pi_\gamma(\mathrm{d}\theta) &= \gamma(\Sigma \otimes I + I \otimes \Sigma - \gamma \Sigma \otimes \Sigma)^{-1} \mathbb{E}_{\varepsilon_1}\left[\int_{\mathbb{R}^d} \varepsilon_1(\theta)^{\otimes 2} \pi_\gamma(\mathrm{d}\theta)\right], \end{aligned}$$

where we denote, for any $\theta \in \mathbb{R}^d$, $\theta^{\otimes 2} := \theta\theta^\top$, where for any matrices $M, N \in \mathbb{R}^{d\times d}$, $M \otimes N$ is defined as the following operator from $\mathbb{R}^{d\times d}$ into $\mathbb{R}^{d\times d}$ such that $M \otimes N : P \mapsto MPN$.

The first part of the result, which highlights the crucial fact that for a quadratic function, the mean under the limit distribution is the optimal point, is easy to prove. Indeed, since π_γ is invariant for $(\theta_k^{(\gamma)})_{k\geqslant 0}$, if $\theta_0^{(\gamma)}$ is distributed according to π_γ, then $\theta_1^{(\gamma)}$ is distributed according to π_γ as well. Thus as $\theta_1^{(\gamma)} = \theta_0^{(\gamma)} - \gamma f'(\theta_0^{(\gamma)}) + \gamma\varepsilon_1(\theta_0^{(\gamma)})$ taking expectations on both sides, we get $\int_{\mathbb{R}^d} f'(\vartheta)\mathrm{d}\pi_\gamma(\vartheta) = 0$. For a quadratic function, its gradient is linear: $\int_{\mathbb{R}^d} f'(\vartheta)\mathrm{d}\pi_\gamma(\vartheta) = f'(\bar{\theta}_\gamma) = 0$ and thus that $\bar{\theta}_\gamma = \theta_*$. This implies that the averaged iterate converges to θ_*, see e.g. Bach and Moulines (2013). The proof for the second expression is given in Section C.2.3.

General case. While the quadratic case led to particularly simple exact expressions, in general, we can only get a first order development of these expectations as $\gamma \to 0$ (proofs are given in Section C.2.3). Note that it improved on (Pflug, 1986), which shows a similar expansion but an error of order of $O(\gamma^{3/2})$.

Theorem 4.3 (Properties under stationarity, general case)**.** *Let $\gamma < 1/L$ and assume A1-A2-A3-A4(4). Then*

$$\begin{aligned} \bar{\theta}_\gamma - \theta_* &= \gamma f''(\theta_*)^{-1} f'''(\theta_*)\Big(\big[f''(\theta_*) \otimes \mathrm{I} + \mathrm{I} \otimes f''(\theta_*)\big]^{-1} \mathbb{E}_\varepsilon\left[\int_{\mathbb{R}^d} \varepsilon(\theta)^{\otimes 2} \pi_\gamma(\mathrm{d}\theta)\right]\Big) + O(\gamma^2) \\ \int_{\mathbb{R}^d} (\theta - \theta_*)^{\otimes 2} \pi_\gamma(\mathrm{d}\theta) &= \gamma\big[f''(\theta_*) \otimes \mathrm{I} + \mathrm{I} \otimes f''(\theta_*)\big]^{-1} \mathbb{E}_\varepsilon\left[\int_{\mathbb{R}^d} \varepsilon(\theta)^{\otimes 2} \pi_\gamma(\mathrm{d}\theta)\right] + O(\gamma^2)\,, \end{aligned}$$

where π_γ is the stationary distribution of the Markov chain $(\theta_k^{(\gamma)})_{k\geqslant 0}$ defined by the recursion (4.1) and $\bar{\theta}_\gamma$ is given by (4.3). We denote $f'''(\theta_*)$ the third order derivative, which is a third order tensor (thus such that for any matrix, $M \in \mathbb{R}^{d\times d}$, $f'''(\theta_*)M$ is a vector in $\mathbb{R}^d$ such that $(f'''(\theta_*)M)_k = \sum_{i,j=1}^d M_{i,j} \frac{\partial^3 f}{\partial x_i \partial x_j \partial x_k}$, for any $k \in [\![1; n]\!]$).

Proof. The proof is postponed to Section C.2.3. □

This shows that $\gamma \mapsto \bar{\theta}_\gamma$ is a differentiable function at $\gamma = 0$. The "drift" $\bar{\theta}_\gamma - \theta_*$ can be understood as an additional error occurring because the function is non quadratic and the step sizes are not decaying to zero. The mean under the limit distribution is at distance γ from θ_* while the final iterate oscillates in a sphere of radius proportional to $\sqrt{\gamma}$, as $\int_{\mathbb{R}^d} \|\theta - \theta_*\| \pi_\gamma(\mathrm{d}\theta) \leqslant \sqrt{\gamma}\,\mathrm{tr}^{1/2}(\big[f''(\theta_*) \otimes \mathrm{I} + \mathrm{I} \otimes f''(\theta_*)\big]^{-1} \int_{\mathbb{R}^d} \varepsilon(\theta)^{\otimes 2} \pi_\gamma(\mathrm{d}\theta))$, where for any matrix $M \in R^{d\times d}$, $\mathrm{tr}(M)$ is the trace of M, *i.e.*, the sum of diagonal elements of the matrix M.

4.3.2 Expansion for a given $\gamma > 0$ when k tends to $+\infty$

In this Section, we analyze the convergence of $\bar{\theta}_k^{(\gamma)}$ to $\bar{\theta}_\gamma$, when $k \to \infty$, and the convergence of $\mathbb{E}\left[\left\|\bar{\theta}_k^{(\gamma)} - \bar{\theta}_\gamma\right\|^2\right]$ to 0. Under suitable conditions (Meyn and Tweedie, 1993; Jones, 2004), $\bar{\theta}_k^{(\gamma)}$ satisfies a central limit theorem: $\sqrt{k}\left(\bar{\theta}_k^{(\gamma)} - \bar{\theta}_\gamma\right) \xrightarrow{d} \mathcal{N}(0, \sigma_\varphi^2)$, where $\sigma_\varphi^2 \geqslant 0$. However, this result is purely asymptotic; we propose a new tighter development that describes how the initial conditions are forgotten: we prove that the convergence behaves similarly to the convergence in the quadratic case, where the expected squared distance decomposes as a sum of a bias term, that scales as k^{-2}, and a variance term, that scales as k^{-1}, plus linearly decaying residual terms. We also describe how the asymptotic bias and variance can be expressed easily as moments of solutions to several *Poisson equations*.

Poisson equation. For any Lipschitz function $\varphi : \mathbb{R}^d \to \mathbb{R}$, the convergence speed of $k^{-1}\sum_{i=0}^{k-1}\varphi(\theta_i^{(\gamma)})$ towards $\int_{\mathbb{R}^d}\varphi(\vartheta)\mathrm{d}\pi_\gamma(\vartheta)$ can be decomposed as a sum of two main terms, that can be expressed as moments of two Poisson solutions associated with φ which we now described. It shows in Section C.2.2 that the sequence of functions $\{\theta \mapsto \sum_{i=1}^k R_\gamma^i\phi(\theta) - \pi_\gamma(\phi)\}_{k\geqslant 0}$ converges uniformly on all compact sets of $\mathbb{R}^d$. Define then $\psi_\gamma = \sum_{i=0}^{+\infty}\{R_\gamma^i\phi - \pi_\gamma(\phi)\}$. Note that ψ_γ satisfies $\pi_\gamma(\psi_\gamma) = 0$, $(I - R_\gamma)\psi_\gamma = \varphi$ and is Lipschitz, see Section C.2.2. ψ_γ will be referred to as the Poisson solution associated with φ.

For the convergence of $\bar{\theta}_k^{(\gamma)}$ to $\bar{\theta}_\gamma$, we thus introduce ψ_γ, the Poisson solution associated to $\varphi : \theta \mapsto \theta - \theta_*$, χ_γ^1 the Poisson solution associated to $\theta \mapsto \psi_\gamma(\theta)\psi_\gamma^\top(\theta)$, and finally χ_γ^2 the Poisson solution associated to $\theta \mapsto ((\psi_\gamma - \varphi)(\theta))^{\otimes 2}$. We then have:

Theorem 4.4 (Convergence of the Markov chain)**.** *Let $\gamma \in (0, 1/(2L))$ and assume A1-A2-A3-A4(4). Then for any starting point $\theta_0 \in \mathbb{R}^d$, setting $\rho := (1 - \gamma\mu)^{1/2}$:*

$$\begin{aligned}
\mathbb{E}\left[\bar{\theta}_k^{(\gamma)} - \bar{\theta}_\gamma\right] &= (1/k)\psi_\gamma(\theta_0) + O(\rho^k)\,, \\
\mathbb{E}\left[\left(\bar{\theta}_k^{(\gamma)} - \bar{\theta}_\gamma\right)^{\otimes 2}\right] &= (1/k)\int_{\mathbb{R}^d}\left[\psi_\gamma(\theta)\psi_\gamma(\theta)^\top - (\psi_\gamma - \varphi)(\theta)(\psi_\gamma - \varphi)(\theta)^\top\right]\mathrm{d}\pi_\gamma(\theta) \\
&\quad + (1/k^2)\left[\psi_\gamma(\theta_0)\psi_\gamma(\theta_0)^\top + \chi_\gamma^1(\theta_0) - \chi_\gamma^2(\theta_0)\right] + O(\rho^k)\,,
\end{aligned}$$

where $(\bar{\theta}_k^{(\gamma)})_{k\geqslant 0}$ is given by (4.2) and π_γ is its unique stationary distribution of the Markov chain defined by the recursion (4.1).

Proof. This result is a consequence of Theorem C.5, proved in Section C.2.4. □

This bound for the second order moment decomposes as a sum of two terms: (i) a variance term, that scales as $1/k$, and does not depend on the initial distribution (but only on the asymptotic distribution π_γ), and (ii) a bias term, which scales as $1/k^2$, and depends on the initial distribution ν_0.

In order to give the intuition of the proof and to underline how the associated Poisson solutions are introduced, we here sketch the proof of the first result:

$$\begin{aligned}
\mathbb{E}\left[\bar{\theta}_k^{(\gamma)}\right] - \theta_* &= \frac{1}{k}\sum_{i=0}^{k-1}(R_\gamma^i\varphi)(\theta_0) \\
&= \pi_\gamma\varphi + \frac{1}{k}\psi_\gamma(\theta_0) + R_\gamma^k\psi_\gamma(\theta_0),
\end{aligned}$$

where we have used $R_\gamma^i \pi_\gamma(\varphi) = \pi_\gamma \varphi$, and

$$\begin{aligned}\sum_{i=0}^{k-1} R_\gamma^i(\varphi - \pi_\gamma(\varphi)) &= \sum_{i=0}^{\infty} R_\gamma^i(\varphi - \pi_\gamma(\varphi)) - R_\gamma^k \sum_{i=0}^{\infty} R_\gamma^i(\varphi - \pi_\gamma(\varphi)) \\ &= \psi_\gamma - R_\gamma^k \psi_\gamma\end{aligned}$$

Finally, we have that $R_\gamma^k \psi_\gamma(\theta_0)$ converges to 0 at linear speed, using Proposition 4.1.

This result gives an exact closed form for the asymptotic bias and variance, for a fixed γ, and as $k \to \infty$. Unfortunately, in the general case, it is neither possible to compute the Poisson solutions exactly, nor is it possible to prove a first order development of the limits as $\gamma \to 0$. Indeed, part of the difficulty comes from the fact that as γ goes to zero, the Markov chain does not mix fast enough.

When f_Σ is a quadratic function, it is possible, for any $\gamma > 0$, to compute ψ_γ and $\chi_\gamma^{1,2}$ explicitly; we get the following decomposition of the error, which exactly recovers the result of Défossez and Bach (2015).

Corollary 4.5. *Assume that f is a quadratic function f_Σ, **A**3 and **A**4(4). Consider the least mean squares algorithm iterates $(\theta_k^{(\gamma)})_{k \geqslant 0}$ starting from $\theta_0 \in \mathbb{R}^d$ with $\gamma L \leqslant 1/2$. Then*

$$\begin{aligned}\mathbb{E}\left[(\bar{\theta}_k^{(\gamma)} - \theta_*)^{\otimes 2}\right] &= \frac{1}{k^2\gamma^2}\Sigma^{-1}\Omega(\theta_0 - \theta_*)^{\otimes 2}\Sigma^{-1} + \frac{1}{k}\Sigma^{-1}\left[\mathbb{E}_{\varepsilon_1,\pi_\gamma}(\varepsilon_1^{\otimes 2}(\theta))\right]\Sigma^{-1} \\ &\quad - \frac{1}{k^2\gamma}\Sigma^{-1}\Omega\left[\Sigma \otimes \mathrm{I} + \mathrm{I} \otimes \Sigma - \gamma T\right]^{-1}\left[\mathbb{E}\xi_1^{\otimes 2}\right]\Sigma^{-1} + O(\rho^k)\,,\end{aligned}$$

where $\rho = (1 - \gamma\mu)^{1/2}$, $\Omega := (\Sigma \otimes I + I \otimes \Sigma - \gamma\Sigma \otimes \Sigma)(\Sigma \otimes I + I \otimes \Sigma - \gamma T)^{-1}$, $T : A \mapsto \mathbb{E}\left[(x^\top A x)xx^\top\right]$ and ξ_1 is given by (4.4).

4.3.3 Continuous interpretation of SGD and weak error expansion

In this section, we propose a new decomposition of $\bar{\theta}_k^{(\gamma)} - \theta_*$. It is comparable to the decomposition used in classical proofs, e.g., by Nemirovsky and Yudin (1983), that we recall bellow, yet is more powerful to understand the behavior when γ is small. It uses the link with the continuous interpretation of SGD.

For the sake of comparison, we first informally present the two decompositions, then introduce the necessary tools and assumptions rigorously.

Two decompositions. For smooth and strongly convex functions, classical proofs of the convergence of SGD rely on the following decomposition (Nemirovsky and Yudin, 1983; Bach and Moulines, 2011), which comes from a Taylor expansion of $f'(\theta_{k+1}^{(\gamma)})$ around θ_*. For any $k \in \mathbb{N}$,

$$f'(\theta_k^{(\gamma)}) = f''(\theta_*)(\theta_k^{(\gamma)} - \theta_*) + O\left(\left\|\theta_k^{(\gamma)} - \theta_*\right\|^2\right).$$

As a consequence, using the definition of the SGD recursion,

$$\begin{aligned}\theta_{k+1}^{(\gamma)} - \theta_k^{(\gamma)} &= -\gamma f'(\theta_k^{(\gamma)}) - \gamma\varepsilon_{k+1}(\theta_k^{(\gamma)}) \\ &= -\gamma f''(\theta_*)(\theta_k^{(\gamma)} - \theta_*) - \gamma\varepsilon_{k+1}(\theta_k^{(\gamma)}) + \gamma O\left(\left\|\theta_k^{(\gamma)} - \theta_*\right\|^2\right)\end{aligned}$$

Thus

$$f''(\theta_*)(\theta_k^{(\gamma)} - \theta_*) = \frac{1}{\gamma}(-\theta_{k+1}^{(\gamma)} + \theta_k^{(\gamma)}) - \varepsilon_{k+1}(\theta_k^{(\gamma)}) + O\left(\left\|\theta_k^{(\gamma)} - \theta_*\right\|^2\right)$$

Averaging over the first k iterates yields:

$$(k+1)(\bar{\theta}_k^{(\gamma)} - \theta_*) = \frac{1}{\gamma} f''(\theta_*)^{-1}(\theta_0^{(\gamma)} - \theta_{k+1}^{(\gamma)}) - \sum_{i=0}^{k} f''(\theta_*)^{-1}\varepsilon_{i+1}(\theta_i^{(\gamma)}) + \sum_{i=0}^{k} O\left(\left\|\theta_i^{(\gamma)} - \theta_*\right\|^2\right). \tag{4.7}$$

The term on the right-hand part of Equation 4.7 is composed of a bias term (which clearly depends on the initial condition), a variance term which is the average of noise, and a residual term. This residual term is of course highly important, as it differentiates the general setting from the quadratic one (in which it simply does not appear, as the first order Taylor expansion of f' is exact). While this decomposition is powerful and has been used in many proofs, it does not allow for a tight decomposition in powers of γ when $\gamma \to 0$, because the residual $\theta_i^{(\gamma)} - \theta_*$ simply does not go to 0 when $\gamma \to 0$: on the contrary, the chain becomes ill-conditioned when $\gamma = 0$.

To better understand the behavior when $\gamma \to 0$, we here propose another decomposition. The idea is that, when γ tends to 0, we can compare the recursion to the gradient flow. For a function $g : \mathbb{R}^d \to \mathbb{R}^q$ regular enough, we show that there exists a function $h_g : \mathbb{R}^d \to \mathbb{R}^q$ such that, for any $\theta \in \mathbb{R}^d$:

$$h_g'(\theta) f'(\theta) = g(\theta) - g(\theta_*),$$

where $h_g'(\theta) \in \mathbb{R}^{q \times d}$, and $f'(\theta) \in \mathbb{R}^d$. This function h_g will be the solution to the *continuous Poisson equation*. We then use its first order Taylor development of $h_g(\theta_{k+1}^{(\gamma)})$ around $\theta_k^{(\gamma)}$. For any $k \in \mathbb{N}$,

$$\begin{aligned}
h_g(\theta_{k+1}^{(\gamma)}) =& h_g(\theta_k^{(\gamma)}) + h_g'(\theta_k^{(\gamma)})(\theta_{k+1}^{(\gamma)} - \theta_k^{(\gamma)}) + O\left(\left\|\theta_{k+1}^{(\gamma)} - \theta_k^{(\gamma)}\right\|^2\right) \\
=& h_g(\theta_k^{(\gamma)}) - \gamma h_g'(\theta_k^{(\gamma)}) f'(\theta_k^{(\gamma)}) - \gamma h_g'(\theta_k^{(\gamma)})\varepsilon_{k+1}(\theta_k^{(\gamma)}) + O\left(\left\|\theta_{k+1}^{(\gamma)} - \theta_k^{(\gamma)}\right\|^2\right) \\
=& h_g(\theta_k^{(\gamma)}) - \gamma(g(\theta_k^{(\gamma)}) - g(\theta_*)) - \gamma h_g'(\theta_k^{(\gamma)})\varepsilon_{k+1}(\theta_k^{(\gamma)}) + O\left(\left\|\theta_{k+1}^{(\gamma)} - \theta_k^{(\gamma)}\right\|^2\right).
\end{aligned}$$

Thus reorganizing terms,

$$g(\theta_k^{(\gamma)}) - g(\theta_*) = \frac{1}{\gamma}(h_g(\theta_k^{(\gamma)}) - h_g(\theta_{k+1}^{(\gamma)})) + h_g'(\theta_k^{(\gamma)})\varepsilon_{k+1}(\theta_k^{(\gamma)}) + \frac{1}{\gamma} O\left(\left\|\theta_{k+1}^{(\gamma)} - \theta_k^{(\gamma)}\right\|^2\right).$$

Finally, averaging over the first k iterations, for g the identity function,

$$\begin{aligned}
(k+1)(\bar{\theta}_k^{(\gamma)} - \theta_*) =& \frac{1}{\gamma}\left(h_{\mathrm{id}}(\theta_0^{(\gamma)}) - h_{\mathrm{id}}(\theta_{k+1}^{(\gamma)})\right) + \sum_{i=0}^{k} h_{\mathrm{id}}'(\theta_i^{(\gamma)})\varepsilon_{i+1}(\theta_i^{(\gamma)}) \\
&+ \frac{1}{\gamma}\sum_{i=0}^{k} O\left(\left\|\theta_{i+1}^{(\gamma)} - \theta_i^{(\gamma)}\right\|^2\right).
\end{aligned} \tag{4.8}$$

This expansion is the root of the proof of Theorem 4.6, which formalizes the expansion as powers of γ. The key difference between decomposition (4.7) and (4.8) is that in the latter, the residual term $\theta_{i+1}^{(\gamma)} - \theta_i^{(\gamma)}$ tends to 0 and can naturally be controlled when $\gamma \to 0$.

We now formally introduce the function h_g for a real valued function g (generalization to functions with values in $\mathbb{R}^q$ is not difficult as everything can be defined on projections onto coordinates), some sufficient conditions for its existence, and assumptions for Theorem 4.6 to be valid.

Detailed setting. We here describe how this recursion can be seen as a noisy discretization of the following gradient flow equation, with now $t \in \mathbb{R}$:

$$\dot{\theta}_t = -f'(\theta_t) \; . \tag{4.9}$$

Note that since $f'(\theta_*) = 0$ by definition of θ_* and **A**1, then θ_* is an equilibrium point of (4.9), *i.e.*, $\theta_t = \theta_*$ for all $t \geqslant 0$ if $\theta_0 = \theta_*$. Under **A**2, (4.9) admits a unique solution on $\mathbb{R}_+$ for any starting point $\theta \in \mathbb{R}^d$. Denote by $(\phi_t)_{t \geqslant 0}$ the flow of (4.9), defined for all $\theta \in \mathbb{R}^d$ by $(\phi_t(\theta))_{t \geqslant 0}$ as the solution of (4.9) starting at θ.

Denote by $(\mathcal{A}, D(\mathcal{A}))$, the *infinitesimal generator* associated with the flow $(\phi_t)_{t \geqslant 0}$ defined by

$$\begin{aligned} D(\mathcal{A}) &= \left\{ h : \mathbb{R}^d \to \mathbb{R} \; : \; \text{for all } \theta \in \mathbb{R}^d, \; \lim_{t \to 0} \frac{h(\phi_t(\theta)) - h(\theta)}{t} \text{ exists } \right\} \\ \mathcal{A}h(\theta) &= \lim_{t \to 0} t^{-1} \left\{ h(\phi_t(\theta)) - h(\theta) \right\} \text{ for all } h \in D(\mathcal{A}) \; , \quad \theta \in \mathbb{R}^d \; . \end{aligned} \tag{4.10}$$

Note that for all $h \in C^1(\mathbb{R}^d)$, $h \in D(\mathcal{A})$, $\mathcal{A}h = - \langle f', h' \rangle$.

Under **A**1 and **A**2, for any locally Lipschitz function $g : \mathbb{R}^d \to \mathbb{R}$, denote by h_g the solution of the continuous Poisson equation defined for all $\theta \in \mathbb{R}^d$ by $h_g(\theta) = \int_0^\infty (g(\phi_s(\theta)) - g(\theta_*)) ds$. Note that h_g is well-defined by Lemma C.9-b) in Section C.4, since g is assumed to be locally Lipschitz. Note that by (4.10), we have for all $g : \mathbb{R}^d \to \mathbb{R}$, locally Lipschitz,

$$\mathcal{A}h_g(\theta) = -g(\theta) + g(\theta_*) \; . \tag{4.11}$$

Under regularity assumptions on g (see Theorem C.11), h_g is twice continuously differentiable and therefore satisfies $- \left\langle f', h_g' \right\rangle = \mathcal{A}h_g$. As described in the second expansion above, the idea is then to make a Taylor expansion of $h_g(\theta_{k+1}^{(\gamma)})$ around $\theta_k^{(\gamma)}$ to express $k^{-1} \sum_{i=1}^k g(\theta_i^{(\gamma)}) - g(\theta_*)$ as convergent terms imvolving the derivatives of h_g. For $g : \mathbb{R}^d \to \mathbb{R}$ and $k_1, k_2 \in \mathbb{N}$, $k_1 \geqslant 1$ we consider the following assumptions on the regularity of g.

A5 (k_1, k_2). *There exist $a_g, b_g \in \mathbb{R}_+$ such that $g \in C^{k_1}(\mathbb{R}^d)$ and for all $x \in \mathbb{R}^d$ and $i \in \{1, \cdots, k_1\}$, $\|D^i g(\theta)\| \leqslant a_g \left\{ \|\theta - \theta_*\|^{k_2} + b_g \right\}$, where $D^i g$ is the differential of order i of g.*

We then have the following result.

Theorem 4.6. *Assume **A**1-**A**2-**A**3-**A**4$(2(q+3))$, for $q \in \mathbb{N}$. Let $g : \mathbb{R}^d \to \mathbb{R}$ be a function satisfying **A**5$(5, q)$. Then there exists a constant $C_{2(q+3)}$ only depending on q such that for all $\gamma \in \left(0, C_{2(q+3)}\right)$, $k \in \mathbb{N}^*$ and $\theta_0 \in \mathbb{R}^d$ it holds*

$$\begin{aligned} \mathbb{E}\left[k^{-1} \sum_{i=1}^k \left\{ g(\theta_i^{(\gamma)}) - g(\theta_*) \right\} \right] &= \frac{h_g(\theta_0) - \mathbb{E}\left[h_g(\theta_{k+1}^{(\gamma)}) \right]}{k\gamma} \\ &+ (\gamma/2) \operatorname{tr}\left(h_g''(\theta_*) \mathbb{E}\left[\{\varepsilon(\theta_*)\}^{\otimes 2} \right] \right) + \frac{\gamma}{k} A_1(\theta_0) + \gamma^2 A_2(\theta_0, k) \; , \end{aligned} \tag{4.12}$$

where $\theta_k^{(\gamma)}$ is the Markov chain starting from θ_0 and defined by the recursion (4.1). In addition for some constant $C \geqslant 0$ independent of γ and n, we have

$$A_1(\theta_0) \leqslant C \left\{ 1 + \|\theta_0 - \theta_*\|^{q+2} \right\} \; , \; A_2(\theta_0, k) \leqslant C \left\{ 1 + \|\theta_0 - \theta_*\|^{q+3} / k \right\} \; .$$

First in the case where f' is linear, choosing for g the identity function, then $h_{\mathrm{Id}} = \int_0^{+\infty} \{\phi_s - \theta_*\} \mathrm{d}s = \Sigma^{-1}$, and we get that the first term in (4.12) vanishes which is natural since in that case $\bar{\theta}_\gamma = \theta_*$. Second by Lemma C.10-c), we recover the first expansion of Theorem 4.3 for arbitrary objective functions f. Finally note that for all $q \in \mathbb{N}$, under appropriate conditions, Theorem 4.6 implies that there exist $C_1, C_2(\theta_0) \geqslant 0$ such that $\mathbb{E}\left[k^{-1} \sum_{i=1}^k \left\|\theta_i^{(\gamma)} - \theta_*\right\|^{2q}\right] = C_1 \gamma + C_2(\theta_0)/n + O(\gamma^2)$.

4.4 Experiments

We performed experiments on simulated data, for logistic regression, with $n = 10^7$ observations, for $d = 10$ and 25. Results are presented in Figure 4.2. We consider SGD with constant step-sizes $1/R^2$, $1/2R^2$ (and $1/4R^2$) with or without averaging, with $R^2 = L$. Without averaging, the chain saturates with an error proportional to γ (as $\|\theta_k^{(\gamma)} - \theta_*\| = O(\sqrt{\gamma})$). Note that the ratio between the convergence limits of the two sequences is roughly 2 in the un-averaged case, and 4 in the averaged case, which confirms the predicted limits. We consider Richardson Romberg iterates, which saturate at a much lower level, and performs much better than decaying step sizes (as $1/\sqrt{n}$) on the first iterations, as it forgets the initial conditions faster. Finally, we run the online-Newton (Bach and Moulines, 2013), which performs very well but has no convergence guarantee. On the Right plot, we also propose an estimator that uses 3 different step sizes to perform a higher order interpolation. More precisely, we compute $\tilde{\theta}_k^3 := \frac{8}{3}\bar{\theta}_k^{(\gamma)} - 2\bar{\theta}_k^{(2\gamma)} + \frac{1}{3}\bar{\theta}_k^{(4\gamma)}$. With such an estimator, the *first 2* terms in the expansion, scaling as γ and γ^2, should vanish, which explains that it does not saturate. We also perform an experiment on a the covertype[1] data-set: the Richardson-Romberg iterate improves on simple recursions.

4.5 Conclusion

In this chapter, we have used and developed Markov chain tools to analyze the behavior of constant step-size SGD, with a complete analysis of its convergence, outlining the effect of initial conditions, noise and step-sizes. For machine learning problems, this allows us to extend known results from least-squares to all loss functions. This analysis leads naturally to using Romberg-Richardson extrapolation, that provably improves the convergence behavior of the averaged SGD iterates.

The proofs of the results given in this chapter are given in the next chapter (Ch. C): it might be skipped at first reading.

[1] https://archive.ics.uci.edu/ml/datasets/covertype

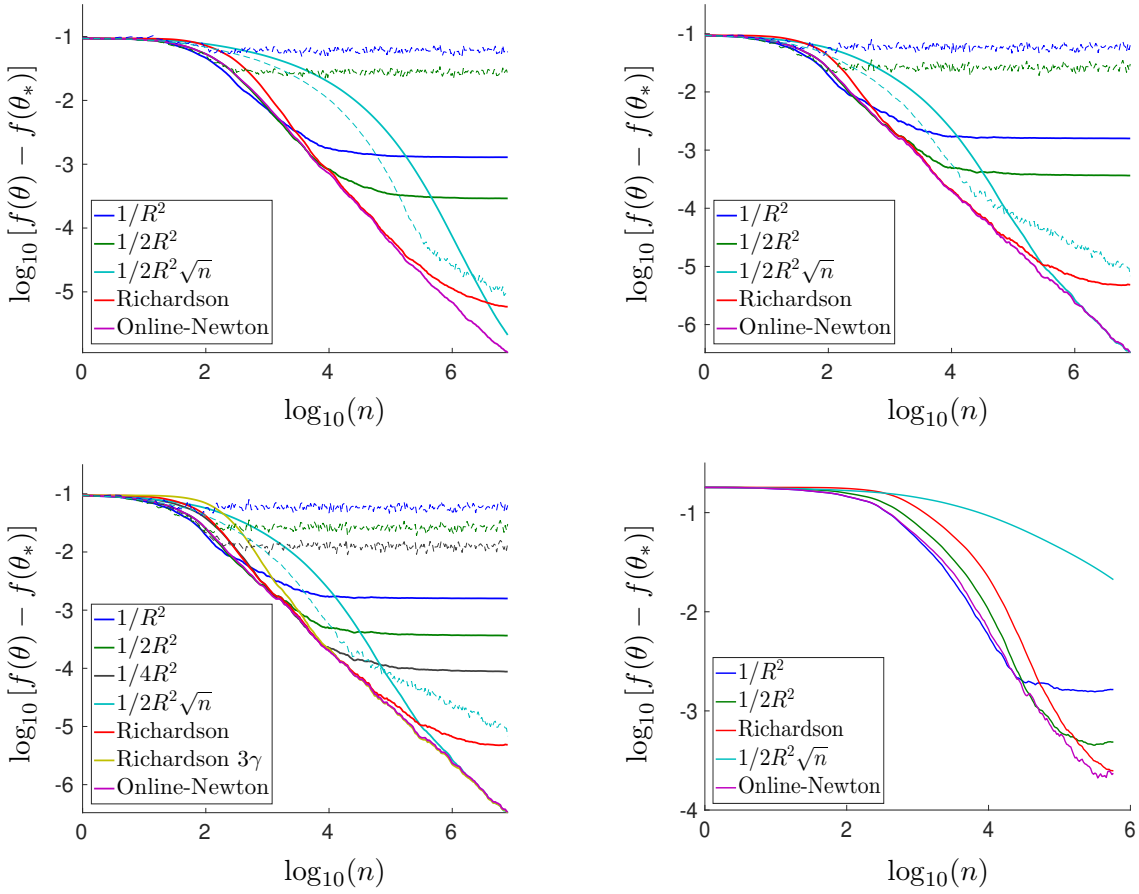


Figure 4.2: Plot of the excess risk as a function of n, logarithmic scales. Upper-left: synthetic data, logistic regression, $d = 12$, with averaged SGD with step-size $1/R^2$, $1/2R^2$, decaying step sizes as $1/2R^2\sqrt{n}$ (averaged (plain) and non-averaged (dashed)), Richardson Romberg extrapolated iterates, and online Newton iterates. Upper-right: same in lower dimension ($d = 4$). Bottom: same but with three different step sizes and an estimator built using Richardson on 3 different sequences: $\tilde{\theta}_k^3 = \frac{8}{3}\bar{\theta}_k^{(\gamma)} - 2\bar{\theta}_k^{(2\gamma)} + \frac{1}{3}\bar{\theta}_k^{(4\gamma)}$, with $\gamma = 1/4R^2$. Bottom-right: experiment on the *covertype* data-set, $d = 55$, $n = 581012$, logistic regression.

C

Appendix to Bridging the Gap between Constant Step Size Stochastic Gradient Descent and Markov Chains

Notation

Denote by $\{\mathbf{e}_1, \dots, \mathbf{e}_d\}$ the canonical basis of $\mathbb{R}^d$. Let E and F be two vector spaces, denote by $E \otimes F$ the tensor product of E and F. For all $x \in E$ and $y \in F$ denote by $x \otimes y \in E \otimes F$ the tensor product of x and y. Let $n \in \mathbb{N}^*$, denote by $C^n(\mathbb{R}^d)$ the set of n times continuously differentiable functions from $\mathbb{R}^d$ to $\mathbb{R}$. Let $f \in C^n(\mathbb{R}^d)$, denote by $D^n f$ the n^{th} differential of f. Let $f \in C^1(\mathbb{R}^d)$, denote by ∇f the gradient of f. Let $f \in C^2(\mathbb{R}^d)$, denote by Δf the Laplacian of f. Denote by $\lfloor \cdot \rfloor$ and $\lceil \cdot \rceil$ the floor and ceiling function respectively. For $a, b \in \mathbb{R}$, denote by $a \vee b$ and $a \wedge b$ the maximum and the minimum of a and b respectively. Denote $\mathcal{S}_{L,\mu}$ the set of μ-strongly convex and L-smooth functions on $\mathbb{R}^d$. By abuse of notation, we will denote sometimes $x^{\otimes 2} = xx^\top$.

In the next sections mainly devoted to proofs, we first introduce definitions and generalities about convex functions in Section C.1.1, then discuss extra different possible assumptions on the noise in Section C.1.3. We prove the existence of a limit distribution in Section C.2.1, and address asymptotic properties when $\gamma \to 0$ in Section C.1.1. We prove the convergence of the Markov chain in Section C.2.4, and study the relationship with the gradient flow in Section C.4.

C.1 Generalities on convex and strongly convex functions

C.1.1 Definitions

Most of the following definitions can be found in Nesterov (2004). A continuously differentiable function f is **convex** if there exists for any $\theta, \eta \in \mathbb{R}^d$ we have:

$$f(\eta) \geqslant f(\theta) + \left\langle f'(\theta), \eta - \theta \right\rangle .$$

A continuously differentiable function f is L**-smooth** if its gradient is L-Lipschitz, *i.e.*, if there exists a constant $L > 0$, such that for any $\theta, \eta \in \mathbb{R}^d$ we have:

$$\|f'(\eta) - f'(\theta)\| \leqslant L\|\eta - \theta\| .$$

A continuously differentiable function f is μ-**strongly convex** if there exists a constant $\mu > 0$, such that for any $\theta, \eta \in \mathbb{R}^d$ we have:

$$f(\eta) \geqslant f(\theta) + \left\langle f'(\theta), \eta - \theta \right\rangle + \frac{\mu}{2}\|\theta - \eta\|^2 .$$

Recall that θ_* refers to as $\arg\min_{\theta \in \mathbb{R}^d} f$, which is unique when f is strongly convex.

Let f be a L-smooth and μ-strongly convex function. Then for all $\theta, \eta \in \mathbb{R}^d$, it holds

$$f(\theta) - f(\theta_*) \geqslant \frac{\mu}{2}\|\theta - \theta_*\|^2 \tag{C.1}$$

$$f(\theta_n^{(\gamma)}) - f(\theta_*) \leqslant L\|\theta_n^{(\gamma)} - \theta_*\|^2 \tag{C.2}$$

$$\left\langle f'(\theta) - f'(\eta), \theta - \eta \right\rangle \geqslant \mu\|\theta - \eta\|^2 \tag{C.3}$$

$$\left\langle f'(\theta) - f'(\eta), \theta - \eta \right\rangle \geqslant \frac{1}{L}\|f'(\theta) - f'(\eta)\|^2 \tag{C.4}$$

$$\left\langle f'(\theta) - f'(\eta), \theta - \eta \right\rangle \geqslant \frac{L\mu}{L+\mu}\|\theta - \eta\|^2 + \frac{1}{L+\mu}\|f'(\theta) - f'(\eta)\|^2 . \tag{C.5}$$

The first two inequalities are direct consequences of the definition and the fact that $f'(\theta_*) = 0$. (C.3) is shown in (Nesterov, 2004, Chapter 2, (2.1.24)). (C.4) is the co-coercivity equation in (Zhu and Marcotte, 1996). (C.5) is a combination of the co-coercivity equation and of (C.3). It can be found in (Nesterov, 2004, Chapter 2, (2.1.24)),

C.1.2 Quadratic case

Consider the following assumption on f.

Q1. *There exists a positive definite matrix* Σ *such that* $f = f_\Sigma := (\theta \mapsto \left\|\Sigma^{1/2}(\theta - \theta_*)\right\|^2)$.

If there exists a positive definite matrix Σ such that $f = f_\Sigma := (\theta \mapsto \left\|\Sigma^{1/2}(\theta - \theta_*)\right\|^2)$, then **A**1 and **A**2 are satisfied, with μ the smallest eigenvalue of Σ, L its largest eigenvalue, and $M = 0$.

C.1.3 Discussion on assumptions on the noise

Assumption **A**4, made in the text, can be weakened in order to apply to settings where input observations are un-bounded (typically, Gaussian inputs would not satisfy Assumption **A**4). Especially, for most situations, we only need Assumption **A**6 below.

A6. *(i) There exists* $\tau \geqslant 0$ *such that* $\{\mathbb{E}^{1/4}[\|\varepsilon_1(\theta_*)\|^4]\} \leqslant \tau$.

(ii) For all $\theta_1, \theta_2 \in \mathbb{R}^d$*, there exists* $L \geqslant 0$ *such that, for* $p = 2, \ldots, 4$*,*

$$\mathbb{E}\left\|f_n'(\theta_1) - f_n'(\theta_2)\right\|^p \leqslant L^{p-1}\left\|\theta_1 - \theta_2\right\|^{p-2}\left\langle \theta_1 - \theta_2, f'(\theta_1) - f'(\theta_2)\right\rangle , \tag{C.6}$$

We can also make the stronger assumption that the noise is independent of θ (the "semi-stochastic" setting in Chapter 3), or more generally that the noise has a uniformly bounded fourth order moment.

A7. *There exists* $\tau \geqslant 0$ *such that* $\sup_{\theta \in \mathbb{R}^d}\{\mathbb{E}^{1/4}[\|\varepsilon_1(\theta)\|^4]\} \leqslant \tau$.

Assumption **A**6 is the weakest, as it is satisfied for random design least mean squares and logistic regression with bounded fourth moment of the inputs. Note that we do not assume that gradient or gradient estimates are a.s. bounded, to avoid the need for a constraint on the space where iterates live. Of course Assumption **A**4 implies Assumption **A** 6. Moreover, in the special case of Assumption **A**7 where the noise is independent of θ, then Assumption **A**4 is clearly satisfied under Assumption **A**2.

C.2 Results on the Markov chain defined by SGD

C.2.1 Proof of Proposition 4.1

Let λ_1, λ_2 be two probability measures on $\mathcal{B}(\mathbb{R}^d)$ with finite second moment and $\gamma > 0$. Let $\theta_0^{(1)}, \theta_0^{(2)}$ be independent and distributed according to λ_1, λ_2 respectively, and $(\theta_k^{(1)})_{\geqslant 0}$,$(\theta_k^{(2)})_{k\geqslant 0}$ the SGD iterates associated with the step size γ, starting from $\theta_0^{(1)}$ and $\theta_0^{(2)}$ respectively and sharing the same noise, *i.e.*, for all $k \geqslant 0$,

$$\begin{cases} \theta_{k+1}^{(1)} &= \theta_k^{(1)} - \gamma[f'(\theta_k^{(1)}) + \varepsilon_{k+1}(\theta_k^{(1)})] \\ \theta_{k+1}^{(2)} &= \theta_k^{(2)} - \gamma[f'(\theta_k^{(2)}) + \varepsilon_{k+1}(\theta_k^{(2)})] \ . \end{cases} \tag{C.7}$$

Therefore for all $k \geqslant 0$, the distribution of $(\theta_k^{(1)}, \theta_k^{(2)})$ belongs to $\Pi(\lambda_1 R_\gamma, \lambda_2 R_\gamma)$ defined in Section 4.3 in the main document. Then by definition of the Wasserstein distance,

$$\begin{aligned} W_2^2(\lambda_1 R_\gamma, \lambda_2 R_\gamma) &\leqslant \mathbb{E}\left[\|\theta_1^{(1)} - \theta_1^{(2)}\|^2\right] \\ &\leqslant \mathbb{E}\left[\|\theta^{(1)} - \gamma f_1'(\theta^{(1)}) - (\theta^{(2)} - \gamma f_1'(\theta^{(2)}))\|^2\right] \\ &\overset{i)}{\leqslant} \mathbb{E}\left[\left\|\theta^{(1)} - \theta^{(2)}\right\|^2 - 2\gamma\left\langle f'(\theta^{(1)}) - f'(\theta^{(2)}), \theta^{(1)} - \theta^{(2)}\right\rangle\right] \\ &\qquad + \gamma^2 \mathbb{E}\left[\left\|f_1'(\theta^{(1)}) - f_1'(\theta^{(2)})\right\|^2\right] \\ &\overset{ii)}{\leqslant} \mathbb{E}\left[\left\|\theta^{(1)} - \theta^{(2)}\right\|^2 - 2\gamma(1-\gamma L)\left\langle f'(\theta^{(1)}) - f'(\theta^{(2)}), \theta^{(1)} - \theta^{(2)}\right\rangle\right] \\ &\overset{iii)}{\leqslant} (1 - 2\mu\gamma(1-\gamma L))\mathbb{E}\left[\left\|\theta^{(1)} - \theta^{(2)}\right\|^2\right] \ , \end{aligned}$$

using **A**3 for $i)$, **A**6 for $ii)$, and finally **A**1 for $iii)$.

Thus by a straightforward induction, we get setting $\rho = (1 - 2\mu\gamma(1-\gamma L))$

$$\begin{aligned} W_2^2(\lambda_1 R_\gamma^n, \lambda_2 R_\gamma^n) &\leqslant \mathbb{E}\left[\|\theta_n^{(1)} - \theta_n^{(2)}\|^2\right] \\ &\leqslant \rho \mathbb{E}\left[\|\theta_{n-1}^{(1)} - \theta_{n-1}^{(2)}\|^2\right] \leqslant \rho^n \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \, \mathrm{d}\lambda_1(x) \mathrm{d}\lambda_2(y) \ , \end{aligned} \tag{C.8}$$

By (Villani, 2009, Theorem 6.16), the space $\mathcal{P}_2(\mathbb{R}^d)$ of probability measures with second order moment on $\mathbb{R}^d$ endowed with W_2 is a Polish space. As a consequence of (C.8) for $\lambda_2 = \lambda_1 R_\gamma^p$, for $p \in \mathbb{N}$, and Picard fixed point theorem, $(\lambda_1 R_\gamma^n)_{n\geqslant 0}$ is a Cauchy sequence and converges to a limit $\pi_\gamma^{\lambda_1} \in \mathcal{P}_2(\mathbb{R}^d)$:

$$\lim_{n\to+\infty} W_2(\lambda_1 R_\gamma^n, \pi_\gamma^{\lambda_1}) = 0 \ . \tag{C.9}$$

In addition by the triangle inequality

$$W_2(\pi_\gamma^{\lambda_1}, \pi_\gamma^{\lambda_2}) \leqslant W_2(\pi_\gamma^{\lambda_1}, \lambda_1 R_\gamma^n) + W_2(\lambda_1 R_\gamma^n, \lambda_2 R_\gamma^n) + W_2(\pi_\gamma^{\lambda_2}, \lambda_2 R_\gamma^n) \ .$$

Thus taking the limits as $n \to +\infty$, we get $W_2(\pi_\gamma^{\lambda_1}, \pi_\gamma^{\lambda_2}) = 0$ and $\pi_\gamma^{\lambda_1} = \pi_\gamma^{\lambda_2}$. The limit is thus the same for all initial distributions and is denoted by π_γ.

Moreover, π_γ is invariant for R_γ. Indeed for all $n \in \mathbb{N}$, $n \geqslant 1$,

$$W_2(\pi_\gamma R_\gamma, \pi_\gamma) \leqslant W_2(\pi_\gamma R_\gamma, \pi_\gamma R_\gamma^n) + W_2(\pi_\gamma R_\gamma^n, \pi_\gamma) \; .$$

Using (C.8) and (C.9), we get taking $n \to +\infty$, $W_2(\pi_\gamma R_\gamma, \pi_\gamma) = 0$ and $\pi_\gamma R_\gamma = \pi_\gamma$. The fact that π_γ is the unique stationary distribution can be shown by contradiction and using (C.8).

Thus finally for $\lambda_1 = \delta_\theta$, $\lambda_2 = \pi_\gamma$, using the invariance of π_γ and (C.8), we get:

$$W_2^2(R_\gamma^n(\theta, \cdot), \pi_\gamma) \quad \leqslant \quad (1 - 2\mu\gamma(1 - \gamma L))^n \int \|\theta - \vartheta\|^2 \mathrm{d}\pi_\gamma(\vartheta) \; .$$

C.2.2 Existence of Poisson solutions

Using the process $(\theta_{k,\gamma}^{(1)})_{\geqslant 0}$,$(\theta_{k,\gamma}^{(2)})_{k \geqslant 0}$ defined by (C.7) with $\lambda_1 = \delta_\theta$ and $\lambda_2 = \pi_\gamma$ and (C.8), we have if h is L_h-Lipschitz, for any $x \in \mathbb{R}^d$, any $n \in \mathbb{N}^*$:

$$\begin{aligned} \left| R_\gamma^n(h - \pi_\gamma(h))(\theta) \right| &\leqslant L_h W_2^2(R_\gamma^n(\theta, \cdot), \pi_\gamma) \\ &\leqslant L_h (1 - 2\mu\gamma(1 - \gamma L))^{n/2} \left(\int \|\theta - \vartheta\|^2 \mathrm{d}\pi_\gamma(\vartheta) \right)^{1/2} \; . \end{aligned} \tag{C.10}$$

In addition, for any $(\theta, \vartheta) \in \mathbb{R}^d \times \mathbb{R}^d$, $n \in \mathbb{N}^*$, using (C.7):

$$\begin{aligned} \left\| R_\gamma^n h(\theta) - R_\gamma^n h(\vartheta) \right\| &\leqslant L_h W_2^2(R_\gamma^n(\theta, \cdot), R_\gamma^n(\vartheta, \cdot)) \\ &\leqslant L_h (1 - 2\mu\gamma(1 - \gamma L))^{n/2} \|\theta - \vartheta\| \; . \end{aligned} \tag{C.11}$$

As a consequence by (C.10), for any Lipschitz continuous function φ and any $\theta \in \mathbb{R}^d$, $\{\theta \mapsto \sum_{i=1}^k (R_\gamma^i \varphi(\theta) - \pi_\gamma(\varphi))\}_{k \geqslant 0}$ converges absolutely on all compact sets of $\mathbb{R}^d$. Denote by ψ_γ the limit associated with this sequence: $\psi_\gamma : \theta \mapsto \sum_{i=1}^\infty (R_\gamma^i \varphi(\theta) - \pi_\gamma(\varphi))$. By (C.11), ψ_γ is also Lipschitz continuous. This function is called the solution to the Poisson equation since it satisfies $(I - R_\gamma)\psi_\gamma = \varphi - \pi_\gamma(\phi)$. Moreover, $\pi_\gamma(\psi_\gamma) = 0$.

C.2.3 Asymptotic properties of the chain, behavior under equilibrium, and drift.

In the following, we consider the function $\varphi_1 : \theta \mapsto \theta - \theta_* \in \mathbb{R}^d$, and the function $\varphi_2 : \theta \mapsto (\theta - \theta_*)(\theta - \theta_*)^\top \in \mathbb{R}^{d \times d}$. In the quadratic case, we give an exact formula for the expectation under the limit distribution of these two terms. For the general case, we propose a first order development of these expectations.

The most important quantity, as we are eventually interested in the behavior of the averaged iterate $\bar{\theta}_n^{(\gamma)}$, is the expectation of the identity function under the limit distribution, $\bar{\theta}_\gamma$ defined by (4.3).

This part extends existing ideas from the literature to prove that $\gamma^{-1/2}(\pi_\gamma - \theta_*)$ converges in distribution to a normal law when $\gamma \to 0$. See for example (Pflug, 1986; Ljung et al., 1992). We consider the Markov chain under the limiting stationary distribution, together with a Taylor expansion of the function around the optimal point θ_*, in order to analyze how the average under the stationary distribution $\bar{\theta}_\gamma$ deviates from θ_*.

Analysis is carried through the dynamic (4.1) at stationarity, *i.e.*, we assume that θ_0 is distributed according to π_γ given by the study of the equilibrium equation: under stationarity, *i.e.*, if $\theta_n^{(\gamma)} \sim \pi_\gamma$,

$$\theta_{n+1}^{(\gamma)} \stackrel{d}{=} \theta_n^{(\gamma)} - \gamma f'(\theta_n^{(\gamma)}) - \gamma \varepsilon_{n+1}(\theta_n^{(\gamma)}) \stackrel{d}{=} \pi_\gamma. \tag{C.12}$$

In order to get a first order development of $\bar{\theta}_\gamma$ around θ_*, we use the definition of the stationary distribution. We are going to use this equality several times to obtain information on θ's first moments under π_γ. The first consequence of this equation is that, taking expectations on both sides,

$$\int_{\mathbb{R}^d} f'(\theta)\pi_\gamma(\mathrm{d}\theta) \quad = \quad 0\,. \tag{C.13}$$

Lemma C.1 (Properties under stationarity, Quadratic case)**.**
We consider, the stochastic gradient descent algorithm (4.1), for the quadratic function $f_\Sigma(\theta) := \left\|\Sigma^{1/2}(\theta - \theta_*)\right\|^2$. Then the mean value under the stationary distribution of the iterate is the optimal point:

$$\begin{aligned}
\bar{\theta}_\gamma &= \int_{\mathbb{R}^d} \theta \pi_\gamma(\mathrm{d}\theta) = \theta_* \\
\int_{\mathbb{R}^d} (\theta - \theta_*)^{\otimes 2} \pi_\gamma(\mathrm{d}\theta) &= \gamma(\Sigma \otimes I + I \otimes \Sigma - \gamma \Sigma \otimes \Sigma)^{-1} \int_{\mathbb{R}^d} \varepsilon_1(\theta)^{\otimes 2} \pi_\gamma(\mathrm{d}\theta)\,.
\end{aligned}$$

Moreover, for the least mean squares algorithm, as defined described in the examples in Section 4.2.1,

$$\begin{aligned}
\theta_n^{(\gamma)} - \theta_* &= (I - \gamma\Sigma)\left(\theta_{n-1}^{(\gamma)} - \theta_*\right) + \gamma \varepsilon_n(\theta_{n-1}^{(\gamma)}) \\
\varepsilon_n(\theta_{n-1}^{(\gamma)}) &= (\Sigma - x_n \otimes x_n)(\theta_{n-1}^{(\gamma)} - \theta_*) + (y_n - \langle \theta_*, x_n \rangle) x_n\,,
\end{aligned}$$

we have another formula:

$$\int_{\mathbb{R}^d} (\theta - \theta_*)^{\otimes 2} \pi_\gamma(\mathrm{d}\theta) \quad = \quad \gamma(\Sigma \otimes I + I \otimes \Sigma - \gamma M)^{-1} \mathbb{E}[\xi_1^{\otimes 2}]\,,$$

where in the last equation, M is an operator on matrices such that $M : A \mapsto \mathbb{E}[x_n x_n^\top A x_n x_n^\top]$, and $\xi_n = (y_n - \langle \theta_*, x_n \rangle) x_n$ is the additive part of the noise (the part that does not depend on θ).

Proof. The first part directly comes from Equation (C.13) and the fact that gradients of f_Σ are linear: $\int_{\mathbb{R}^d} f'(\theta)\pi_\gamma(\mathrm{d}\theta) = \Sigma \int_{\mathbb{R}^d} \theta - \theta_* \pi_\gamma(\mathrm{d}\theta) = 0$, thus $\int_{\mathbb{R}^d} \theta \pi_\gamma(\mathrm{d}\theta) = \theta_*$.

The second part comes from the development of Equation (C.12):

$$\begin{aligned}
(\theta_1^{(\gamma)} - \theta_*)^{\otimes 2} &\stackrel{d}{=} \left((I - \gamma\Sigma)\left(\theta_0^{(\gamma)} - \theta_*\right) + \gamma \varepsilon_1(\theta_0^{(\gamma)})\right)^{\otimes 2} \\
\mathbb{E}(\theta_1^{(\gamma)} - \theta_*)^{\otimes 2} &= (I - \gamma\Sigma)\mathbb{E}\left(\theta_0^{(\gamma)} - \theta_*\right)^{\otimes 2}(I - \gamma\Sigma) + \gamma^2 \mathbb{E}\left(\varepsilon_1(\theta_0^{(\gamma)})\right)^{\otimes 2} \\
\mathbb{E}(\theta_1^{(\gamma)} - \theta_*)^{\otimes 2} &= (I - \gamma\Sigma \otimes I - \gamma I \otimes \Sigma + \gamma^2 \Sigma \otimes \Sigma)\mathbb{E}\left(\theta_0^{(\gamma)} - \theta_*\right)^{\otimes 2} \\
&\qquad + \gamma^2 \mathbb{E}\left(\varepsilon_1(\theta_0^{(\gamma)})\right)^{\otimes 2}\,,
\end{aligned} \tag{C.14}$$

Thus as if $\theta_0^{(\gamma)} \sim \pi_\gamma$, then $\theta_1^{(\gamma)} \sim \pi_\gamma$:

$$\int_{\mathbb{R}^d} (\theta - \theta_*)^{\otimes 2} \pi_\gamma(\mathrm{d}\theta) \quad = \quad \gamma(\Sigma \otimes I + I \otimes \Sigma - \gamma \Sigma \otimes \Sigma)^{-1} \int_{\mathbb{R}^d} \varepsilon_1(\theta)^{\otimes 2} \pi_\gamma(\mathrm{d}\theta)\,.$$

Similarly, starting from:

$$\theta_1^{(\gamma)} - \theta_* \quad = \quad (I - \gamma x_1 \otimes x_1)\left(\theta_0^{(\gamma)} - \theta_*\right) + \gamma\xi_1 \ ,$$

using the fact that $\mathbb{E}[x_n x_n^\top] = \Sigma$ and the definition of M, one gets:

$$\int_{\mathbb{R}^d} (\theta - \theta_*)^{\otimes 2} \pi_\gamma(\mathrm{d}\theta) \quad = \quad \gamma(\Sigma \otimes I + I \otimes \Sigma - \gamma M)^{-1} \mathbb{E}[\xi_1^{\otimes 2}] \ .$$

Which concludes the proof. □

Lemma C.2. *Assume A1, A2, A3, A6. Then*

$$\mathbb{E}\left[-2\gamma\left\langle f'_{n+1}(\theta_n^{(\gamma)}), \theta_n^{(\gamma)} - \theta_*\right\rangle + \gamma^2 \left\|f'_{n+1}(\theta_n^{(\gamma)})\right\|^2 |\mathcal{F}_n\right] \leqslant -2\gamma\mu(1-\gamma L)\left\|\theta_n^{(\gamma)} - \theta_*\right\|^2 + 2\gamma^2\tau^2,$$

where $f'_n = \varepsilon_n + f'$ for all $n \geqslant 1$ and $(\theta_n^{(\gamma)})_{n\geqslant 0}$ is given by (4.1).

Proof. Under **A6**, we have:

$$\begin{aligned}
\mathbb{E}\left[\left\|f'_{n+1}(\theta_n^{(\gamma)})\right\|^2 |\mathcal{F}_n\right] &\leqslant 2\left(\mathbb{E}\left[\left\|f'_{n+1}(\theta_n^{(\gamma)}) - f'_{n+1}(\theta_*)\right\|^2\right] + \mathbb{E}\left[\left\|f'_{n+1}(\theta_*)\right\|^2 |\mathcal{F}_n\right]\right) \\
&\leqslant 2\left(\mathbb{E}\left[\left\|f'_{n+1}(\theta_n^{(\gamma)}) - f'_{n+1}(\theta_*)\right\|^2 |\mathcal{F}_n\right] + \tau^2\right) \\
&\leqslant 2\left(L\mathbb{E}\left[\left\langle f'_{n+1}(\theta_n^{(\gamma)}) - f'_{n+1}(\theta_*), \theta_n^{(\gamma)} - \theta_*\right\rangle |\mathcal{F}_n\right] + \tau^2\right) \\
&\leqslant 2\left(L\left\langle f'(\theta_n^{(\gamma)}) - f'(\theta_*), \theta_n^{(\gamma)} - \theta_*\right\rangle + \tau^2\right) \ .
\end{aligned}$$

Combining this result and **A1** concludes the proof. □

Note that if we instead make Assumption **A7**, we have a slightly different result. We only gives this result as it underlines the difference between a stochastic noise and a semi-stochastic noise, especially the fact that the maximal step size differs depending on this assumption made. This Lemma is not used in the following.

Lemma C.3. *Assume A1, A2, A3, A7. Then, for $\gamma \leqslant \frac{2}{L+\mu}$*

$$\mathbb{E}\left[-2\gamma\left\langle f'_{n+1}(\theta_n^{(\gamma)}), \theta_n^{(\gamma)} - \theta_*\right\rangle + \gamma^2 \left\|f'_{n+1}(\theta_n^{(\gamma)})\right\|^2 |\mathcal{F}_n\right] \leqslant -2\gamma\tilde{\mu}\left\|\theta_n^{(\gamma)} - \theta_*\right\|^2 + \gamma^2\tau^2 \ ,$$

where $f'_n = \varepsilon_n + f'$, for all $n \geqslant 1$ and $(\theta_n^{(\gamma)})_{n\geqslant 0}$ is given by (4.1). We are allowed a larger step size (nearly twice as large), but we slightly degrade μ into $\tilde{\mu} := \frac{\mu L}{L+\mu}$.

Proof. Under **A7**, we have:

$$\begin{aligned}
\mathbb{E}\left[\left\|f'_{n+1}(\theta_n^{(\gamma)})\right\|^2 |\mathcal{F}_n\right] &= \left(\left\|f'(\theta_n^{(\gamma)})\right\|^2 + \mathbb{E}\left[\left\|f'_{n+1}(\theta_n^{(\gamma)}) - f'(\theta_n^{(\gamma)})\right\|^2\right]\right) \\
&\leqslant \left(\left\|f'(\theta_n^{(\gamma)})\right\|^2 + \tau^2\right) \ .
\end{aligned}$$

So that finally, using Equation (C.5), and rearranging terms:

$$\begin{aligned}
\mathbb{E}\left[-2\gamma\left\langle f'_{n+1}(\theta_n^{(\gamma)}), \theta_n^{(\gamma)} - \theta_*\right\rangle + \gamma^2 \left\|f'_{n+1}(\theta_n^{(\gamma)})\right\|^2 |\mathcal{F}_n\right] &\leqslant -2\gamma\tilde{\mu}\left\|\theta_n^{(\gamma)} - \theta_*\right\|^2 + \gamma^2\tau^2 \\
&\quad - 2\frac{\gamma}{L+\mu}\left\|f'(\theta_n^{(\gamma)})\right\| + \gamma^2\left\|f'(\theta_n^{(\gamma)})\right\|^2 \\
&\leqslant -2\gamma\tilde{\mu}\left\|\theta_n^{(\gamma)} - \theta_*\right\|^2 + \gamma^2\tau^2 \ ,
\end{aligned}$$

□

Lemma C.4 (Properties under stationarity, general case)**.**
If f satisfies Assumptions A1, A2, and we study stochastic gradient descent under Assumptions A 3, A6, we have:

$$\bar{\theta}_\gamma - \theta_* = \frac{1}{2}\gamma f''(\theta_*)^{-1} f'''(\theta_*)\Big(\big[f''(\theta_*) \otimes \mathrm{I} + \mathrm{I} \otimes f''(\theta_*) \big]^{-1} \int_{\mathbb{R}^d} \varepsilon(\theta)^{\otimes 2} \pi_\gamma(\mathrm{d}\theta) \Big) + O(\gamma^2)$$

$$\int_{\mathbb{R}^d} (\theta - \theta_*)^{\otimes 2} \pi_\gamma(\mathrm{d}\theta) \quad = \quad \gamma \big[f''(\theta_*) \otimes \mathrm{I} + \mathrm{I} \otimes f''(\theta_*) \big]^{-1} \int_{\mathbb{R}^d} \varepsilon(\theta)^{\otimes 2} \pi_\gamma(\mathrm{d}\theta) + O(\gamma^2) \, .$$

This lemma improves some result of (Pflug, 1986), and proves that the residual term is of order $O(\gamma^2)$ (we first prove that it is of order $O(\gamma^{3/2})$) and then improve on that result.

Proof. As before, the proof relies on the analysis of the recursion under stationarity. That is we consider $\theta_0^{(\gamma)} \sim \pi_\gamma$ (thus $\theta_1^{(\gamma)} \sim \pi_\gamma$), and expand the stochastic gradient recursion:

$$\begin{aligned} \theta_1^{(\gamma)} &= \theta_0^{(\gamma)} - \gamma f_1'(\theta_0^{(\gamma)}) \\ &= \theta_0^{(\gamma)} - \gamma \Big(f'(\theta_0^{(\gamma)}) + \varepsilon_1(\theta_0^{(\gamma)}) \Big) \, . \end{aligned}$$

For simplicity, in the rest of the proof, we skip the explicit dependence in γ in $\theta_i^{(\gamma)}$, for $i \in \{0, 1\}$. We only denote it θ_i.

We first prove that:

$$\bar{\theta}_\gamma - \theta_* = \frac{1}{2}\gamma f''(\theta_*)^{-1} f'''(\theta_*)\Big(\big[f''(\theta_*) \otimes \mathrm{I} + \mathrm{I} \otimes f''(\theta_*) \big]^{-1} \mathbb{E}\varepsilon^{\otimes 2} \Big) + O(\gamma^{3/2}) \, .$$

We first notice that $\mathbb{E}_{\pi_\gamma} \|\theta - \theta_*\| = O(\gamma^{1/2})$, which will be used several times in the following. Indeed, if $\theta_0 \sim \pi_\gamma$:

$$\begin{aligned} \mathbb{E}\left[\|\theta_1 - \theta_*\|^2 \right] &= \mathbb{E}\left[\|\theta_0 - \theta_* - \gamma f_1'(\theta_0)\|^2 \right] \\ &= \mathbb{E}\left[\|\theta_0 - \theta_*\|^2 - 2\gamma \left\langle f_1'(\theta_0), \theta_0 - \theta_* \right\rangle - \gamma^2 \left\| f_1'(\theta_0) \right\|^2 \right] \\ \Leftrightarrow 0 &\leqslant -2\gamma\mu \mathbb{E}\left[\|\theta_0 - \theta_*\|^2 \right] + \gamma^2 \tau^2 \end{aligned}$$

Using Lemma C.2, under Assumption **A**6, with τ^2 the bound on $\mathbb{E}[\|\varepsilon_1(\theta_*)\|^2]$. Thus we have $\mathbb{E}_{\pi_\gamma}[\|\theta - \theta_*\|^2] \leqslant \frac{\gamma\tau^2}{2\mu}$, and by Jensen, $\mathbb{E}_{\pi_\gamma}[\|\theta - \theta_*\|] \leqslant \frac{\gamma^{1/2}\tau}{\sqrt{2\mu}} = O(\gamma^{1/2})$. More generally, we show in Section C.3, in Lemma C.7, that $\mathbb{E}_{\pi_\gamma}[\|\theta - \theta_*\|^4] = O(\gamma^2)$, and thus $\mathbb{E}_{\pi_\gamma}[\|\theta - \theta_*\|^3] = O(\gamma^{3/2})$.

We now use the following expression for the SGD recursion:

$$\theta_1 = \theta_0 - \gamma \left(f'(\theta_0) + \varepsilon_1(\theta_0) \right) \, .$$

For simplicity, in the following, we may denote: $\varepsilon_1 = \varepsilon_1(\theta_0)$. By definition, we have $\bar{\theta}_\gamma = \mathbb{E}_{\pi_\gamma}\theta$, and as it has been seen before, $\mathbb{E}_{\pi_\gamma} f'(\theta) = 0$.

At it has been proved above, $\mathbb{E}_{\pi_\gamma}\|\theta - \theta_*\|^2 = O(\gamma)$, which also implies by Jensen's inequality that $\|\bar{\theta}_\gamma - \theta_*\|^2 = O(\gamma)$. Using a Taylor expansion, we have that:

$$f'(\theta) = f''(\theta_*)(\theta - \theta_*) + \frac{1}{2} f'''(\theta_*)(\theta - \theta_*)^{\otimes 2} + O(\|\theta - \theta_*\|^3).$$

Where $f''(\theta_*)$ is the Hessian matrix of f, and $f'''(\theta_*)$ a third order tensor that acts on the second order tensor $(\theta - \theta_*)^{\otimes 2}$: $f'''(\theta_*)(\theta - \theta_*)^{\otimes 2}$ is a vector in $\mathbb{R}^d$, such that for $k \in [1; d]$, $(f'''(\theta_*)(\theta - \theta_*)^{\otimes 2})_k = \sum_{i,j=1}^{n} \frac{\partial^3 f}{\partial\theta_i \partial\theta_j \partial\theta_k}(\theta - \theta_*)_i (\theta - \theta_*)_j$.

$$0 \quad = \quad \mathbb{E}_{\pi_\gamma}\Big[f''(\theta_*)(\theta - \theta_*) + \frac{1}{2} f'''(\theta_*)(\theta - \theta_*)^{\otimes 2} \Big] + O(\gamma^{3/2}),$$

using the fact that f is $\mathcal{C}^4$, with bounded 4^{-th} derivative, and $\mathbb{E}_{\pi_\gamma}[\|\theta-\theta_*\|^3] = O(\gamma^{3/2})$. This leads to

$$f''(\theta_*)(\bar{\theta}_\gamma - \theta_*) + \frac{1}{2}f'''(\theta_*)[\mathbb{E}_{\pi_\gamma}(\theta-\theta_*)^{\otimes 2}] = O(\gamma^{3/2}) \ . \tag{C.15}$$

Moreover, we have:

$$\begin{aligned}\theta_1 - \theta_* &= \theta_0 - \theta_* - \gamma[f''(\theta_*)(\theta_0-\theta_*) + \varepsilon_1 + O(\|\theta_0-\theta_*\|)] \\ &= (\mathrm{I} - \gamma f''(\theta_*))(\theta_0-\theta_*) - \gamma\varepsilon_1 + \gamma O(\|\theta_0-\theta_*\|).\end{aligned}$$

Taking the second order moment of this equation, and using the fact that $\mathbb{E}_{\pi_\gamma}[\varepsilon_1(\theta_0-\theta_*)^\top] = \mathbb{E}_{\pi_\gamma}[\mathbb{E}[\varepsilon_1(\theta_0-\theta_*)^\top|\mathcal{F}_0]] = \mathbb{E}_{\pi_\gamma}[\mathbb{E}[\varepsilon_1|\mathcal{F}_0](\theta_0-\theta_*)^\top] = 0$, we get:

$$\mathbb{E}_{\pi_\gamma}(\theta-\theta_*)^{\otimes 2} = (\mathrm{I}-\gamma f''(\theta_*))\mathbb{E}_{\pi_\gamma}(\theta-\theta_*)^{\otimes 2}(\mathrm{I}-\gamma f''(\theta_*)) + \gamma^2\mathbb{E}_{\pi_\gamma}[\varepsilon_1^{\otimes 2}] + O(\gamma^{5/2}).$$

This leads to:

$$\mathbb{E}_{\pi_\gamma}(\theta-\theta_*)^{\otimes 2} = \gamma\big[f''(\theta_*)\otimes\mathrm{I} + \mathrm{I}\otimes f''(\theta_*)\big]^{-1}\mathbb{E}_{\pi_\gamma}[\varepsilon_1^{\otimes 2}] + O(\gamma^{3/2}). \tag{C.16}$$

And combining Equation (C.15) and Equation (C.16), we get:

$$\bar{\theta}_\gamma - \theta_* = \frac{1}{2}\gamma f''(\theta_*)^{-1}f'''(\theta_*)\Big(\big[f''(\theta_*)\otimes\mathrm{I} + \mathrm{I}\otimes f''(\theta_*)\big]^{-1}\mathbb{E}_{\pi_\gamma}[\varepsilon_1^{\otimes 2}]\Big) + O(\gamma^{3/2}) \ .$$

The rest of the proof is devoted to showing that the residual term is of order $O(\gamma^2)$. At that point, we have also proved that $\mathbb{E}[\theta-\theta_*] = O(\gamma)$. To find the next term in the development, we develop further each of the terms. We introduce the 4^{-th} order tensor $f^{(4)} \in \mathbb{R}^{d\times d\times d\times d}$, which acts on $\mathbb{R}^{d\times d\times d}$ to give a vector of $\mathbb{R}^d$. Using the following Taylor expansion, with f assumed to be $\mathcal{C}^5$:

$$\begin{aligned}\theta_1 - \theta_* &= \theta_0 - \theta_* - \gamma\big[f''(\theta_*)(\theta_0-\theta_*) + \frac{1}{2}f^{(3)}(\theta_*)(\theta_0-\theta_*)^{\otimes 2} \\ &\quad + \frac{1}{6}f^{(4)}(\theta_*)(\theta_0-\theta_*)^{\otimes 3} + \varepsilon_1 + O(\|\theta_0-\theta_*\|^4)\big] \ .\end{aligned} \tag{C.17}$$

Thus if $\theta_0 \sim \pi_\gamma$:

$$\begin{aligned}\mathbb{E}_{\pi_\gamma}[\theta-\theta_*] &= \mathbb{E}_{\pi_\gamma}[\theta-\theta_*] - \mathbb{E}_{\pi_\gamma}\Big[\gamma\big[f''(\theta_*)(\theta-\theta_*) + \frac{1}{2}f^{(3)}(\theta_*)(\theta-\theta_*)(\theta-\theta_*)^\top \\ &\quad + \frac{1}{6}f^{(4)}(\theta_*)(\theta-\theta_*)^{\otimes 3} + \varepsilon_1\big]\Big] + \gamma O(\gamma^2) \\ f''(\theta_*)\mathbb{E}_{\pi_\gamma}[\theta-\theta_*] &= -\mathbb{E}_{\pi_\gamma}\left[\frac{1}{2}f^{(3)}(\theta_*)(\theta-\theta_*)^{\otimes 2} + \frac{1}{6}f^{(4)}(\theta_*)(\theta-\theta_*)^{\otimes 3} + \varepsilon_1\right] + O(\gamma^2) \\ f''(\theta_*)(\bar{\theta}_\gamma-\theta_*) &= -\frac{1}{2}f^{(3)}(\theta_*)\mathbb{E}_{\pi_\gamma}[(\theta-\theta_*)^{\otimes 2}] - \frac{1}{6}f^{(4)}(\theta_*)\mathbb{E}_{\pi_\gamma}[(\theta-\theta_*)^{\otimes 3}] + O(\gamma^2) \ .\end{aligned} \tag{C.18}$$

Using Assumption 3 (implying $\mathbb{E}[\varepsilon_1(\theta_0)] = 0$). To get the next term in the development, we need to

- Expand $\mathbb{E}_{\pi_\gamma}[\theta-\theta_*]^{\otimes 2} = \square\gamma + \triangle\gamma^2 + o(\gamma^2)$;
- Expand $\mathbb{E}_{\pi_\gamma}[(\theta-\theta_*)^{\otimes 3}] = \blacksquare\gamma^2 + o(\gamma^2)$.

First, we have, squaring Equation (C.17) and taking expectations:

$$\begin{aligned}\mathbb{E}[\theta_1-\theta_*]^{\otimes 2} &= \mathbb{E}\Big[\left(I-\gamma f''(\theta_*)\right)(\theta_0-\theta_*)+\frac{\gamma}{2}f^{(3)}(\theta_*)(\theta_0-\theta_*)^{\otimes 2}+\gamma\varepsilon_1\\ &\quad +O(\gamma\left\|\theta_0-\theta_*\right\|^3)\Big]^{\otimes 2}\\ &= \mathbb{E}[\theta_0-\theta_*]^{\otimes 2}-\gamma(I\otimes f''(\theta_*)+f''(\theta_*)\otimes I)\mathbb{E}\big[(\theta-\theta_*)^{\otimes 2}\big]+O(\gamma^3)\\ &\quad +\frac{\gamma}{2}\Big((\theta_0-\theta_*)f^{(3)}(\theta_*)(\theta_0-\theta_*)^{\otimes 2}+[(\theta_0-\theta_*)f^{(3)}(\theta_*)(\theta_0-\theta_*)^{\otimes 2}]^\top\Big)\\ &\quad +\gamma^2\mathbb{E}\varepsilon_1^{\otimes 2}+\gamma\mathbb{E}[\left(I-\gamma f''(\theta_*)\right)(\theta_0-\theta_*)\varepsilon_1^\top]\,.\end{aligned}$$

Where we have used:

- $\gamma^2\mathbb{E}\big[(\theta-\theta_*)^{\otimes 2}\big]=O(\gamma^3)$.
- $\mathbb{E}[\left(I-\gamma f''(\theta_*)\right)(\theta_0-\theta_*)\varepsilon_1^\top]=0$ (Assumption 3 again).

Under $\theta_0\overset{d}{=}\theta_1\sim\pi_\gamma$, and simplifying by $\mathbb{E}_{\pi_\gamma}[\theta-\theta_*]^{\otimes 2}$ left and right and dividing by γ:

$$\begin{aligned}(I\otimes f''(\theta_*)+f''(\theta_*)\otimes I)\mathbb{E}_{\pi_\gamma}\big[(\theta-\theta_*)^{\otimes 2}\big]=&O(\gamma^2)-\mathbb{E}\frac{1}{2}f^{(3)}(\theta_*)(\theta-\theta_*)^{\otimes 3}-\\ &\mathbb{E}[\frac{1}{2}f^{(3)}(\theta_*)(\theta-\theta_*)^{\otimes 3}]^\top-\gamma\mathbb{E}\varepsilon_1^{\otimes 2}.\end{aligned}\tag{C.19}$$

We now show that $\mathbb{E}_{\pi_\gamma}[(\theta-\theta_*)^{\otimes 3}]=O(\gamma^2)$. It can then be used in both (C.19) and (C.18), to prove that the next leading term is indeed or order $O(\gamma^2)$ and not $\gamma^{3/2}$. To compute $\mathbb{E}_{\pi_\gamma}[(\theta-\theta_*)^{\otimes 3}]$ we use the second order development again:

$$\begin{aligned}\theta_1-\theta_* &= \theta_0-\theta_*-\gamma\big[f''(\theta_*)(\theta_0-\theta_*)+\varepsilon_1+O(\gamma)\big]\\ &= (\mathrm{I}-\gamma f''(\theta_*))(\theta_0-\theta_*)-\gamma\varepsilon_1+O(\gamma^2).\end{aligned}$$

$$\mathbb{E}_{\pi_\gamma}(\theta-\theta_*)^{\otimes 2}=(\mathrm{I}-\gamma f''(\theta_*))\mathbb{E}_{\pi_\gamma}(\theta-\theta_*)^{\otimes 2}(\mathrm{I}-\gamma f''(\theta_*))+\gamma^2\mathbb{E}\varepsilon^{\otimes 2}+O(\gamma^{5/2}).$$

Let us denote in the following $\eta_i=\theta_i-\theta_*$, $i\in\{1,2\}$:

$$\begin{aligned}\mathbb{E}[\eta_1^{\otimes 3}] &= \mathbb{E}(\theta_1-\theta_*)^{\otimes 3}\\ &= \mathbb{E}\left((\mathrm{I}-\gamma f''(\theta_*))\eta_0-\gamma\varepsilon_1+O(\gamma^2)\right)^{\otimes 3}\\ &= \mathbb{E}((I-(\gamma f''(\theta_*)\otimes I\otimes I+I\otimes\gamma f''(\theta_*)\otimes I+\ I\otimes I\otimes\gamma f''(\theta_*))(\eta_0)^{\otimes 3}\\ &\quad +O((\gamma^{2+3/2}))+\gamma^2\mathbb{E}[(\mathrm{I}-\gamma f''(\theta_*))\eta_0\otimes\varepsilon_1^{\otimes 2}+\varepsilon_1\otimes(\mathrm{I}-\gamma f''(\theta_*))\eta_0\otimes\varepsilon_1\\ &\quad +\varepsilon_1^{\otimes 2}\otimes(\mathrm{I}-\gamma f''(\theta_*))\eta_0]+\gamma^3\mathbb{E}[\varepsilon_1^{\otimes 3}]+0+O(\gamma^3)\,.\end{aligned}$$

Using the fact that $\mathbb{E}[\varepsilon_1]=0$, and the fact that $\mathbb{E}[O(\gamma^2)\otimes((\mathrm{I}-\gamma f''(\theta_*))\eta)^{\otimes 2}]=O(\gamma^3)$ as $\mathbb{E}[\eta^{\otimes 2}]=O(\gamma)$. Thus, if $\theta_0\overset{d}{=}\theta_1$, simplifying by $\mathbb{E}[\eta_i^{\otimes 3}]$:

$$\begin{aligned}\gamma\mathbf{M}\mathbb{E}[\eta_0^{\otimes 3}] &= \gamma^2\mathbb{E}[(\mathrm{I}-\gamma f''(\theta_*))\eta_0\otimes\varepsilon_1^{\otimes 2}+\varepsilon_1\otimes(\mathrm{I}-\gamma f''(\theta_*))\eta\otimes\varepsilon_1\\ &\quad +\varepsilon_1^{\otimes 2}\otimes(\mathrm{I}-\gamma f''(\theta_*))\eta_0]+\gamma^3\mathbb{E}[\varepsilon_1^{\otimes 3}]+0+O(\gamma^3)\,.\end{aligned}$$

With $\mathbf{M}=(f''(\theta_*)\otimes I\otimes I+I\otimes f''(\theta_*)\otimes I+\ I\otimes I\otimes f''(\theta_*))\,:\,\mathbb{R}^{d\times d\times d}\to\mathbb{R}^{d\times d\times d}$. We need to bound the term $\mathbb{E}[(\mathrm{I}-\gamma f''(\theta_*))\eta_0\otimes\varepsilon_1^{\otimes 2}]$ and its symmetric counterparts. We recall that ε_1 stands for $\varepsilon_1(\theta_0)$ and decompose it as the sum of an additive noise

(independent on θ_0) and a multiplicative one: $\varepsilon_1(\theta_0) = \varepsilon_1(\theta_0) - \varepsilon_1(\theta_*) + \varepsilon_1(\theta_*)$. For the multiplicative part, under Assumption 6, $\mathbb{E}[\|\varepsilon_1(\theta_0) - \varepsilon_1(\theta_*)\|^2 \,|\mathcal{F}_0] \leqslant L[\|\theta_0 - \theta_*\|^2]$, and thus $\mathbb{E}[(\mathrm{I} - \gamma f''(\theta_*))\eta_0 \otimes (\varepsilon_1(\theta_0) - \varepsilon_1(\theta_*))^{\otimes 2}] = O(\gamma^{3/2})$. For the additive part,

$$\begin{aligned}\mathbb{E}[(\mathrm{I} - \gamma f''(\theta_*))\eta_0 \otimes \varepsilon_1(\theta_*)^{\otimes 2}] &= \mathbb{E}[(\mathrm{I} - \gamma f''(\theta_*))\eta_0 \otimes \mathbb{E}[\varepsilon_1(\theta_*)^{\otimes 2}|\mathcal{F}_0]] \\ &= \mathbb{E}[(\mathrm{I} - \gamma f''(\theta_*))\eta_0 \otimes C] \\ &= (\mathrm{I} - \gamma f''(\theta_*))(\bar{\theta}_\gamma - \theta_*) \otimes C \,,\end{aligned}$$

with $C = \mathbb{E}[\varepsilon_1(\theta_*)^{\otimes 2}] = \mathbb{E}[\varepsilon_1(\theta_*)^{\otimes 2}|\mathcal{F}_0]$ as $\varepsilon_1(\theta_*)^{\otimes 2}$ is independent of $\mathcal{F}_0$, and thus $\mathbb{E}[(\mathrm{I} - \gamma f''(\theta_*))\eta_0 \otimes \varepsilon_1(\theta_*)^{\otimes 2}] = O(\gamma)$. Finally, for the crossed term, we use the fact that the multiplicative noise is Lipschitz to get the same result. Overall

$$\begin{aligned}\mathbf{M}\mathbb{E}_{\pi_\gamma}\left[(\theta - \theta_*)^{\otimes 3}\right] &= \gamma^2\left(\mathbb{E}_{\pi_\gamma}[\varepsilon_1^{\otimes 3}] + \frac{1}{\gamma}\mathbb{E}_{\pi_\gamma}[\eta_0 \otimes \varepsilon_1^{\otimes 2} + \varepsilon_1 \otimes \eta_0 \otimes \varepsilon_1 + \varepsilon_1^{\otimes 2} \otimes \eta_0]\right) \\ &= O(\gamma^2) \qquad \text{(C.20)}\end{aligned}$$

Combining (C.20) and the previously established results, we get the Lemma. □

C.2.4 Convergence of second order moments

Poisson equation

We now introduce the Poisson equation; for a function $\varphi : \mathbb{R}^d \to \mathbb{R}^q$ locally-Lipschitz, let $\psi : \mathbb{R}^d \to \mathbb{R}^q$ be a function such that $\pi_\gamma(\psi) = 0$ and the following equations:

$$(I - R_\gamma)\psi_f = \varphi - \pi_\gamma(\varphi) \qquad \text{(C.21)}$$

$$\psi_f = \sum_{i=0}^{\infty} R_\gamma^i(\varphi - \pi_\gamma(\varphi)) \,, \qquad \text{(C.22)}$$

such that for any $x \in \mathbb{R}^d$, $\psi_f(x) = \sum_{i=0}^{\infty} R_\gamma^i(\varphi - \pi_\gamma(\varphi))(x) = \sum_{i=0}^{\infty} \mathbb{E}\left[\varphi(\theta_i^{(\gamma)}(x))\right] - \pi_\gamma(\varphi)$. The convergence of this sum has already been proved for Lipschitz functions, using the contraction in Wasserstein distance between the law of iterates. More generally, for any locally Lipschitz function, Theorem C.8, proved in Section C.3, shows that the solution to the Poisson equation exists, and is locally Lipschitz. As a consequence, we can consider recursively consider the solution to a Poisson equation associated to the solution of a Poisson equation.

Convergence theorem

Theorem C.5. *Let $\varphi : \mathbb{R}^d \to \mathbb{R}^q$ be a locally Lipschitz function, let ψ be the solution of the Poisson Equation (C.21). We assume that $\theta_0 \sim \nu_0$ for some initial distribution ν_0. We study Φ defined as the following random variable in $\mathbb{R}^q$.*

$$\Phi := \frac{1}{n}\sum_{i=0}^{n-1} \varphi(\theta_i^{(\gamma)}(\nu_0)) \,,$$

Then:

$$\mathbb{E}\Phi = \pi_\gamma(\varphi) + \frac{1}{n}\nu_0(\psi) + O(\rho^n) \,.$$

And if $\pi_\gamma(\varphi) = 0$*:*

$$\begin{aligned}\mathbb{E}(\Phi\Phi^\top) - (\mathbb{E}\Phi)(\mathbb{E}\Phi)^\top &= \frac{1}{n}\int_{\mathbb{R}^d}\left[\psi_\gamma(\theta)\psi_\gamma(\theta)^\top - (\psi_\gamma-\varphi)(\theta)(\psi_\gamma-\varphi)(\theta)^\top\right]\mathrm{d}\pi_\gamma(\theta) \\ &\quad + \frac{1}{n^2}\int_{\mathbb{R}^d}\left[\psi_\gamma(\theta)\psi_\gamma(\theta)^\top + \chi^1_\gamma(\theta) - \chi^2_\gamma(\theta)\right]\mathrm{d}\nu_0(\theta) + O(\rho^n)\ ,\end{aligned}$$

where:

1. $\rho := (1-2\mu\gamma(1-\gamma L))^{1/2}$.
2. ψ_γ *is the solution to the Poisson equation associated with* φ.
3. χ^1_γ *is the solution to the Poisson equation associated with* $\psi_\gamma\psi_\gamma^\top$.
4. χ^2_γ *is the solution to the Poisson equation associated with* $(R_\gamma\psi_\gamma)(R_\gamma\psi_\gamma^\top)$.

Proof. In the following proof, in order to improve readability, we skip the dependance on γ for $\theta_n^{(\gamma)}$, which is thus simply denoted θ_n. We have:

$$\begin{aligned}\mathbb{E}\Phi &= \frac{1}{n}\sum_{i=0}^{n-1}\mathbb{E}\left[\varphi(\theta_i^{\nu_0})\right] = \frac{1}{n}\sum_{i=0}^{n-1}\nu_0(R_\gamma^n(\varphi)) \\ &= \pi_\gamma(\varphi) + \frac{1}{n}\sum_{i=0}^{n-1}\nu_0(R_\gamma^n(\varphi-\pi_\gamma(\varphi))) \\ &= \pi_\gamma(\varphi) + \frac{1}{n}\nu_0(\psi_\gamma) + \nu_0(R_\gamma^n(\psi_\gamma)) \\ &= \pi_\gamma(\varphi) + \frac{1}{n}\nu_0(\psi) + O(\rho^n)\ ,\end{aligned}$$

with $\rho := (1-2\mu\gamma(1-\gamma L))^{1/2}$, and using the fact that $\nu_0(R_\gamma^n(\psi_\gamma)) = \nu_0(R_\gamma^n(\psi_\gamma - \pi(\psi_\gamma)))$. We now consider:

$$\begin{aligned}\mathbb{E}\Phi\Phi^\top &= \frac{1}{n^2}\sum_{i,j=0}^{n-1}\mathbb{E}\varphi(\theta_i^{\nu_0})\varphi(\theta_j^{\nu_0})^\top \\ &= \frac{1}{n^2}\sum_{i=0}^{n-1}\left(\mathbb{E}\varphi(\theta_i^{\nu_0})\varphi(\theta_i^{\nu_0})^\top + \sum_{j=i+1}^{n-1}\left[\mathbb{E}\varphi(\theta_i^{\nu_0})\varphi(\theta_j^{\nu_0})^\top + \mathbb{E}\varphi(\theta_j^{\nu_0})\varphi(\theta_i^{\nu_0})^\top\right]\right) \\ &= -\frac{1}{n^2}\sum_{i=0}^{n-1}\nu_0(R_\gamma^i(\varphi(\cdot)\varphi(\cdot)^\top)) \\ &\quad + \frac{1}{n^2}\sum_{i=0}^{n-1}\left(\sum_{j=i+1}^{n-1}\left[\mathbb{E}\varphi(\theta_i^{\nu_0})\varphi(\theta_j^{\nu_0})^\top + \mathbb{E}\varphi(\theta_j^{\nu_0})\varphi(\theta_i^{\nu_0})^\top\right]\right) \\ &= -\frac{1}{n}\pi_\gamma(\varphi(\cdot)\varphi(\cdot)^\top) - \frac{1}{n^2}\nu_0\left(\sum_{i=0}^{\infty}R_\gamma^i\left((\varphi(\cdot)\varphi(\cdot)^\top) - \pi_\gamma(\varphi(\cdot)\varphi(\cdot)^\top\right)\right) \\ &\quad + O(\rho^n) + \frac{1}{n^2}\sum_{i=0}^{n-1}\sum_{j=i}^{n-1}\left[\mathbb{E}\varphi(\theta_i^{\nu_0})(R_\gamma^{j-i}\varphi(\theta_i^{\nu_0}))^\top + \mathbb{E}(R_\gamma^{j-i}\varphi(\theta_i^{\nu_0}))\varphi(\theta_i^{\nu_0})^\top\right] \\ &= -\frac{1}{n}\pi_\gamma(\varphi(\cdot)\varphi(\cdot)^\top) - \frac{1}{n^2}\nu_0\left(\chi^3_\gamma\right) \\ &\quad + \frac{1}{n^2}\sum_{i=0}^{n-1}\left(\sum_{j=0}^{n-1-i}\left[\mathbb{E}\varphi(\theta_i^{\nu_0})(R_\gamma^j\varphi(\theta_i^{\nu_0}))^\top + \mathbb{E}(R_\gamma^j\varphi(\theta_i^{\nu_0}))\varphi(\theta_i^{\nu_0})^\top\right]\right)\ .\end{aligned}$$

With χ^3 the solution to the Poisson equation associated with $\varphi\varphi^\top$. Thus:

$$\begin{aligned}\mathbb{E}\Phi\Phi^\top &= -\frac{1}{n}\pi_\gamma(\varphi(\cdot)\varphi(\cdot)^\top) - \frac{1}{n^2}\nu_0\left(\chi^3_\gamma\right) + O(\rho^n) \\ &\quad + \frac{1}{n^2}\sum_{i=0}^{n-1}\nu_0\left(R^i_\gamma\left[\varphi(\cdot)\psi_\gamma(\cdot) - \varphi(\cdot)R^{n-i}_\gamma\psi(\cdot)^\top\right] + \text{ symmetric term}\right) .\end{aligned}$$

Using that $\frac{1}{n^2}\sum_{i=0}^{n-1}\nu_0\left(R^i_\gamma\left[\varphi(\cdot)R^{n-i}_\gamma\psi(\cdot)^\top\right]\right) = O(\rho^n)$, we get:

$$\begin{aligned}\mathbb{E}\Phi\Phi^\top &= -\frac{1}{n}\pi_\gamma(\varphi(\cdot)\varphi(\cdot)^\top) - \frac{1}{n^2}\nu_0\left(\chi^3_\gamma\right) \\ &\quad + \frac{1}{n}\pi_\gamma\left(\varphi(\cdot)\psi_\gamma(\cdot)^\top\right) + \frac{1}{n^2}\nu_0(\chi^4_\gamma) \\ &\quad + \text{symmetric terms} + O(\rho^n) .\end{aligned}$$

With χ^4 the solution to the Poisson equation associated with $\varphi\psi_\gamma^\top$.
For the first order terms, which scale as $\frac{1}{n}$, we have:

$$\begin{aligned}\mathbb{E}(\Phi\Phi^\top) - (\mathbb{E}\Phi)(\mathbb{E}\Phi)^\top &= \frac{1}{n}\pi_\gamma\left(-\varphi(\cdot)\varphi(\cdot)^\top + \varphi(\cdot)\psi_\gamma(\cdot)^\top + \psi_\gamma(\cdot)\varphi(\cdot)^\top\right) \\ &= \frac{1}{n}\pi_\gamma\left(-\varphi(\cdot)\varphi(\cdot)^\top + \varphi(\cdot)\psi(\cdot)^\top + \psi(\cdot)\varphi(\cdot)^\top\right) \\ &= \frac{1}{n}\pi_\gamma\left(-(\varphi-\psi)(\cdot)(\varphi-\psi)(\cdot)^\top + \psi(\cdot)\psi(\cdot)^\top\right) \\ &= \frac{1}{n}\pi_\gamma\left(-(R_\gamma\psi)(\cdot)(R_\gamma\psi)(\cdot)^\top + \psi(\cdot)\psi(\cdot)^\top\right) ,\end{aligned}$$

using the fact that for the solution to the Poisson equation: $\psi - R_\gamma\psi = \varphi$, *i.e.*, $\psi - \varphi = R_\gamma\psi$.
This can also be written:

$$\mathbb{E}(\Phi\Phi^\top) - (\mathbb{E}\Phi)(\mathbb{E}\Phi)^\top = \frac{1}{n}\int_{\mathbb{R}^d}\left[\psi_\gamma(\theta)\psi_\gamma(\theta)^\top - (\psi_\gamma-\varphi)(\theta)(\psi_\gamma-\varphi)(\theta)^\top\right]\mathrm{d}\pi_\gamma(\theta) .$$

For the following order in $O(1/n^2)$, we have:

$$\begin{aligned}\mathbb{E}(\Phi\Phi^\top) - (\mathbb{E}\Phi)(\mathbb{E}\Phi)^\top - \frac{\text{term}}{n} &= \frac{-1}{n^2} + \frac{1}{n^2}\nu_0(-\chi^3_\gamma + \chi^4_\gamma) + \text{ symmetric term} \\ &= \frac{1}{n^2}\nu_0(\chi^1_\gamma - \chi^2_\gamma) ,\end{aligned}$$

using the linearity of R_γ and the fact that: $-\varphi\varphi^\top + \psi_\gamma\varphi^\top + \varphi\psi_\gamma^\top = -(\varphi-\psi)(\cdot)(\varphi-\psi)(\cdot)^\top + \psi(\cdot)\psi(\cdot)^\top$, thus: $\nu_0(-\chi^3_\gamma + \chi^4_\gamma) = \nu_0(\chi^1_\gamma - \chi^2_\gamma)$.
This is the expected result. □

Application in the quadratic case ($f = f_\Sigma$), for $\varphi = \mathrm{I}$

We consider, the stochastic gradient descent algorithm (4.1), for the quadratic function $f_\Sigma(\theta) := \left\|\Sigma^{1/2}(\theta-\theta_*)\right\|^2$. We consider the classical stochastic approximation noise oracle of the least mean squares (LMS) algorithm:

$$\begin{aligned}\theta_{n,\gamma} - \theta_* &= (I - \gamma\Sigma)(\theta_{n-1,\gamma} - \theta_*) + \gamma\varepsilon_n(\theta_{n-1,\gamma}) \\ \varepsilon_n(\theta_{n-1,\gamma}) &= (\Sigma - x_n\otimes x_n)(\theta_{n-1,\gamma} - \theta_*) + (y_n - \langle\theta_*, x_n\rangle)x_n .\end{aligned}$$

We first recall the observation made in Section C.2.3: for quadratic functions, under the stationary distribution, the mean value of the iterate is the optimal point. According to Lemma C.1, we have $\pi_\gamma(\varphi) = 0$. The following Lemma recovers result from Défossez and Bach (2015), as a corollary of our more general theorem.

Lemma C.6. *If f is a quadratic function f_Σ, and we consider the LMS algorithm with $\gamma L \leqslant 1/2$, then with $\rho \leqslant (1-\gamma\mu)$, we have:*

$$\begin{aligned}\mathbb{E}\left[(\bar{\theta}_n^{(\gamma)} - \theta_*)^{\otimes 2}\right] &= \frac{1}{n^2\gamma^2}\Sigma^{-1}\Omega(\theta_0-\theta_*)^{\otimes 2}\Sigma^{-1} + \frac{1}{n}\Sigma^{-1}\big[\mathbb{E}_{\pi_\gamma}\varepsilon^{\otimes 2}\big]\Sigma^{-1} \\ &\quad - \frac{1}{n^2\gamma}\Sigma^{-1}\Omega\big[\Sigma\otimes \mathrm{I} + \mathrm{I}\otimes\Sigma - \gamma T\big]^{-1}\big[\mathbb{E}\xi^{\otimes 2}\big]\Sigma^{-1}\,.\end{aligned}$$

With $\Omega := (\Sigma\otimes I + I\otimes\Sigma - \gamma\Sigma\otimes\Sigma)(\Sigma\otimes I + I\otimes\Sigma - \gamma T)^{-1}$.

Moreover, the value of ρ is known: $\rho = (1-2\gamma\mu(1-\gamma L)) \leqslant (1-\gamma\mu)$ if $\gamma L \leqslant 1/2$, with $\mu = \lambda_{\min}(\Sigma)$.

Proof. We consider the linear function φ which is $\varphi(\theta) = \theta - \theta_*$. We then have that $\psi(\theta) = (\gamma\Sigma)^{-1}(\theta-\theta_*)$. Indeed from Equation (C.22), for any θ_0:

$$\psi(\theta_0) = \sum_{i=0}^{\infty}\mathbb{E}(\theta_{i,\gamma}^{(\theta_0)}) - \theta_* = \sum_{i=0}^{\infty}(I-\gamma\Sigma)^i(\theta_0-\theta_*) = (\gamma\Sigma)^{-1}(\theta_0-\theta_*)\,.$$

We can thus apply Theorem 4.4 to get a bound on $\mathbb{E}\left((\bar{\theta}_n^{(\gamma)} - \theta_*)(\bar{\theta}_n^{(\gamma)} - \theta_*)^\top\right)$. Indeed, with the previous notations, $\varphi = \bar{\theta}_n^{(\gamma)} - \theta_*$. We recall that:

$$\begin{aligned}\mathbb{E}(\Phi\Phi^\top) - (\mathbb{E}\Phi)(\mathbb{E}\Phi)^\top &= \frac{1}{n}\int_{\mathbb{R}^d}\left[\psi_\gamma(\theta)\psi_\gamma(\theta)^\top - (\psi_\gamma-\varphi)(\theta)(\psi_\gamma-\varphi)(\theta)^\top\right]\mathrm{d}\pi_\gamma(\theta) \\ &\quad + \frac{1}{n^2}\int_{\mathbb{R}^d}\left[\psi_\gamma(\theta)\psi_\gamma(\theta)^\top + \chi_\gamma^1(\theta) - \chi_\gamma^2(\theta)\right]\mathrm{d}\nu_0(\theta) + O(\rho^n)\,.\end{aligned}$$

Term proportional to $1/n$.
We need to compute the expectation under the stationary distribution of $\varphi(\theta)^{\otimes 2}$. For simplicity, we here denote $\mathbb{E}\varepsilon^{\otimes 2} = \int_{\mathbb{R}^d}\varepsilon_1(\theta)^{\otimes 2}\pi_\gamma(\mathrm{d}\theta)$. We have, according to Lemma C.1:

$$\int_{\mathbb{R}^d}(\theta-\theta_*)^{\otimes 2}\pi_\gamma(\mathrm{d}\theta) = \gamma\big[\Sigma\otimes\mathrm{I} + \mathrm{I}\otimes\Sigma - \gamma\Sigma\otimes\Sigma\big]^{-1}\mathbb{E}\varepsilon^{\otimes 2}.$$

The expectation of $\psi(\theta)\psi(\theta)^\top$ under the stationary is

$$\begin{aligned}\int_{\mathbb{R}^d}\psi(\theta)\psi(\theta)^\top\pi_\gamma(\mathrm{d}\theta) &= (\gamma\Sigma)^{-1}\gamma\big[\Sigma\otimes\mathrm{I} + \mathrm{I}\otimes\Sigma - \gamma\Sigma\otimes\Sigma\big]^{-1}\mathbb{E}\varepsilon^{\otimes 2}(\gamma\Sigma)^{-1} \\ &= \frac{1}{\gamma}(\Sigma^{-1}\otimes\Sigma^{-1})\big[\Sigma\otimes\mathrm{I} + \mathrm{I}\otimes\Sigma - \gamma\Sigma\otimes\Sigma\big]^{-1}\mathbb{E}\varepsilon^{\otimes 2}\,.\end{aligned}$$

Moreover,

$$\int_{\mathbb{R}^d}(\varphi(\theta)-\psi(\theta))(\varphi(\theta)-\psi(\theta))^\top\pi_\gamma(\mathrm{d}\theta) = \big[\mathrm{I}-(\gamma\Sigma)^{-1}\big]\gamma\big[\Sigma\otimes\mathrm{I}+\mathrm{I}\otimes\Sigma\big]^{-1}\mathbb{E}\varepsilon^{\otimes 2}\big[\mathrm{I}-(\gamma\Sigma)^{-1}\big].$$

Adding both these results and simplifying by $\big[\Sigma\otimes\mathrm{I} + \mathrm{I}\otimes\Sigma - \gamma\Sigma\otimes\Sigma\big]$, we get the following $1/n$-term:

$$\frac{1}{n}\mathbb{E}_{\theta\sim\pi_\gamma}\left[\psi(\theta)\psi(\theta)^\top - (R_\gamma\psi)(\theta)(R_\gamma\psi)(\theta)^\top\right] = \frac{1}{n}\Sigma^{-1}\Big[\int_{\mathbb{R}^d}\left(\varepsilon_1(\theta)^{\otimes 2}\right)\pi_\gamma(\mathrm{d}\theta)\Big]\Sigma^{-1}\,.$$

Term proportional to $1/n^2$.
We assume $\nu_0 = \delta_{\theta_0}$. This term is composed of three terms:

$$\begin{aligned} T_1 &:= -\mathbb{E}_{\theta_0\sim\nu_0}\left[\psi(\theta_0)\right]\mathbb{E}_{\theta_0\sim\nu_0}\left[\psi(\theta_0)\right]^\top \\ \psi(\theta_0) &= (\gamma\Sigma)^{-1}(\theta_0-\theta_*) \\ T_1 &= -\frac{1}{\gamma^2}\Sigma^{-1}\left[(\theta_0-\theta_*)^{\otimes 2}\right]\Sigma^{-1} . \end{aligned}$$

We note that, using $\psi = (\gamma\Sigma)^{-1}\varphi$, and $R_\gamma\psi = \psi - \varphi = -(I-(\gamma\Sigma)^{-1})\varphi$ that:

$$\begin{aligned} T_2 &:= \nu_0(\chi^1_\gamma) \\ &= (I-(\gamma\Sigma)^{-1})\nu_0(\chi^3_\gamma)(I-(\gamma\Sigma)^{-1}) . \end{aligned}$$

Similarly:

$$\begin{aligned} T_2 &:= \nu_0(\chi^1_\gamma) \\ &= (\gamma\Sigma)^{-1}\nu_0(\chi^3_\gamma)(\gamma\Sigma)^{-1} . \end{aligned}$$

Where we recall that denote χ^3_γ the solution to the Poisson equation associated with $\theta\mapsto\varphi(\theta)^{\otimes 2}$. We can compute explicitly this solution, indeed, following Equation C.14:

$$\begin{aligned} \mathbb{E}\left[(\theta^x_{n,\gamma}-\theta_*)^{\otimes 2}\right] &= (I-\gamma\Sigma\otimes I-\gamma I\otimes\Sigma+\gamma^2 M)\mathbb{E}\left[(\theta^x_{n-1,\gamma}-\theta_*)^{\otimes 2}\right]+\mathbb{E}[\xi_n^{\otimes 2}] \\ \chi^3_\gamma(x) &:= \sum_{i=1}^{\infty}\mathbb{E}\left[(\theta^x_{n,\gamma}-\theta_*)^{\otimes 2}\right]-\pi_\gamma(\varphi(\theta)^{\otimes 2}) \\ &= (\gamma\Sigma\otimes I+\gamma I\otimes\Sigma-\gamma^2 M)^{-1}\left[\mathbb{E}\left[(\theta^x_{0,\gamma}-\theta_*)^{\otimes 2}\right]-\pi_\gamma(\varphi(\theta)^{\otimes 2}))\right] \\ \mathbb{E}_{\theta\sim\nu_0}\left[\chi^3_\gamma\right] &:= (\gamma\Sigma\otimes I+\gamma I\otimes\Sigma-\gamma^2 M)^{-1}\left[(\theta_0-\theta_*)^{\otimes 2}-\pi_\gamma(\varphi(\theta)^{\otimes 2}))\right] . \end{aligned}$$

Simplification comes from the fact that we study an arithmetico-geometric recursion of the form $w_{n+1} = aw_n + b$, $a<1$, and study $\sum_{i=0}^{\infty} w_n - w_\infty = (1-a)^{-1}(w_0-w_\infty)$. Here we cannot apply the recursion with $(\Sigma\otimes I + I\otimes\Sigma-\gamma\Sigma\otimes\Sigma)$ because then b would depend on n. Finally,

$$\begin{aligned} T_2+T_3 &= \frac{1}{\gamma}(\Sigma^{-1}\otimes\Sigma^{-1})(\Sigma\otimes I+I\otimes\Sigma-\gamma\Sigma\otimes\Sigma)\mathbb{E}_{\theta\sim\nu_0}\left[\chi(x)\right] \\ &= (\Sigma^{-1}\otimes\Sigma^{-1})\Omega\left[(\theta_0-\theta_*)^{\otimes 2}-\gamma(\Sigma\otimes I+I\otimes\Sigma-\gamma M)^{-1}\mathbb{E}[\xi_1^{\otimes 2}])\right] . \end{aligned}$$

With: $\Omega = (\Sigma\otimes I+I\otimes\Sigma-\gamma\Sigma\otimes\Sigma)(\Sigma\otimes I+I\otimes\Sigma-\gamma T)^{-1}$.
Overall, we get that:

$$\begin{aligned} \mathbb{E}\bar{\theta}_n-\theta_* &= \frac{1}{n}(\gamma\Sigma)^{-1}(\theta_0-\theta_*) \\ \mathrm{cov}(\bar{\theta}_n) &= \frac{1}{n}\Sigma^{-1}\left[\mathbb{E}\varepsilon^{\otimes 2}\right]\Sigma^{-1}-\frac{1}{n^2\gamma}\left[\Sigma^{-1}\otimes\Sigma^{-1}\right]\Omega\left[\Sigma\otimes\mathrm{I}+\mathrm{I}\otimes\Sigma-\gamma T\right]^{-1}\left[\mathbb{E}\xi^{\otimes 2}\right] \\ &\quad +\frac{1}{n^2}(\Sigma^{-1}\otimes\Sigma^{-1})(\Omega-I)(\theta_0-\theta_*)^{\otimes 2} . \end{aligned}$$

Finally:

$$\begin{aligned} \mathbb{E}\left[(\bar{\theta}^{(\gamma)}_n-\theta_*)^{\otimes 2}\right] &= \frac{1}{n^2\gamma^2}(\Sigma^{-1}\otimes\Sigma^{-1})(\Omega)(\theta_0-\theta_*)^{\otimes 2}+\frac{1}{n}\Sigma^{-1}\left[\mathbb{E}\varepsilon^{\otimes 2}\right]\Sigma^{-1} \\ &\quad -\frac{1}{n^2\gamma}\left[\Sigma^{-1}\otimes\Sigma^{-1}\right]\Omega\left[\Sigma\otimes\mathrm{I}+\mathrm{I}\otimes\Sigma-\gamma T\right]^{-1}\left[\mathbb{E}\xi^{\otimes 2}\right] . \end{aligned}$$

In the semi stochastic setting, we would get:

$$\mathbb{E}\left[(\bar{\theta}_n^{(\gamma)}-\theta_*)^{\otimes 2}\right] = \frac{1}{n^2\gamma^2}(\Sigma^{-1}\otimes\Sigma^{-1})(\theta_0-\theta_*)^{\otimes 2}+\frac{1}{n}\Sigma^{-1}\left[\mathbb{E}\varepsilon^{\otimes 2}\right]\Sigma^{-1} - \frac{1}{n^2\gamma}\left[\Sigma^{-1}\otimes\Sigma^{-1}\right]\left[\Sigma\otimes \mathrm{I}+\mathrm{I}\otimes\Sigma-\gamma\Sigma\otimes\Sigma\right]^{-1}\left[\mathbb{E}\xi^{\otimes 2}\right] .$$

□

C.3 Further properties of the Markov chain $(\theta_k^{(\gamma)})_{k\geqslant 0}$

We give uniform bound on the moments of the chain $(\theta_k^{(\gamma)})_{k\geqslant 0}$ for $\gamma > 0$. We denote $\delta_n = \|\theta_n - \theta_*\|$. Denote by

$$\kappa = 2\mu L/(\mu+L) . \tag{C.23}$$

For $p \geqslant 1$ define

$$\mathrm{m}_p = \mathbb{E}^{1/p}\left[\|\varepsilon_1(\theta_*)\|^p\right] \text{ , for } p \geqslant 1 . \tag{C.24}$$

We give a bound on the p-order moment of the chain, under the assumption that the noise has a moment of order $2p$.

Lemma C.7 (Final iterate)**.** *Under Assumptions A1,A2, A3, A6, one has the following bound on the $\mathbb{E}^{1/p}[\delta_{n+1}^{2p}]$, $p=1,2$. For the 2^{nd} order moment,*

$$\mathbb{E}[\delta_{n+1}^2] \leqslant (1-2\gamma\mu(1-\gamma L))^n\,\delta_0^2+\frac{\gamma\sigma^2}{\mu} . \tag{C.25}$$

For th 4^{th}-order moment, for $\gamma \leqslant \frac{1}{18L}$

$$\mathbb{E}^{1/2}[\delta_{n+1}^4] \leqslant (1-2\gamma\mu(1-9\gamma L))\,\mathbb{E}^{1/2}[\delta_n^4]+20\gamma^2\tau^2$$
$$\mathbb{E}^{1/2}[\delta_n^4] \leqslant (1-2\gamma\mu(1-9\gamma L))^n\mathbb{E}^{1/2}[\delta_0^4]+\frac{20\gamma\tau^2}{\mu} .$$

More generally, assume A1-A2-A3-A4(2p), for $p \geqslant 1$. There exist numerical constants C_p, D_p that only depend on p, such that, if $\gamma L \leqslant 1/2C_p$,

$$\mathbb{E}_\theta^{1/p}\left[\left\|\theta_n^{(\gamma)}-\theta_*\right\|^{2p}\right] \leqslant (1-2\gamma\mu(1-C_p\gamma L))^n\mathbb{E}_\theta^{1/p}\left[\|\theta_0-\theta_*\|^{2p}\right]+\frac{D_p\gamma\mathrm{m}_{2p}^2}{\mu} .$$

Moreover, under stationary distribution π_γ, under the Assumptions above, one has:

$$\mathbb{E}_{\pi_\gamma}\left[\|\delta_n\|^{2p}\right] \leqslant \left(\frac{D_p\gamma\mathrm{m}_{2p}^2}{\mu}\right)^p . \tag{C.26}$$

Remark: Note that there is no contradiction between Equation (C.26) and Theorem 4.6, as for any $p \geqslant 2$, one has for $g(\theta) = \|\theta-\theta_*\|^2$ and h_g the solution to the Poisson equation, that $h_g''(\theta_*) = 0$, so that the first term in the development (of order γ) is indeed 0.

Lemma C.7. We only prove the result for $p = 1, 2$ as it then naturally extends for any p.

The proof for the 2nd moment is very close to the one from (Needell et al., 2014) but we extend it without a.s. Lipschitzness (Assumption **A**4) but with Assumption **A**6. We recall that $\theta_{n+1} = \theta_n - \gamma f'(\theta_n) + \gamma\varepsilon_{n+1}$.

We have that

$$\|\theta_{n+1}-\theta_*\|^2 \;=\; \|\theta_n-\theta_*-\gamma f'(\theta_n)+\gamma\varepsilon_{n+1}\|^2\,. \tag{C.27}$$

According to assumption **A**3, we have θ_n is $\mathcal{F}_n$ measurable, and $\mathbb{E}[\varepsilon_{n+1}|\mathcal{F}_n]=0$. Thus $\mathbb{E}[\langle\theta_n-\theta_*,\varepsilon_{n+1}\rangle|\mathcal{F}_n]=0$.

$$\begin{aligned}\mathbb{E}[\|\theta_{n+1}-\theta_*\|^2|\mathcal{F}_n] &= \mathbb{E}[\|\theta_n-\theta_*\|^2|\mathcal{F}_n]-2\gamma\mathbb{E}[\langle f'(\theta_n),\theta_n-\theta_*\rangle|\mathcal{F}_n]\\ &\quad+\gamma^2\mathbb{E}[\|f_n'(\theta_n)-f_n'(\theta_*)\|^2|\mathcal{F}_n]+2\gamma^2\mathbb{E}[\|f_n'(\theta_*)\|^2|\mathcal{F}_n]\,.\end{aligned} \tag{C.28}$$

Moreover, under Assumption **A**6, one has that $\mathbb{E}[\|f_n'(\theta_*)\|^2|\mathcal{F}_n]=\mathbb{E}[\|\varepsilon_1(\theta_*)\|^2]\leqslant\tau^2$ (using Hölder's inequality), and $\mathbb{E}[\|f_n'(\theta_n)-f_n'(\theta_*)\|^2|\mathcal{F}_n]\leqslant L\langle f'(\theta_n)-f'(\theta_*),\theta_n-\theta_*\rangle$. Thus:

$$\begin{aligned}\mathbb{E}[\delta_{n+1}^2|\mathcal{F}_n] &\leqslant \mathbb{E}[\delta_n^2|\mathcal{F}_n]-2\gamma\langle f'(\theta_n)-f'(\theta_*),\theta_n-\theta_*\rangle+2\gamma^2L\langle f'(\theta_n)-f'(\theta_*),\theta_n-\theta_*\rangle\\ &\quad+\gamma^2\tau^2\\ &\leqslant (1-2\gamma\mu(1-\gamma L))\,\delta_n^2+2\gamma^2\tau^2\,.\end{aligned} \tag{C.29}$$

Thus if $\gamma\leqslant\frac{1}{L}$, we have

$$\mathbb{E}[\delta_{n+1}^2] \;\leqslant\; (1-2\gamma\mu(1-\gamma L))\,\mathbb{E}[\delta_n^2]+2\gamma^2\tau^2\,. \tag{C.30}$$

Thus if $\gamma L\leqslant 1$.

$$\mathbb{E}[\delta_{n+1}^2] \;\leqslant\; (1-2\gamma\mu(1-\gamma L))^n\,\delta_0^2+\gamma^2\tau^2\sum_{i=0}^{n-1}(1-2\gamma\mu)^i \tag{C.31}$$

$$\;=\; (1-2\gamma\mu(1-\gamma L))^n\,\delta_0^2+\frac{\gamma\tau^2}{\gamma\mu(1-\gamma L)}\,. \tag{C.32}$$

□

Lemma C.7. We have that

$$\begin{aligned}\delta_{n+1}^4 &= \left(\|\theta_n-\theta_*\|^2-2\gamma\langle f_n'(\theta_n),\theta_n-\theta_*\rangle+\gamma^2\|f_n'(\theta_n)\|^2\right)^2\\ &= \left(\delta_n^2-2\gamma\langle f_n'(\theta_n),\theta_n-\theta_*\rangle+\gamma^2\|f_n'(\theta_n)\|^2\right)^2\\ &= \delta_n^4-4\gamma\delta_n^2\langle f_n'(\theta_n),\theta_n-\theta_*\rangle+4\gamma^2\langle f_n'(\theta_n),\theta_n-\theta_*\rangle^2+2\gamma^2\delta_n^2\|f_n'(\theta_n)\|^2\\ &\quad-4\gamma^3\langle f_n'(\theta_n),\theta_n-\theta_*\rangle\|f_n'(\theta_n)\|^2+\gamma^4\|f_n'(\theta_n)\|^4.\end{aligned}$$

Moreover:

$$\begin{aligned}\mathbb{E}[\|f_n'(\theta_n)\|^p|\mathcal{F}_n] &\leqslant 2^{p-1}\left(\mathbb{E}[\|f_n'(\theta_n)-f_n'(\theta_*)\|^p|\mathcal{F}_n]+\mathbb{E}[\|f_n'(\theta_*)\|^p|\mathcal{F}_n]\right)\\ &\leqslant 2^{p-1}\left(\|f_n'(\theta_n)-f_n'(\theta_*)\|^p+\mathbb{E}[\|\varepsilon_1(\theta_*)\|^p|\mathcal{F}_n]\right)\\ &\leqslant 2^{p-1}\left(\|f_n'(\theta_n)-f_n'(\theta_*)\|^p+\tau^p\right),\end{aligned} \tag{C.33}$$

using at the first line Minkowski's inequality and the fact that $x\mapsto x^p$ is convex on $\mathbb{R}^+$ for $p=1,\dots,4$ thus $(x+y)^p\leqslant 2^{p-1}(x^p+y^p)$, and at the last line the Assumption **A**6 on the noise: $\mathbb{E}[\|\varepsilon_1(\theta_*)\|^p|\mathcal{F}_n]\leqslant\tau^p$.

Thus,

$$\mathbb{E}[\delta_{n+1}^4|\mathcal{F}_n] \;\leqslant\; \delta_n^4-4\gamma\delta_n^2\mathbb{E}[\langle f_n'(\theta_n),\theta_n-\theta_*\rangle|\mathcal{F}_n]+4\gamma^2\mathbb{E}[\langle f_n'(\theta_n),\theta_n-\theta_*\rangle^2|\mathcal{F}_n]$$

$$
\begin{aligned}
&+2\gamma^2\delta_n^2\mathbb{E}[\|f_n'(\theta_n)\|^2|\mathcal{F}_n] - 4\gamma^3\mathbb{E}[\langle f_n'(\theta_n), \theta_n-\theta_*\rangle\|f_n'(\theta_n)\|^2|\mathcal{F}_n]\\
&+\gamma^4\mathbb{E}[\|f_n'(\theta_n)\|^4|\mathcal{F}_n]\\
\leqslant\ & \delta_n^4 - 4\gamma\delta_n^2\langle f'(\theta_n), \theta_n-\theta_*\rangle + 4\gamma^2\mathbb{E}[\|f_n'(\theta_n)\|^2\delta_n^2|\mathcal{F}_n]\\
&+2\gamma^2\delta_n^2\mathbb{E}[\|f_n'(\theta_n)\|^2|\mathcal{F}_n] + 4\gamma^3\delta_n\mathbb{E}[\|f_n'(\theta_n)\|^3|\mathcal{F}_n] + \gamma^4\mathbb{E}[\|f_n'(\theta_n)\|^4|\mathcal{F}_n]\\
\leqslant\ & \delta_n^4 - 4\gamma\delta_n^2\langle f'(\theta_n), \theta_n-\theta_*\rangle + 12\gamma^2\delta_n^2\mathbb{E}[\|f_n'(\theta_n)-f_n'(\theta_*)\|^2|\mathcal{F}_n]\\
&+16\gamma^3\delta_n\mathbb{E}[\|f_n'(\theta_n)-f_n'(\theta_*)\|^3|\mathcal{F}_n] + 8\gamma^4\mathbb{E}[\|f_n'(\theta_n)-f_n'(\theta_*)\|^4|\mathcal{F}_n]\\
&+12\gamma^2\tau^2\delta_n^2 + 16\gamma^3\delta_n\tau^3 + 8\gamma^4\tau^4,
\end{aligned}
$$

using Cauchy Schwartz several times for the second inequality and equation (C.33) for the third one.

Then, using part (ii) of Assumption **A**6:

$$
\begin{aligned}
\mathbb{E}[\delta_{n+1}^4|\mathcal{F}_n] \leqslant\ & \delta_n^4 - 4\gamma\delta_n^2\langle f'(\theta_n), \theta_n-\theta_*\rangle + 12\gamma^2L\delta_n^2\langle f'(\theta_n), \theta_n-\theta_*\rangle\\
&+16\gamma^3L^2\delta_n^2\langle f'(\theta_n), \theta_n-\theta_*\rangle + 8\gamma^4L^3\delta_n^2\langle f'(\theta_n), \theta_n-\theta_*\rangle\\
&+12\gamma\tau^2\delta_n^2 + 8\gamma^2\tau^2\delta_n^2 + 8\gamma^4\tau^4 + 8\gamma^4\tau^4\\
=\ & \delta_n^4 + (-4\gamma + 12\gamma^2L + 16\gamma^3L^2 + 8\gamma^4L^3)\delta_n^2\langle f'(\theta_n), \theta_n-\theta_*\rangle\\
&+(12\gamma^2\tau^2 + 8\gamma^2\tau^2)\delta_n^2 + 16\gamma^4\tau^4\\
\leqslant\ & \delta_n^4 - 4\gamma(1-9\gamma L)\delta_n^2\langle f'(\theta_n), \theta_n-\theta_*\rangle + 20\gamma^2\tau^2\delta_n^2 + 16\gamma^4\tau^4,
\end{aligned}
$$

using $\gamma L\leqslant 1$ at the last line. Finally, using the smooth and strong convexity equation (C.5), we have:

$$
\mathbb{E}[\delta_{n+1}^4|\mathcal{F}_n] \leqslant (1-4\gamma\mu(1-9\gamma L))\,\delta_n^4 + 20\gamma^2\tau^2\delta_n^2 + 16\gamma^4\tau^4,
$$

Thus finally:

$$
\begin{aligned}
\mathbb{E}[\delta_{n+1}^4] &\leqslant (1-4\gamma\mu(1-9\gamma L))\,\mathbb{E}[\delta_n^4] + 20\gamma^2\tau^2\mathbb{E}[\delta_n^2] + 16\gamma^4\tau^4\\
&\leqslant \left((1-4\gamma\mu(1-9\gamma L))^{1/2}\,\mathbb{E}[\delta_n^4]^{1/2} + 20\gamma^2\tau^2\right)^2.
\end{aligned}
$$

Using that $20\gamma^2\tau^2\mathbb{E}[\delta_n^2]\leqslant(1-4\gamma\mu(1-9\gamma L))^{1/2}\mathbb{E}[\delta_n^4]^{1/2}40\gamma^2\tau^2$ *i.e.*, $\mathbb{E}[\delta_n^2]\leqslant\mathbb{E}[\delta_n^4]^{1/2}$, and $(1-4\gamma\mu(1-9\gamma L))^{1/2}\geqslant 1/2$ which is true if $\gamma\leqslant\frac{1}{9L}$ and $(1-4\gamma\mu(1-9\gamma L))\geqslant(1-4/9)^{1/2}\geqslant 1/2$.

$$
\mathbb{E}^{1/2}[\delta_{n+1}^4] \leqslant (1-2\gamma\mu(1-9\gamma L))\,\mathbb{E}^{1/2}[\delta_n^4] + 20\gamma^2\tau^2.
$$

If $9\gamma L\leqslant 1$.

Which concludes the proof. □

Theorem C.8. *Assume **A**1-**A**2-**A**3-**A**4($2k_2$)-**A**8(k_1)- for $k_1,k_2\in\mathbb{N}$, $k_1\geqslant 1$. Let $g:\mathbb{R}^d\to\mathbb{R}$ satisfying **A**5(k_1,k_2) for $k_2\in\mathbb{N}$. Then, there exists $C_{k_2}\geqslant 0$ only depending on k_2 such that for all $\gamma\in(0,C_{k_2}/L)$, for all initial point $\theta\in\mathbb{R}^d$, there exists C such that for all $n\geqslant 1$:*

$$
\left|\mathbb{E}_\theta\left[n^{-1}\sum_{i=1}^n\left\{g(\theta_i^{(\gamma)})\right\}\right] - \int_{\mathbb{R}^d} g(\theta)\pi_\gamma(\mathrm{d}\theta)\right| \leqslant Cn^{-1}\,.
$$

Proof.

$$
\left|\sum_{i=1}^n\left(\mathbb{E}_\theta\left[g(\theta_{i,\gamma}^\theta)\right] - \int_{\mathbb{R}^d} g(\theta)\pi_\gamma(\mathrm{d}\theta)\right)\right| = \sum_{i=1}^n\left|\left(\int_{y\in\mathbb{R}^d}\mathbb{E}_\theta\left[g(\theta_{i,\gamma}^\theta) - g(\theta_{i,\gamma}^y)\right]\pi_\gamma(y)\right)\right|
$$

$$= \sum_{i=1}^{n} \left(\int_{y \in \mathbb{R}^d} \mathbb{E}_\theta \left[\left\| g(\theta^\theta_{i,\gamma}) - g(\theta^y_{i,\gamma}) \right\| \right] \pi_\gamma(y) \right) .$$

Using Lemma C.10, a.s.,

$$\left\| g(\theta^\theta_{i,\gamma}) - g(\theta^y_{i,\gamma}) \right\| \leqslant a_g \left\| \theta^\theta_{i,\gamma} - \theta^y_{i,\gamma} \right\| \left((b_g + \left\| \theta^\theta_{i,\gamma} - \theta_* \right\|^{k_2} + \left\| \theta^y_{i,\gamma} - \theta_* \right\|^{k_2}) \right) .$$

By Cauchy Schwartz, then Minkowski:

$$\begin{aligned}
\mathbb{E}_\theta \left[\left\| g(\theta^\theta_{i,\gamma}) - g(\theta^y_{i,\gamma}) \right\| \right] \leqslant & a_g \mathbb{E}_\theta^{1/2} \left[\left\| \theta^\theta_{i,\gamma} - \theta^y_{i,\gamma} \right\|^2 \right] \mathbb{E}_\theta^{1/2} \left[(b_g + \left\| \theta^\theta_{i,\gamma} - \theta_* \right\|^{k_2} + \left\| \theta^y_{i,\gamma} - \theta_* \right\|^{k_2})^2 \right] \\
\leqslant & a_g \left(W_2(R^n_\gamma(\theta, .), R^n_\gamma(y, .)) \right)^{1/2} \\
& \times \left(b_g + \mathbb{E}_\theta^{1/2} \left[\left\| \theta^\theta_{i,\gamma} - \theta_* \right\|^{2k_2} \right] + \mathbb{E}_\theta^{1/2} \left[\left\| \theta^y_{i,\gamma} - \theta_* \right\|^{2k_2} \right] \right) .
\end{aligned}$$

With $\rho = (1 - \gamma\mu(1 - \gamma L))$, we have, using Lemma C.7 , which implies that:

$$\mathbb{E}_\theta^{1/2} \left[\left\| \theta^{(\gamma)}_n - \theta_* \right\|^{2p} \right] \leqslant 2^{p/2-1} \mathbb{E}_\theta^{1/2} \left[\left\| \theta^{(\gamma)}_0 - \theta_* \right\|^{2p} \right] + 2^{p/2} \left(\frac{D_p \gamma \mathrm{m}^2_{2p}}{\mu} \right)^{p/2} .$$

$$\begin{aligned}
\mathbb{E}_\theta \left[\left\| g(\theta^\theta_{i,\gamma}) - g(\theta^y_{i,\gamma}) \right\| \right] \leqslant & a_g \rho^{n/2} \left\| \theta - y \right\| \left(b_g + 2^{p/2-1} \mathbb{E}_\theta^{1/2} \left[\left\| \theta^{(\gamma)}_0 - \theta_* \right\|^{2k_2} \right] \right. \\
& \left. + 2^{p/2-1} \left\| y - \theta_* \right\|^{k_2} 2^{p/2+1} \left(\frac{D_p \gamma \mathrm{m}^2_{2p}}{\mu} \right)^{p/2} \right) .
\end{aligned}$$

Thus

$$\left| \mathbb{E}_\theta \left[n^{-1} \sum_{i=1}^{n} \left\{ g(\theta^{(\gamma)}_i) \right\} \right] - \int_{\mathbb{R}^d} g(\theta) \pi_\gamma(\mathrm{d}\theta) \right| \leqslant \frac{C}{n} \sum_{i=1}^{n} \rho^{n/2} \leqslant \frac{C}{\gamma\mu n}$$

$$\begin{aligned}
C = & a_g \int_{\mathbb{R}^d} \left(\left\| \theta - y \right\| \left(b_g + 2^{p/2-1} \mathbb{E}_\theta^{1/2} \left[\left\| \theta^{(\gamma)}_0 - \theta_* \right\|^{2k_2} \right] + 2^{p/2-1} \left\| y - \theta_* \right\|^{k_2} \right. \right. \\
& \left. \left. 2^{p/2+1} \left(\frac{D_p \gamma \mathrm{m}^2_{2p}}{\mu} \right)^{p/2} \right) \mathrm{d}\pi_\gamma(y) \right) .
\end{aligned}$$

□

C.4 Regularity of the gradient flow and estimates on Poisson solution

Let $k \in \mathbb{N}^*$ and consider the following assumption.

A8 (k). *$f \in C^k(\mathbb{R}^d)$ and there exists $M \geqslant 0$ such that for all $i \in \{2, \ldots, k\}$, $\sup_{\theta \in \mathbb{R}^d} \left\| D^i f(\theta) \right\| \leqslant \bar{L}$.*

Lemma C.9. *Assume A1 and A8$(k+1)$ for $k \in \mathbb{N}$, $k \geqslant 1$.*

a) For all $t \geqslant 0$, $\phi_t \in C^k(\mathbb{R}^d)$. In addition for all $\theta \in \mathbb{R}$, $\phi^{(k)}_t(x) : t \mapsto D^k \phi_t(\theta)$ satisfies the following ordinary differential equation,

$$\dot{\phi}^{(k)}_t(x) = D^k \left\{ \nabla f(\phi_t(\theta)) \right\} \text{ , for all } t \geqslant 0 ,$$

with $\phi^{(2)}_0(x) = \mathrm{Id}$ and $\phi^{(k)}_0(x) = 0$ for $k \geqslant 2$.

b) For all $t \geqslant 0$ *and* $\theta \in \mathbb{R}^d$, $\|\phi_t(\theta) - \theta_*\|^2 \leqslant \mathrm{e}^{-2\mu t} \|\theta - \theta_*\|^2$.

c) If $k \geqslant 2$, *for all* $t \geqslant 0$,

$$\nabla \phi_t(\theta_*) = \mathrm{e}^{-\nabla^2 f(\theta_*)} .$$

d) If $k \geqslant 3$, *for all* $t \geqslant 0$ *and* $i, j, k \in \{1, \dots, d\}$,

$$\left\langle D^2 \phi_t(\theta_*) \left\{ \mathbf{v}_i, \mathbf{v}_j \right\}, \mathbf{v}_k \right\rangle = \frac{\mathrm{e}^{-\lambda_i t} - \mathrm{e}^{-(\lambda_k + \lambda_j)t}}{\lambda_i - \lambda_k - \lambda_j} ,$$

where $\{\mathbf{v}_1, \dots, \mathbf{v}_d\}$ *and* $\{\lambda_1, \dots, \lambda_d\}$ *are the eigenvectors and the eigenvalues of* $\nabla^2 f(\theta_*)$ *respectively satisfying for all* $i \in \{1, \dots, d\}$, $\nabla^2 f(\theta_*) \mathbf{v}_i = \lambda_i \mathbf{v}_i$.

Proof. a) This is a fundamental result on the regularity of flows of autonomous differential equations, see e.g. (Hartman, 1982, Theorem 4.1 Chapter V)

b) Let $\theta \in \mathbb{R}^d$. Differentiate $\|\phi_t(\theta)\|^2$ with respect to t and using **A**1, that f is at least continuously differentiable and Grönwall's inequality concludes the proof.

c) By Lemma C.9-a) and since θ_* is an equilibrium point, for all $x \in \mathbb{R}^d$, $\xi_t^x(\theta_*) = D\phi_t(\theta_*)\{x\}$ satisfies the following ordinary differential equation

$$\dot{\xi}_s^x(\theta_*) = -\nabla^2 f(\phi_s(\theta_*)) \xi_s^x(\theta_*) \mathrm{d}s = -\nabla^2 f(\theta_*) \xi_s^x(\theta_*) \mathrm{d}s . \tag{C.34}$$

with $\xi_0^x(\theta_*) = x$. The proof then follows from uniqueness of the solution of (C.34).

d) By Lemma C.9-a), for all $x_1, x_2 \in \mathbb{R}^d$, $\xi_t^{x_1,x_2}(\theta_*) = D^i \phi_t(\theta_*)\{x_1 \otimes x_2\}$ satisfies the ordinary stochastic differential equation:

$$\frac{\mathrm{d}\xi_s^{x_1,x_2}}{\mathrm{d}s}(\theta_*) = -D^3 f(\phi_s(\theta_*)) \{\nabla \phi_s(\theta_*) x_1 \otimes \nabla \phi_s(\theta_*) x_2 \otimes \mathbf{e}_i\} - D^2 f(\phi_s(\theta_*)) \{\xi_s^{x_1,x_2}\} \mathbf{e}_i .$$

By c) and since θ_* is an equilibrium point we get that $\xi_t^{x_1,x_2}(\theta_*)$ satisfies

$$\frac{\mathrm{d}\xi_s^{x_1,x_2}}{\mathrm{d}s}(\theta_*) = -D^3 f(\theta_*) \left\{ \mathrm{e}^{-\nabla^2 f(\theta_*)t} x_1 \otimes \mathrm{e}^{-\nabla^2 f(\theta_*)t} x_2 \otimes \mathbf{e}_i \right\} - D^2 f(\theta_*) \{\xi_s^{x_1,x_2}\} \mathbf{e}_i .$$

Therefore we get for all $i, j, k \in \{1, \dots, d\}$,

$$\frac{\mathrm{d} \left\langle \xi_s^{\mathbf{v}_i, \mathbf{v}_j}, \mathbf{v}_k \right\rangle}{\mathrm{d}s} = -D^3 f(\theta_*) \left\{ \mathrm{e}^{-\lambda_i t} \mathbf{v}_i \otimes \mathrm{e}^{-\lambda_j t} \mathbf{v}_j \otimes \mathbf{v}_k \right\} - \lambda_k \left\langle \xi_s^{\mathbf{v}_i, \mathbf{v}_j}, \mathbf{v}_k \right\rangle .$$

This ordinary differential equation can be solved analytically which finishes the proof. □

Under **A**1 and **A**8(k), $k \in \mathbb{N}$, $k \geqslant 1$, for any function $g : \mathbb{R}^d \to \mathbb{R}^q$, locally Lipschitz, denote by h_g the solution of the continuous Poisson equation defined for all $\theta \in \mathbb{R}^d$ by

$$h_g(\theta) = \int_0^\infty (g(\phi_s(\theta)) - g(\theta_*)) dt . \tag{C.35}$$

Note that h_g is well-defined by Lemma C.9-b) and since g is assumed to be locally-Lipschitz. Note that by (4.10), we have for all $g : \mathbb{R}^d \to \mathbb{R}$, locally Lipschitz,

$$\mathcal{A} h_g(\theta) = -g(\theta) + g(\theta_*) . \tag{C.36}$$

In addition define $h_{\mathrm{Id}} : \mathbb{R}^d \to \mathbb{R}^d$ for all $x \in \mathbb{R}^d$ by

$$h_{\mathrm{Id}}(\theta) = \int_0^\infty \{\phi_s(\theta) - \theta_*\} dt . \tag{C.37}$$

Note that h_{Id} is also well-defined by Lemma C.9-b).

Lemma C.10. *Let $g : \mathbb{R}^d \to \mathbb{R}$ satisfying A5(k_1, k_2) for $k_1, k_2 \in \mathbb{N}$, $k_1 \geqslant 1$.*

a) Then for all $\theta_1, \theta_2 \in \mathbb{R}^d$,

$$|g(\theta_1) - g(\theta_2)| \leqslant a_g \|\theta_1 - \theta_2\| \left\{ b_g + \|\theta_1 - \theta_*\|^{k_2} + \|\theta_2 - \theta_*\|^{k_2} \right\} .$$

Assume in addition A1 and A8$(k_1 + 1)$.

b) Then for all $\theta \in \mathbb{R}^d$,

$$|h_g|(\theta) \leqslant a_g \left\{ (b_g/\mu) \|\theta - \theta_*\| + (k_2\mu)^{-1} \|\theta - \theta_*\|^{k_2} \right\} .$$

c) If $k_1 \geqslant 2$, then $\nabla h_{\mathrm{Id}}(\theta_*) = (\nabla^2 f(\theta_*))^{-1}$. If $k_1 \geqslant 3$, then for all $i, j \in \{1, \ldots, d\}$,

$$\frac{\partial^2 h_{\mathrm{Id}}}{\partial\theta_i \partial\theta_j}(\theta_*) = -D^3 f(\theta_*) \left\{ \left[\left(\nabla^2 f(\theta_*) \otimes \mathrm{Id} + \mathrm{Id} \otimes \nabla^2 f(\theta_*) \right)^{-1} \{\mathbf{e}_i \otimes \mathbf{e}_j\} \right] \otimes \mathbf{e}_i \right\} (\nabla^2 f(\theta_*))^{-1} \mathbf{e}_i ,$$

where $\{\mathbf{e}_1, \ldots, \mathbf{e}_d\}$ are the canonical basis of $\mathbb{R}^d$.

Proof. a) Let $\theta_1, \theta_2 \in \mathbb{R}^d$. By the mean value theorem, there exists $s \in [0, 1]$ such that if $\eta_s = s\theta_1 + (1 - s)\theta_2$ then

$$|g(\theta_1) - g(\theta_2)| = Dg(\eta_s) \{\theta_1 - \theta_2\} .$$

The proof is then concluded using **A**5(k_1, k_2) and

$$\|\eta_s - \theta_*\| \leqslant \max \left(\|\theta_1 - \theta_*\| , \|\theta_2 - \theta_*\| \right) .$$

b) For all $\theta \in \mathbb{R}^d$, we have using the first result of the Lemma and (C.35)

$$|h_g(\theta)| \leqslant a_g \int_0^{+\infty} \|\phi_s(\theta) - \theta_*\| \left\{ b_g + \|\phi_s(\theta) - \theta_*\|^{k_2} \right\} \mathrm{d}s .$$

The proof then follows from Lemma C.9-b).

c) The proof is a direct consequence of Lemma C.9-c)-d) and (C.35).

□

Theorem C.11. *Assume A1-A8$(k_1 + 1)$ for $k_1, k_2 \in \mathbb{N}$, $k_1 \geqslant 2$. Let $g : \mathbb{R}^d \to \mathbb{R}$ satisfying A 5(k_1, k_2) for $k_2 \in \mathbb{N}$.*

a) For all $t \geqslant 0$, $\phi_t \in C^{k_1}(\mathbb{R}^d)$ and for all $i \in \{1, \ldots, k\}$, there exists $C_i \geqslant 0$ such that for all $\theta \in \mathbb{R}^d$ and $t \geqslant 0$,

$$\left\| D^i \phi_t(\theta) \right\| \leqslant C_i \mathrm{e}^{-\mu t} .$$

b) Let $g \in C^{k_1}(\mathbb{R}^d)$. Then $h_g \in C^{k_1}(\mathbb{R}^d)$ and for all $i \in \{0, \ldots, k_1\}$, there exists $C_i \geqslant 0$ such that for all $\theta \in \mathbb{R}^d$,

$$\left\| D^i h_g(\theta) \right\| \leqslant C_i \left\{ 1 + \|\theta - \theta_*\|^{k_2} \right\} .$$

Proof. a) The proof is by induction on k_1. By Lemma C.9-a), for all $x \in \mathbb{R}^d$, and $\theta \in \mathbb{R}^d$, $\xi_t^x(\theta) = D\phi_t(\theta)\{x\}$ satisfies

$$\frac{\mathrm{d}\xi_s^x}{\mathrm{d}s}(\theta) = -\nabla^2 f(\phi_s(\theta))\xi_s^x(\theta)\mathrm{d}s \,. \tag{C.38}$$

with $\xi_0^x(\theta) = x$. Now differentiating $s \to \|\xi_s^x(\theta)\|^2$, using **A**1 and Grünwall's inequality, we get $\|\xi_s^x(\theta)\|^2 \leqslant \mathrm{e}^{-2ms}\|x\|^2$ which implies the result for $k_1 = 2$.

Let now $k_1 > 2$. Using again Lemma C.9-a), Faà di Bruno's formula (Levy, 2006, Theorem 1) and since (4.9) can be written on the form

$$\frac{\mathrm{d}\phi_t}{\mathrm{d}s}(\theta) = -\sum_{j=1}^d Df(\phi_t(\theta))\{e_i\}\, e_i \,,$$

for all $i \in \{2, \dots, k_1\}$, $\theta \in \mathbb{R}^d$ and $x_1, \cdots, x_i \in \mathbb{R}^d$, $\xi_t^{x_1,\cdots,x_i}(\theta) = D^i\phi_t(\theta)\{x_1 \otimes \cdots \otimes x_i\}$ satisfies the ordinary differential equation:

$$\frac{\mathrm{d}\xi_s^{x_1,\cdots,x_i}}{\mathrm{d}s}(\theta) = -\sum_{j=1}^d \sum_{\Omega \in \mathsf{P}(\{1,\dots,i\})} D^{|\Omega|+1} f(\phi_s(\theta)) \left\{ e_i \otimes \bigotimes_{l=1}^{i} \bigotimes_{j_1,\cdots,j_l \in \Omega} \xi_s^{x_{j_1},\cdots,x_{j_l}}(\theta) \right\} e_i \,, \tag{C.39}$$

where $\mathsf{P}(\{1,\dots,i\})$ is the set of partitions of $\{1,\dots,i\}$, which does not contain the empty set and $|\Omega|$ is the cardinal of $\Omega \in \mathsf{P}(\{1,\dots,i+1\})$. We now show by induction on i that for all $i \in \{1,\dots,k_1\}$, there exists a universal constant C_i such that for all $t \geqslant 0$ and $\theta \in \mathbb{R}^d$,

$$\sup_{x \in \mathbb{R}^d} \left\| D^i\phi_t(\theta) \right\| \leqslant C_i \mathrm{e}^{-\mu t} \,. \tag{C.40}$$

For $i = 1$, the result follows from the case $k_1 = 1$. Assume that the result is true for $\{1,\dots,i\}$ for $i \in \{1,\dots,k_1 - 1\}$. We show the result for $i+1$. By (C.39), we have for all $\theta \in \mathbb{R}^d$ and $x_1, \cdots, x_i \in \mathbb{R}^d$,

$$\frac{\left\|\xi_t^{x_1,\cdots,x_{i+1}}(\theta)\right\|^2}{\mathrm{d}t} = \\ -\int_0^t \sum_{\Omega \in \mathsf{P}(\{1,\dots,i+1\})} D^{|\Omega|+1} f(\phi_s(\theta)) \left\{ \xi_t^{x_1,\cdots,x_{i+1}}(\theta) \otimes \bigotimes_{l=1}^{i+1} \bigotimes_{j_1,\dots,j_l \in \Omega} \xi_s^{x_{j_1},\cdots,x_{j_l}}(\theta) \right\} \mathrm{d}s \,.$$

Isolating the term corresponding to $\Omega = \{\{1,\dots,i+1\}\}$ in the sum above and using Young's inequality, **A**1, Grönwall's inequality and the induction hypothesis, we get that there exists a universal constant C_{i+1} such that for all $t \geqslant 0$ and $x \in \mathbb{R}^d$ (C.40) holds for $i+1$.

b) The proof is a consequence of a), (C.35), **A**5(k_1, k_2) and Leibniz's rule.

□

C.5 Proof of Theorem 4.6

We preface the proof of the Theorem by two fundamental first estimates.

Theorem C.12. *Assume **A**1-**A**2-**A**3-**A**4$(2(k_2+3))$, for $k_1, k_2 \in \mathbb{N}$, $k_1 \geqslant 1$. Let $g : \mathbb{R}^d \to \mathbb{R}$ satisfying **A**5$(3, k_2)$. Then, there exists $C_{k_2} \geqslant 0$ only depending on k_2 such that for all $\gamma \in (0, C_{k_2}/L)$, $n \in \mathbb{N}^*$, $\gamma > 0$ and $\theta \in \mathbb{R}^d$,*

$$-\mathbb{E}_\theta\left[n^{-1}\sum_{i=1}^n\left\{g(\theta_i^{(\gamma)})-g(\theta_*)\right\}\right]=\frac{\mathbb{E}_\theta\left[h_g(\theta_{n+1}^{(\gamma)})\right]-h_g(\theta)}{n\gamma}$$
$$-(\gamma/2)\int_{\mathbb{R}^d}D^2h_g(\tilde\theta)\mathbb{E}\left[\left\{\varepsilon(\tilde\theta)\right\}^{\otimes 2}\right]\mathrm{d}\pi_\gamma(\tilde\theta)+(\gamma/n)\tilde A_1(\theta)+\gamma^2\tilde A_2(\theta,n)\;,$$

where

$$\tilde A_1(\theta)\leqslant C\left\{1+\|\theta-\theta_*\|^{k_2+2}\right\}\;,\tilde A_2(\theta,n)\leqslant C\left\{1+\|\theta-\theta_*\|^{k_2+3}/n\right\}\;,$$

for some constant $C\geqslant 0$ independent of γ and n.

Proof. Let $n\in\mathbb{N}^*$, $\gamma>0$ and $\theta\in\mathbb{R}^d$. Consider the sequence $(\theta_k^{(\gamma)})_{k\geqslant 0}$ defined by the stochastic gradient recursion (4.1) and starting at θ. Theorem C.11 shows that $h_g\in C^3(\mathbb{R}^d)$. Therefore using (4.1) and the Taylor expansion formula, we have for all $i\in\{1,\dots,n\}$

$$h_g(\theta_{i+1}^{(\gamma)})=h_g(\theta_i^{(\gamma)})+\gamma Dh_g(\theta_i^{(\gamma)})\left\{-\nabla f(\theta_i^{(\gamma)})+\varepsilon_{i+1}(\theta_i^{(\gamma)})\right\}$$
$$+(\gamma^2/2)D^2h_g(\theta_i^{(\gamma)})\left\{-\nabla f(\theta_i^{(\gamma)})+\varepsilon_{i+1}(\theta_i^{(\gamma)})\right\}^{\otimes 2}$$
$$+(\gamma^3/(3!))D^3h_g(\theta_i^{(\gamma)}+s_i^{(\gamma)}\Delta\theta_{i+1}^{(\gamma)})\left\{-\nabla f(\theta_i^{(\gamma)})+\varepsilon_{i+1}(\theta_i^{(\gamma)})\right\}^{\otimes 3}\;,$$

where $s_i^{(\gamma)}\in[0,1]$ and $\Delta\theta_{i+1}^{(\gamma)}=\theta_{i+1}^{(\gamma)}-\theta_i^{(\gamma)}$. Therefore by (C.36), we get

$$-n^{-1}\sum_{i=1}^n\left\{g(\theta_i^{(\gamma)})-g(\theta_*)\right\}=\frac{h_g(\theta_{n+1}^{(\gamma)})-h_g(\theta)}{n\gamma}-n^{-1}\sum_{i=1}^n Dh_g(\theta_{i-1}^{(\gamma)})\varepsilon_{i+1}(\theta_i^{(\gamma)})$$
$$-(\gamma/(2n))\sum_{i=1}^n D^2h_g(\theta_i^{(\gamma)})\left\{-\nabla f(\theta_i^{(\gamma)})+\varepsilon_{i+1}(\theta_i^{(\gamma)})\right\}^{\otimes 2}$$
$$-(\gamma^2/(3!n))\sum_{i=1}^n D^3h_g(\theta_i^{(\gamma)}+s_i^{(\gamma)}\Delta\theta_{i+1}^{(\gamma)})\left\{-\nabla f(\theta_i^{(\gamma)})+\varepsilon_{i+1}(\theta_i^{(\gamma)})\right\}^{\otimes 3}\;.$$

Taking the expectation and using **A**3, we have

$$-\mathbb{E}_\theta\left[n^{-1}\sum_{i=1}^n\left\{g(\theta_i^{(\gamma)})-g(\theta_*)\right\}\right]=\frac{\mathbb{E}_\theta\left[h_g(\theta_{n+1}^{(\gamma)})\right]-h_g(\theta)}{n\gamma}$$
$$-(\gamma/2)\int_{\mathbb{R}^d}D^2h_g(\tilde\theta)\mathbb{E}\left[\left\{\varepsilon(\tilde\theta)\right\}^{\otimes 2}\right]\mathrm{d}\pi_\gamma(\tilde\theta)+\tilde A_1+\tilde A_2\;,$$

where

$$\tilde A_1=(\gamma/(2n))\mathbb{E}_\theta\left[\sum_{i=1}^n\left(D^2h_g(\theta_*)\left\{\varepsilon_{i+1}(\theta_*)\right\}^{\otimes 2}-D^2h_g(\theta_i^{(\gamma)})\left\{-\nabla f(\theta_i^{(\gamma)})+\varepsilon_{i+1}(\theta_i^{(\gamma)})\right\}^{\otimes 2}\right)\right]$$
$$\tilde A_2=-(\gamma^2/(3!n))\mathbb{E}_\theta\left[\sum_{i=1}^n D^3h_g(\theta_i^{(\gamma)}+s_i^{(\gamma)}\Delta\theta_{i+1}^{(\gamma)})\left\{-\nabla f(\theta_i^{(\gamma)})+\varepsilon_{i+1}(\theta_i^{(\gamma)})\right\}^{\otimes 3}\right]\;.$$

The proof is then concluded using Theorem C.11, Lemma C.7 and Theorem C.8. □

Corollary C.13. *Assume **A**1-**A**2-**A**3-**A**4$(2(k_2+3))$, for $k_1,k_2\in\mathbb{N}$, $k_1\geqslant 1$. Let $g:\mathbb{R}^d\to\mathbb{R}$ satisfying **A**5$(3,k_2)$. Then there exists $C_{k_2}\geqslant 0$ only depending on k_2 such that for all $\gamma\in(0,C_{k_2}/L)$, there exists $C\geqslant 0$ independent of γ such that*

$$\left|\int_{\mathbb{R}^d}g(\tilde\theta)\pi_\gamma(\mathrm{d}\tilde\theta)-g(\theta_*)+(\gamma/2)\int_{\mathbb{R}^d}D^2h_g(\tilde\theta)\mathbb{E}\left[\left\{\varepsilon(\tilde\theta)\right\}^{\otimes 2}\right]\mathrm{d}\pi_\gamma(\tilde\theta)\right|\leqslant C\gamma^2\;.$$

Proof. The proof is a direct consequence of Theorem C.8 and Theorem C.12. □

Proof of Theorem 4.6. Under the stated assumptions, $\theta \mapsto D^2 h_g(\theta)\mathbb{E}\left[\{\varepsilon(\theta)\}^{\otimes 2}\right]$ satisfies the conditions of Corollary C.13. The proof then follows from combining Corollary C.13 applied to this function and Theorem C.12. □

5

Conclusion and Future Work

5.1 Summary of the thesis

In this thesis, we investigate several aspects of stochastic approximation for machine learning, especially in the non-parametric setting.

In the opening chapter, we introduce supervised machine learning, the convex optimization framework, stochastic approximation theory, and the non-parametric regression setting. We pave the way for the following chapters by describing the related literature and illustrating some of the fundamental concepts.

In our first contribution, we provide statistically optimal rates of convergence for learning in a reproducing kernel Hilbert space, under both the source condition (which quantifies the smoothness of the optimal prediction function) and the capacity condition (related to the eigenvalue decay of the covariance operators). This problem was stated as an open problem by Rosasco et al. (2014) and Ying and Pontil (2008). We show that averaging, combined with larger step sizes than traditional approaches, allows to get this optimal behavior. We give results in both the finite horizon setting and the online setting, describing the regimes in which the optimal rate is reached. We also present minimal assumptions under which the problem can be analyzed, removing un-necessary topological assumptions and strong conditions (e.g., uniform bounds) on the kernel. The optimal rate underlines the power of averaging in stochastic gradient descent, which substantially improves the convergence in comparison to un-averaged methods. It also shows that the statistical performance can be achieved by an online algorithm performing a single pass on observations, an important and insightful property in practice.

In our second contribution, we propose a new algorithm, achieving simultaneously the best possible bias and variance term in the parametric regime: the optimal variance term matches the statistical rate of convergence as $\frac{\sigma^2 d}{n}$, while the bias term (the speed at which initial conditions are forgotten) matches the best possible rate for a first order algorithm as $\frac{L\|\theta_0-\theta_*\|^2}{n^2}$. This results in a theoretical improvement in the non-parametric regime for certain situations in which the function is particularly not regular with respect to the reproducing kernel Hilbert space. While the bias term was systematically dominating for the averaged recursion, the acceleration allows to recover the optimal tradeoff and

statistical rate of convergence.

Finally, in our last contribution, we consider a more general setting where the objective function is no longer quadratic. We analyze the averaged stochastic gradient descent with constant step as a Markov chain. We give a complete analysis of its convergence, outlining the effect of initial conditions, noise and step-sizes. While we mainly consider the general minimization framework, these results directly apply for supervised machine learning, extending some results known for least-squares to all loss functions. This analysis naturally leads to using Romberg-Richardson extrapolation, that provably improves the convergence behavior of the averaged SGD iterates.

5.2 Perspectives

Our work has triggered a few questions, which are still open.

(i) First, we showed in Chapter 2 that averaged stochastic gradient descent was partially adaptive to the difficulty of the problem. For any constant step, the bias and variance terms decay faster when the problem is easier (*i.e.*, satisfies a stronger source condition or capacity condition). However, the choice of the optimal learning rate is not fully automatic and generally requires to use a cross validation approach. Proposing a data dependent rule to find an optimal step size, or a fully adaptive algorithm reaching the optimal rate, without cross validation, would be of major interest. Orabona (2014) describes a parameter-free algorithm that adapts to the source condition; and Raskutti et al. (2014) propose a data dependent early stopping rule, achieving optimal convergence rate with respect to the capacity condition assumption. To the best of our knowledge, getting a simple algorithm that would adapt to both parameters is still an open problem. The most promising direction to get such a result seems to be the use of Lepski's method (Lepski et al., 1997).

(ii) While the optimal rate of convergence is reached, the complexity of the averaged stochastic gradient descent remains sub-optimal in the non-parametric setting: in a reproducing kernel Hilbert space, after n iterations, the computational complexity is $O(n^2)$. Several approaches have been proposed to reduce this complexity, especially with random features (Rudi et al., 2016) and Nyström approximation (El Alaoui and Mahoney, 2014; Lin and Rosasco, 2016). Roughly speaking, these methods solve a linear system in a lower dimension d_n, corresponding to an "implicit" dimension of the problem. If this dimension is carefully chosen, these methods can achieve the statistical rate. Stochastic algorithms can also be used in such a setting: while an analysis combining iterative algorithm and Nyström approximation was recently proposed by Rudi et al. (2017), a general analysis of stochastic algorithms together with random features would be interesting.

(iii) Our results in Chapter 3 underline the differences between additive and multiplicative noise oracles. Understanding to which extent these results can be generalized to more general noise oracles would be very interesting.

(iv) In the analysis of SGD as a Markov chain proposed in Chapter 4, the convergence results are derived for strongly convex functions. This analysis opens several directions and potential extensions: providing an analysis for non-strongly convex

functions, a complete analysis for decreasing step sizes, and extensions of our results under self-concordance condition, would be valuable. Moreover, Richardson-Romberg interpolations methods could also be used on other parameters, especially the regularization parameter when regularization is used.

(v) Finally, non-parametric estimation goes beyond prediction: density estimation, its twin brother, is of equal importance (Tsybakov, 2008). *Shape constrained density estimation* has been a stimulating topic these last years (Kim and Samworth, 2014; Kim et al., 2016). In this setting, one generally considers the maximum likelihood estimator, which has good statistical properties. However, computing this estimator is challenging in dimension bigger than one; Cule et al. (2010) suggest using Shor's R-algorithm, but this algorithm scales badly with the dimension. Improving optimization techniques in this field would be interesting. Moreover, an important open question remains: similarly to the regression tasks, is it possible to propose an online algorithm (performing a single pass on input points), that achieves the statistical rate of convergence ?

Bibliography

A. Abdulle, G. Vilmart, and K. C. Zygalakis. High order numerical approximation of the invariant measure of ergodic SDEs. *SIAM Journal on Numerical Analysis*, 52(4): 1600–1622, 2014.

M. Abramowitz and I. Stegun. *Handbook of mathematical functions*. Dover publications, 1964.

A. Agarwal and L. Bottou. A lower bound for the optimization of finite sums. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2015.

A. Agarwal, P. L. Bartlett, P. Ravikumar, and M. J. Wainwright. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *IEEE Transactions on Information Theory*, 58(5):3235–3249, 2012.

Z. Allen-Zhu and L. Orecchia. Linear Coupling: An Ultimate Unification of Gradient and Mirror Descent. In *Proceedings of the Conference on Innovations in Theoretical Computer Science (ITCS)*, 2017.

N. S. Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185, 1992.

N. Aronszajn. Theory of Reproducing Kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

F. Bach. Sharp analysis of low-rank kernel matrix approximations. In *Proceedings of the International Conference on Learning Theory (COLT)*, 2012.

F. Bach. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *Journal of Machine Learning Research*, 15(1):595–627, 2014.

F. Bach and E. Moulines. Non-asymptotic Analysis of Stochastic Approximation Algorithms for Machine Learning. In *Advances in Neural Information Processing Systems (NIPS)*, 451–459. Curran Associates Inc., USA, 2011.

F. Bach and E. Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate O(1/n). In *Advances in Neural Information Processing Systems (NIPS)*. 2013.

P. L. Bartlett, O. Bousquet, and S. Mendelson. Local rademacher complexities. *Annals of Statistics*, 33(4):1497–1537, 2005.

F. Bauer, S. Pereverzev, and L. Rosasco. On regularization algorithms in learning theory. *Journal of complexity*, 23(1):52–72, 2007.

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

M. Benaim and M. W. Hirsch. Mixed equilibria and dynamical systems arising from fictitious play in perturbed games. *Games and Economic Behavior*, 29(1-2):36–72, 1999.

A. Benveniste, M. Métivier, and P. Priouret. *Adaptive algorithms and stochastic approximations*, 22. Springer Science & Business Media, 2012.

S. Bergmann. Über die entwicklung der harmonischen funktionen der ebene und des raumes nach orthogonalfunktionen. *Mathematische Annalen*, 86(3):238–271, 1922.

A. Berlinet and C. Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*, 3. Springer, 2004.

D. Bertsekas. *Nonlinear programming*. Athena Scientific, 1995.

P. Billingsley. *Probability and measure*. John Wiley & Sons, 2008.

L. Birgé. An Alternative Point of View on Lepski's Method. *Lecture Notes-Monograph Series*, 36:113–133, 2001.

G. Blanchard and N. Krämer. Optimal learning rates for kernel conjugate gradient regression. In *Advances in Neural Information Processing Systems (NIPS)*, 226–234. 2010.

A. Bordes, S. Ertekin, J. Weston, and L. Bottou. Fast kernel classifiers with online and active learning. *Journal of Machine Learning Research*, 6:1579–1619, 2005.

K. M. Borgwardt and H.-P. Kriegel. Shortest-path kernels on graphs. In *Proceedings of the IEEE International Conference on Data Mining*, 8–pp, 2005.

L. Bottou and O. Bousquet. The tradeoffs of large scale learning. In *Advances in Neural Information Processing Systems (NIPS)*. 2008.

S. Boucheron and P. Massart. A high-dimensional Wilks phenomenon. *Probability Theory and Related Fields*, 150(3):405–433, Aug 2011.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004.

H. Brezis. *Analyse fonctionnelle, Théorie et applications*. Masson, 1983.

H. D. Brunk. Maximum likelihood estimates of monotone parameters. *Annals of Mathematical Statistics*, 26(4):607–616, 1955.

H. D. Brunk. Estimation of isotonic regression. *Nonparametric Techniques in Statistical Inference*, 177–195, 1970.

S. Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

S. Bubeck and N. Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

S. Bubeck, Y. T. Lee, and M. Singh. A geometric alternative to nesterov's accelerated gradient descent. In *Proceedings of the International Conference on Learning Theory (COLT)*, 2015.

A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

A. Caponnetto and Y. Yao. Adaptation for regularization operators in learning theory. Technical report, Computer Science and Artificial Intelligence Laboratory, 2006.

N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

C. Chen, N. Ding, and L. Carin. On the convergence of stochastic gradient MCMC algorithms with high-order integrators. In *Advances in Neural Information Processing Systems (NIPS)*, 2269–2277. 2015.

H.-F. Chen. Asymptotically efficient stochastic approximation. *Stochastics: An International Journal of Probability and Stochastic Processes*, 45(1-2):1–16, 1993.

A. Cotter, O. Shamir, N. Srebro, and K. Sridharan. Better mini-batch algorithms via accelerated gradient methods. In *Advances in Neural Information Processing Systems (NIPS)*. 2011.

M. Cule, R. Samworth, and M. Stewart. Maximum likelihood estimation of a multi-dimensional log-concave density. *Journal of the Royal Statistical Society: Series B*, 72(5), 2010.

M. Cuturi, K. Fukumizu, and J.-P. Vert. Semigroup kernels on measures. *Journal of Machine Learning Research*, 6:1169–1198, 2005.

A. Dalalyan. Theoretical guarantees for approximate sampling from a smooth and log-concave density. submitted 1412.7392, arXiv, December 2014.

A. d'Aspremont. Smooth optimization with approximate gradient. *SIAM Journal on Optimization*, 19(3):1171–1183, 2008.

E. De Vito, A. Caponetto, and L. Rosasco. Model selection for regularized least-squares algorithm in learning theory. *Foundations of Computational Mathematics*, 5:59–85, 2005.

A. Défossez and F. Bach. Averaged least-mean-squares: bias-variance trade-offs and optimal sampling distributions. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2015.

O. Dekel, S. Shalev-Shwartz, and Y. Singer. The Forgetron: A kernel-based perceptron on a fixed budget. In *Advances in Neural Information Processing Systems (NIPS)*. 2005.

O. Dekel, R. Gilad-Bachrach, O. Shamir, and L. Xiao. Optimal distributed online prediction using mini-batches. *Journal of Machine Learning Research*, 13(1):165–202, 2012.

B. Delyon and A. Juditsky. Stochastic optimization with averaging of trajectories. *Stochastics: An International Journal of Probability and Stochastic Processes*, 39(2-3):107–118, 1992.

O. Devolder, F. Glineur, and Y. Nesterov. First-order methods of smooth convex optimization with inexact oracle. *Mathematical Programming*, 146(1-2, Ser. A):37–75, 2014.

A. Dieuleveut and F. Bach. Non-parametric stochastic approximation with large step-sizes. *Annals of Statistics*, 44(4):1363–1399, 2016.

A. Dieuleveut, N. Flammarion, and F. Bach. Harder, Better, Faster, Stronger Convergence Rates for Least-Squares Regression. *ArXiv e-prints*, 2016.

A. Dieuleveut, A. Durmus, and F. Bach. Bridging the Gap between Constant Step Size Stochastic Gradient Descent and Markov Chains. *arxiv*, 2017.

M. Duflo. *Random Iterative Models*. Springer, 1st edition, 1997.

A. Durmus, U. Şimşekli, E. Moulines, R. Badeau, and G. Richard. Stochastic Gradient Richardson-Romberg Markov Chain Monte Carlo. In *Advances in Neural Information Processing Systems (NIPS)*, 2047–2055. 2016.

A. El Alaoui and M. W. Mahoney. Fast Randomized Kernel Methods With Statistical Guarantees. *ArXiv e-prints*, 2014.

H. W. Engl, M. Hanke, and N. A. Regularization of Inverse Problems. *Klüwer Academic Publishers*, 1996.

V. Fabian. On asymptotic normality in stochastic approximation. *Annals of Mathematical Statistics*, 1327–1332, 1968.

N. Flammarion and F. Bach. From averaging to acceleration, there is only a step-size. In *Proceedings of the International Conference on Learning Theory (COLT)*, 2015.

C. F. Gauss. *Theoria motus corporum coelestium in sectionibus conicis solem ambientium*. Sumtibus F. Perthes et IH Besser, 1809.

G. H. Golub and C. F. Van Loan. *Matrix Computations*. J. Hopkins University Press, 1996.

C. Gu. *Smoothing Spline ANOVA Models*, 297. Springer, 2002.

L. Györfi and H. Walk. On the averaged stochastic approximation for linear regression. *SIAM Journal on Control and Optimization*, 34(1):31–61, 1996.

L. Györfi, M. Kohler, A. Krzyzak, and H. Walk. *A Distribution-free Theory of Nonparametric Regression*. Springer, 2002.

D. Hanson and G. Pledger. Consistency in concave regression. *Annals of Statistics*, 1038–1050, 1976.

Z. Harchaoui and F. Bach. Image classification with segmentation graph kernels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1–8, 2007.

P. Hartman. *Ordinary Differential Equations: Second Edition*. Classics in Applied Mathematics. SIAM, 3600 Market Street, Floor 6, Philadelphia, PA 19104, 1982.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., NY, USA, 2001.

E. Hazan and S. Kale. Beyond the regret minimization barrier: an optimal algorithm for stochastic strongly-convex optimization. In *Proceedings of the International Conference on Learning Theory (COLT)*, 2011.

E. Hazan, A. Agarwal, and S. Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2):169–192, 2007.

D. Hilbert. Grundzüge einer allgemeinen theorie der linearen integralgleichungen. vierte mitteilung. *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse*, 1906:157–228, 1904.

C. Hildreth. Point estimates of ordinates of concave functions. *Journal of the American Statistical Association*, 49(267):598–619, 1954.

H. Hochstadt. *Integral equations*. John Wiley & Sons, 1973.

D. Hsu, S. M. Kakade, and T. Zhang. Random design analysis of ridge regression. *Foundations of Computational Mathematics*, 14(3):569–600, 2014.

I. A. Ibragimov and R. Z. Has' Minskii. *Asymptotic theory of estimation*. Nauka, Moscow, 1979.

I. A. Ibragimov and R. Z. Has' Minskii. Bounds for the risks of non-parametric regression estimates. *Theory of Probability & Its Applications*, 27(1):84–99, 1982.

T. S. Jaakkola, M. Diekhans, and D. Haussler. Using the Fisher kernel method to detect remote protein homologies. In *Proceedings of the Conference on Intelligent Systems for Molecular Biology (ISMB)*, 99, 149–158, 1999.

P. Jain, S. M. Kakade, R. Kidambi, P. Netrapalli, and A. Sidford. Parallelizing Stochastic Approximation Through Mini-Batching and Tail-Averaging. *ArXiv e-prints*, 2016.

P. Jain, S. M. Kakade, R. Kidambi, P. Netrapalli, and A. Sidford. Accelerating Stochastic Gradient Descent. *arXiv preprint*, 2017.

I. M. Johnstone. Minimax Bayes, asymptotic minimax and sparse wavelet priors. In *Proceedings of the Conference on Statistical Decision Theory and Related Topics V*, 303–326. Springer, New York, NY, 1994.

G. L. Jones. On the Markov chain central limit theorem. *Probability Surveys*, 1:299–320, 2004.

A. Juditsky and A. Nemirovski. First order methods for nonsmooth convex large-scale optimization, i: general purpose methods. *Optimization for Machine Learning*, 121–148, 2011.

A. K. H. Kim and R. J. Samworth. Global rates of convergence in log-concave density estimation. *Annals of Statistics*, 2014.

A. K. H. Kim, A. Guntuboyina, and R. J. Samworth. Adaptation in log-concave density estimation. *ArXiv e-prints*, 2016.

G. Kimeldorf and G. Wahba. Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, 33(1):82–95, 1971.

G. S. Kimeldorf and G. Wahba. Spline functions and stochastic processes. *Sankhyā: The Indian Journal of Statistics, Series A*, 173–180, 1970.

J. Kivinen, S. A.J., and R. C. Williamson. Online learning with kernels. *IEEE Transactions on Signal Processing*, 52(8):2165–2176, 2004.

A. N. Kolmogorov and S. V. Fomin. *Elements of the theory of functions and functional analysis*, 1. Courier Dover Publications, 1999.

V. Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47(5):1902–1914, 2001.

V. Koltchinskii. Local rademacher complexities and oracle inequalities in risk minimization. *Annals of Statistics*, 34(6):2593–2656, 2006.

H. Kushner and G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, 2003.

H. J. Kushner and J. Yang. Stochastic approximation with averaging of the iterates: Optimal asymptotic rate of convergence for general processes. *SIAM Journal on Control and Optimization*, 31(4):1045–1062, 1993.

S. Lacoste-Julien, M. Schmidt, and F. Bach. A simpler approach to obtaining an O(1/t) convergence rate for the stochastic projected subgradient method. Arxiv e-prints, technical report, INRIA, 2012.

G. Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1-2, Ser. A):365–397, 2012.

Y. Le Cun, Y. Bengio, and G. Hinton. Deep Learning. *Nature*, 521(7553):436–444, 2015.

G. Lecué and S. Mendelson. Performance of empirical risk minimization in linear aggregation. *Bernoulli*, 22(3):1520–1534, 2016.

M. Ledoux and M. Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer, Berlin, May 1991.

A. M. Legendre. *Nouvelles méthodes pour la détermination des orbites des comètes*. F. Didot, 1805.

O. V. Lepski, E. Mammen, and V. G. Spokoiny. Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors. *The Annals of Statistics*, 929–947, 1997.

C. S. Leslie, E. Eskin, and W. S. Noble. The spectrum kernel: A string kernel for SVM protein classification. In *Proceedings of Pacific symposium on biocomputing*, 7, 564–575. Hawaii, USA, 2002.

E. Levy. Why do partitions occur in Faa di Bruno's chain rule for higher derivatives? Technical Report 0602183, arXiv, 2006.

J. Lin and L. Rosasco. Optimal Learning for Multi-pass Stochastic Gradient Methods. In *Advances in Neural Information Processing Systems (NIPS)*. 2016.

L. Ljung, G. C. Pflug, and H. Walk. *Stochastic approximation and optimization of random systems*. DMV Seminar. Birkhauser Verlag, Basel, Boston, 1992.

H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins. Text classification using string kernels. *Journal of Machine Learning Research*, 2:419–444, 2002.

M. W. Mahoney. Randomized Algorithms for Matrices and Data. *Foundations and Trends in Machine Learning*, 3(2):123–224, 2011.

A. Marcet and T. J. Sargent. Convergence of least-squares learning in environments with hidden state variables and private information. *Journal of Political Economy*, 97(6): 1306–1322, 1989.

P. Massart. *Concentration Inequalities and Model Selection*. Lecture Notes in Mathematics. Springer, 2007.

J. C. Mattingly, A. M. Stuart, and D. J. Higham. Ergodicity for SDEs and approximations: locally Lipschitz vector fields and degenerate noise. *Stochastic Processes and their Applications*, 101(2):185–232, 2002.

P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Monographs on Statistics and Applied Probability. Chapman & Hall, London, second edition, 1989.

S. Mendelson. Learning without concentration. In *Proceedings of the International Conference on Learning Theory (COLT)*, 2014.

J. Mercer. Functions of positive and negative type, and their connection with the theory of integral equations. *Philosophical transactions of the royal society of London. Series A, containing papers of a mathematical or physical character*, 209:415–446, 1909.

S. Meyn and R. Tweedie. *Markov Chains and Stochastic Stability*. Cambridge University Press, New York, NY, USA, 2nd edition, 2009.

S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Springer-Verlag Inc, Berlin; New York, 1993.

C. A. Micchelli, Y. Xu, and H. Zhang. Universal kernels. *The Journal of Machine Learning Research*, 7:2651–2667, 2006.

P. Mikusinski and E. Weiss. The Bochner Integral. *ArXiv e-prints*, 2014.

E. Moore. On properly positive hermitian matrices. *Bulletin American Mathematical Society*, 59(23):66–67, 1916.

E. Moore. General analysis, part i. *Memoirs American Philosophical Society*, 1935.

E. A. Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9(1): 141–142, 1964.

A. Nedić and D. Bertsekas. Convergence rate of incremental subgradient algorithms. In *Stochastic optimization: algorithms and applications*, 223–264. Springer, 2001.

D. Needell, R. Ward, and N. Srebro. Stochastic Gradient Descent, Weighted Sampling, and the Randomized Kaczmarz algorithm. In *Advances in Neural Information Processing Systems (NIPS)*, 1017–1025. Curran Associates, Inc., 2014.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust Stochastic Approximation Approach to Stochastic Programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

A. S. Nemirovski, B. T. Polyak, and A. B. Tsybakov. Estimators of maximum likelihood type for nonparametric regression. *Soviet Mathematics Doklady*, 28(3):788–92, 1983.

A. S. Nemirovski, B. T. Polyak, and A. B. Tsybakov. Signal processing by the nonparametric maximum-likelihood method. *Problemy peredachi informatsii*, 20(3):29–46, 1984.

A. S. Nemirovski, B. T. Polyak, and A. B. Tsybakov. Convergence rate of nonparametric estimates of maximum-likelihood type. *Problemy peredachi informatsii*, 21(4):17–33, 1985.

A. S. Nemirovsky and D. B. Yudin. *Problem complexity and method efficiency in optimization*. A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York, 1983. Translated from the Russian and with a preface by E. R. Dawson, Wiley-Interscience Series in Discrete Mathematics.

Y. Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Soviet Mathematics Doklady*, 27(2):372–376, 1983.

Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Applied Optimization. Springer, 2004.

Y. Nesterov and J. P. Vial. Confidence Level Solutions for Stochastic Programming. *Automatica*, 44(6):1559–1568, 2008.

J. Neter, M. H. Kutner, C. J. Nachtsheim, and W. Wasserman. *Applied linear statistical models*, 4. Irwin Chicago, 1996.

B. O'Donoghue and E. Candès. Adaptive restart for accelerated gradient schemes. *Foundations of Computational Mathematics*, 1–18, 2013.

R. I. Oliveira. The lower tail of random quadratic forms with applications to ordinary least squares. *Probability Theory and Related Fields*, 166(3-4):1175–1194, 2016.

F. Orabona. Simultaneous Model Selection and Optimization through Parameter-free Stochastic Learning. In *Advances in Neural Information Processing Systems (NIPS)*. 2014.

E. Parzen. On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.

F. Paulin. *Topologie, analyse et calcul différentiel*. Notes de cours, École Normale Supérieure, 2009.

G. C. Pflug. Stochastic minimization with constant step-size: asymptotic laws. *SIAM Journal on Control and Optimization*, 24(4):655–666, 1986.

B. T. Polyak. Some methods of speeding up the convergence of iteration methods. *Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

B. T. Polyak. *Introduction to Optimization*. Translations Series in Mathematics and Engineering. Optimization Software, Inc., Publications Division, New York, 1987.

B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

A. Rahimi and B. Recht. In *Random features for large-scale kernel machines*. 2007.

A. Rahimi and B. Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In *Advances in Neural Information Processing Systems (NIPS)*. 2008.

A. Rakhlin, O. Shamir, and K. Sridharan. Making Gradient Descent Optimal for Strongly Convex Stochastic Optimization. *ArXiv e-prints*, 2011.

G. Raskutti, M. J. Wainwright, and B. Yu. Early stopping and non-parametric regression: an optimal data-dependent stopping rule. *Journal of Machine Learning Research*, 15(1): 335–366, 2014.

H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22(3):400–407, 1951.

H. Robbins and D. Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In *Herbert Robbins Selected Papers*, 111–135. Springer, 1985.

R. T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, New-Jersey, 1970.

L. Rosasco, A. Tacchetti, and S. Villa. Regularization by Early Stopping for Online Learning Algorithms. *ArXiv e-prints*, 2014.

M. Rosenblatt. Remarks on some non-parametric estimates of a density function. *Annals of Mathematical Statistics*, 27(3):832–837, 1956.

A. Rudi, R. Camoriano, and L. Rosasco. Less is More: Nyström Computational Regularization. In *Advances in Neural Information Processing Systems (NIPS)*. 2015.

A. Rudi, R. Camoriano, and L. Rosasco. Generalization Properties of Learning with Random Features. *ArXiv e-prints*, Feb. 2016.

A. Rudi, L. Carratino, and L. Rosasco. Falkon: An optimal large scale kernel method. *arXiv preprint arXiv:1705.10958*, 2017.

D. Ruppert. Efficient estimations from a slowly convergent Robbins-Monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988.

M. Schmidt, N. Le Roux, and F. Bach. Convergence rates of inexact proximal-gradient methods for convex optimization. In *Advances in Neural Information Processing Systems (NIPS)*. 2011.

B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, 2002.

G. A. Seber and A. J. Lee. *Linear regression analysis*, 936. John Wiley & Sons, 2012.

S. Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2011.

S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 1 edition, May 2014.

S. Shalev-Shwartz, Y. Singer, and N. Srebro. PEGASOS: Primal Estimated sub-GrAdient SOlver for SVM. In *Proceedings of the International Conference on Machine Learning (ICML)*, 807–814. ACM, 2007.

S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Stochastic convex optimization. In *Proceedings of the International Conference on Learning Theory (COLT)*, 2009.

O. Shamir. Without-replacement sampling for stochastic gradient methods: Convergence results and application to distributed optimization. In *Advances in Neural Information Processing Systems (NIPS)*. 2016.

O. Shamir and T. Zhang. Stochastic Gradient Descent for Non-smooth Optimization: Convergence Results and Optimal Averaging Schemes. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2013.

C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:349–423, 1948.

C. E. Shannon. Communication in the presence of noise. In *Proceedings of the Institute of Radio Engineers (IRE)*, 37, 10–21, 1949.

J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

S. Smale and F. Cucker. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39(1):1–49, 2001.

S. Smale and F. Cucker. Best choices for regularization parameters in learning theory: on the bias-variance problem. *Foundations of Computational Mathematics*, 2:413–418, 2002.

S. Smale and Y. Yao. Online learning algorithms. *Foundations of Computational Mathematics*, 6(2):145–170, 2006.

S. Smale and D.-X. Zhou. Learning Theory Estimates via Integral Operators and Their Approximations. *Constructive Approximation*, 26(2):153–172, 2007.

K. Sridharan, S. Shalev-Shwartz, and N. Srebro. Fast rates for regularized objectives. In *Advances in Neural Information Processing Systems (NIPS)*, 1545–1552. 2008.

B. K. Sriperumbudur, K. Fukumizu, and G. R. Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *The Journal of Machine Learning Research*, 12: 2389–2410, 2011.

I. Steinwart and A. Christmann. *Support Vector Machines*. Springer Series in Information Science and Statistics. Springer, 2008.

I. Steinwart, D. Hush, and C. Scovel. Optimal rates for regularized least squares regression. In *Proceedings of the International Conference in Learning Theory (COLT)*, 2009.

J. Stoer and R. Bulirsch. *Introduction to numerical analysis*, 12. Springer Science & Business Media, 2013.

C. J. Stone. Consistent nonparametric regression. *Annals of Statistics*, 595–620, 1977.

C. J. Stone. Optimal global rates of convergence for nonparametric regression. *Annals of Statistics*, 1040–1053, 1982.

W. Su, S. Boyd, and E. Candès. A differential equation for modeling Nesterov's accelerated gradient method: theory and insights. In *Advances in Neural Information Processing Systems (NIPS)*. 2014.

D. Talay and L. Tubaro. Expansion of the global error for numerical schemes solving stochastic differential equations. *Stochastic Analysis and Applications*, 8(4):483–509, 1990.

P. Tarrès and Y. Yao. Online learning as stochastic approximation of regularization paths. *IEEE Transactions in Information Theory*, 60(99):5716–5735, 2014.

B. Thomson, J. Bruckner, and A. M. Bruckner. *Elementary real analysis*. Pearson, 2000.

A. B. Tsybakov. Optimal rates of aggregation. In *Proceedings of the International Conference on Computational Learning Theory (COLT)*, 2003.

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition, 2008.

A. van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, Nov. 2000.

A. W. Van der Vaart. *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

A. W. Van der Vaart and J. A. W. Wellner. *Empirical processes indexed by estimated functions*, 55 of *Lecture Notes–Monograph Series*, 234–252. Institute of Mathematical Statistics, Beachwood, Ohio, USA, 2007.

V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995.

J.-P. Vert. *Kernel Methods*. Master M2 "Mathematique, Vision, Apprentissage", Ecole normale superieure de Cachan, 2014.

C. Villani. *Optimal transport : old and new*. Grundlehren der mathematischen Wissenschaften. Springer, Berlin, 2009.

G. Wahba. *Spline Models for observational data*. SIAM, 1990.

L. Wasserman. *All of statistics: a concise course in statistical inference*. Springer Science & Business Media, 2006.

G. S. Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, 26(4):359–372, 1964.

M. Welling and Y. W. Teh. Bayesian learning via Stochastic Gradient Langevin Dynamics. In *Proceedings of the International Conference on Machine Learning (ICML)*, 681–688, 2011.

C. Williams and M. Seeger. Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems (NIPS)*. 2001.

J. Wishart. The generalised product moment distribution in samples from a normal multivariate population. *Biometrika*, 20A(1-2):32–52, 1928.

L. Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 11:2543–2596, 2010.

Y. Yao. *A dynamic theory of learning*. PhD thesis, University of California at Berkeley, 2006.

Y. Yao, L. Rosasco, and A. Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007.

G. Yin. On extensions of polyak's averaging approach to stochastic approximation. *Stochastics: An International Journal of Probability and Stochastic Processes*, 36(3-4):245–264, 1991.

Y. Ying and M. Pontil. Online gradient descent learning algorithms. *Foundations of Computational Mathematics*, 2008.

S. Zaremba. L'équation biharmonique et une classe remarquable de fonctions fondamentales harmoniques. *Bulletin International de l'Académie des Sciences de Cracovie, Classe des Sciences Mathématiques et Naturelles*, 3:147–196, 1907.

S. Zaremba. Sur le calcul numérique des fonctions demandées dans le problème de dirichlet et le problème hydrodynamique. *Bulletin International de l'Académie des Sciences de Cracovie, Classe des Sciences Mathématiques et Naturelles*, 2:125–195, 1909.

T. Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2004.

T. Zhang and B. Yu. Boosting with early stopping: convergence and consistency. *Annals of Statistics*, 33(4):1538–1579, 2005.

D. L. Zhu and P. Marcotte. Co-coercivity and its role in the convergence of iterative schemes for solving variational inequalities. *SIAM Journal on Optimization*, 6(3):714–726, 1996.

List of Figures

1.1 Risk decomposition 8
1.2 Gradient descent and Accelerated Gradient Descent 14
1.3 Stochastic gradient descent as a Markov Chain 26
1.4 Source condition 35
1.5 Capacity condition 36
1.6 Upper bound on the variance term as a function of α 37

2.1 Regions of optimal convergence 57
2.2 Observed optimal learning rate 64
2.3 Experimental error decay of different algorithms. 65
2.4 Regions of optimal convergence for tail averaging 66

3.1 Regions of optimal convergence for acceleration 130

4.1 Convergence of iterates $\theta_k^{(\gamma)}$, and Richardson-Romberg extrapolation 162
4.2 Plot of the excess risk for synthetic and real data-sets. 172

List of Tables

1.1 Interplay of the different domains in the thesis 3
1.2 Organization of the Section 1.3 16
1.3 Summary of rates for Stochastic Approximation. 27

2.1 Summary of assumptions and related results 61
2.2 Rates of convergence for various choices of the parameters α, r 63
2.3 Predicted and effective rates 66

A.1 Error decomposition in the finite horizon setting. 72
A.2 Error decomposition in the finite horizon setting. 89
A.3 Sketch of the proof, on-line setting. 96

3.1 Organization of the Chapter 4 116

Résumé

Le but de l'apprentissage supervisé est d'inférer des relations entre un phénomène que l'on souhaite prédire et des variables "explicatives". À cette fin, on dispose d'observations de multiples réalisations du phénomène, à partir desquelles on propose une règle de prédiction. L'émergence récente de sources de données à très grande échelle, a fait émerger deux difficultés : d'une part, il devient difficile d'éviter l'écueil du sur-apprentissage lorsque le nombre de variables explicatives est très supérieur au nombre d'observations ; d'autre part, l'aspect algorithmique devient déterminant, car la seule résolution d'un système linéaire peut devenir une difficulté majeure.

Des algorithmes issus des méthodes d'approximation stochastique, qui sont au coeur de cette thèse, proposent une réponse simultanée à ces deux difficultés : l'utilisation d'une méthode stochastique réduit drastiquement le coût algorithmique, sans dégrader la qualité de la règle de prédiction proposée, en évitant naturellement le sur-apprentissage.

Les méthodes paramétriques proposent comme prédictions des fonctions linéaires d'un ensemble choisi de variables explicatives, mais aboutissent souvent à une approximation de la structure statistique sous-jacente. Dans le cadre non-paramétrique, qui est un des thèmes centraux de cette thèse, la restriction aux prédicteurs linéaires est levée. Ces méthodes sont cruciales pour de nombreuses applications.

Cette thèse présente d'abord une analyse détaillée de l'approximation stochastique dans le cadre non-paramétrique, en particulier dans le cadre des espaces à noyaux reproduisants. Cette analyse permet d'obtenir des taux de convergence optimaux pour l'algorithme de descente de gradient stochastique moyennée.

Ensuite, un algorithme basé sur un principe d'accélération est présenté. Il converge à une vitesse optimale, tant du point de vue de l'optimisation que du point de vue statistique. Cela permet, dans le cadre non-paramétrique, d'améliorer la convergence jusqu'au taux optimal, dans certains régimes pour lesquels le premier algorithme analysé restait sous-optimal.

Enfin, la troisième contribution de la thèse consiste en l'extension du cadre étudié au delà de la perte des moindres carrés : l'algorithme de descente de gradient stochastique est analysé comme une chaine de Markov. Cette approche résulte en une interprétation intuitive, et souligne les différences entre le cadre quadratique et le cadre général. Une méthode simple permettant d'améliorer substantiellement la convergence est également proposée.

Mots Clés

Approximation stochastique, optimisation convexe, apprentissage supervisé, estimation non-paramétrique, espaces de Hilbert à noyaux reproduisants.

Abstract

The goal of supervised machine learning is to infer relationships between a phenomenon one seeks to predict and "explanatory" variables. To that end, multiple occurrences of the phenomenon are observed, from which a prediction rule is constructed. The last two decades have witnessed the apparition of very large data-sets. This has raised two challenges: first, avoiding the pitfall of over-fitting, especially when the number of explanatory variables is much higher than the number of observations; and second, dealing with the computational constraints, such as when the mere resolution of a linear system becomes a difficulty of its own.

Algorithms that take their roots in stochastic approximation methods tackle both of these difficulties simultaneously: these stochastic methods dramatically reduce the computational cost, without degrading the quality of the proposed prediction rule, and they can naturally avoid over-fitting.

The popular parametric methods give predictors which are linear functions of a set of explanatory variables. However, they often result in an imprecise approximation of the underlying statistical structure. In the non-parametric setting, which is paramount in this thesis, this restriction is lifted. The class of functions from which the predictor is proposed depends on the observations. In practice, these methods have multiple purposes.

The first contribution of this thesis is to provide a detailed analysis of stochastic approximation in the non-parametric setting, precisely in reproducing kernel Hilbert spaces. This analysis proves optimal convergence rates for the averaged stochastic gradient descent algorithm.

The second contribution is an algorithm based on acceleration, which converges at optimal speed, both from the optimization point of view and from the statistical one. In the non-parametric setting, this can improve the convergence rate up to optimality, even in particular regimes for which the first algorithm remains sub-optimal.

Finally, the third contribution of the thesis consists in an extension of the framework beyond the least-square loss. The stochastic gradient descent algorithm is analyzed as a Markov chain. This point of view leads to an intuitive and insightful interpretation, that outlines the differences between the quadratic setting and the more general setting. A simple method resulting in provable improvements in the convergence is then proposed.

Keywords

Stochastic approximation, convex optimization, supervised learning, non-parametric estimation, reproducing kernel Hilbert spaces.