Learning with Reproducing Kernel Hilbert spaces: Stochastic Gradient Descent and Laplacian Estimation

Loucas Pillaud-Vivien

Learning with Reproducing Kernel Hilbert Spaces: Stochastic Gradient Descent and Laplacian Estimation

Loucas Pillaud-Vivien

Under the supervision of Francis Bach and Alessandro Rudi

2020

ENS - Inria Paris - PSL

Maths et Poésie,
Pour les unes, tu trouves et ça commence,
Pour l'autre, tu trouves et ça finit.

— Marc Yor : "Les Unes et l'autre"

Abstract

Machine Learning has received a lot of attention during the last two decades both from industry for data-driven decision problems and from the scientific community in general. This recent attention is certainly due to its ability to efficiently solve a wide class of high-dimensional problems with fast and easy-to-implement algorithms. What is the type of problems machine learning tackles ? Generally speaking, answering this question requires to divide it into two distinct topics: *supervised* and *unsupervised learning*. The first one aims to infer relationships between a phenomenon one seeks to predict and "explanatory" variables leveraging *supervised information*. On the contrary, the second one does not need any supervision and aims at extracting some structure, information or significant features of the variables.

These two main directions find an echo in this thesis. On the one hand, the supervised learning part theoretically studies the cornerstone of all optimization techniques for these problems: stochastic gradient methods. For their versatility, they are the workhorses of the recent success of ML. However, despite their simplicity, their efficiency is not yet fully understood. Establishing some properties of this algorithm is one of the two important questions of this thesis. On the other hand, the part concerned with unsupervised learning is more problem-specific: we design an algorithm to find reduced order models in physically-based dynamics addressing an crucial question in computational statistical physics (also called molecular dynamics).

Even if the two problems are of different nature, these two directions share an important feature: they leverage the use of Reproducing Kernel Hilbert Spaces, which have two nice properties: (i) they naturally adapt to this stochastic framework on a computational-friendly manner, (ii) they display a great expressivity as a class of *test functions*.

More precisely, the first contribution of this thesis is to prove the exponential convergence of stochastic gradient descent of the binary test loss in the case where the classification task is well specified. This work establishes also fine theoretical bounds on stochastic gradient descent in reproducing kernel Hilbert spaces that are a result on their own.

The second contribution focuses on optimality of stochastic gradient descent in the non-parametric setting for regression problems. Remarkably, this work is the first to show that multiple passes over the data allow to reach optimality in certain cases where the Bayes optimum is hard to approximate. This work tries to reconcile theory and practice as common knowledge on stochastic gradient descent always stated that one pass over the data is optimal.

In computational statistical physics as in Machine Learning, the question of finding *low-dimensional representations* (main degrees of freedom) is crucial. This is the question tackled by the third contribution of this thesis. We show, more precisely, how it is possible to estimate the Poincaré constant of a distribution through samples of it. Then, we exploit this estimate to design an algorithm looking for reaction coordinates which are the cornerstones of accelerating dynamics in the context of molecular dynamics.

Detailing, refining and improving this result is the forth contribution of this manuscript. This current work is still not completely finished, but gives some deeper theoretical and empirical insights on the diffusion operator estimation. It was therefore natural that it should be part of this thesis.

Keywords: stochastic approximation, supervised learning, non-parametric estimation, reproducing kernel Hilbert spaces, dimensionality reduction, Langevin dynamics, Poincaré inequality.

Résumé

L'apprentissage automatique a reçu beaucoup d'attention au cours des deux dernières décennies, à la fois de la part de l'industrie pour des problèmes de décision basés sur des données et de la communauté scientifique en général. Cette attention récente est certainement due à sa capacité à résoudre efficacement une large classe de problèmes en grande dimension grâce à des algorithmes rapides et faciles à mettre en oeuvre. Plus spécifiquement, quel est le type de problèmes abordés par l'apprentissage automatique ? D'une manière générale, répondre à cette question nécessite de le diviser en deux thèmes distincts: *l'apprentissage supervisé* et *l'apprentissage non supervisé*. Le premier vise à déduire des relations entre un phénomène que l'on cherche à prédire et des variables "explicatives" exploitant des informations qui ont fait l'objet d'une *supervision*. Au contraire, la seconde ne nécessite aucune supervision et son but principal est de parvenir à extraire une structure, des informations ou des caractéristiques importantes relative aux données.

Ces deux axes principaux trouvent un écho dans cette thèse. Dans un premier temps, la partie concernant l'apprentissage supervisé étudie théoriquement la pierre angulaire de toutes les techniques d'optimisation liées à ces problèmes: les méthodes de gradient stochastique. Grâce à leur polyvalence, elles participent largement au récent succès de l'apprentissage. Cependant, malgré leur simplicité, leur efficacité n'est pas encore pleinement comprise. L'étude de certaines propriétés de cet algorithme est l'une des deux questions importantes de cette thèse. Dand un second temps, la partie consacrée à l'apprentissage non supervisé est liée à un problème plus spécifique : nous concevons dans cette étude un algorithme pour trouver des modèles réduits pour des dynamiques empruntées à la physique. Cette partie aborde une question cruciale en physique statistique computationnelle (également appelée dynamique moléculaire).

Même si les deux problèmes sont de nature différente, ces deux directions partagent une caractéristique commune : elles tirent parti de l'utilisation d'espaces à noyau reproduisant, qui possèdent deux propriétés essentielles : (i) ils s'adaptent naturellement au cadre stochastique tout en préservant une certaine efficacité numérique, (ii) ils montrent une grande expressivité en tant que classe de *fonctions de test*.

La première contribution de cette thèse est de montrer la convergence exponentielle de la descente de gradient stochastique pour la perte binaire dans le cas où la tâche de classification est "facile". Ce travail établit également des bornes théoriques fines sur la descente de gradient stochastique dans les espaces à noyau reproduisant, ce qui peut être considéré comme un résultat en lui-même.

La deuxième contribution se concentre sur l'optimalité de la descente de gradient stochastique dans le cadre non paramétrique pour des problèmes de régression. Plus précisément, ce travail est le premier à montrer que de multiples passages sur les données permettent d'atteindre l'optimalité dans certains cas où l'optimum de Bayes est difficile à approcher. Ce travail tente de réconcilier la théorie et la pratique car les travaux actuels sur la descente de gradient stochastique ont toujours montré qu'il suffisait d'un passage sur les données.

En physique statistique computationnelle comme en apprentissage automatique, la question de trouver des *représentations de faible dimension* (principaux degrés de liberté) est cruciale. Telle est la question abordée par la troisième contribution de cette thèse. Nous montrons plus précisément comment il est possible d'estimer la constante de Poincaré d'une distribution à travers des échantillons de celle-ci. Ensuite, nous exploitons cette estimation pour concevoir un algorithme à la recherche de coordonnées de réaction qui sont les pierres angulaires des techniques d'accélération dans le contexte de la dynamique moléculaire.

Détailler, affiner et améliorer ce résultat est la quatrième contribution de ce manuscrit. Ce travail actuel n'est pas encore complètement terminé, mais il donne de la profondeur aux analyses théorique et empirique de l'estimation des opérateurs de diffusion. Il était donc naturel qu'il fasse partie de cette thèse.

Mots-clés: approximation stochastique, apprentissage supervisé, estimation non-paramétrique, espaces à noyau reproduisant, réduction de dimension, dynamique de Langevin, inégalité de Poincaré.

Contents

I Introduction 13

1 ML framework 14
- 1.1 General Framework of statistical learning . . . 14
- 1.2 Why we use optimization in ML . . . 20
- 1.3 Solving the Least-squares problem . . . 26

2 Stochastic gradient descent 29
- 2.1 Setting . . . 29
- 2.2 SGD as a Markov chain and continuous time limit . . . 31
- 2.3 Analysis of SGD in the ML setting . . . 35

3 Reproducing Kernel Hilbert Spaces 38
- 3.1 Definition - Construction - Examples . . . 38
- 3.2 The versatility of RKHS . . . 42
- 3.3 Promise and pitfalls of kernels in ML . . . 46

4 Langevin Dynamics 50
- 4.1 What is Langevin Dynamics ? . . . 50
- 4.2 Sampling with Langevin dynamics . . . 52
- 4.3 The metastability problem . . . 54

II Non-parametric Stochastic Gradient descent 58

1 Exponential convergence of testing error for stochastic gradient methods 60
- 1.1 Introduction . . . 60
- 1.2 Problem Set-up . . . 61
- 1.3 Concrete Examples and Related Work . . . 63
- 1.4 Stochastic Gradient descent . . . 64
- 1.5 Exponentially Convergent SGD for Classification error . . . 68
- 1.6 Conclusion . . . 69

A Appendix of Exponential convergence of testing error for stochastic gradient descent 71
- A.1 Experiments . . . 71
- A.2 Probabilistic lemmas . . . 73
- A.3 From $\mathcal{H}$ to 0-1 loss . . . 74
- A.4 Exponential rates for Kernel Ridge Regression . . . 75
- A.5 Proofs and additional results about concrete examples . . . 77
- A.6 Preliminaries for Stochastic Gradient Descent . . . 80
- A.7 Proof of stochastic gradient descent results . . . 81
- A.8 Exponentially convergent SGD for classification error . . . 91
- A.9 Extension of Corollary 1 and Theorem 4 for the full averaged case. . . . 93
- A.10 Convergence rate under weaker margin assumption . . . 97

2 Statistical Optimality of SGD on Hard Learning Problems through Multiple Passes **100**
2.1 Introduction . . . 100
2.2 Least-squares regression in finite dimension . . . 101
2.3 Averaged SGD with multiple passes . . . 103
2.4 Application to kernel methods . . . 104
2.5 Experiments . . . 106
2.6 Conclusion . . . 108

B Appendix of Statistical Optimality of SGD on Hard Learning Problems through Multiple Passes **110**
B.1 A general result for the SGD variance term . . . 110
B.2 Proof sketch for Theorem 8 . . . 114
B.3 Bounding the deviation between SGD and batch gradient descent . . . 115
B.4 Convergence of batch gradient descent . . . 117
B.5 Experiments with different sampling . . . 128

III Statistical estimation of Laplacian 132

1 Statistical estimation of the Poincaré constant and application to sampling multimodal distributions **134**
1.1 Introduction . . . 134
1.2 Poincaré Inequalities . . . 135
1.3 Statistical Estimation of the Poincaré Constant . . . 137
1.4 Learning a Reaction Coordinate . . . 140
1.5 Numerical experiments . . . 142
1.6 Conclusion and Perspectives . . . 144

C Appendix of Statistical estimation of the Poincaré constant and application to sampling multimodal distributions **145**
C.1 Proofs of Proposition 11 and 12 . . . 145
C.2 Analysis of the bias: convergence of the regularized Poincaré constant to the true one . . . 146
C.3 Technical inequalities . . . 150
C.4 Calculation of the bias in the Gaussian case . . . 154

2 Statistical estimation of Laplacian and application to dimensionality reduction **164**
2.1 Introduction . . . 164
2.2 Diffusion operator . . . 165
2.3 Approximation of the diffusion operator in the RKHS . . . 166
2.4 Analysis of the estimator . . . 171
2.5 Conclusion and further thoughts . . . 174

IV Conclusion and future work 177

1 Summary of the thesis **177**

2 Perspectives **178**

Contributions and thesis outline

Part I. This manuscript is based on the publications that were accepted during this thesis. Hence, a significant effort in the writing of this manuscript has been spent in this Part. It introduces the main ideas and questions that we will address in the rest of the manuscript. This introduction has two main purposes. First, this part sets the stage for the rest of the thesis by justifying its framework, the use of stochastic gradient descent, RKHS and the study of Langevin dynamics. Secondly, and perhaps more importantly, it gives a personal point of view on the topics under study and defines what are the main interests and foci of future research.

Part II. This Part gathers two results for the non-parametric stochastic gradient descent in two different settings:

- **SGD for classification.** Here, we consider binary classification problems with positive definite kernels and square loss, and study the convergence rates of stochastic gradient methods. We show that while the excess testing loss (squared loss) converges slowly to zero as the number of observations (and thus iterations) goes to infinity, the testing error (classification error) converges exponentially fast if low-noise conditions are assumed.
- **SGD for the Least-squares problem.** We consider stochastic gradient descent (SGD) for least-squares regression with potentially several passes over the data. While several passes have been widely reported to perform practically better in terms of predictive performance on unseen data, the existing theoretical analysis of SGD suggests that a single pass is statistically optimal. While this is true for low-dimensional easy problems, we show that for hard problems, multiple passes lead to statistically optimal predictions while single pass does not; we also show that in these hard models, the optimal number of passes over the data increases with sample size. In order to define the notion of hardness and show that our predictive performances are optimal, we consider potentially infinite-dimensional models and notions typically associated to kernel methods, namely, the decay of eigenvalues of the covariance matrix of the features and the complexity of the optimal predictor as measured through the covariance matrix. We illustrate our results on synthetic experiments with non-linear kernel methods and on a classical benchmark with a linear model.

Part III. In this part we propose a way to estimate Laplacian operators through Poincaré inequalities. Poincaré inequalities are ubiquitous in probability and analysis and have various applications in statistics (concentration of measure, rate of convergence of Markov chains). The Poincaré constant, for which the inequality is tight, is related to the typical convergence rate of diffusions to their equilibrium measure. This part is divided in two blocks:

- **Poincaré constant and reaction coordinates.** We show both theoretically and experimentally that, given sufficiently many samples of a measure, we can estimate its Poincaré constant. As a by-product of the estimation of the Poincaré constant, we derive an algorithm that captures a low dimensional representation of the data by finding directions which are difficult to sample. These directions are of crucial importance for sampling or in fields like molecular dynamics, where they are called reaction coordinates. Their knowledge can leverage, with a simple conditioning step, computational bottlenecks by using importance sampling techniques.
- **Laplacian Estimation and dimensionality reduction.** Here, we extend the previous results on Poincaré constant estimation by proving that the same procedure gives, without additional cost, all the spectrum of the diffusion operator and not only the first eigenvalue. This work highlights the fact that the use of positive definite kernels allows to estimate Laplacian operators with possibly circumventing the curse of dimensionality unlike local methods –which are currently used.

Part IV. This Part concludes the thesis by summarizing our contributions and describing future directions.

Publications. Published articles related to this manuscript are listed below:

- Part II is based on two articles published during the thesis:
  - ⋆⋆ **Exponential convergence of testing error for stochastic gradient methods**, L. Pillaud-Vivien, A. Rudi and F. Bach, published in the *Conference On Learning Theory* in 2018.
  - ⋆⋆ **Statistical Optimality of Stochastic Gradient Descent on Hard Learning Problems through Multiple Passes**, L. Pillaud-Vivien, A. Rudi and F. Bach, published in the *Advances in Neural Information Processing Systems* in 2018.
- Part III is based on a published article and a work in preparation:
  - ⋆⋆ **Statistical Estimation of the Poincaré constant and Application to Sampling Multimodal Distributions**, L. Pillaud-Vivien, F. Bach, T. Lelievre, A. Rudi, G. Stoltz, published in the *International Conference on Artificial Intelligence and Statistics* in 2020.
  - ⋆⋆ **Statistical estimation of Laplacian and application to dimensionality reduction**, L. Pillaud-Vivien and F. Bach, *in preparation*, 2020.

⋆

⋆ ⋆

Forewords and precautions

Before the reader dives into this manuscript, I would like to take the time to write a few words about how I wanted it to be presented. Hopefully, these precautions could guide the reader throughout this work, softening its judgment and luckily putting light on several of its important features.

First, let us begin by saying that this thesis gathers the articles published during the time of PhD. Hence, as this will be the case in Part III of this thesis, several unsolved questions may be stated as *assumptions* in one part and showed later. This approach is deliberate: the will behind this manuscript is to restore what has been done in this thesis, both its questions and its evolution. This is the reason why we decided to leave Part III, Section 2 as an unfinished contribution and preferred to explain the main ideas behind what has to be finished rather than writing a self-contained complete project omitting important pieces of the whole story. In this context, the only newly written *contributions* of this thesis are the introduction, the conclusion and the discussion conducted in the final Section of the thesis (Part III, Section 2).

Second, let me comment briefly on how the introduction has been thought of. Besides, as tradition, recalling the context of this thesis including the general ML framework in supervised learning or the presentation of the less-known dynamics studied in statistical physics, we have tried to think of this introduction as a natural story that has lead us to the studies involved in Part II and III. This is the reason why a particular attention has been paid to motivate deeply the use of stochastic gradient descent or reproducing kernel Hilbert spaces together with their possible pitfalls and future promises. The use of transitions under the form of questions, remarks or developments concluding each subsection of the introduction is the unifying thread of this way of thinking. The reader will certainly remark the following patterns:

⋆
⋆ ⋆

Motivation / Transition / Conclusion / Guideline.

They are breathes in the thesis and their goals are to link, motivate and create a common story line to all this introduction.

Going further, I would like to stress that my personal background on partial differential equations and probability (I had never seen a Machine Learning problem before the beginning of my thesis) drives me naturally to theoretical and modelling questions and to try as much as possible to build bridges with other fields of applied mathematics. This is a personal inclination that hopefully will enrich my future research and be a pleasant guide when reading this thesis.

I would also like to put a particular emphasis on the fact that all what I could say during this introduction or during further developments are *personal point of views*. Even though mathematical theorems are *always true by their logical nature*, (at least my) way of tackling a problem is always subjective and personal. In this thesis, I tried to motivate why certain questions have a particular relevance and why some directions or ways to think could convince me more than others. Nonetheless, these ways of facing a problem are only *personal interpretations* and do not, in any manner, claim to indisputable truth. As a matter of fact, given my young age and inexperience, I am always thrilled to change, sharpen or refine my point of views on many subjects when convinced by good arguments.

Finally, people have often warn me that the PhD was the last moment of the academic life where we could take the time to explore ideas freely. I truly thank my PhD advisor Francis that let me take this freedom. This thesis, and especially the introduction, is, in a way, the presentation of this *other work* that I accomplished during my PhD: gathering, looking into new ideas and building my own personal sensibility.

⋆

⋆ ⋆

Part I

Introduction

1 ML framework **14**
- 1.1 General Framework of statistical learning . . . 14
- 1.2 Why we use optimization in ML . . . 20
- 1.3 Solving the Least-squares problem . . . 26

2 Stochastic gradient descent **29**
- 2.1 Setting . . . 29
- 2.2 SGD as a Markov chain and continuous time limit . . . 31
- 2.3 Analysis of SGD in the ML setting . . . 35

3 Reproducing Kernel Hilbert Spaces **38**
- 3.1 Definition - Construction - Examples . . . 38
- 3.2 The versatility of RKHS . . . 42
- 3.3 Promise and pitfalls of kernels in ML . . . 46

4 Langevin Dynamics **50**
- 4.1 What is Langevin Dynamics ? . . . 50
- 4.2 Sampling with Langevin dynamics . . . 52
- 4.3 The metastability problem . . . 54

1. ML FRAMEWORK

In this part, we will try to introduce the main questions raised in the thesis and we will try to define and motivate the natural setting of this manuscript. We will begin in Section 1.1 with standard definitions, introducing the standard Machine Learning framework from the last or three two decades. Then, as this is the main point of view of this thesis, we will show in Section 1.2 why and how optimization is of crucial importance in common Machine Learning problems. We finally illustrate all these ideas in Section 1.3 in the Least-squares setting.

1.1. General Framework of statistical learning

1.1.1. What is Machine Learning ?

Due to its recent successes in the industry and the phantasms associated to it, Machine Learning (ML) is nowadays often invoked every time data are concerned. However, ML is not the only field dealing with data: other and perhaps older applied mathematics fields such as optimization, statistics or signal processing have tackle numerous problems during the past decades. Obviously, ML is deeply linked to all of them, but a more interesting question is *how* they are related and what are the main differences ? *What* is ML *proper focus* ? Considering my youth in the field I cannot claim that I can sharply define ML, but I will try to pinpoint what is my vision of it. The aim of this manuscript is to guide the reader throughout all the questions I have ask myself during these three years and the answers I tried to give.

Let us begin with one definition: in my opinion, ML is an *high-dimensional* look at statistics that take the current *computational framework* into account.

High-dimensionality. Because all along this thesis two important quantities related to the data will be considered as huge:

- *The size of the samples*: $\boldsymbol{d}$. Examples such that Natural Language Processing (words), vision (pixels) or biological systems (genome) are often embedded in spaces of more that one million dimensions.
- *The number of samples*: $\boldsymbol{n}$. To face the large dimensionality of the data, engineers have built huge data bases, so that $\boldsymbol{n}$ can be also considered as large as million.

To handle well these two large numbers, we will try to focus on *non-asymptotic* results: this will have the benefit to stress the dependence into these two important parameters of the problem. Indeed, asymptotic results can sometimes hide large constants preventing from clear phenomenological explanations. Note that another way to apprehend high-dimensionality may be to give results with respect to a certain function of both n and d going to infinity (*e.g.* n/d). This is not the case in this thesis. We refer to the introduction of [Wai19] for a remarkably clear presentation of the setting of non-asymptotic statistical analysis.

Computational framework. We will also draw a particular attention to the computational complexity of the algorithms analyzed or designed. In ML, as both $\boldsymbol{n}$ and $\boldsymbol{d}$ can be very large, we have to take into account that easy mathematical expressions can be very expensive and thus very time-consuming to compute. Operations such as matrix inversion or multiplication must be avoided as they could lead to unpractical algorithms for real problems in such a high-dimensional setting. Let us add that even if the algorithms designed and analyzed in this thesis carry no memory issue, this could be also a serious computer-related limitation for other procedures.

This being said, we now describe mathematically what is the common setting of our different works.

1.1.2. Supervised learning

Distinction between supervised and unsupervised learning. Learning, as its name states, is all about *learning* from the data. One piece of information that we may want to retrieve from raw data could be to understand its structure or extract representative and understandable features of it. In this case we talk about *unsupervised learning* [STC04, HTF09]. Another task that we may want to do is to infer the outcome of a system leveraging the access of some known input/output pairs of it. Because it often requires the *supervision* of the system by a human being (labeling data for classification is one of the most important example of this), we call this *supervised learning* [Vap13, SSBD14, HTF09]. Roughly speaking, Part II analyzes the workhorse algorithm of *supervised learning* whereas Part III designs and analyzes an algorithm for *unsupervised learning* tasks.

As the work of this thesis on unsupervised learning is very related to some particular task I refer to Part III and to Section 4 for the description of the mathematical setting. I now describe the mathematical framework behind supervised learning whose versatility and wide range apply in Part II.

Supervised learning. In supervised ML, the aim is to predict output $Y \in \mathcal{Y}$ from input(s) $X \in \mathcal{X}$ given that we have access to n input/outputs pairs $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$. The usual ML framework states that there exists some distribution ρ on $\mathcal{X} \times \mathcal{Y}$ such that (x_i, y_i) are independent and identically distributed according to ρ. Note that even if (x_i, y_i) are random variables, we will not use capital letters to denote them, emphasizing on the fact that they are samples. We can here decompose the problem into two sources of randomness:

- **Randomness in the inputs.** They are given according the some law ρ_X (marginal of ρ along $\mathcal{X}$). In this case the fixed design setting arises when ρ_X is a sum of diracs on the x_i. Note also that unsupervised learning techniques with respect to the samples given according to ρ_X can be used as pre-processing on the dataset.
- **Randomness in the outputs and noise model.** A common modelling of the randomness in the outputs is to write that there exists f_* such that

$$Y = f_*(X) + \varepsilon \tag{1}$$

  where ε is the noise of the model. Hence the randomness hypothesis on the output can be instead cast into a random noise on the model. This can be caused by mistakes in the labeling or some errors coming from experiments when collecting the data. Note that when we assume ε independent of X, we often say that the model is well-specified.

Remark 1 (*Support of ρ_X*)

> *It is really important to note that ρ carries all the information of the problem and that we do not have access to it. Even finding the support of ρ_X is a problem in itself and a very difficult task. To understand this, let us take the example of face recognition on images with $d \sim 10^6$ pixels. The marginal ρ_X lives naturally in the space of vectorized images $\mathbb{R}^d$. However* only a few *images are faces and the support ρ_X would exactly be the* sub-manifold of images constituted of faces. *Sampling from this manifold is actually a very hard task and a problem in itself.*

Remark 2 (*Hypothesis on ρ*)

> *Making hypothesis on ρ changes dramatically the problem under study. One example already given is the difference between the random and the fixed design settings. But we can also make hypothesis on the noise through ρ. For example, what we can place ourselves in the* interpolation regime $\varepsilon = 0$, *corresponding to the case where the marginal along $\mathcal{Y}$ is a dirac in $f_*(X)$.*

Considering Eq. (1), the problem of supervised learning is to *learn* the function f_*. For this, quantifying the precision of a predictor will be necessary: this is what we do in the following section.

1.1.3. Losses and Generalization error

Let us define our predictor, f: this is simply is a measurable function from $\mathcal{X}$ to $\mathcal{Y}$, we denote the set of such functions $\mathcal{M}(\mathcal{X} \times \mathcal{Y})$. Quantifying the accuracy of the output $Y = f(X)$ is the first task that we want to do to try solve our model. For this we define naturally a loss

$$\ell : ((\mathcal{X} \times \mathcal{Y}), \mathcal{M}(\mathcal{X} \times \mathcal{Y})) \to \mathbb{R}_+, \tag{2}$$

where we say that ℓ is a suited loss for the problem if

$$\ell((X, Y), f) \text{ is small } \Leftrightarrow f(X) \text{ is a good predictor of } Y. \tag{3}$$

Here, we also want our predictor to show some good performances not only on the n samples we have access to but also on all the possible data coming from ρ. Hence, the good quantity to consider is the risk, also called generalization error or test error:

$$\mathcal{R}(f) := \mathbb{E}_{(X,Y)\sim\rho}\left[\ell((X, Y), f)\right]. \tag{4}$$

Rephrase mathematically, the aim of supervised learning is to find f such that $\mathcal{R}(f)$ is the smallest possible. We will denote with the subscript "$*$" the fact that we reach the minimum value or the argument that minimize this value. We define here the best predictor of our learning problem and the minimum risk associated to it:

$$f_* = \underset{f\in\mathcal{M}(\mathcal{X}\times\mathcal{Y})}{\operatorname{argmin}} \mathcal{R}(f) \tag{5}$$

$$\mathcal{R}^* = \mathcal{R}(f_*) \tag{6}$$

The choice of the loss is determinant and has to be made thoroughly and according to the problem under consideration. Besides the obvious requirement stated in (3), we will see later that other issues such as the need of convexity or robustness will come into consideration. But first let us stress out two important classes of problem and their commonly associated losses.

Regression. When $\mathcal{Y}$ is some interval of $\mathbb{R}$, we call the problem regression. For this and throughout Part II the typical loss will be the square loss: $\ell((X, Y), f) = \frac{1}{2}(Y - f(X))^2$.

Classification. It arises when the output space is binary. To set ideas, we can take $\mathcal{Y} = \{-1, +1\}$. Yes-No decisions or the well-known cat and dog problems are instances of this type of problem. To tackle this, the more natural loss is the binary loss $\ell((X, Y), f) = \mathbb{1}_{Y\neq\text{sign}(f(X))}$ that penalized by 1 each time a wrong prediction in made. However, as the binary loss lacks some good mathematical property (convexity, smoothness) we often use surrogates losses for the problem: the logistic loss $\ell((X, Y), f) = \log(1 + \exp(-Yf(X)))$ or the hinge loss used in support vector machines $\ell((X, Y), f) = \max\{0, 1 - Yf(X)\}$ [SC08]. The classification problem is tackled in the Section 1 of Part III.

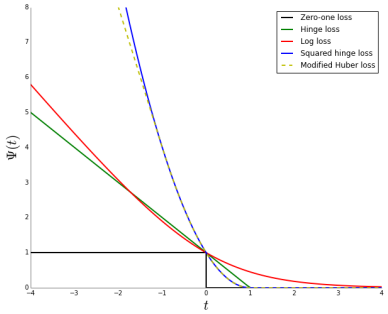

Figure 1: Usual losses in ML

1.1.4. Choosing the space of functions: pros and cons

Bayes predictor. Now that Eq. (3) give us a good measure of how good our predictor is, we can try to solve the problem Eq. (5). Mathematically speaking this is an infinite-dimensional optimization problem over the space of measurable functions $\mathcal{M}(\mathcal{X} \times \mathcal{Y})$ which is obviously intractable as $\mathcal{M}(\mathcal{X} \times \mathcal{Y})$ is a very hard space to apprehend. Yet, it is quite remarkable that in the case of the square loss, when X, Y are square-integrable real random variables, we can compute exactly the optimum: f_* is the orthogonal projection from X to the linear subspace of Y-measurable functions:

$$f_*(X) = \mathbb{E}\left[Y|X\right]. \tag{7}$$

This function is called the Bayes predictor, and even if we have a closed form in this case, it remains to approximate it properly. Recall here that *we do not have access to the joint distribution ρ but only to samples of it.* Approximating directly the Bayes predictor is possible by local averaging techniques [Tsy08] but it is very expensive in terms of samples even if moderate dimensions. Thus it is not the path we follow during this thesis.

How to choose the space of function $\mathcal{H}$. Recall that one of the focus of this work is to be able to *compute numerically* good predictors. When dealing directly with infinite dimensional spaces such as $\mathcal{M}(\mathcal{X} \times \mathcal{Y})$ or even smaller like $L^2(\mathcal{X} \times \mathcal{Y})$, it seems impossible to design computational-friendly routines to solve Eq. (5). Hence, a good idea is to parametrize the space of functions by some parameter θ living in a finite dimensional space $\mathbb{R}^s$ and that encodes a dictionary of functions on which we can solve the problem Eq. (5). We call these parametric spaces $\mathcal{H}$. One of the most basic yet powerful ideas is to define $\mathcal{H}$ as the linear functions from $\mathbb{R}^s$ to $\mathbb{R}$: $\mathcal{H}_\phi = \{f \mid f(x) = \langle \theta, \phi(x) \rangle,\ \theta \in \mathbb{R}^s\}$, where $\phi(x)$ is a vector of $\mathbb{R}^s$ containing features of x. However this parametrization comes at a cost: when restricting all the possible predictors to a smaller class, we may be far away from the best predictor possible. Note that ϕ is not necessarily itself linear, and can be learned, for example using deep learning techniques [LBH15] (that we will introduce in few lines). In fact, we need two ingredients to choose properly the space $\mathcal{H}$ of possible predictors:

(i) $\mathcal{H}$ has to make the problem (5) solvable with a computer.

(ii) $\mathcal{H}$ has to be large enough to approximate well the Bayes predictor. We often call this the *expressivity* of the function space.

Other classes of function spaces satisfy (i) and (ii) without being parametric such as Reproducing Kernel Hilbert Spaces (RKHS) [SS02, SC08]. As they are the core of this thesis, we decided to postpone a little bit the description of RKHS in Section 3 of this introduction.

Another class of functions satisfying (i) and (ii) that I will only introduce are function spaces represented by Neural Networks. Their construction is not new and date back to the 60s [IL67]: they are parametric function spaces simply built as successive compositions of linear functions and non-linear activations (such that the rectified linear activation $x \to \max(0, x)$). It is worthy to say that, from a very high-level point of view, their ability to solve well ML problems comes from their expressivity and easy computational framework (even if they still carry some mysteries).

Remark 3 (*Splines*)

> *An example of function spaces that are not well suited for our framework, and yet can solve very well the problem (5) are* splines *[Wah90]. These are function spaces defined by piece-wise polynomials. On the one hand, they show a great expressivity but are computationally demanding on the other hand (especially in the high-dimensional setting).*

1.1.5. Solving the ML problem: statistical issues, overfitting and minimax rates

Empirical Risk Minimization (ERM). As we already said a few times before, we do not have access to the distribution ρ and thus to the true risk defined in Eq. (4). As we only have access to n samples from ρ a good idea is to substitute the risk defined by an expectation over ρ to an expectation over the associated empirical measure of ρ: $\widehat{\rho}_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$. This defines the empirical risk:

$$\widehat{\mathcal{R}}_n(f) := \mathbb{E}_{(X,Y)\sim\widehat{\rho}_n}\left[\ell((X,Y),f)\right] = \frac{1}{n}\sum_{i=1}^n \ell((x_i,y_i),f). \tag{8}$$

Now we have all the tools to define the cornerstone of supervised learning, *Empirical Risk Minimization*, which is simply reformulating the true problem (5) with respect to the empirical measure associated to the samples:

$$\widehat{f}_n = \operatorname*{argmin}_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n \ell((x_i,y_i),f). \tag{9}$$

Remark 4 (*Statistical point of view on least-squares and logistic regression*)

The ERM framework described above can be viewed as Maximum Likelihood Estimation (MLE) for (at least two) statistical models on the distribution ρ.

- *In the case of Gaussian linear regression, when we want to fit a Gaussian of mean $\langle\theta, X\rangle$ as the law that generated Y, the maximum likelihood estimation is exactly the least-squares empirical risk minimization.*
- *We can also cast a MLE setting to a classification problem with the logistic loss when considering a statistical model on the joint distribution ρ such that $\mathbb{P}(Y|X,\theta) = \mathsf{B}\left(\frac{\exp\langle\theta,X\rangle}{1+\exp\langle\theta,X\rangle}\right)$, where $\mathsf{B}(p)$ is a Bernoulli law with parameter p.*

Note that the main difference with our work is that we never specify a priori *a statistical model on the distribution and do not assume that the model is well-specified.*

Overfitting and regularization. Solving directly and perfectly the ERM in Eq. (9) seems a good idea. But, actually, there is no guarantee that solving the empirical problem will generalize well when we want to solve the true one Eq. (5). In fact, solving perfectly without further considerations will lead to a bad estimation of the true predictor. Indeed, if the space of test function is large enough one always can find a predictor such that $f(x_i) = y_i$ but generalized very poorly outside of the x_i: you can picture yourself this with degree n Lagrange polynomials on $\mathbb{R}$ that will interpolate perfectly inputs and outputs but behave very badly outside of the interpolated points. This phenomenon is known as *overfitting*. One way to avoid this is whether to *regularize* the problem by some penalty term forcing a certain regularity of the estimator (see Figure 2 for an illustration), this is an old idea in statistics that occurs for example in smoothing splines [Gu13]. Another way to do this is to restrict the space of function to a regular one to avoid chaotic behavior outside the dataset. Note that both approaches are in fact equivalent [HTF09].

Approximation and estimation errors. As said above, regularizing to avoid overfitting is equivalent to work in a smaller and smoother space. But the smaller the space of predictors we look for the less expressive our model get and the more we fail to approximate the best achievable predictor f_*. To formulate this fact more formally let us call $f_{\mathcal{H}}$ the best estimator of our class of function $\mathcal{H}$ and $\widehat{f}_n$ the estimator based on the empirical risk. What we want to control is the *excess risk*, which is the best achievable risk considering our model. It can be decomposed into two terms:

$$\mathcal{R}(\widehat{f}_n) - \mathcal{R}(f_*) = \underbrace{\mathcal{R}(\widehat{f}_n) - \mathcal{R}(f_{\mathcal{H}})}_{\text{estimation error}} + \underbrace{\mathcal{R}(f_{\mathcal{H}}) - \mathcal{R}(f_*)}_{\text{approximation error}} \tag{10}$$

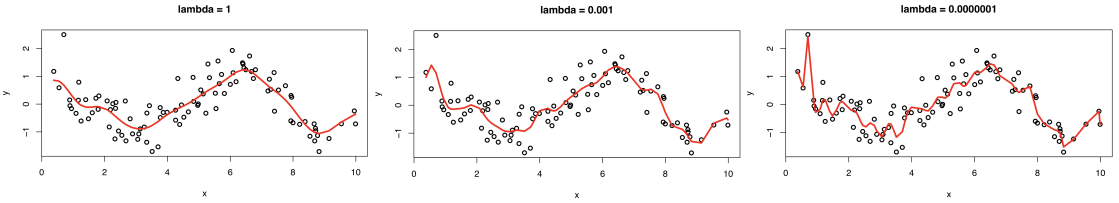

Figure 2: Showing the regularization to overfitting phases when decreasing the regularization parameter in a regression task. These plots come from the slides of J.-P. Vert and J. Mairal lessons on Kernel methods.

The *approximation error* only depends on the class of function $\mathcal{H}$ chosen for our problem. This is a deterministic term that gets smaller as $\mathcal{H}$ get bigger.

The *estimation error* comes from the fact that our estimator $\widehat{f}_n$ comes from the minimization of the empirical risk and not of the true one. In fact, one can show that we can upper bound it by a certain uniform distance between the two functions $\widehat{\mathcal{R}}_n$ (which is a random function) and $\mathcal{R}$:

$$\begin{aligned}\mathcal{R}(\widehat{f}_n) - \mathcal{R}(f_{\mathcal{H}}) &= \mathcal{R}(\widehat{f}_n) - \widehat{\mathcal{R}}_n(\widehat{f}_n) + \underbrace{\widehat{\mathcal{R}}_n(\widehat{f}_n) - \widehat{\mathcal{R}}_n(f_{\mathcal{H}})}_{\leqslant 0} + \widehat{\mathcal{R}}_n(f_{\mathcal{H}}) - \mathcal{R}(f_{\mathcal{H}}) \\ &\leqslant \mathcal{R}(\widehat{f}_n) - \widehat{\mathcal{R}}_n(\widehat{f}_n) + \widehat{\mathcal{R}}_n(f_{\mathcal{H}}) - \mathcal{R}(f_{\mathcal{H}}) \\ &\leqslant 2 \sup_{f\in\mathcal{H}} \left|\widehat{\mathcal{R}}_n(f) - \mathcal{R}(f)\right|.\end{aligned}$$

A little taste of *empirical process theory*. Bounding uniformly the deviation between $\mathcal{R}$ and its corresponding average is the key point of *empirical process theory* [VDVW96, Tal94]. Let us now put emphasis on the fact that this kind of development *is not* the point of view taken in this thesis as our estimator will come from an optimization procedure and will benefit from implicit forms of regularization (see next section for more details). However, let us try to summarize what are the main ideas and results behind this. An important quantity is the *Rademacher complexity* associated to the loss and the function space $\mathcal{H}$. It measures richness of a class of real-valued functions with respect to a probability distribution:

$$\mathsf{Rad}_{\mathsf{n}} = \mathbb{E}_{\sigma,\rho}\left[\sup_{f\in\mathcal{H}}\left(\frac{1}{n}\sum_{i=1}^{n}\sigma_i\ell((x_i,y_i),f)\right)\right],$$

where σ are i.i.d. Rademacher variables $\mathbb{P}(\sigma_i = \pm 1) = 1/2$. We can show from a symmetrization argument that

$$\mathbb{E}\sup_{f\in\mathcal{H}}\left|\widehat{\mathcal{R}}_n(f) - \mathcal{R}(f)\right| \leqslant 2\mathsf{Rad}_{\mathsf{n}}.$$

Hence, controlling the Rademacher complexity allows to bound the excess error for a wide range of classes of losses and $\mathcal{H}$. For examples, if the loss is L-Lipschitz, inputs are bounded by R and the functional space in formed of κ-bounded functions, one has $\mathsf{Rad}_{\mathsf{n}} \leqslant \kappa RL/\sqrt{n}$ [HTF09]. Note that these bounds could be tighten with finer assumptions using localized version of Rademacher complexities [BBM05]. But, as already stated, this is not the line of search of our work as our analysis relies on direct and straight computations. However, I truly believe that knowing and summing up this beautiful and deeply rooted theory of statistical learning was worth the detour and could at any point complement my point of view.

Minimax rates of convergence. Throughout the thesis, we will focus only on upper bounds of our estimators like in the precedent paragraph. However each time we find such an upper bound we may immediately ask the following questions: is the analysis tight ? Given the level of information I have on the problem (number of samples n, *a priori* on ρ, level of noise...), can I build a different estimator will generalize better ? In what way is my result or my estimator impossible to improve ?

These questions raise the fundamental concept of *optimality* of the result (in the sense that it cannot be improved). Minimax rates of convergence are exactly the good mathematical tool to embrace this concern: they give the best possible level of precision we can reach considering the problem we have. More formally, let $\Theta \subset L^2(\rho_X)$ be a space (parametric or non-parametric at this stage) of function where *a priori* we expect the target function, f_ρ, to lie. Let $\mathcal{M}(\Theta)$, be the associated classes of measure such that $f_\rho \in \Theta$. The best we know is that $\rho \in \mathcal{M}(\Theta)$. Our goal is to have a lower bound on the best estimator over the set of all the estimators $\mathbb{E}_n : \mathbf{z} \to \mathbf{f_z}$ where $\mathbf{z}$ stands for the data set.

$$\mathsf{Minimax}_n(\Theta) := \inf_{\mathbb{E}_n} \sup_{\mu \in \mathcal{M}(\Theta)} \mathbb{E}\left(\|f_\mu - f_{\mathbf{z}}\|^2_{L^2(\mu_X)} \right). \tag{11}$$

Even if for some classical settings, such minimax bounds can be derived [Tsy08] (we refer also to Section 1.3 for least-squares and Part II, Section 2 in non-parametric settings), the reader can imagine how difficult the problem of finding such a quantity can be: we need to construct monstrous functions that are the *less learnable ones* over a class of distributions.

From statistical learning to optimization. All this theory seems satisfying to solve supervised machine learning problems and gives guarantees for the estimators. But as stated in the first subsection 1.1.1 our concern is end-to-end: we really want to be able to compute numerically our estimators. And here is the big elephant in the room. For general problems

$$\widehat{f}_n = \operatorname*{argmin}_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell((x_i, y_i), f)$$

is not computable in closed form. We will see that even when it is (e.g. for least-squares), numerical computations can be an important limitation. This is why the point of view of this thesis is the optimization one. We will try to give intuition and explanations behind its efficiency in the next section.

1.2. Why we use optimization in ML

As we have seen in the previous section, finding a good estimator for supervised learning tasks is naturally cast into the ERM optimization problem:

$$\text{Find} \qquad \operatorname*{argmin}_{f \in \mathcal{H}} \widehat{\mathcal{R}}_n(f) = \frac{1}{n} \sum_{i=1}^{n} \ell((x_i, y_i), f). \tag{12}$$

The questions addressed in this section is how to solve concretely such a problem and what are the main safeguards and elements that we have to pay attention for.

1.2.1. Advantages of optimization: numerical cost, implicit regularization and bias, and eventually a bit of magic

When it comes to numerically solve optimization problems like (12), the first idea that should come into mind is one of the more versatile approach of applied mathematics: *gradient descent* [BBV04]. Besides its

simplicity, it is actually the cornerstone of (almost) all the optimization techniques used to solve supervised learning problems. Of course, one has to require some Hilbert structure of $\mathcal{H}$ and some smoothness and convexity property to solve well this problem. However note that smoothness is not always necessary if replacing gradients by sub-gradients [Boy04] and that escaping from the convexity imperative might be the next important question –we will come back to this later on.

Numerical cost. Even when the problem (12) is well posed and has a solution, there exist only a few cases for which we can explicitly build such an estimator. Worse, as we will show later, even for one of the simplest setting that is least-squares regression (linear space of functions with square loss): to compute the estimator (12) requires a matrix inversion which is not compatible with our high-dimensional computationally-friendly framework. On the contrary, gradient descent methods are based on a certain number of low cost iterations: even if there exist important variants of it as we will see in the next section, basically the cost of one iteration only requires to compute one gradient.

Versatility. As said earlier, as long as there is some Hilbert structure on the space $\mathcal{H}$ and some very mild assumption on the second variable of the loss ℓ, gradient descent techniques can always be used. Computationally speaking we may add at this point that the successes of Neural Networks is partly due to the automatic differentiation [G+, PGC+17] (at the heart of the back-propagation in Neural Networks [HN92]): this is a very user-friendly framework for computing automatically derivatives and thus implementing gradient descent.

Implicit regularization and implicit bias. As we have seen earlier with the overfitting phenomenon, the space of function $\mathcal{H}$ may be too large and solving exactly (12) could lead to poor generalization. However there are two widely studied effect that can prevent overfitting to occur:

- *Implicit regularization by early stopping.* The first ingredient that can prevent optimization to overfit the data is the fact that it is not necessary to optimize (12) until the end. More importantly, we can show that stopping the gradient descent before it has fully optimized the empirical risk is a way to regularize the problem [YRC07]. In practice, one can use the criterion that when test error (on the validation set) is going up again overfitting is starting to appear and one should stop the gradient descent.
- *Implicit bias by norm minimization.* The second ingredient is more subtle in a way. First let us recall that we say that a problem is *overparametrized* when we have enough degrees of freedom in our model to perfectly fit the data. Hence, the question becomes: if there are plenty of estimators minimizing the training risk, then which one should I select to generalize well ? This is where gradient descent comes into play: we can show in certain settings that gradient descent has the property to *select good estimator*. Here are two examples showing the implicit bias of gradient descent:
  - ⋆⋆ One can show that for least-squares regression, gradient descent converges to the interpolating estimator that has to the minimum $\|\cdot\|_2$ norm solution [SHN+18].
  - ⋆⋆ Similarly, one of the success of SVMs classifiers in the case where the data is fully separable is that gradient descent on the empirical risk problem for the logistic loss converges to the maximum margin solution [SC08] (for the norm induced by the space of function $\mathcal{H}$). Recently and quite remarkably similar results have even been shown for Neural Networks in the case of gradient flow [CB20].

Bit of magic. There have been quite some efforts spent to show how optimization procedures provide good estimators in overparametrized and non-convex systems. Many works invoke the ability of such algorithms to find wide and flat regions of the empirical risk that have the ability to generalize well [CCS+19].

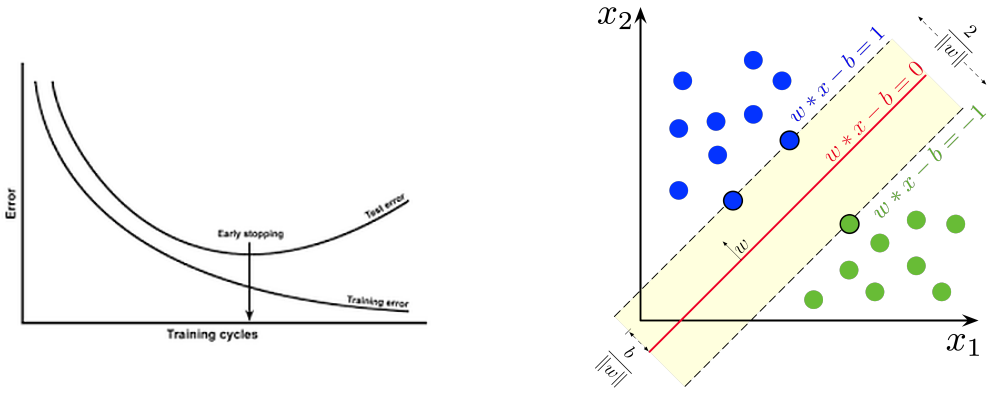

Figure 3: (Left) Showing earling stopping strategy as a regularization procedure (Right) Showing max-margin effect in classification with logistic loss (*Wikipedia image*).

Another line of work supports that such algorithms avoid naturally bad regions and escape from local minima thanks to their momentum and/or their stochasticity. All those directions are very promising, yet, it seems that none of these ideas have fully convinced enough people to establish a form of consensus. We will conclude by saying that this is still an exciting line of research to discover what makes gradient descent and all its variants perform so well in these tasks.

1.2.2. Gradient based algorithms: which one is the most suited for ML ?

There exists a large bestiary of gradient-based algorithms to solve optimization problems. The purpose of this section is not to give a precise and exhaustive description of such techniques but rather to give an intuition behind the use of certain algorithms. For a more mathematical perspective on such algorithms (see [BBV04] for detail analysis), we refer to subsection 1.2.3 for gradient methods and for section 2 for stochastic gradient methods.

Gradient descent algorithms. As already said all the algorithms that we will define are based on the standard gradient descent algorithm.

- *Gradient descent.* You cannot be simpler than gradient descent principle: if you want to find the minimum of a function, just follow the line of its steepest descent. More formally, and if we use a notation that rings with risk minimization, to minimize $\mathcal{R}(\theta)$ over θ, the gradient descent is an iterative process that chooses γ_t as step-size, $\theta_{t=0} = \theta_0$ at initial time and writes at times $t > 0$:

$$\theta_t = \theta_{t-1} - \gamma_t \, \nabla_\theta \mathcal{R}(\theta_{t-1}). \tag{13}$$

- *Newton's method.* Newton method can be seen as a way to choose optimally the step-size γ_t. In fact, if we perform a Taylor expansion of order 2 of the function and find the step-size that optimize such a local parabola, then, the optimal step-size is remarkably the inverse of the Hessian $\nabla^2\mathcal{R}(\theta)$. Sometimes called natural gradient in Bayesian learning, this algorithm has the nice idea to leverage the local geometry of the function around the current iterate to speed-up the convergence.

$$\theta_t = \theta_{t-1} - \left[\nabla^2\mathcal{R}(\theta_{t-1})\right]^{-1} \nabla_\theta \mathcal{R}(\theta_{t-1}). \tag{14}$$

  Note that when $\mathcal{R}(\theta)$ is quadratic then Newton's method converges in one iteration. Many algorithms are inspired by this very efficient method of order two and try to approximate the inverse of the Hessian (which is the bottleneck of the computation as we will see later).

- *Stochastic gradient descent and mini-batch Gradient descent.* When $\mathcal{R}$ has a sum structure such as in supervised learning problems, it is possible to leverage this structure by taking only a minibatch B of the whole gradient.

$$\nabla_\theta \mathcal{R}(\theta) = \frac{1}{n}\sum_{i\leqslant n} \nabla_\theta \mathcal{R}_i(\theta) \longrightarrow \frac{1}{|B|}\sum_{i\in B} \nabla_\theta \mathcal{R}_i(\theta).$$

  A limit case that we will study throughout Part II of this thesis is the limit case where $|B| = 1$, we can thus write with the above notation replacing $\mathcal{R}$ by $\mathcal{R}_t$ in Eq. (15):

$$\theta_t = \theta_{t-1} - \gamma_t \, \nabla_\theta \mathcal{R}_t(\theta_{t-1}). \tag{15}$$

- *Acceleration methods.* There are different ways to accelerate such procedures and we will not dwell into these techniques as they are not very relevant for this thesis. Up to my knowledge, almost every acceleration methods boil down to adding some extra inertial term on top of the classical gradient descent [Pol64, Nes83]. One personal remark about them: even though they can be widely used in practice, it seems to me after many discussions with practitioners that for ML applications they can be unstable and do not offer very different performances of a properly tuned basic stochastic gradient descent algorithm. However, note that acceleration can perform well in certain settings: it is the case for the randomized coordinate gradient descent as shown in [Nes12] where the only source of noise is multiplicative (see Section 2 of the introduction for more details).

The Bottou-Bousquet lessons. In a celebrated article [BB08], Bottou and Bousquet analyzed the relevance of the different algorithms presented above in the context of Machine Learning when the function to minimize is the population risk (yet we have only access to the empirical risk). The two main ideas given by the article have influenced largely the optimization framework for Machine Learning in the past decade.

- *First idea: we should really be concerned about minimizing the true risk and not the training one.* This naive idea has the following consequence: as the train risk is not exactly the true one (typical distance is of order $1/\sqrt{n}$) it is useless optimize under a certain radius (of typical size $1/\sqrt{n}$).
- *Second idea: for large-scale optimization, "bad" optimization algorithms can perform better.* For the large-scale optimization framework where we are in (large n and d), some operations are very costly to perform as recalled earlier in this thesis. Note that computing the whole gradient of the empirical risk cost $O(nd)$ computing the Hessian costs the square of this price and inverting it is extremely expensive and unstable ! Gradient descent and Newton methods need only a few iterations to converge but each iteration costs a lot. This is the reason why, as far a the *time cost* is concerned, stochastic gradient descent is preferable is such settings in comparison to full gradient descent.

1.2.3. General Optimization

The first thing that one has to know about gradient descent is that for smooth functions it always converges to a critical point of the function to minimize. Convexity is then the good way to turn the set of critical points to global minimizers of the function. In all this section, let us call f such a function for simplicity. Note that, as deterministic optimization is not our particular concern, all the theorems stated below will be stated in user-friendly settings. Note that all the hypothesis can be weakened. We refer to [BBV04] for further details on the topic.

Some definitions. Let us suppose that the function to minimize is continuously differentiable: $f \in \mathcal{C}^1(\mathbb{R}^d)$. We say that f is *convex* if it satisfies the following inequality for all $x,\ y \in \mathbb{R}^d$,

$$f(x) \geqslant f(y) + \langle \nabla f(y) | x - y \rangle \,, \tag{16}$$

which only traduces the fact that at any point $x \in \mathbb{R}^d$ the affine approximation of f is below it. We will also need some smoothness of the gradient (L-Lipschitz) to ensure stability of the convergence with respect to the step-size. We say that f is L-smooth if for all $x,\ y \in \mathbb{R}^d$,

$$\|\nabla f(x) - \nabla f(y)\| \leqslant L \|x - y\| . \tag{17}$$

Finally, we say that f is μ-strongly convex if there exists a constant $\mu > 0$ such that for all $x,\ y \in \mathbb{R}^d$,

$$f(x) \geqslant f(y) + \langle \nabla f(y) | x - y \rangle + \frac{\mu}{2}\|x - y\|^2, \tag{18}$$

which is a stronger statement than convexity. In fact if f is twice continuously derivable all of the above properties turn into Hessian conditions: (i) Convexity (16) is equivalent to $\nabla^2 f \succcurlyeq 0$, (ii) Smoothness (17) is equivalent to $\nabla^2 f \preccurlyeq L$ and (iii) Strong convexity (18) is equivalent to $\nabla^2 f \succcurlyeq \mu$. As already stated before, the geometry of the function f is given by its Hessian so that it seems quite natural that such hypothesis are the cornerstone of convergence guarantees.

Gradient descent. Consider the simple optimization problem over L-smooth and convex functions f on $\mathbb{R}^d$:

$$\text{Find} \quad \min_{\theta \in \mathbb{R}^d} f(\theta). \tag{19}$$

Let us call as usual θ_* the unique minimizer of f (we suppose for clarity that f has a unique minimizer, note that it is true when f is strongly convex and that it does not change the idea behind gradient descent to suppose this). As said earlier gradient descent corresponds to making a step towards the steepest direction for the $\|\cdot\|_2$-norm. Note that changing the norm will change the direction, for example choosing the $\|\cdot\|_1$-norm will lead to another descent algorithm called *coordinate descent*. Let us recall the iteration scheme of gradient descent: it begins at θ_0 and for $t > 0$,

$$\theta_t = \theta_{t-1} - \gamma_t \nabla_\theta f(\theta_{t-1}).$$

As Newton's method shows, the choice of the step-size, also called learning rate in ML is of crucial importance. For L-smooth functions, we can chose uniformly the step-size as $\gamma_t = \frac{1}{L}$ to make the algorithm converge to the optimal solution $f(\theta_*)$, this is the meaning of the following proposition:

Proposition 1 (*Convergence of gradient descent*)

Let f be convex and L-smooth, let $\gamma_t = \frac{1}{L}$. The sequence of gradient descent $(\theta_t)_{t \geqslant 0}$ initialized at θ_0 satisfies at time $t > 0$ the following inequality:

$$f(\theta_t) - f(\theta_*) \leqslant \frac{2L\|\theta_0 - \theta_*\|^2}{t + 4}.$$

Moreover, if f is μ-strongly convex, we have the following upper bound:

$$f(\theta_t) - f(\theta_*) \leqslant \left(1 - \frac{\mu}{L}\right)^t \left(f(\theta_0) - f(\theta_*)\right).$$

Note that this choice of the step-size is fairly adaptive since without changing the step-size we have acceleration from linear to exponential convergence when f is strongly convex.

Lower bound for first order algorithms. One natural question to ask is whether this algorithm achieves the best possible rate. Is this possible to accelerate it only using gradients of the function ? Actually the answer to this question is negative, and one step forward to understand this is the fact that usual lower bounds are faster than the rates achieved in Proposition 1. More precisely, one can design functions such that the convergence over all first order methods is lower bounded by $1/t^2$ for convex function and $\sim (1 - \sqrt{\frac{\mu}{L}})^t$ for strongly convex ones. To tighten this lower bound, Nesterov remarkably designed an eponymous acceleration [Nes83], this is the object of the next paragraph.

Accelerated gradient descent. As said earlier for ML learning, numerous accelerated methods can be seen as first order methods where we add some inertia to increase the speed of the procedure. This idea dates back to the seminal work of Polyak [Pol64] with the heavy ball algorithm. Nesterov's acceleration, even if very similar, compute the gradient in a extrapolated step whereas Polyak's heavy ball compute it on the current point *then* apply some inertia. This little difference seems to stabilize the acceleration as in some cases the heavy ball does not converge. More precisely, Nesterov adds an extra sequence η_t and momentum δ_t following

$$\theta_t = \eta_t - \gamma_t \nabla f(\eta_{t-1}) \qquad \text{gradient step} \tag{20}$$
$$\eta_t = \theta_t + \delta_t(\theta_t - \theta_{t-1}) \qquad \text{momentum step.} \tag{21}$$

The following proposition shows that this procedure is in fact optimal for first-order methods.

Proposition 2 (*Convergence of accelerated gradient descent*)

Let f be convex and L-smooth, let $\gamma_t = \frac{1}{L}$ and $\delta_t = \frac{t-1}{t+2}$. The sequence of accelerated gradient descent $(\theta_t)_{t\geqslant 0}$ initialized at θ_0 satisfies at time $t > 0$ the following inequality:

$$f(\theta_t) - f(\theta_*) \leqslant \frac{2L\|\theta_0 - \theta_*\|^2}{(t+1)^2}.$$

Moreover, if f is μ-strongly convex, change $\delta_t = \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$ we have the following upper bound:

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

Let us give two remarks about this acceleration. First, this acceleration keeps some algebraic mysteries: the way the momentum is chosen and the resulting acceleration has received a large attention but there does not seem to be a consensus to explain its miraculous behavior. Different interpretations have been given, heavy-ball-like effect, coupling with mirror descent [AZO14], geometric reasons [BLS15], second-order ODE [SBC16], but none of them seems to have convinced the entire community. Second, as we are concerned by ML optimization and stochastic counterparts of gradient methods, it is notable to see that accelerated methods are not very robust to noise and hence not much employed. In a word, when it comes to accelerating stochastic algorithms, other ideas could be better than Nesterov's acceleration.

⋆
⋆ ⋆

Conclusion of optimization for ML. From this part we conclude two important things about supervised learning. First, optimization algorithm are very well suited for solving the empirical risk minimization associated to ML problems. Second, even if there are some variants built from it, stochastic gradient descent (SGD), is the most adapted algorithm to solve these problems. We will generously detail the performance of SGD later in Section 2. We try to illustrate all what we have discussed above in the canonical example of supervised learning: the least-squares problem.

1.3. Solving the Least-squares problem

1.3.1. Precise setting of the Least-squares problem

Let us illustrate all the above ideas by solving one of the most simple (yet rich) problem of supervised learning. Let us consider

$$\min_{f\in\mathcal{H}} \mathcal{R}(f) := \mathbb{E}_\rho\left[\ell((X,Y),f)\right].$$

And let parametrize the problem in the simplest way:

- $\mathcal{H}$ is space of linear functions of feature vectors Φ of $\mathbb{R}^d$: $\mathcal{H}_\theta := \{x \to \langle\theta, \Phi(x)\rangle,\ \theta\in\mathbb{R}^d\}$.
- Take ℓ the square loss, $\ell((X,Y),f) = \frac{1}{2}(f(X)-Y)^2$.

The problems of risk minimization and its empirical counterpart given n i.i.d. samples $(x_i,y_i)_{i\leqslant n}$ writes:

$$\text{Find} \qquad \min_{\theta\in\mathbb{R}^d} \mathcal{R}(\theta) := \frac{1}{2}\,\mathbb{E}_\rho\left(\langle\theta,\Phi(X)\rangle - Y\right)^2 \tag{22}$$

$$\text{Find} \qquad \min_{\theta\in\mathbb{R}^d} \widehat{\mathcal{R}}_n(\theta) := \frac{1}{2n}\sum_{i=1}^n \left(\langle\theta,\Phi(x_i)\rangle - y_i\right)^2. \tag{23}$$

Let us note $\Sigma = \mathbb{E}_{\rho_X}\left[\Phi(X)\Phi(X)^\top\right]$ the $d\times d$ covariance matrix. Let us describe the optimal predictor θ_*. It satisfies first order optimality condition: $\nabla_\theta\mathcal{R}(\theta_*) = 0$, i.e., $\mathbb{E}\left[\langle\theta_*,\Phi(X)\rangle - Y)\Phi(X)\right] = 0$, which is equivalent to:

$$\Sigma\theta_* = \mathbb{E}(Y\Phi(X)).$$

This means that if Σ is invertible there exists a unique minimizer $\theta_* = \Sigma^{-1}\mathbb{E}(Y\Phi(X))$. Using the optimality condition of θ_*, we can now write in a closed form the excess risk:

$$\mathcal{R}(\theta) - \mathcal{R}(\theta_*) = \langle\Sigma(\theta-\theta_*),\theta-\theta_*\rangle = \left\|\Sigma^{1/2}(\theta-\theta_*)\right\|^2$$

1.3.2. Review of the classical methods to solve the Least-squares problem

Here, we try to review the classical methods to solve this problem considering the high-dimensional setting we are in. Like since the beginning of the thesis we will draw a particular attention to the possible limitations in terms of computation. Keep in mind as Bottou and Bousquet recalled that the best solutions for the empirical risk minimization problem are not the one we shall prefer in ML.

Ordinary least-squares and ridge regression estimators. We denote Φ the $n\times d$ matrix of features whose i-th row is the feature vector $\Phi(x_i)^\top$, for $i\leqslant n$ and Y the vector of $\mathbb{R}^d$ containing all the y_i. The empirical risk writes:

$$\widehat{\mathcal{R}}_n(\theta) = \frac{1}{2n}\sum_{i=1}^n \left(\langle\theta,\Phi(x_i)\rangle - y_i\right)^2 = \frac{1}{2n}\left\|\Phi\,\theta - \mathsf{Y}\right\|^2,$$

and the *ordinary least-squares* (OLS) estimator is $\theta_{\mathsf{ols}} = \left(\Phi^\top\Phi\right)^{-1}\Phi^\top\mathsf{Y}$. Note that this estimator can be computed if and only if the $n\times n$ gram matrix $\Phi^\top\Phi$ is invertible. In other cases it is always possible to add some regularization term λ that makes the matrix invertible: this is called the *ridge regression estimator* (RR). The OLS estimator can be seen as the RR estimator when the regularization goes to 0. In the sequel,

we assume that the gram matrix is invertible to avoid a deep (and certainly rich!) discussion on ridge regression assuming that every thing behave the same in this case.

To go further, we can simplify the model without losing intuition on it by considering that the features $(\phi(x_i))_i$ are deterministic: this is called the *fixed design setting*. In this case, the calculations are straightforward: $\theta_* = \left(\Phi^\top \Phi\right)^{-1} \Phi^\top \mathbb{E}\left[\mathsf{Y}\right]$, $\Sigma = \frac{1}{n}\Phi^\top\Phi$ and denoting $\varepsilon = \mathsf{Y} - \mathbb{E}\left[\mathsf{Y}\right]$ and considering that $\Phi(\Phi^\top\Phi)^{-1}\Phi^\top$ is a projector, we have

$$\begin{aligned}\mathcal{R}(\theta_{\mathsf{ols}}) - \mathcal{R}(\theta_*) = \frac{1}{n}\left\langle \Sigma\left(\Phi^\top\Phi\right)^{-1}\Phi^\top\varepsilon, \left(\Phi^\top\Phi\right)^{-1}\Phi^\top\varepsilon\right\rangle = \frac{1}{n}\left\langle \Phi\left(\Phi^\top\Phi\right)^{-1}\Phi^\top\varepsilon, \varepsilon\right\rangle \\ = \frac{1}{n}\mathrm{tr}\left(\Phi\left(\Phi^\top\Phi\right)^{-1}\Phi^\top\varepsilon\varepsilon^\top\right).\end{aligned}$$

Let suppose a isotropic noise assumption $\mathbb{E}[\varepsilon\varepsilon^\top] = \sigma^2 I$, then $\mathbb{E}\,\mathcal{R}(\theta_{\mathsf{ols}}) - \mathcal{R}(\theta_*) = \frac{\sigma^2\mathrm{rank}(\Phi)}{n}$, and more generally, for uniformly bounded covariance noise, i.e., $\mathbb{E}[\varepsilon\varepsilon^\top] \preccurlyeq \sigma^2 I$, $\mathbb{E}\mathcal{R}(\theta_{\mathsf{ols}}) - \mathcal{R}(\theta_*) \leqslant \frac{\sigma^2 d}{n}$.

The last paragraph was to show the error bound $O(\frac{\sigma^2 d}{n})$ for the OLS estimator in the fixed design setting. Even in the random design setting where the $(\phi(x_i))_i$ are no longer deterministic, such a bound (involving more calculations) is still valid and is in fact optimal for this problem!. We refer to [LM+16, VDVW96] for more details. However, note that this closed form estimator requires large d, n matrix multiplications and a $n \times n$ matrix inversion which can be prohibitive for large scale problems. We will see in the two next paragraph that gradient descent and stochastic gradient descent perform similarly but have the advantage to do it at lower cost.

Gradient descent. First we can write the gradient descent iterations to solve the empirical risk minimization problem. For $t > 0$, we simply derive the empirical risk with respect to θ,

$$\theta_t = \theta_{t-1} - \frac{\gamma_t}{n}\sum_{i=1}^n \left(\langle\theta_{t-1}, \Phi(x_i)\rangle - y_i\right)\Phi(x_i).$$

This recursion corresponds to gradient descent for a strongly convex function. The main problem is that the strongly convex parameter controlling the exponential convergence is the smallest eigenvalue of the empirical covariance matrix $\widehat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^n \phi(x_i)\phi(x_i)^\top$, and in the large scale setting, it could be extremely small. In other words, as the matrix is badly conditioned, the exponential convergence can be arbitrarily slow. To bypass this, a classic idea could be to regularize by some parameter λ: in this case the function is λ-strongly convex, but when applying gradient descent results with $\lambda \sim 1/n$, these approaches fail to give exploitable bounds. On the contrary, Yao and co-authors have shown [YRC07] by early stopping that the gradient descent performs optimally on the true risk at rate $O(\frac{\sigma^2 d}{n})$.

As said earlier, the complexity per iteration of gradient descent is overpassed by the one of stochastic gradient descent, which is at the core of this thesis.

Stochastic gradient descent. We recall here for the subsection to be self-contained what are stochastic gradient descent iterations in this setting. Note however that this algorithm will be explained at length in the next section and throughout Part II. SGD follows the same principle that GD but selects only one sample to optimize instead of computing the whole sum over the dataset (that can be costly). For $t > 0$, the iterations read:

$$\theta_t = \theta_{t-1} - \gamma_t\left(\left\langle\theta_{t-1}, \Phi(x_{i(t)})\right\rangle - y_{i(t)}\right)\Phi(x_{i(t)}),$$

where $i(t) \sim \mathcal{U}\left(\{1,\dots,n\}\right)$ selects uniformly at random an input/output pair in the dataset. Note that there are other types of sampling as the cycle sampling often used in practice that has the rule $i(t) = t \mod n$. Even if the complexity per iteration is very low, SGD achieves the same $O(\frac{\sigma^2 d}{n})$ error bounds after n step [BM11].

Minimax rates for least-squares. Lower bounds in the case of uniformly bounded covariance of the noise have been derived in the least-squares setting. In these works people have shown that the rate $\frac{\sigma^2 d}{n}$ found for our three ways to solve Empirical risk minimization is in fact optimal! We will see other minimax rates for non-parametric settings further in this thesis in Part II, Section 2.

⋆

⋆ ⋆

Introducing why should we use SGD. When it comes to inverting a matrix standard SVD or QR are very efficient for medium-scaled problems $d \leqslant 10^4$. However, in large-scaled settings like in ML, gradient based algorithm are often preferred instead: indeed, solving $Ax = b$ is equivalent to solving the optimization problem : $\min \|Ax - b\|^2$. To solve this problem the conjugate gradient algorithm is one of the most efficient algorithm as it is defined as the best momentum algorithm built under gradient descent. Furthermore, in this setting, conjugate gradient descent has the same complexity as gradient descent. How to be better than such an algorithm ? For deterministic first-order method it is impossible, but why not try stochastic versions ? This is what Strohmer and Vershynin have considered by designing the randomized Kaczmarz algorithm [SV09], which is simply a version of importance sampling SGD for this problem. Hence, the important question: which of conjugate gradient descent and randomized Kaczmarz algorithm is better ? The answer is that it depends on the problem at stake, and I cannot explain this better that in the seminal paper:

"It is known that the CG method may converge faster when the singular values of A are clustered. For instance, take a matrix whose singular values, all but one, are equal to one, while the remaining singular value is very small, say 10^8. While this matrix is far from being well-conditioned, CGLS will nevertheless converge in only two iterations, due to the clustering of the spectrum of A. In comparison, the proposed Kaczmarz method will converge extremely slowly in this example by Theorem 3, since $\kappa(A) \sim 10^8$ [κ is the condition number of A]. On the other hand, [the randomized Kaczmarz algorithm] can outperform CG on problems for which CG is actually quite well suited, in particular for random Gaussian matrices A..."

In a word, randomized gradient descent techniques can be very powerful to solve large and possibly random optimization problems. They are at the core of this thesis and introduced in the next section.

2. Stochastic gradient descent

In the previous section we have tried to motivate why Stochastic gradient descent is nowadays the workhorse of every large-scale ML problems. However stochastic approximations have an older history than ML. It has been studied at first by Robbins and Monroe in [RM51] to find the roots of a function that we only have noisy measurements from. Stochastic gradient descent, in its most general definition, is simply the application of the Robbins and Monroe's procedure the roots a the *gradient* of a function f. First in Section 2.1, we will describe the general setting of SGD and see that it can be exploited to analyze many used algorithms. Then in Section 2.2 we will talk about two different points of view that help getting intuition about the dynamics of SGD: see it as a Markov chain and consider a diffusion associated to it. Finally, in Section 2.3, we will shortly review the known convergence results of SGD in different settings.

2.1. Setting

2.1.1. Stochastic approximation

We have already written the stochastic gradient descent used to minimize the empirical risk. However, let us see how SGD is defined in a more general setting. At each time $t \in \mathbb{N}$ of the procedure, let us suppose that we only have access to an unbiased estimate of the gradient, ∇f_t, of the function f we want to minimize (it is sometimes called a first-order *oracle*). More formally the unbiased esimate means that for a filtration $(\mathcal{F}_t)_{t\in\mathbb{N}}$ such that θ_t is $\mathcal{F}_t$-measurable: $\mathbb{E}\left[\nabla f_t(\theta_{t-1})|\mathcal{F}_{t-1}\right] = \nabla f(\theta_{t-1})$. Then, the SGD iterates with step-size $(\gamma_t)_{t\in\mathbb{N}}$, and initialized at $\theta_{t=0} = \theta_0$, writes

$$\theta_t = \theta_{t-1} - \gamma_t \nabla f_t(\theta_{t-1}). \tag{24}$$

To put the emphasis on the noise induced by the noisy estimates of the true gradient, we prefer sometimes to rephrase the recursion (24) in term of the zero-mean noise sequence $\varepsilon_t = \nabla f - \nabla f_t$.

$$\theta_t = \theta_{t-1} - \gamma_t \nabla f(\theta_{t-1}) + \gamma_t \varepsilon_t(\theta_{t-1}).$$

Note that $\eta_t := \mathbb{E}\theta_t$ verifies the classical deterministic gradient descent recursion and hence under mild assumptions, η_t converges to the minimum argument of f (as described in section 1.2.3). However, handling the variance of the recursion will necessitate two ingredients: (i) some assumptions on the noise, typically of bounded variance: $\mathbb{E}\left[\|\varepsilon\|^2|\mathcal{F}_{t-1}\right] \leqslant \sigma^2$ (ii) assumption on the step size as we will see later in subsection 2.3.

2.1.2. The versatility of Stochastic gradient descent

The general recursion stated in Eq. (24) can be applied to many settings, but as we have already seen, it fits particularly well in the supervised learning framework. We then briefly introduce other SGD-type procedures particularly useful in the high-dimensional regime.

Supervised learning reformulation. We have already seen that SGD can be seen as taking only one element in the sum in Empirical risk minimization. However, one of the real power of SGD, as described above, is that it can be seen as a direct gradient method to optimize the true risk [BCN18]. Indeed, recall that the true risk is $\mathcal{R}(\theta) = \mathbb{E}_\rho\, \ell((X,Y),\theta)$. Now consider a input/output sample pair (x_i, y_i) drawn from ρ. Now, $\ell((x_i,y_i),\theta)$ is an unbiased estimate of the true risk $\mathcal{R}(\theta)$ such that $\nabla_\theta \ell((x_i,y_i),\theta)$ is an unbiased estimate of the true gradient of the risk. Hence, if we denote $\mathcal{F}_t = \sigma((x_i,y_i),\ i \leqslant t)$, then the stochastic gradient descent optimizes the true risk $\mathcal{R}(\theta)$ as long as new points (x,y) are added in the data set.

$$\theta_t = \theta_{t-1} - \gamma_t \nabla_\theta \ell((x_t,y_t),\theta_{t-1}),$$

and θ_t is $\mathcal{F}_t$-measurable. This reveals the real strength of SGD against other type of gradient descent algorithm: as long as we consider $t < n$ iterations, SGD optimize *directly* the true risk although it is an *a priori* unknown function. As a consequence, the SGD algorithm when $t < n$ cannot overfit the dataset and does not need any regularization.

Finite sum. As, we have already seen SGD can be seen as a stochastic optimization method that optimizes the empirical risk. In this case the function to minimize is $\widehat{\mathcal{R}}_n(\theta) = \mathbb{E}_{(X,Y)\sim\widehat{\rho}_n}\left[\ell((X,Y),\theta)\right] = \frac{1}{n}\sum_{i=1}^n \ell((x_i,y_i),\theta)$, where $\widehat{\rho}_n$ is the empirical measure associated to the samples $\widehat{\rho}_n = \frac{1}{n}\sum_{i=1}^n \delta_{(x_i,y_i)}$. Once again, we can derive an unbiased estimate $\nabla_\theta \ell((x_{i(t)},y_{i(t)}),\theta)$ of the true gradient of the empirical risk where $(i(t))_t$ is the sequence of uniformly sampled indices over $\{1,\dots,n\}$. For this, we define the adapted filtration: $\mathcal{F}_t = \sigma((x_k,y_k)_{k\leqslant n},\ i(l)_{l\leqslant t})$. The recursion reads:

$$\theta_t = \theta_{t-1} - \gamma_t \nabla_\theta \ell((x_{i(t)},y_{i(t)}),\theta_{t-1}).$$

Note also that for this problem, a sequence of works, SAG [RSB12], SVRG [JZ13], SAGA [DBLJ14], have shown explicit exponential convergence. However, these results, once applied to ML say nothing about the loss on unseen data.

Randomized coordinate descent. Coordinate Descent (CD) [Wri15] is a popular algorithm based on picking according to cycles one by one each coordinate of the gradient. Even if as standard SGD it reduces the cost of computing the gradient in all the directions, some of them might be not useful to follow and could represent a waste of time. This is why some randomized strategies (with possible importance sampling techniques to accelerate them) have complemented the study of Coordinate Descent. Randomized CD [Wri15, RT14] is another example of Stochastic gradient descent. Indeed, let us minimize the function f over $\mathbb{R}^d$. Let $\mathcal{F}_t = \sigma(i(l)_{l\leqslant t})$ be the adapted filtration of the problem where $(i(t))_t$ is the sequence of uniformly sampled indices over $\{1,\dots,d\}$ (we could replace the uniform law with other laws to resort to importance sampling techniques), the iterations write:

$$\theta_t^{i(t)} = \theta_{t-1}^{i(t)} - \gamma_t \partial_{i(t)} f(\theta_{t-1}),$$

where $\forall i \leqslant d,\ \theta^i$ is the i-th coordinate of the vector θ.

Randomized Kaczmarz algorithm. We already introduced earlier in this thesis the Randomized Kaczmarz as being exactly a stochastic gradient method to solve largely overparametrized linear systems. Quite remarkably, the Randomized Kaczmarz algorithm [SV09] can be seen as the dual of randomized coordinate descent with proper importance sampling [Wri15, Section 1.4]. Let us simply recall it in two lines with natural ML notations. To solve the system $X\theta = y$, where X is a $n \times d$ matrix, $\theta \in \mathbb{R}^d$ and $y \in \mathbb{R}^n$, the idea is to solve the optimization problem: $\min_\theta \|X\theta - y\|^2$ with stochastic gradient descent. Let $(x_i)_i$ be the n row vectors containing X, the iterations write

$$\theta_t = \theta_{t-1} - \frac{1}{\|x_{i(t)}\|^2}\left(\langle x_{i(t)},\theta_{i(t)}\rangle - y_{i(t)}\right) x_{i(t)},$$

where $i(t) = k \in \{1,\dots,n\}$ with probability $\frac{\|x_k\|^2}{\|A\|^2}$. The two main differences with the other algorithms above are that the step-size $\gamma_t = \frac{1}{\|x_{i(t)}\|^2}$ and the law of $i(t)$ are chosen according to the problem to perform optimally.

⋆

⋆ ⋆

A systematic analysis of random algorithms. All of these stochastic algorithms come from the same SGD framework described at the beginning of this section. Hence, whether we are studying SGD to minimize the true or empirical risk, whether we use randomized coordinate descent or Karczmarz algorithm, the same general results will apply. The only difference between them is the way that we get the unbiased estimate: this can be seen mathematically in the filtration used to define the noise in SGD. And this could lead a systematic statistical study of SGD. Indeed, at first order all of these algorithm are the same as their first moment are the same: this is the SGD definition,

$$\mathbb{E}\left[\nabla \mathsf{F}_t(\theta_{t-1})|\mathcal{F}_{t-1}\right] = \nabla f(\theta_{t-1}),$$

where we have put a particular emphasis on the fact that F is a random variable. Note that what may control the speed of convergence in expectation at this point is hence the Hessian of the problem: for example, it would be in the case of least-squares for the true risk $\mathbb{E}[XX^\top]$ and $n^{-1}\sum_{i=1}^n x_i x_i^\top$ for the finite sum problem. As they share the same expectation, properties like rates, possible acceleration can be the same. But to go deeper and study their difference, in a very first-principled way of thinking of a random problem, we can look at the second moment of the random variable:

$$\mathbb{E}\left[\nabla \mathsf{F}_t(\theta_{t-1})\nabla \mathsf{F}_t(\theta_{t-1})\top|\mathcal{F}_{t-1}\right].$$

This is where the multiplicative noise of SGD come from and may differ from one setting to another.

2.2. SGD as a Markov chain and continuous time limit

In this section, let us go back to the general stochastic gradient descent algorithm. In mathematics it is always useful to understand and cast problems into known and developed theory. Indeed, when it comes to apprehend abstract objects, mental images and representations are often the key to really understand them. We try to develop in this section two of these intuitions on SGD by giving a Markov chain interpretation of it in subsection 2.2.1 and providing a high-level comprehension of how we can model SGD by a continuous time diffusion in subsection 2.2.2.

2.2.1. Markov chain interpretation of SGD

The first key to understand the behavior of SGD is that when the step-size is constant, i.e. $\gamma_t = \gamma$, the iterates define an homogeneous Markov chain [MT12]. The case where the step-size is constant is widely used in practice as it allows to forget initial conditions rapidly. Let us recall the SGD recursion to keep it next to us.

$$\theta_t = \theta_{t-1} - \gamma \nabla f(\theta_{t-1}) + \gamma \varepsilon_t(\theta_{t-1}).$$

When γ is constant, we see that SGD is a time-homogeneous Markov chain as the distribution of θ_t depends only on the one of the previous iterate θ_{t-1}. What can we deduce from this point of view ? The first thing is that under mild technical conditions, the distribution of the iterates $(\theta_t)_t$ converges to an invariant distribution that depends on γ and we call π_γ. From this fact, we can initiate two important reflections.

(i) Closeness to θ_*. Note that our aim is to two show that $(\theta_t)_t$ is close to θ_*. Rephrased in the Markov chain language, this is to show that the distribution of θ_t, call it $\pi_\gamma(t)$ is not far from the target distribution δ_{θ_*}. However, we know that $\pi_\gamma(t)$ converges to π_γ. Hence the question:

How far π_γ is from δ_{θ_*} ?

Giving an answer to this question is not trivial, and more importantly, it depends heavily on the noise $\varepsilon(\theta)$. We will try to give some intuition about this question in the ML setting in the next subsection. For now, let us stick with some common modeling to study the recursion: make the assumption that the noise does not depend on the current iterate θ. In this case, it has been shown that the iterates of SGD oscillates around θ_* on a ball of radius that scales like $\gamma^{1/2}$ [Pfl86]. From this point of view, the smaller the gamma, the closer to θ_* we get.

(ii) The need to average. In this case, ergodic theorems for Markov chains [MT12, Section 13] give us an important insight: they often give a way for a time-mean to converge to a deterministic quantities. As a matter of fact, under mild assumption, we can effectively show that the time-mean $\bar{\theta}_t = \frac{1}{t+1}\sum_{i=0}^{t}\theta_i$ converges to $\bar{\theta}_\gamma = \mathbb{E}_{\pi_\gamma}[\theta]$. Furthermore, some magic happens with the quadratic case, as we can show that $\bar{\theta}_\gamma = \theta_*$: this justify the fact that throughout Part II, we will consider the averaged SGD estimator as we have justified that almost surely $\bar{\theta}_t \to \theta_*$. However, note that Markov chains are not the point of view of our methods as we want to derive non-asymptotic rates of convergence. We will prefer direct calculations instead. Note also that the same analysis can be done in the non-quadratic case, indeed, in [DDB17] the authors showed that the order of magnitude of the distance between $\bar{\theta}_\gamma$ and θ_* is of order γ^2. This discussion is illustrated in Figure 4 taken from the well written thesis of A. Dieuleuveut [Die17].

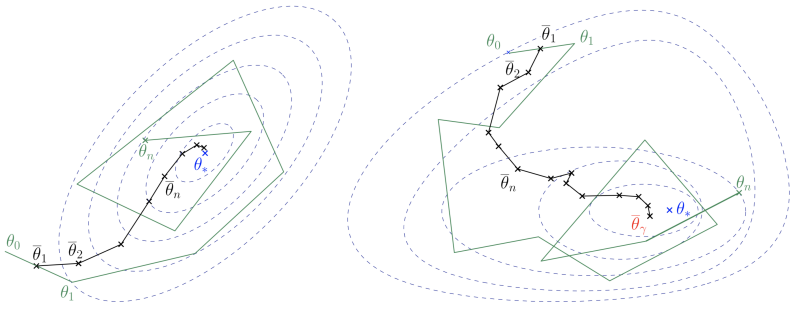

Figure 4: Stochastic Gradient Descent with constant learning rate. Dashed lines are the level lines of the objective function f, green points correspond to the main recursion, and black to the averaged one. (Left) Quadratic case, the limit is the optimal point θ_*. (Right) General case, the limit is a different point $\bar{\theta}_\gamma$ at distance γ^2 form θ_*.

2.2.2. Continuous time limit of SGD

Continuous time counterparts of (distrete) numerical optimization methods and Markov chain are well-worn subject in applied mathematics.

Gradient flow. A simple example is the gradient-flow associated to the gradient descent method. Indeed, one can see the gradient descent method as a discretized approximation (called Euler scheme) with time step $\Delta T = \gamma$. More precisely, if we define Θ some function of time t such that it corresponds to gradient descent at each multiple of time γ: $\theta_n = \Theta(n\gamma)$ and $\theta_n = \theta_{n-1} - \gamma\nabla f(\theta_{n-1})$, then for $t = n\gamma$, $\frac{\Theta(t+\gamma)-\Theta(t)}{\gamma} = \frac{\theta_{n+1}-\theta_n}{\gamma} = -\nabla f(\Theta(t))$ such that as $\gamma \to 0$

$$\frac{d}{dt}\Theta(t) = -\nabla f(\Theta(t)). \tag{25}$$

The analysis of the gradient flow (25) is already rich to understand the gradient descent. The fact that there are no step-sizes to take care of, the use of differential calculus and the centuries of applied mathematics studying such models make the analysis often more straightforward and always give powerful insights on its discretized counterpart.

Modelling stochastic gradient descent. Let us try to do the same with SGD. First recall the SGD iterations with fixed step-size and i.i.d. noise. Let us suppose that the noise can be put under the form $\varepsilon_t(\theta_{t-1}) = \sigma(\theta_{t-1})^{1/2}\mathcal{G}_t$, where $\mathcal{G}_t$ is a standard Gaussian $\mathcal{N}(0, I_d)$ and $\sigma(\theta_{t-1})$ encodes the covariance of the noise. It writes,

$$\theta_t = \theta_{t-1} - \gamma\nabla f(\theta_{t-1}) + \gamma\sigma(\theta_{t-1})^{1/2}\mathcal{G}_t. \tag{26}$$

Now, a little bit of knowledge of Itô calculus and SDE discretization [KP13] will show that the noise should be of scale $\gamma^{1/2}$ to model well a diffusion. This shows two things: (i) that the first order approximation of SGD is not a diffusion but the gradient flow itself (25) since the noise in SGD is too large; (ii) if we want a second order approximation of Eq.(26), we should include $\gamma^{1/2}$ in the covariance matrix $\sigma(\theta)$. The continuous time diffusion modelling SGD is thus:

$$d\Theta_t = -\nabla f(\Theta_t)dt + \sqrt{\gamma}\sigma(\Theta_t)^{1/2}dB_t, \tag{27}$$

where B_t is a standard Brownian motion. Note the presence of $\gamma^{1/2}$ in the covariance of the noise of the diffusion (27). In fact, in [LT19] it is shown that the SGD sequence Eq. (26) is a first order approximation with time-step γ of the SDE (27). Let us emphasize that two independent quantities are the same here: it is of crucial importance that *both the time-step and the noise* in (27) depend on γ. Even if the computation is not very complicated, this continuous time version of SGD has not been largely studied. It could be because diffusion experts that could enlighten us are far away from the ML community. It could also be because the continuous time model is too difficult to study or too far from helping us building solid intuition on SGD behavior. However I really wanted to write a paragraph discussing it because it seems to me that this could be a interesting open line of research.

A toy continuous time example. To show how this continuous model can give insights about what happen during SGD, let us illustrate this with a slight improvement of a nice example drawn from [ADT20]. Consider a very simple least-squares responseless 1-d setting where we want to solve the minimization problem:

$$\min_{\theta\in\mathbb{R}} \frac{1}{2}\mathbb{E}_\rho(X\cdot\theta)^2,$$

Note that, in this case, obviously, $\theta_* = 0$ and there is no noise at optimum. We say that the noise is purely multiplicative. Assume that the moment of order 4 of X exists and denote $V = \mathbb{E}[X^2]$ and $V_2 = \mathrm{Var}[X^2]^{1/2}$. For samples $x_1, \dots, x_n$ distributed according to ρ, the SGD recursion writes

$$\theta_n = \theta_{n-1} - \gamma x_n^2\theta_{n-1} = \prod_{i=1}^n \left(1-\gamma x_i^2\right)\theta_0 \leqslant \prod_{i=1}^n \exp\left(-\gamma x_i^2\right)\theta_0 = \exp\left(-\gamma\sum_{i=1}^n x_i^2\right)\theta_0 \sim e^{-\gamma V n}\theta_0,$$

where we used that for large n, at the first order, $\sum_{i=1}^n x_i^2 \sim Vn$, by the law of large numbers. Obviously, the sequence of estimators $(\theta_n)_n$ converges to 0 exponentially fast at rate γV (at least in expectation from which the approximation is exact). Now, let us see what gives the continuous time SGD (27). We have $\varepsilon_t(\theta) = \nabla f - \nabla f_t = (V - x_t^2)\theta$ whose square root of (co)variance (matrix) is $\sigma(\Theta_t)^{1/2} = V_2\theta$. Following (27), the SGD-SDE writes:

$$d\Theta_t = -V\Theta_t dt + \sqrt{\gamma}V_2\Theta_t dB_t.$$

The acute reader will certainly recognize a geometric Brownian motion for which a close form solution can be written along the whole trajectory. From this, we can infer the expectation $\mathbb{E}[\Theta_t] = \Theta_0 e^{-Vt}$ and the variance $\mathrm{Var}[\Theta_t] = \Theta_0^2 e^{-2Vt}(e^{\gamma V_2^2 t} - 1)$. Two conclusions can be drawn from this modelling:

(i) Convergence in expectation. With the equivalence $t = \gamma n$ we find the exact same rate of convergence e^{-Vt} for the two models.

(ii) Fluctuations. Remarkably, we can go further: if $\gamma \leqslant \frac{2V}{V_2^2}$, the variance of the continuous model will go to zero exponentially fast. And with a precise look of the fluctuations in the SGD iterates, we can certainly write an expression similar to $\mathrm{Var}[\Theta_t]$ showing that at the fluctuation level the two system match! Note the important fact that in this case, the law of SGD converges to a degenerate distribution: $\delta_{\theta_*=0}$ and does not oscillate as predicted with the Markov chain model.

Note also here that the fact that the noise depends on Θ (through its covariance) is absolutely fundamental to see the convergence to θ_*. Indeed, if we model the noise without this θ-dependence (additive noise perspective), one gets the following diffusion:

$$d\Theta_t = -V\Theta_t dt + \sqrt{\gamma} V_2 dB_t,$$

which is the well-known Ornstein-Uhlenbeck process: it converges to a Gaussian law of mean 0 and of variance $\gamma \frac{e^{-V} V_2^2}{2V}$. Hence Θ_t will oscillate around zero on a typical ball of radius $\sqrt{\gamma}$ like described in the precedent paragraph when we modeled SGD by a Markov chain! Note also that the quantity $\frac{\gamma V_2^2}{2V}$ seems to govern, in any case, the behavior of the algorithm.

⋆

⋆ ⋆

Modelling SGD: a noise issue. To conclude, we hope that we convinced the reader that the SDE model is a very good approximation to SGD recursion and that for ML settings, the noise in SGD can depend strongly on the iterates θ. This last fact changes dramatically the behavior of the algorithm as illustrated in Figure 5.

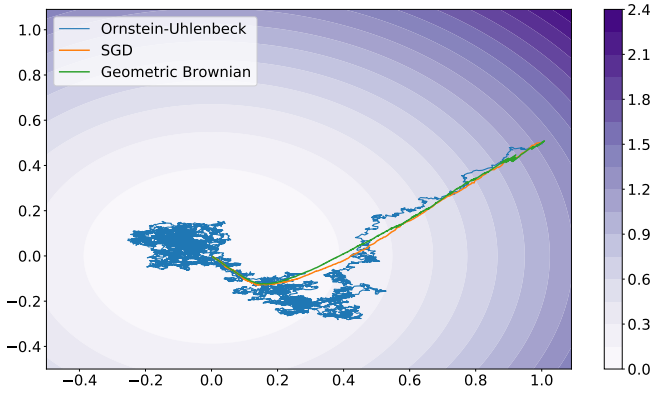

Figure 5: Comparison of the different trajectories for the SGD diffusion models. In this case ρ is the uniform distribution over the square $[0, 1]^2$. The color map shows the value of the function to optimize.

2.3. Analysis of SGD in the ML setting

2.3.1. Properties of SGD in the ML setting

Before stating the known results for the convergence of Stochastic Gradient Descent in different settings let us briefly explore further certain properties of SGD in the ML setting. First, recall that the general SGD recursion can be written under the following form

$$\theta_t = \theta_{t-1} - \gamma(\nabla f(\theta_{t-1}) + \varepsilon_t(\theta_{t-1})),$$

where $\varepsilon_t(\theta_{t-1})$ is a sequence of noise.

Decomposition of the noise. The noise can be decomposed as the sum of two very different noises, that will have two different effects on the dynamics. More precisely, we can write

$$\varepsilon_t(\theta_{t-1}) = \underbrace{\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_*)}_{\text{Multiplicative noise}} + \underbrace{\varepsilon_t(\theta_*)}_{\text{Additive noise}} .$$

As illustrated in the previous section, one expect the multiplicative noise to eventually shrink to 0 (see the geometric Brownian motion in Figure 5) as soon as it is smooth enough (say Lipschitz), whereas the additive noise, totally independent of the iterates, is a residual noise that will eventually make the iterates turn around the optimum θ_* (see the Ornstein-Ulhenbeck process in Figure 5). Let us calculate it explicitly in the quadratic case:

$$\begin{aligned}
\varepsilon_t(\theta_{t-1}) &= \nabla_t f(\theta_{t-1}) - \nabla f(\theta_{t-1}) \\
&= (\langle \phi(x_t), \theta_{t-1}\rangle - y_t)\phi(x_t) - \mathbb{E}\left[(\langle \phi(X), \theta_{t-1}\rangle - Y)\phi(X)\right] \\
&= \underbrace{\left(\langle \phi(x_t)\phi(x_t)^\top - \mathbb{E}\left[\phi(X)\phi(X)^\top\right]\right)(\theta_{t-1} - \theta_*)}_{\text{Multiplicative noise}} + \underbrace{\left(\langle \phi(x_t), \theta_*\rangle - y_k\right)\phi(x_t)}_{\text{Additive noise}} .
\end{aligned}$$

The two noises here have a second modelling difference: the multiplicative noise quantifies how far the covariance matrix is from picking only one sample of the inputs, whereas the additive noise can model the amount of noise we have when collecting the input/output pairs.

Smoothness and strong convexity. If we want to apply directly the results of convex optimization, we have to know what are the main properties of our test risk $\mathcal{R}$ or empirical risk $\widehat{\mathcal{R}}_n$. As they can both be expressed as an expectation of the same function, with distribution ρ and $\widehat{\rho}_n$ respectively, the properties detailed below will stand for both functions. In this paragraph, for the sake of clarity, let us assume a parametric model for our risk : $\mathcal{R}(\theta) = \mathbb{E}_\rho\left[\ell(\langle \theta, \phi(X)\rangle, Y)\right]$.

(i) Convexity. Given that ℓ is convex is θ, then by integration $\mathcal{R}$ is convex.

(ii) Strong Convexity and smoothness. $\mathcal{R}$ will be as differentiable as ℓ is and

$$\nabla^2\mathcal{R}(\theta) = \mathbb{E}_\rho\left[\nabla^2_\theta \ell(\langle \theta, \phi(X)\rangle, Y)\phi(X)\phi(X)^\top\right].$$

Hence, L-smoothness and μ-strong convexity of ℓ will extend to $\mathcal{R}$ through the covariance matrix:

$$\mu\lambda_{\min}(\Sigma)\,\mathrm{Id} \preccurlyeq \mu\Sigma \preccurlyeq \nabla^2\mathcal{R}(\theta) \preccurlyeq L\Sigma \preccurlyeq L\lambda_{\max}(\Sigma)\,\mathrm{Id}$$

However, even if it is quite reasonable to assume that $\lambda_{\max}(\Sigma)$ is not to big, $\lambda_{\min}(\Sigma)$ can be arbitrarily small in high dimension as $\lambda_{\min}(\Sigma) \leqslant \frac{\mathrm{Tr}\Sigma}{d} = \frac{\mathbb{E}\left[\|\phi(X)\|^2\right]}{d}$. This is one of the main reasons of regularizing the problem by a parameter λ.

2.3.2. Convergence of Stochastic Gradient Descent

General behavior. The convergence of stochastic gradient descent can almost always be decomposed in two parts:

(i) The bias term. It corresponds to the forgetting of the initial conditions (this is really the gradient descent part on the noiseless model). For this to append is common settings, the minimal assumption on the step size sequence is that $\sum_{t\geqslant 1}\gamma_t = \infty$. Note that it covers the constant step size and all step sizes going slower that $1/t$ at infinity. Note also that the bigger the step-size the fastest the convergence to zero.

(ii) The variance term. It corresponds to the robustness to the two types of noise seen in the previous subsection. We have already seen that constant step-sizes is not a good idea to unsure convergence as it could lead for general noises to iterates turning around the optimum on a scale $\gamma^{1/2}$. More precisely, in this case, handling the noise needs that the step-size sequence goes strictly faster to infinity than $1/\sqrt{n}$, as one needs: $\sum_{t\geqslant 1}\gamma_t^2 < \infty$. Note also that, as already stated above, handling the variance term can be made by averaging SGD and consider the average estimate [PJ92]: $\bar{\theta}_t = \frac{1}{t+1}\sum_{i=0}^{t}\theta_i$.

The need of Lyapounov function. Convergence of θ_t to θ_* is difficult to prove, instead, as it is classical in the optimization community, we leverage the knowledge of a Lyapounov function that ends up decreasing to zero and helps to quantify the convergence of SGD. In ML, the Lyapounov function is almost always the true excess risk: $\mathcal{R}(\theta_t) - \mathcal{R}(\theta_*)$. And under bounded noise assumption: $\mathbb{E}[\|\varepsilon_t\|^2|\mathcal{F}_{t-1}] \leqslant \sigma^2$ and with step-sizes as described above, we can show almost sure convergence of $\mathcal{R}(\theta_t)$ to $\mathcal{R}(\theta_*)$ [RM51]. As we are concerned with non-asymptotic results, let us review some of these.

Rates of convergence. Even if the difference is less striking than in deterministic optimization, the rates are very different for convex and strongly-convex functions. Note also that whether we average or not will lead to two different choices of step sizes: constant ones if we average and decreasing if we consider the final iterate. Finally, because of the variance term related to the noise, convergence for the last iterate was more difficult to show than convergence of the averages despite being more used in practice.

It has been proven in [NY83] that the *optimal rate* of convergence for SGD in the convex case is $O\left(1/\sqrt{t}\right)$, whereas in the μ-strongly convex case the optimal value is $O\left(1/\mu t\right)$.

- ⋆ **Last iterate**. When f in only convex, [Sha11] proved that with step-sizes of order $1/\sqrt{t}$, the expected excess risk was of order $O\left(\frac{\log t}{\sqrt{t}}\right)$, which is near the optimal rate in the convex case. Similarly, for strongly convex functions, [Sha11] proved that with step-sizes of order $1/\mu t$, the expected excess risk was of order $O\left(\frac{\log t}{\mu t}\right)$. With non-practical step sizes [JNN19] show that they can remove the $\log t$ to achieve optimal rates for the last iterate of SGD in both cases.
- ⋆ **Averaging**. As shown in the previous section, averaging techniques enable to take larger step-sizes and even constant step-sizes in the case of a quadratic cost. To show optimality of the convergence rates, non-uniform averaging or tail-averaging have been proposed, but simple averages match almost the same bounds and add only logarithmic terms. In [LJSB12, RSS12], it is proven that respectively for $1/\sqrt{t}$ and $1/(\mu t)$ step-sizes, (tail-)averaged SGD converges at optimal rates for convex and strongly-convex respectively. Remarkably, for smooth strongly convex functions, in [BM11], Bach and Moulines showed that for various step-sizes $n^{-\xi}$ with $\xi \in [1/2; 1]$, the averaged sequence (and not the excess risk as previously) converges at rate $1/\mu t$. Also, for a class of non-strongly convex functions (self-concordant) including the logistic loss, Bach showed in [Bac14] that with $1/\sqrt{t}$ step-sizes, average SGD achieves the optimal rate $1/(\mu t)$ where μ is the local strong convexity at optimum. Note finally that for least-squares, for constant step-sizes, the convergence achieves fast rates of $O(1/t)$ with no prior knowledge about the strong convexity constant [BM13]. This enables the right to go derive fast-rates for non-parametric settings where there is a priori no strong-convexity. This will be the case throughout Part II.

⋆
⋆ ⋆

Conclusion on SGD and the need of non-parametric models. We have seen in this section that SGD is a versatile algorithm and that, under different names, it is used for many problems in high-dimension. Several questions can be raised by the noise model and it is crucial matter to understand what is typically the behavior of the noise induced by SGD for ML problems. In the previous part, we have always taken examples of SGD in parametric settings so that the class of functions is pretty restrictive. In the next section we illustrate how non-parametric function spaces can be built preserving the low-cost computational aspects of ML algorithms.

3. Reproducing Kernel Hilbert Spaces

When we want to choose appropriate classes of function for supervised and unsupervised tasks, the first thinking goes to simple finite-dimensional parametric classes of functions equipped with natural scalar products. Going further to infinite dimensional spaces (for larger spaces of *test functions*), it is quite natural the require some structure to preserve the important properties needed for ML tasks. Hilbert spaces of functions are perfectly suited for these as they are complete linear spaces equipped with a dot-product: one can define the gradient for optimization, one can define projections... As in all this thesis, an important requirement is to be computationally friendly: this is why Reproducing Kernel Hilbert Spaces (RKHS) are so well adapted for numerical applied mathematics in general. Describing their use and properties is the purpose of this part. In Section 3.1 we will define RKHS, derive their main properties and show how to build such spaces, then we will see how they can be useful for many different problems (including ML) in Section 3.2. Finally in Section 3.3, we explain some of the current limitations of RKHS but also some future exciting possibilities to circumvent these limitations.

3.1. Definition - Construction - Examples

3.1.1. Definition, construction and properties of RKHS

As we will see there are many ways to define Reproducing Kernel Hilbert spaces (RKHS), from concrete computationally oriented to abstract functional analysis based ways [Aro50]. Following [SW06], we will not adopt an abstract point of view for the construction of such spaces, focusing on the intuition behind their use. We refer also to [SS00, SC08, Tsy08] for generous introductions to RKHS.

Kernels and feature maps. First, recall that *kernels* are one of the most important tools in modern applied mathematics as they occur in harmonic analysis (with integral based transformations such as the Fourier transform), partial differential equations, mathematical physics in general (to define fundamental solutions) and signal processing (to deal with convolutions). Hence, the first thing we have to define is the kernel itself which is in all its generality a function over $\mathcal{X} \times \mathcal{X}$:

$$K \ : \ \mathcal{X} \times \mathcal{X} \to \mathbb{R}. \tag{28}$$

The kernel is the building block of RKHS and will be often used in ML as a measure of similarity between two inputs $x, x' \in \mathcal{X}$. In certain situation where the space of inputs has some structure, say $\mathcal{X} = \mathbb{R}^d$, one can define explicitly the kernel, as the *Gaussian kernel* [SHS06], that the reader can always have in mind throughout this part:

$$\forall x, x' \in \mathbb{R}^d, \quad K(x, x') = \exp(-\|x - x'\|^2). \tag{29}$$

However, one of the strength of kernels is that $\mathcal{X}$ does not need to have so much structure to measure correlations between its elements. One can define an application-dependent *feature map*, $\phi : \mathcal{X} \to \mathcal{F}$ with values in a feature Hilbert space $\mathcal{F}$. With this feature map ϕ, one can define the kernel:

$$\forall x, x' \in \mathbb{R}^d, \quad K(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{F}}. \tag{30}$$

Note that these examples and the intuition behind the fact that $K(x, x')$ represents a similarity between x and x' suggest the kernel to be symmetric, i.e., for all x and x' in $\mathcal{X}$, $K(x, x') = K(x', x)$. Note also that feature maps allow to apply linear technique in a structure space $\mathcal{F} = \mathrm{range}(\phi)$, while their domain is a possibly non-structured space $\mathcal{X}$ (text, graphs, images) (see Figure 6).

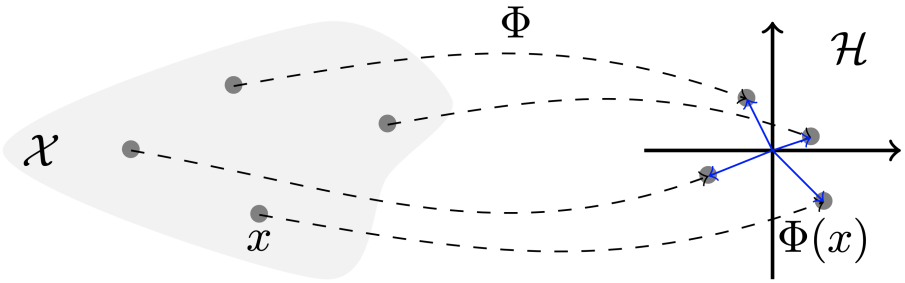

Figure 6: Representation of the feature that maps data in a non-structure space to a Hilbert space. This representation is taken from the slides of J.-P. Vert and J. Mairal lessons on Kernel methods.

Space of trial functions. Kernels can also define spaces of functions by linearly composing a canonical basis kernel functions $K_x := K(x, \cdot) = \{x' \to K(x, x')\}$ for $x \in \mathcal{X}$:

$$\mathcal{H}_0 := \text{span}\ \{K_x,\ \text{for}\ x \in \mathcal{X}\}.$$

When define from features, this gives, for $x \in \mathcal{X}$, linear combinations of $x' \to \langle \phi(x), \phi(x') \rangle_{\mathcal{F}}$. However, even if $\mathcal{H}_0$ is a convenient space of function for computations, an important property fails short: it is not *complete*. This is why we turn it to an Hilbert space via the following construction. We first define symmetric positive (semi-) definite kernels (psd):

We say that a symmetric kernel is psd if, for all finite subsets $(x_1, \ldots, x_n)$ of $\mathcal{X}$, the symmetric matrices with entries $K_{ij} := K(x_i, x_j)$ are psd.

We define also the following scalar product on $\mathcal{H}_0$:

$$\forall x, x' \in \mathcal{X},\ \ \langle K_x, K_{x'} \rangle_{\mathcal{H}_0} := K(x, x'). \tag{31}$$

This define a numerically accessible norm from which we can construct the RKHS by closing $\mathcal{H}_0$ with the norm defined in Eq.(31):

$$\mathcal{H} := \text{closure}\ (\mathcal{H}_0) = \overline{\text{span}\ \{K_x,\ \text{for}\ x \in \mathcal{X}\}}. \tag{32}$$

Note that we can show that the RKHS is uniquely defined and much larger than $\mathcal{H}_0$. It can be hard to deduce from the kernel function K the space of functions $\mathcal{H}$. For example, Sobolev spaces $\mathcal{H} = W_2^s(\mathbb{R}^d)$, with $s > d/2$ (we will come back to this $s > d/2$ after), are RKHS with kernel: $K(x, x') = \|x - y\|_2^{s-d/2} B_{s-d/2}(\|x - y\|_2)$, where B_ν is the Bessel function of third kind. Is is hard to see why this defines a psd kernel, and even harder to see that Sobolev spaces are linked with these kernels (we will come back to this later). Hence, even if the kernel alone encodes constructively the RKHS, it is somehow difficult to infer $\mathcal{H}$ from K. Finally note that, from this construction, we can always define the feature map $\phi(x) = K_x$ that satisfies $\mathcal{F} = \mathcal{H}$. However, even if this precise choice of feature map gives the right space $\mathcal{H}$, note that there can be several features maps that will lead to the same RKHS with $\mathcal{F} \neq \mathcal{H}$.

Reproduction property and abstract definitions of RKHS. As it is clear from our construction, RKHS have a nice *reproduction property*:

$$\forall f \in \mathcal{H},\ \forall x \in \mathcal{H},\ \ f(x) = \langle f, K_x \rangle_{\mathcal{H}}. \tag{33}$$

This reproduction property gives all the strength of RKHS: we can think of it as an analogous to the fact that the dirac is the reproducing function of L^2: $\langle f, \delta_x \rangle_{L^2} = f(x)$. But on the contrary of diracs that do not belong the L^2, the "dirac" of RKHS, K_x, belongs to the native space! Actually, one can define abstractly RKHS from this important property:

Let K be a symmetric psd kernel. Let $\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}}$ be a Hilbert space containing all the $(K_x)_x$ s.t. the reproducing property $f(x) = \langle f, K_x \rangle_{\mathcal{H}}$ holds. Then, $\mathcal{H}$ is the unique RKHS associated with K.

Even more abstractly, one can define RKHS without explicitly define the kernel K. In fact, with Riesz's representation theorem, the reproducing property states that the evaluation functional L_x in a continuous linear form on $\mathcal{H}$ (or equivalently bounded): $\|L_x\| \leqslant \|K_x\|_{\mathcal{H}}$. With this point of view, one can define RKHS as Hilbert spaces of functions such that for all $x \in \mathcal{X}$, L_x in continuous [Aro50]. This point of view enlightens clearly why of all Sobolev spaces, $W_l^s(\mathbb{R}^d)$, the only ones that are RKHS are necessarily: (i) $l = 2$ to have an Hilbert structure and (ii) $s > d/2$ so that the Sobolev space is injected in the space of continuous functions.

Finally, as already said, note that the reproducing property (33) gives automatically the feature map $\phi(x) = K_x$ which implies that the kernel K is psd as $K(x, x') = K_x(x') = \langle K_x, K_{x'} \rangle_{\mathcal{H}}$. So that for any collection of points $(x_i)_{i \leqslant n}$:

$$\sum_{i,j=1}^{n} a_i a_j K_{ij} = \left\| \sum_{i=1}^{n} a_i K_{x_i} \right\|^2 \geqslant 0 \tag{34}$$

3.1.2. Classical examples of RKHS

Let us begin by saying that if $\mathcal{X}$ carries some additional structure, then it may be possible to construct kernels respecting this structure by being invariant under some geometric transformations. With a slight abuse of notation, classical cases are of the form:

- *Translation-invariant kernels*: if $\mathcal{X}$ is an abelian group then we can have: $K(x, y) = K(x - y)$.
- *Zonal kernels*: they only depend on the scalar product in $\mathcal{X}$: $K(x, y) = K(\langle x, y \rangle_{\mathcal{X}})$.
- *Radial kernels*: they only depend on the norm of the difference: $K(x, y) = K(\|x - y\|)$.

Translation-invariant and radial kernels. They are a special class of kernels for which Fourier analysis can give us insights and practical tools to deal with team. In this case, as $K(x, y) = K(x - y)$, we can define the Fourier transform $\widehat{K}(\xi)$ of the kernel. The fact that K is psd is then equivalent to having a non-negative Fourier transform (which is very different to be itself non-negative as required for standard kernels in non-parametric estimation [Tsy08]). Furthermore, we can calculate explicitly the norm of $f \in \mathcal{H}$ as

$$\|f\|_{\mathcal{H}}^2 = \int \frac{f(\xi)^2}{\widehat{K}(\xi)} d\xi.$$

Note also that in this case, we see that the space $\mathcal{H}$ is composed of function that show a regularity related to K: if K is very regular, then $\widehat{K}$ will decrease fast and for f to have a bounded norm in $\mathcal{H}$ will require that it decreases fast to. Note also that the rescaling of K by a factor σ will have important impact on the norm of the functions $\|f\|_{\mathcal{H}}$ belonging to the RKHS as the norm will be multiplied by σ^d. Hence, choosing well the scale of radial kernel is a crucial task. We will come back to this fact later in this section.

A few concrete examples. When it comes to using kernels in practical settings, it is very important to use or design adapted kernels. However, to give some concrete examples here are three classical kernels.

(i) Linear and polynomial kernels. They are respectively defined as $K(x, x') = \langle x, x' \rangle_{\mathbb{R}}^d$ and $K(x, x') = \langle x, x' \rangle_{\mathbb{R}^d}^m$ for some $m \in \mathbb{N}^*$. They lead to finite dimensional RKHS.

(ii) Gaussian kernel. This is one of the most typical radial kernel: $K(x, x') = \exp(\|x - x'\|_2^2 / \sigma^2)$. Non-trivially, it can be shown that the associated space $\mathcal{H}$ is the space of analytical functions [SHS06]. As said before, the space-norm and all the properties of such a kernel depend heavily on the bandwidth parameter σ.

(iii) Laplace or exponential kernel. This is another radial kernel $K(x,x') = \exp(\|x-x'\|/\sigma)$. It looks like the Gaussian kernel but it is in fact very different as it produces a bigger and less smooth space of functions which can be consider as an equivalent of the Sobolev space $W_2^d(\mathbb{R}^d)$. Note that its Fourier transform is the psd Cauchy kernel $(1+(x-x')^2/\sigma^2)^{-1}$.

3.1.3. Constructing new kernels

Expressivity and adaptivity to the problem are keys in the difficult task of *kernel-engineering*. Understanding well how to adapt the kernel for a specified ML problem is an active research field (see for example [MKHS14]) and may be the next important task in this community. To do this, let us give some classical ways to create new kernels.

New kernels from old. There are plenty of ways of constructing new kernels from old ones but the most basics are sum, products and compositions of kernels. Indeed, when K_1 and K_2 are kernels on $\mathcal{X}$, every positive sum and product between these are kernels for which the RKHS can be described. Finally, if A is a mapping from $\mathcal{X}$ to $\mathcal{X}'$, then $K(A(x),A(x'))$ is also a kernel on $\mathcal{X}'$. These tools are the main ideas behind Multiple Kernel Learning [BLJ04] and hierarchical kernels [STS16] that we will discuss later in Section 3.3.

Kernels based on feature maps. If we construct kernels directly from feature maps as in the introduction of this part: $K(x,x') = \langle \phi(x), \phi(x')\rangle_{\mathcal{F}}$, then we know that K will automatically be a psd kernel.

(i) Mercer kernels. Let $(\psi_i)_{i\in\mathbb{N}^*}$ be a sequence of function of $\mathcal{X}$ associated to weights $(w_i)_{i\in\mathbb{N}^*}$. Take $\phi(x) := \{\psi_i(x)\}_{i\in\mathbb{N}^*}$ and assume that $\psi_i(x) \in \ell_2(w)$, then

$$K(x,x') := \sum_{i\in\mathbb{N}^*} w_i\psi_i(x)\psi_i(x') \tag{35}$$

defines a kernel. Such kernels are often called *Mercer kernels* due to their connection with Mercer theorem for integral compact operators. In this case $(w_i,\psi_i)_i$ would be the eigenelements of such a compact operator. Note however that we can define kernels as above for general settings: they only require a basis of functions to show some expressivity (sin and cos or spherical harmonics for example). An interesting class of expansion-type kernel (35) are *multiscale kernels* [Opf06] where the functions $(\psi)_i$ are scaled shifts of compactly supported refinable functions from wavelet theory. Adapting this to a computer-friendly ML setting is something I really would like to further dig in the future.

(ii) Convolutional kernels. We can of course generalize (35) to convolution type kernels:

$$K(x,x') := \int \psi_i(x,t)\psi_i(x',t)w(t)dt = \mathbb{E}_{T\sim w}[\psi_i(x,T)\psi_i(x',T)]. \tag{36}$$

Note that the integral structure of the kernel gave birth to random features in ML [RR08] (see Section 3.3 for more details).

Compactly supported kernels. When the kernels have full support, the gram matrix associated to the dataset is dense, hence for efficient numerical analysis it could be convenient to define compactly supported kernels. This is the case of Wendland kernels [Wen95] that are radial kernels define from a basis of polynomial on the unitary ball and produce RKHS that are equivalent to Sobolev spaces. Once again, despite their use in other fields of applied mathematics, their use in ML suffer from the fact that it can be numerically slow to compute them in high-dimensions.

⋆

⋆ ⋆

User-friendly RKHS and the art of kernel engineering. We have defined constructively RKHS from kernels by putting a real emphasis on their easy usability nature, leaving abstract aspects as comments. Being either used as *test* or *trial* functions, RKHS are today widely used in almost all applied mathematics communities. One the main important aspects is to know how to construct kernels *adapted* to the specified problem. This crucial art of *kernel engineering* is too often put aside in the ML community and we will try to discuss a bit if this later.

3.2. The versatility of RKHS

In this section, we go further in subsection 3.2.1 to one of the most straightforward application of kernel methods for ML: kernel ridge regression. We then explore briefly other problems for which kernels methods are efficiently used (subsection 3.2.2).

3.2.1. Empirical risk minimization

Dimensionless bounds. Let us first rephrase the problem of supervised learning with a RKHS as *space of test function*. We will see that, except from working in an infinite dimensional space, all the previous results of empirical risk minimization will stand. This represents the strength of kernel methods: once the features are correctly defined, everything is as if we perform linear regression in an infinite dimensional space. The only concern is then to derive dimensionless bounds that can be adapted to this setting.

Kernel Ridge regression. Recall that we want to minimize the generalization error $\mathcal{R}(f)$. Throughout this part we will adopt the functional notation f and not θ to put emphasis on the non-parametric nature of the RKHS $\mathcal{H}$. As in RKHS we have the reproducing property $f(x) = \langle f, K_x\rangle_{\mathcal{H}}$, the problems writes:

$$\text{Find} \quad \inf_{f\in\mathcal{H}} \mathcal{R}(f) = \frac{1}{2}\mathbb{E}_\rho\left[\left(\langle f, K_X\rangle_{\mathcal{H}} - Y\right)^2\right]. \tag{37}$$

To solve this, as stated before, we solve the empirical counterpart of the above function and regularize to avoid overfitting. One of the main difference is that one of the most natural way to regularize is to penalize the problem by the induced norm in $\mathcal{H}$. This is also called Tikhonov regularization in inverse problems [Bis95]. Knowing that the norm in the RKHS, as seen before, encodes the regularity of the function, we have here a way to really address the possible *a priori* that we can have on the Bayes optimum by penalizing the empirical risk with an appropriate norm! The Kernel ridge regression is then:

$$\text{Find} \quad \inf_{f\in\mathcal{H}} \widehat{\mathcal{R}}_n(f) = \frac{1}{2n}\sum_{i=1}^{n}\left(\langle f, K_{x_i}\rangle_{\mathcal{H}} - y_i\right)^2 + \frac{\lambda}{2}\|f\|^2_{\mathcal{H}}. \tag{38}$$

It can be easily seen that if the function f is in the orthogonal of $\operatorname{span}\{K_{x_i}, 1 \leqslant i \leqslant n\}$, the value of the empirical risk will only increase. Hence, f can be looked for as a sum of the basis functions K_{x_i}: this is the *representer theorem*. The problem can be rewritten as finding the coefficients $\alpha = (\alpha_i)_{i\leqslant n}$ of

$f = \sum_{i=1}^n \alpha_i K_{x_i}$:

$$\text{Find} \quad \inf_{\alpha \in \mathcal{H}} \frac{1}{2n} \sum_{i=1}^n \|K\alpha - Y\|^2 + \frac{\lambda}{2} \alpha^\top K \alpha, \tag{39}$$

$$\text{And} \quad f = \sum_{i=1}^n \alpha_i K_{x_i}, \tag{40}$$

where K is the usual kernel gram matrix. Note that the problem above has a unique solution given by $f_{\mathsf{erm}} = \sum_{i=1}^n \alpha_i^\lambda K_{x_i}$, where $\alpha^\lambda = (K + n\lambda I)^{-1} Y$.

Analysis of KRR. Let us give a taste of the classical analysis when the $(x_i)_i$ are fixed. First, note that in non-parametric regression, the approximation error is zero as soon as the space $\mathcal{H}$ is dense in $L^2_{\rho_X}$ (for the $L^2_{\rho_X}$-norm). Hence, noting f_ρ the optimal Bayes risk $f_\rho(x) = \mathbb{E}[Y|X = x]$, the excess risk is only:

$$\mathcal{R}(\widehat{f}_n) - \mathcal{R}(f_\rho) = \|\widehat{f}_n - f_\rho\|^2_{L^2_{\rho_X}}. \tag{41}$$

In the case of the fixed design setting ($\rho_X = \frac{1}{n}\sum \delta_{x_i}$), a straightforward calculation gives that:

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

where C is the covariance matrix of the noise $\varepsilon = Y - \mathbb{E}[Y]$. This is the classical bias-variance tradeoff: the first term is the bias term that increases with λ and the second term depends on the noise. Note that a central quantity describes the rate of convergence: the eigenelements of the matrix K. Indeed, for the bias term, the way $\mathbb{E}[Y]$ is going to decompose on the eigenvectors of K is primordial, and for the variance term, the quantity that controls the convergence is a function of the eigenvalues of K. Those two quantities: (i) how the Bayes optimum decomposes into the spectrum of the covariance matrix and (ii) how fast the eigenvalues of the covariance matrix decreases are central in kernel regression. Introducing them is the purpose of the main paragraph.

Mercer theorem, source and capacity conditions. In the random design analysis, even if the calculations are more involving, the exact same decomposition occurs and the bias-variance tradeoff can be analyzed when quantifying both (i) and (ii). However, the operator that arises in such an analysis is no longer a finite matrix but an integral operator from $\mathcal{H}$ to $\mathcal{H}$ instead:

$$\Sigma f = \int f(x) K_x d\rho(x). \tag{42}$$

Note that is can be seen as the restriction on $\mathcal{H}$ to the well known kernel integral of $L^2_{\rho_X}$:

$$(Tf)(x) = \int f(x') K(x, x') d\rho(x'). \tag{43}$$

Leveraging this fact, we can apply the same analysis thanks to Mercer theorem [Aro50]. It tells us that since Σ is a compact self-adjoint operator, it admits an orthonormal basis of functions of $\mathcal{H}$, $(\phi_i)_i$, associated to vanishing eigenvalues $(\mu_i)_i$ such that

$$K(x, x') = \sum_{i \geqslant 0} \mu_i \phi(x) \phi(x') \tag{44}$$

$$\mathcal{H} = \left\{ f = \sum_{i \geqslant 0} a_i \phi_i \ \text{ such that } \ \sum_{i \geqslant 0} \frac{a_i^2}{\mu_i} < \infty \right\} \tag{45}$$

In Section 2 of Part II, we will introduce such parameters to control the bias and variance of non-parametric regression in RKHS [CDV07, SHS09, RCR17, LR17].

(i) Source condition. This first quantity controlling the bias, will quantify the difficulty of the learning problem through $\|\Sigma^{1/2-r} f_\rho\|_{\mathcal{H}}$, for $r \in [0,1]$. Indeed, the source condition is an assumption on the bigger $r \in [0,1]$ such that

$$\|\Sigma^{1/2-r} f_\rho\|_{\mathcal{H}} < +\infty \tag{46}$$

It represents how far in the closure of $\mathcal{H}$ the Bayes optimum stands. Note that $r = 0$ is always true since f_ρ always belong to $L^2_{\rho_X}$.

(ii) Capacity condition. This second quantity controls the variance and will characterize the decay of eigenvalues of Σ through the quantity $\mathrm{tr}\Sigma^{1/\alpha}$. Indeed, the capacity condition is an assumption on the bigger $\alpha \in [0,1]$ such that

$$\mathrm{tr}\Sigma^{1/\alpha} < +\infty \tag{47}$$

This is related to the α-decay of the *intrinsic dimension* $\mathrm{tr}\left[(\Sigma + \lambda I)^{-1}\Sigma\right]$.

In the finite-dimensional case, these quantities can always be defined, are finite, but may be very large compared to sample size. Note that they both depend on the law ρ and on the choice of the kernel K. As it is shown in Section 2 of Part II, an example one can have in mind is the following. When the distribution ρ_X is uniform over a compact set, the kernel is of Sobolev type of order s and the Bayes predictor is in a Sobolev of order s_*, we have $\alpha = \frac{2s}{d}$ and $r = \frac{2s_*}{s}$. Note also that there is a hidden curse of dimensionality here: the Bayes optimum need to be $O(d)$-times differentiable to recover rates independent of the dimension. Those two quantities represent assumptions on the learning problem and allow to derive dimensionless bounds on it.

To conclude this part, one can show that generally speaking (it will be addressed more precisely in Section 2 of Part II), the kernel ridge estimator achieves the minimax rate for this class of problem which is of order $n^{-\frac{2\alpha r}{2\alpha r+1}}$ (for rates of KRR and minimax rates in this setting, see [CDV07]).

3.2.2. Other uses

As it is our concern in Part II of this manuscript, we have detailed how RKHS are good spaces to perform regression in supervised ML settings. We will now present more succinctly ideas of application that one can have in mind when considering kernel methods in statistics and computational applied mathematics in general.

Going non-linear. In the supervised setting, we have seen that RKHS provide rich trial functions spaces that allow for computations. Parametric spaces of functions, for which we can have theoretical guarantees are often much poorer or necessitate some very good a priori on the problem. In a word, RKHS allows without much pain to go beyond linear models. And supervised learning is not the only case where RKHS are interesting. In unsupervised learning many problems can benefit from the non-linearity of RKHS spaces. This is why there has been a huge interest in developing the analogous of Principal Component Analysis (PCA), Canonical Correlation Analysis (CCA) and Independent Component Analysis (ICA) in the RKHS framework giving birth to *kernelized* versions of it: kernel-PCA, kernel-CCA, kernel-ICA. We will see in Part III Section 2 that this kernelization trick will be used to derive another dimensionality reduction algorithm. All of these leverage the fact that once the data is put (in a non-linear way) in the feature space, then, every thing goes as if it were linear. See Figure 7 for an example in the case of polynomial RKHS.

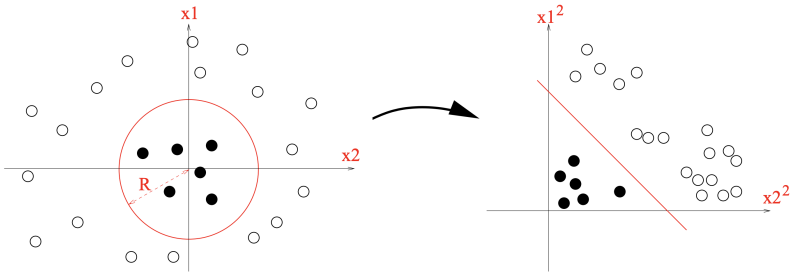

Figure 7: Representation of the feature that maps in the case of order 2 polynomial RKHS in 2-d. We see that in the feature space the data points are linearly separated. Once again, this representation is taken from the slides of J.-P. Vert and J. Mairal lessons on Kernel methods.

Inducing a metric. Kernels also allow to construct some metric in some *a priori* non-structured set $\mathcal{X}$. In fact we can always define a proper metric associated to $\mathcal{X}$ which is

$$d_K^2(x,x') := \|K_x - K_{x'}\|_{\mathcal{H}}^2 = K(x,x) + K(x,x) - 2K(x,x').$$

This allows to compare elements in unstructured sets by comparing their associated canonical features in an Hilbert space $\mathcal{H}$.

Kernel mean embedding. The theory of kernel mean embedding [SGSS07, MFSS17] cannot be summed up in such a short paragraph but let us try to explain in a few words what it is and why it can be useful. The first thing to understand is that probability measures can be embedded in the RKHS: this is the analogous of the reproduction property for the probability measures. Indeed, one can show that

$$\mathbb{E}_\rho(f(X)) = \langle \mu_\rho, f\rangle_{\mathcal{H}} \qquad \text{with} \qquad \mu_\rho := \mathbb{E}_\rho[K_X].$$

μ_ρ is the *kernel mean embedding* of the distribution ρ and for characteristic kernels, μ_ρ encodes uniquely the distribution ρ. This embedding allows to quantify differences between distribution by the natural distance in $\mathcal{H}$: this is the Maximum Mean Discrepancy (MMD) distance:

$$\mathsf{MMD}_{\mathcal{H}}(\rho,\rho') := \sup_{\|f\|_{\mathcal{H}}\leqslant 1} \left(\mathbb{E}_\rho\left[f(X)\right] - \mathbb{E}_{\rho'}\left[f(X)\right]\right) = \|\mu_\rho - \mu_{\rho'}\|_{\mathcal{H}}\,.$$

This computationally tractable distance is the cornerstone of many algorithms that use measure comparison [GBR+12].

Meshless methods and approximation theory. Finally note that kernels methods have received a huge interest in the approximation literature for their rich, yet tractable computation properties. Another very interesting application in this point of view is that RKHS have the ability to reproduce derivatives of functions and thus have been widely used in the context of PDE approximation. These are often called *adaptive or meshless methods* in this literature [SW06].

⋆

⋆ ⋆

The idealized picture of kernels. In this section we have seen two important aspects of Kernel methods. First they manage somehow to leverage the regularity of functions to avoid the curse of dimensionality inherent to local methods. Note that this is often the idealized picture given by kernels but there is a catch with this: one need the function to lie in the RKHS to have dimensionless bound which is pretty restrictive if we do not know how efficiently build large an problem-adapted RKHS. Second, they allow to go for non-linear space of *test* functions, argument which we will mitigate in the next section.

3.3. Promise and pitfalls of kernels in ML

Previous sections have shown how kernels address well problems in ML or in numerical applied mathematics in general. In this section, beyond the well-established literature of kernels, we will try to comment on, in a first step, the limitations of such techniques and secondly explain how these limitations can be overcome. In this section, we will not dwell into a rigorous mathematical development but rather focus on key concepts and practical limitations that one has to have in mind when considering kernels methods.

3.3.1. Limitation in the high-dimensional setting

It is important to recall here the high-dimensional setting we are into: both the dimension of the inputs d and the sample size n can be millions.

The price of computations. The first limitation that people have in mind when thinking about kernel techniques in the computational limitation of these. Indeed, in almost all procedures involving kernel methods, a central object is the Gram matrix, K, associated to the data $(x_1, \dots, x_n)$.

$$\forall\, i, j \leqslant n, \quad K_{ij} = K(x_i, x_j).$$

The computation of this matrix is very consuming both on the memory aspect and on the time performance. Indeed, this is a $n \times n$ matrix and each element of this matrix often requires a scalar product in $\mathbb{R}^d$ to compute. When both n and d are very large, this can be prohibitive. Worse, once this matrix has been computed, people often want to do some classical mathematical operations with it: in ridge regression one want to invert the matrix, then to multiply it with some vector, and in kernel-PCA (for example) one want to compute eigenvalues on this matrix. All these operations scale very poorly with n (typically $O(n^3)$ for inversion $O(n^2)$ for eigenelements) and are often not very stable if the matrix is ill-conditioned.

Are kernels really non-linear? We saw in the previous section 3.2.2 that one of the main success of kernel methods in ML during the first decade of this century is that it allows to look for non-linear functions when solving our problem. But is this really the case? This question has been answered by El Karaoui in [EK+10] and is quite surprising. Roughly speaking, it says that when d and n are of equivalent magnitude, and if the kernel as a *zonal* or *radial* structure, then the kernels methods over $\mathbb{R}^d$ are in fact instances of linear methods. More precisely, the result sates that for kernel $K^{\mathsf{zonal}}(x, x') = K(\langle x, x' \rangle_2)$ or $K^{\mathsf{radial}}(x, x') = K(\|x - x'\|_2)$ then the associated Gram matrices that rules the behavior of the algorithms $K^{\mathsf{zonal}}(\langle x_i, x_j \rangle_2)$ or $K^{\mathsf{radial}}(\|x_i - x_j\|_2)$ behave essentially like linear Gram matrices! An intuition behind this fact is that when $d \sim n$, n vectors are almost certainly orthogonal to each other thus their similarity behave almost linearly.

From this we can draw at least one guideline: in high-dimensional settings, radial or zonal kernels have to be avoided in one want a better result that the one given by linear spaces. Or in other words anisotropy is very important in high dimension. This is certainly one of the superiority of neural networks over (at least radial and zonal) kernels in high-dimension.

3.3.2. Choosing the right kernel

Throughout this thesis, we took the point of view of deriving bounds for general kernels leaving the role played by the RKHS in modelling assumptions such that capacity and source conditions (see section 3.2.1). However, we would like to put emphasis on the fact that for real applications *kernel engineering* is a very important task. To put this into perspective, learning the representative features of the problem all along the optimization path is perhaps one of the most important successes of neural networks. Playing the same role as the architecture in neural networks, choosing the kernel has to be a problem-adapted task for the kernel method to perform well. There are two ways to deal with such this problem:

(i) Find a priori *the right kernel adapted to the problem.*

(ii) Learn the kernel similarly to what happens in Neural Networks.

Problem with radial kernels. People have a tendency to think that kernel learning is a very mature field and that in comparison to neural networks almost everything is almost known is this literature. However, I would argue that even some of the most basic and fundamental questions on kernel learning are still unsolved. Without tackling the problem of selecting a full data-driven kernel for a regression task, we may first ask whether there are guidelines to tune hyper-parameters of kernels to best learn from the data. For example, in the simplest setting of kernel ridge regression with regularizer λ and Gaussian kernel with bandwidth σ, both λ and σ play a smoothing role but there are no precise guidelines on how to tune them to get the best accuracy. However, the role played by the bandwidth in the Gaussian kernel is huge as shows the fact that the norm of the induced space is multiplied by σ^d when changing the scale of the bandwidth. To cut a long story short: learning or adapting the bandwidth of a radial kernel to the problem seems to be an unsolved challenging theoretical and practical question.

Multiscale and hierarchical kernels. Many problems in ML have a natural multiscale structure like vision, text, signal processing, chemistry... One way to adapt to this structure can be to build a RKHS that will take it into consideration. Those multiscale methods have shown good performances in signal processing and in vision with wavelets theory [Dau92] and the aim is to reproduce this in RKHS. One way to construct such a RKHS is to sum RKHS induced by radial kernels with different scales. More concretely, consider ϕ a compactly supported function of $[-1, 1]$ and a sequence of decreasing scales $\delta_1 \geqslant \delta_2 \geqslant \ldots \geqslant \delta_n$, then we can define $K_j(x, x') = \delta_j^{-d}\phi(\|x - x'\|/\delta_j)$ and j-scale approximation spaces $\mathcal{H}_j$ induced by K_j. Then the resulting multiscale RKHS would be:

$$H_{\text{multiscale}} := H_1 + \ldots + H_n.$$

Another line of work concerning the will to adapt the kernel architecture to the problem has been tackled by so-called *hierarchical kernels*. One observation is that (isotropic) kernels often do not adapt well to the underlying structure of the data as they treat all features as equal. This is why in [STS16], Steinwart and al. have constructed a hierarchical kernel based on composition of Gaussian kernels (activation) and weighted sums of linear kernels (layers). They also show that it was possible to find a convex optimization procedure able to learn the weights of the corresponding linear kernels. Even if this line of research did not yield yet a performance comparable to the ones of neural networks, building kernel that mimic the behavior of neural nets with preserving kernels guarantees is very promising.

Learning the kernel. Hierarchical kernels was an attempt to find adapted kernel to the problem. In this direction, let us simply mention two other attempts. The first one is the multiple kernel learning algorithm [BLJ04], which, in a nutshell, replaces a single kernel by a weighted sum of kernels. The advantage of this approach is that finding these weights can again be formulated as a convex objective, while the disadvantage is the limited gain in expressive power unless the used dictionary of kernels is really huge. In the same spirit, Dutchi and al. show that it was possible to learn the best random probability that lead to a random feature model [SD16].

3.3.3. Fast numerical computations

One of the major problem with kernel learning is recalled in the first paragraph of the section: storing and computing the Gram matrix of a kernel problem can be very expensive. In this subsection, we will see how techniques from (random) linear algebra and computer science can improve drastically the numerical performance of such algorithms.

Subsampling and features approximations. Fortunately, to avoid the problem of calculating the whole Gram matrix (which costs $O(n^2 d)$), we simply need low-cost approximations of the kernel matrix [SS02, Sec. 10.2]. More surprisingly, these approximations can improve generalization performances as they induce a form of implicit regularization. Let us state here two important techniques related to this, for more details see Chapter 19 of [MT20].

(i) Subsampling methods. One way is to select uniformly at random p rows among the n of the kernel matrix and perform some Nyström approximation. There exist other techniques for the rows selection but in practice, it is often said that uniform selection is already efficient. Note that $p_n \sim \sqrt{n}$ are necessary to recover the performance of plain kernel learning [RCR15].

(ii) Random features. Another popular way to approximate the kernel matrix is to leverage the convolution structure of certain kernels [RR08]. Indeed, for convolutional kernels as in Eq. (36)

$$K(x, x') := \int \psi(x, t)\psi(x', t)w(t)dt,$$

the integral representation gives immediately a feature map $\psi(x, t)$ to $L^2(w)$. Hence, drawing a sample t_j from the probability measure $w(t)dt$ gives $z_j = (\psi(x_1, t_j), \dots, \psi(x_n, t_j))^\top$ where $z_j z_j^\top$ is an unbiased estimate of the kernel matrix K. If we repeat this procedure with r i.d.d. samples, this will give an approximation of the matrix K at Monte-Carlo rate (independent of dimension):

$$K \sim K_r := \frac{1}{r} \sum_{j=1}^{r} z_j z_j^\top.$$

This procedures has only a computational cost of $O(rnd)$, and one can obtain substantial improvements of performance in the case where r is small in comparison to n. Note that with this technique, we have also a direct access to the derivatives of the kernel with respect to x, x' without additional numerical complexity. This will be leveraged in Part III of this thesis.

Fast numerical routines. To conclude this part, let us add that there have been huge efforts spent to speed up the computations relative to kernel methods. The first one is due to the work of Feydy and co-authors that developed efficient C++ routines to fasten kernel computation (of $x \to Kx$) of several order of magnitude [CFG+20]. Building on this, efficient Nyström approximations and efficient method to invert ill-conditioned matrices, Rudi and al. developed a tool box in [MCRR20] that allow kernels to handle

billions of data points efficiently. Progressed in these directions could really unlock severe limitations of kernels and allow to re-discover techniques that were forgotten due to their slowness in the past.

⋆
⋆ ⋆

Kernel hopes and the need to dig deeper. I hope that I convinced the reader that, even if today kernels are not as used as neural networks, there are hopes for the future of kernels that we manage in circumventing their natural limitations. Furthermore, let us add two characteristics that make them worth studying:

(i) They still give rich insights. In the recent literature there have been two examples of kernels giving insight on ML phenomenons. First, [MM19] explains that random features could show a *double descent behavior* that has received a huge interest lately. Second, [JGH18] states that the behavior of the dynamics of neural networks is very similar to the one with an explicit kernel called the *neural tangent kernel*.

(ii) They give guarantees. Proving convergence guarantees for neural networks (even in simple setting) is still ongoing research. One may want solid statistical guarantees for industrial problems, in this case neural networks could be disregarded and kernels preferred.

4. Langevin Dynamics

Langevin dynamics is at the core of Part III of this thesis. This a physically anchored dynamics that is strongly related to sampling techniques and has been "re-discovered" lately in the Machine Learning community. In Section 4.1, we define precisely what is the Langevin dynamics and how it relates to ML problems. Then, we explain in Section 4.2 how this dynamics can be used to sample efficiently distributions in high-dimension if we manage to deal with the metastability problem described in Section 4.3. Note that several paragraphs of this part are extracted from a post-doc proposal of research that I wrote to explain my interests. As it will be clear for the reader, this part presents also my future intention of working on the interplay between Molecular Dynamics and Machine Learning.

4.1. What is Langevin Dynamics ?

4.1.1. Definition and link with Molecular Dynamics

Langevin Dynamics. Langevin dynamics comes from statistical physics and is a system of dynamical equations that governs the speed and positions of particles (typically atoms). To fix notations, let us suppose that N particles composed the system: typically the magnitude of N is the number of Avogadro $\mathsf{N}_{\mathsf{avo}} \sim 10^{23}$, thus positions q and momenta p are both vectors of $\mathbb{R}^{3N}$. A simple, yet rich model is to divide the total energy of the microscopic system in a kinetic and a potential energy.

$$H(q,p) = \frac{1}{2} p^\top M^{-1} p + V(q),$$

where M is the diagonal matrix of the masses of the particles. Roughly speaking, the particles follow an Hamiltonian dynamic in a thermal bath of fixed temperature T with friction γ that cause fluctuations of order $\sqrt{\gamma\beta^{-1}}$ where $\beta = (k_{\mathrm{B}}T)^{-1}$. This leads to the celebrated *Langevin dynamics*:

$$\begin{cases} dq_t &= M^{-1} p_t dt \\ dp_t &= -\nabla V(q_t)dt - \gamma M^{-1} p_t dt + \sqrt{\frac{2\gamma}{\beta}} dW_t \end{cases} \tag{48}$$

As discussed in [SRL10, Section 2.2.4], a simpler reversible equation can be obtained as a limit case of Langevin dynamics by taking the large friction limit $\gamma \to +\infty$, small mass $m \to 0$ and rescaling time as γt, this gives the *Overdamped Langevin dynamics*:

$$dq_t = -\nabla V(q_t)dt + \sqrt{\frac{2}{\beta}} dW_t \tag{49}$$

While the *Langevin dynamics* is a finer modelling adding some kinetic term to the equations, we will focus on the *Overdamped Langevin dynamics* for simplicity, as it leads already to complex and unsolved problems. Note here that studying quantitatively how the add of the kinetic term in Eq.(48) changes the dynamics is an open and exciting route for future research (see subsection 4.2.2 for more details). Let us simply mention here that one of the difficulties to study the kinetic Langevin is the lack of ellipticity of the dynamics [Vil09] (ruled by a degenerate dissipative operator).

Molecular dynamics and sampling. MD, the computational workhorse of statistical physics, is an interdisciplinary field between computer science, applied mathematics and chemistry whose main objective is to infer macroscopic properties of matter from atomistic models via averages with respect to probability measures dictated by the principles of statistical physics [SRL10, FS01]. In a nutshell, the aim is to be able to derive averages with respect to the canonical Boltzmann-Gibbs distribution,

$$\mu(\mathrm{d}q\ \mathrm{d}p) = Z_\mu^{-1} e^{-\beta H(q,p)} \mathrm{d}q\ \mathrm{d}p, \qquad Z_\mu = \int e^{-\beta H}. \tag{50}$$

Note that the real difficulty is to sample according to q. Indeed momenta and positions are independent and the marginal associated to the momenta follows a Gaussian distribution. The aim of molecular dynamics is to calculate macroscopic quantities like the pressure of the system that can be express as averages with respect to the canonical measure:

$$\mathbb{E}_\mu(\phi) = \int \phi(q,p)\mu(\mathrm{d}q\,\mathrm{d}p). \tag{51}$$

What is the link with Langevin dynamics? Under certain conditions, we can show that the law of the processes define by the Langevin dynamics converge to the Gibbs distribution $\mu(\mathrm{d}q\,\mathrm{d}p)$. The same result holds for the overdamped Langevin dynamics when considering only $\mu(\mathrm{d}q) = e^{-V(q)}$. Note that from this point of view, the Langevin dynamics (48) and (49) are only used as *sampling devices* to compute averages. Other dynamics without physical contents could be studied too, as long as they have $\mu(\mathrm{d}q\,\mathrm{d}p)$ as invariant measure!

4.1.2. Parallel MD - ML

Two motivating examples: Bayesian inference and non-convex optimization. Accurately sampling a certain measure $\pi(q)\mathrm{d}q$ in high dimensions is a difficult task that arises in Bayesian machine learning. If we take $V = -\beta^{-1}\log\pi$, sampling in the Bayesian framework can be directly cast into the same problem of Molecular Dynamics.

Furthermore, note that when decreasing the temperature to 0, i.e., setting $\beta \to +\infty$ the Gibbs measure $\mu(\mathrm{d}q) = e^{-\beta^{-1}V(q)}\mathrm{d}q$ will concentrate around the global minima of V. Hence, performing sampling at low temperature can be a way to solve non-convex problems: this formally describes the intuition behind the celebrated *simulated annealing algorithm* [VLA87].

Links between Molecular Dynamics (MD) and Machine Learning (ML). Besides sharing common goals, let us remark that the function f to be minimized in ML and the potential V in MD share three important features:

- The **high-dimensionality** of the underlying measure or cost-function: this is due to the number of atoms in MD and the high-dimensionality of inputs in ML. In terms of notations, the dimension d of the inputs in ML is equal to three times the number N of particles in MD (three parameters to encode positions).
- The **metastability** phenomenon [Lel13] due to the multimodality of the target measure in MD, which corresponds to the non-convexity of the loss function in ML. Indeed, as we said earlier, in MD the target measure can be written $\mu = \mathrm{e}^{-V}$ where V can be interpreted in optimization as the loss function to minimize, hence casting sampling problems in MD to a tempered version of the minimization of V. The metastability comes from the fact that the dynamics can be trapped for long times in certain regions (modes) preventing it from efficient space exploration or finding the global minimum of a non-convex function. This phenomenon is often responsible for the slow convergence of the algorithms. We will come back to this important point in section 4.3.
- In MD and ML, the question of finding **low-dimensional representations** (main degrees of freedom) is crucial. Standard techniques include for example Principal Component Analysis and variants, manifold learning methods such as diffusion map. Recently, ML techniques have proven to be very useful to perform such tasks, thanks to its ability to handle and extract the main features of high dimensional data.

⋆

⋆ ⋆

Possible future directions and Part III's point of view. The main point of view of this part of the thesis is that both fields can benefit from the knowledge and know-how of the other one: MD seems to be a more mature and theoretically-anchored field of study than ML but the recent successes of the latter may be leveraged to solve long-standing problems of MD. More specifically, we would like on the one hand to investigate if the sampling methods developed in MD could help to improve the learning algorithms in ML, and on the other hand to rely on recent ML techniques to build reduced order models in MD.

Questions. The connections highlighted above raise two symmetrical questions:

- How can techniques and principles from MD enlighten theory and practice behind ML algorithms?
- How can the efficiency of recent ML techniques help solving MD problems?

4.2. Sampling with Langevin dynamics

Besides Molecular dynamics, we have seen above that Langevin dynamics could be useful to sample the posterior distribution in the Bayesian framework or getting near minima of non-convex functions in the low-temperature regime. We review in this section non-asymptotic results obtained during the last decade, following the work of Dalalyan in [Dal17].

4.2.1. Discrete time dynamics and convergence results

Convergence of continuous time dynamics. The convergence of the continuous time dynamics is a rather well-studied problem [BGL14]. Let us recall here, with usual optimization notations, the overdamped Langevin dynamics we are focused on (take $\beta = 1$ for simplicity):

$$d\theta_t = -\nabla f(\theta_t)dt + \sqrt{2}dW_t.$$

As $(\theta_t)_t$ is a random process, define its law μ_t. We can show that μ_t converges to the Gibbs distribution $\mu = e^{-f}$ under certain conditions. Remarkably, this rate of convergence depends only on the invariant law μ. As in optimization with Lyapounov functions, we also need to select a convergence norm to show convergence. One of the most common way is to show convergence in the $L^2(\mu)$ metric, in this case the rate of convergence is described by the Poincaré inequality [BGL14, Section 4].

Definition 1 (*Poincaré inequality*)

We say that the probability measure $d\mu$ satisfies a Poincaré inequality if for all $f \in H^1(\mu)$,

$$\mathrm{Var}_\mu\left(f(X)\right) \leqslant \mathcal{P}_\mu \, \mathbb{E}_\mu\left[\|\nabla f(X)\|^2\right]. \tag{52}$$

Poincaré inequalities are the cornerstone of the Part III of this thesis and we refer to longer discussions in Section 1.2.2 of this Part for details. The rate of convergence to equilibrium is encoded by the Poincaré constant associated to μ as we have the following equivalence:

(i) μ satisfies a Poincaré inequality with constant $\mathcal{P}_\mu$;

(ii) For all f smooth and compactly supported, $\mathrm{Var}_\mu(P_t(f)) \leqslant \mathrm{e}^{-2t/\mathcal{P}_\mu}\mathrm{Var}_\mu(f)$ for all $t \geqslant 0$.

Further comments are given in Section 1.2.2, but let us stress that in the case where f is ρ-strongly convex, we can show that $\mathcal{P}_\mu \leqslant 1/\rho$. In a way, the Poincaré constant is the analogous of strong convexity in optimization when it comes to sampling. Note that f need not be convex functions for μ to satisfy a Poincaré inequality.

Discrete time dynamics. First if we want to perform sampling we need to discretize the overdamped Langevin dynamics (49). As we will see, they are as much ways to perform this discretization as methods in optimization to minimize a function. And, as in optimization, a given hypothesis on the measure to sample from will lead to an adapted discretization scheme. This is why entering into this literature can be difficult, time-consuming and puzzling at first sight. Hence, without seeking exhaustiveness, we will try to focus on the main ideas behind the different settings. From now on, let us change the notation of the potential in Gibbs measure from V to f: $\mu(\theta) = e^{-f(\theta)}\mathrm{d}\theta$ to adopt the classical notation of the optimization literature. And keep in mind the hand-waving rule:

If a deterministic optimization algorithm performs well on a function f, then there is great chance that its noisy counterpart will behave almost the same to sample the probability measure e^{-f}.

Having taken this precautions, let us write -only- the most natural discretization of (49) called explicit Euler-Maruyama :

$$\theta_{t+1} = \theta_t - \gamma \nabla f(\theta_t) + \sqrt{2\gamma} z_t, \tag{53}$$

where $\gamma > 0$ is the step-size and $(z_t)_t \sim \mathcal{N}(0, I)^{\mathbb{N}^*}$ is a sequence of i.i.d. Gaussian random variable in $\mathbb{R}^d$. This algorithm is often called the Unajusted Langevin Algorithm. As already said, there exist other discretizations schemes, e.g. implicit Euler leads to the Proximal Langevin Algorithm and is more stable at the cost of more expensive per-iteration complexity [Wib19].

Convergence of discrete time dynamics. As we have seen, in continuous time Poincaré inequality is sufficient for fast sampling as it leads to exponential convergence of the overdamped Langevin dynamics. However, in discrete time, analyzing convergence is more challenging. Indeed, to control the discretization error, some smoothness assumptions are required on f. Worse, the discretization error leads to an asymptotic bias: the law of the discretized dynamics converges to the wrong distribution. To correct this bias, it is possible to apply some Metropolis filter that gives the convergence in total variation norm [BRH13] but cannot give convergence in smoother norms as the Metropolis filter can make the distribution singular. This is why, in an approach pioneered by Dalalyan in [Dal17], all the current analysis make step-size small enough to be as close as wanted to the Gibbs measure μ. This said, there exist as much results as possible settings: convergence in different probability norms, for different discretizations and under different assumptions (some weaker, some stronger than Poincaré inequality). For example, typical number of steps to be ε-close to the Gibbs distribution is of order $O(\mathcal{P}_\mu^2 d^{3/2} \varepsilon^{-1})$ in the case of a Poincaré inequality assumption [Wib19].

4.2.2. What can we do with MD knowledge ?

Analysis of stochastic algorithms. One of the main difference between MD and ML preventing from direct knowledge transfer is that the main foci of the two disciplines are different. Indeed, in MD, people are interested in sampling according to a given target measure, whereas in ML, the focus in on the minimization of some objective function. Note that those two questions coincide in the zero temperature limit. Hence, it seems possible to transfer technical tools to analyze algorithms in computational statistical physics to try to improve non-asymptotic bounds for Langevin-type discretizations algorithms which are often the main focus of ML works [RRT17, DM19].

Accelerating the dynamics. More importantly, one active field of study in MD and ML is to try to accelerate the stochastic dynamics at hand. These techniques rely on changing it without affecting the invariant measure while accelerating the convergence to equilibrium. We could (i) study Kinetic versions of Langevin dynamics [MCC+19], (ii) add some drift term or (iii) some non-reversibility through birth-and-death processes [LLN19].

⋆
⋆ ⋆

Understanding continuous-time SGD. The Langevin dynamics seems very related to stochastic gradient descent with additive noise as we have seen in Section 2 of this introduction. In this same direction, while the ML community has solid knowledge of discrete-time dynamics, understanding continuous versions of the algorithms often leads to a deeper comprehension of the behavior of the systems. Such a paradigm is often used in MD where mathematical tools were designed to tackle such problems. This could lead to the study of the continuous counterpart of the stochastic gradient descent algorithm [LT19]. Knowing how the noise and potentials behave in SGD will be the key to apply the ideas from MD and the study the possible metastability of the dynamics.

4.3. The metastability problem

4.3.1. What is metastability ?

The metastability problem of Langevin dynamics comes from the two scales involved in the physical problem. Indeed, particles behave very differently at the microscopic level and at a macroscopic one where interesting properties emerge from collective behavior. This discrepancy between those two scales is related to the fact that the system remains trapped for long time in restricted regions of the space of configuration. And it may take a large characteristic time to "jump" into another metastable state. In fact at a coarse grained level, people have modeled the behavior of the particle system by a jump process at certain rate from one region to another [Lel13].

Computationally-wise, metastability is an important problem because it prevents from efficient space exploration and implies that the convergence to the stationary distribution can be very slow. One can easily picture two wells separated by some barrier such that it is a rare event to see the particle going from a well to another one. This is a well-known evidence of metastability (arising because of non-convexity) and large deviation theory [FW98] showed that the typical time scale to go from one well to the other grows exponentially with the inverse of temperature. This is what we call an *energetic barrier* and is represented in Figure 8 (a-b). Another type of metastability can occur: *entropic barriers* are due to the fact that the path to go from one region of configuration space to the other may be hard to fine. This is the case with the tiny corridor represented in Figure 8 (c-d).

A quantitative measure of metastability: Poincaré inequality. A natural question then is how to quantify the metastability of the process. Metastability summarizes qualitatively the fact that the dynamic is very slow due to non-convex effect of the landscape or narrow transitions between separated regions of space. We have seen that the Poincaré constant is itself a measure of non-convexity of the potential and we can show that it degrades also linearly with temperature in the presence of entropic barriers.

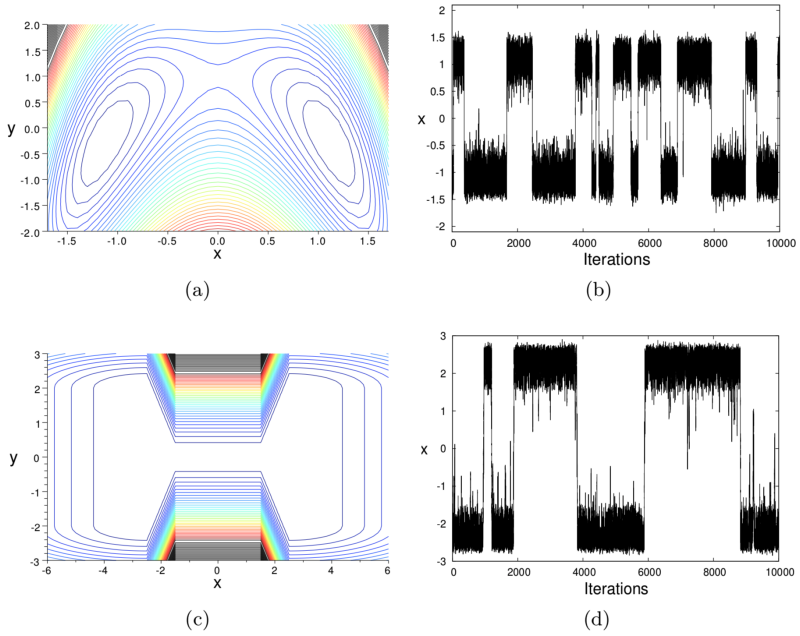

Figure 8: (a,c) Level sets of the two-dimensional potentials in the cases of (a) energetic barrier (c) entropic barrier. (b, d) Evolution along the x-axis of the overdamped Langevin diffusion in these potentials. This figure is extracted from [LS16].

Actually, Poincaré constant or other quantification of convergence (logarithmic Sobolev inequality are the one detailed in [Lel13]) are very well suited to quantify such a phenomenon. Throughout this thesis and especially in Part III, we will keep in mind the following idea:

The larger the Poincaré constant, the more metastable the process is.

Note here that the dependence in the number of step as the square of the Poincaré constant in the discrete time analysis of Langevin dynamics can make the estimation of the invariant measure computationally intractable as soon as it is too big.

4.3.2. Avoiding metastability

Reaction coordinates. Obtaining a good collective variable (i.e. reduced models with fewer degrees of freedom) is a crucial problem in computational statistical physics. It allows to grasp physical behaviors of the systems at a coarse-grained level. This low-dimensional representation, a.k.a. *reaction coordinates*, should index properly transitions between the modes of the probability measure and is the cornerstone in MD to tackle the issue of metastability and accelerate the dynamics.

To illustrate this, let us consider the two examples described in Figure 8. In the two examples, (b) and (d) show cases where the particle stay during long times in restricted regions of the space: in (b) $x \sim \pm 1$ and in (d) $x \sim \pm 2$. Hence, in both cases, there are two metastable states and the transition between the two states can be described by the x-coordinate. This x-coordinate is thus a slow variable of the system compared to the y-coordinate and typical time of changes are much longer than the classical diffusion in the metastable region.

More generally speaking a reaction coordinate is a function

$$\xi : \mathbb{R}^d \to \mathbb{R}^p, \qquad p \leqslant d. \tag{54}$$

such that $(\xi(q_t))_t$ is a metastable process. ξ should encode the transition path between the metastable states. Coming from molecular chemistry, we can understand this denomination as it represents the

coordinate (or path) is configurational space along which chemical reactions occur. It is a reduced order model, hence, the smaller the p the more practical and intuitive the reaction coordinate is. In examples coming from MD, it could be the angle of a protein characterizing the conformation of a molecule, or the position of a material defect.

Finding a good Reaction Coordinate (RC) is of great important both for chemists and computer scientists of the domain. However, today, practitioners need some *a priori* knowledge or deep physical intuition on the system to find RC. One of the aim of the Part III of this manuscript is to automatically find good RC during the sampling procedure leveraging statistical methods. In this context, we proposed in Part III, Section 2 a general algorithm to find reaction coordinates by estimating the spectral gap of the overdamped Langevin dynamics for a given target probability, using samples of this measure.

Using reaction coordinates to accelerate the dynamics. Reaction coordinates are very useful for chemists to understand better the properties of the system. In the eyes of computer scientists, RC are used to accelerate the dynamics by applying free energy biasing methods such that the free energy biased dynamics reaches equilibrium much faster than the original unbiased dynamics [LRS08]. The process is very well described in [LS16, Section 4] and we will only sketch the main principle here. Note also that these techniques have also been studied in the high-dimensional Bayesian framework to accelerate the sampling dynamics in the case of a Gaussian mixture model [CLS12].

Circumventing the difficulty caused by metastability by leveraging the knowledge of reaction coordinate rests upon two ideas:

(i) Using importance sampling strategies by changing the original potential V to $V - F(\xi(\cdot))$ where F is the free energy associated with the RC ξ. Roughly speaking, F is the potential associated with $\mu_F := \xi\#\mu$ the image of the measure μ by ξ.

(ii) Since $\mu_F = \xi\#\mu$, F is not available in practice. The second idea is to use a current estimate F_t that will be better and better along the dynamics. This adaptive feature explains why method of this type are called *free energy adaptive biasing techniques*.

The intuition behind the idea of changing the potential with the free energy is that by doing so, the metastable features of the original potential along ξ will be removed. Indeed, we can show that along ξ the biased measure with potential $V - F(\xi(\cdot))$ is uniform: this allows for fast sampling!

⋆
⋆ ⋆

Conclusion of the introduction. Throughout this introduction we have tried to show how naturally what were the main questions of this thesis. Here are the main messages to take home. In the first section 1, we tried to put a particular emphasis on the fact that optimization was absolutely unavoidable in the high-dimensional ML setting. Going further in Section 2, we tried to convince the reader that stochastic versions of classical first order methods are the cornerstone of these optimization techniques by the low-cost but also their ability to give good estimators. After this, we took a detour in Section 3 to the non-parametric world of RKHS demonstrating that there were still many unsolved questions in this field that may lead them to properly compete with neural networks. We finally show in Section 4 that analyzing high-dimensional algorithms through the lens of their continuous-time counterpart could enlighten their properties, allow to focus on MD-related question such as metastability and accelerate them.
I have now introduced my personal way of thinking in ML and my personal questions, interests and foci. I hope it will enlighten the reading of what consists in the core of this thesis: Part II and III.

Part II

Non-parametric Stochastic Gradient descent

We divide this part into our two contributions for non-parametric Stochastic gradient descent.

The Section 1 together with its Appendix A shows the exponential convergence of Stochastic gradient descent of the binary test loss in the case where the classification is easy: we talk about a hard margin condition. Side results in this work could be of their own interest: among them are derived high probability bounds in $\|\cdot\|_\infty$ norm for non-parametric SGD resting on concentration and fast rates under low-noise conditions *à la Mammen and Tsybakov* are derived. This section is based on our work, **Exponential convergence of testing error for stochastic gradient methods**, L. Pillaud-Vivien, A. Rudi and F. Bach, published in the *Conference On Learning Theory* in 2018.

The Section 2 and its Appendix B are focus on optimality of SGD in the non-parametric setting. More precisely, this work is the first to show that multiple passes over the data allow to reach optimality in certain cases where the Bayes optimum is hard to approximate. This work tries to reconcile theory and practice as common knowledge on SGD always stated that one pass over the data is optimal. This section is based on our work, **Statistical Optimality of Stochastic Gradient Descent on Hard Learning Problems through Multiple Passes**, L. Pillaud-Vivien, A. Rudi and F. Bach, published in the *Advances in Neural Information Processing Systems* in 2018.

Contents

1 Exponential convergence of testing error for stochastic gradient methods 60
- 1.1 Introduction . . . 60
- 1.2 Problem Set-up . . . 61
- 1.3 Concrete Examples and Related Work . . . 63
- 1.4 Stochastic Gradient descent . . . 64
- 1.5 Exponentially Convergent SGD for Classification error . . . 68
- 1.6 Conclusion . . . 69

A Appendix of Exponential convergence of testing error for stochastic gradient descent 71
- A.1 Experiments . . . 71
- A.2 Probabilistic lemmas . . . 73
- A.3 From $\mathcal{H}$ to 0-1 loss . . . 74
- A.4 Exponential rates for Kernel Ridge Regression . . . 75
- A.5 Proofs and additional results about concrete examples . . . 77
- A.6 Preliminaries for Stochastic Gradient Descent . . . 80
- A.7 Proof of stochastic gradient descent results . . . 81
- A.8 Exponentially convergent SGD for classification error . . . 91
- A.9 Extension of Corollary 1 and Theorem 4 for the full averaged case. . . . 93
- A.10 Convergence rate under weaker margin assumption . . . 97

2 Statistical Optimality of SGD on Hard Learning Problems through Multiple Passes 100
- 2.1 Introduction . . . 100
- 2.2 Least-squares regression in finite dimension . . . 101
- 2.3 Averaged SGD with multiple passes . . . 103
- 2.4 Application to kernel methods . . . 104
- 2.5 Experiments . . . 106
- 2.6 Conclusion . . . 108

B Appendix of Statistical Optimality of SGD on Hard Learning Problems through Multiple Passes 110
- B.1 A general result for the SGD variance term . . . 110
- B.2 Proof sketch for Theorem 8 . . . 114
- B.3 Bounding the deviation between SGD and batch gradient descent . . . 115
- B.4 Convergence of batch gradient descent . . . 117
- B.5 Experiments with different sampling . . . 128

1. Exponential convergence of testing error for stochastic gradient methods

1.1. Introduction

Stochastic gradient methods are now ubiquitous in machine learning, both from the practical side, as a simple algorithm that can learn from a single or a few passes over the data [BLC05], and from the theoretical side, as it leads to optimal rates for estimation problems in a variety of situations [NY83, PJ92].

They follow a simple principle [RM51]: to find a minimizer of a function F defined on a vector space from noisy gradients, simply follow the negative stochastic gradient and the algorithm will converge to a stationary point, local minimum or global minimum of F (depending on the properties of the function F), with a rate of convergence that decays with the number of gradient steps n typically as $O(1/\sqrt{n})$, or $O(1/n)$ depending on the assumptions which are made on the problem [PJ92, NV08, NJLS09, SSSS07, Xia10, BM11, BM13, DFB17].

On the one hand, these rates are optimal for the estimation of the minimizer of a function given access to noisy gradients [NY83], which is essentially the usual machine learning set-up where the function F is the expected *loss*, e.g., logistic or hinge for classification, or least-squares for regression, and the noisy gradients are obtained from sampling a single pair of observations.

On the other hand, although these rates as $O(1/\sqrt{n})$ or $O(1/n)$ are optimal, there are a variety of extra assumptions that allow for faster rates, even exponential rates.

First, for stochastic gradient from a finite pool, that is for $F = \frac{1}{k}\sum_{i=1}^{k} F_i$, a sequence of works starting from SAG [RSB12], SVRG [JZ13], SAGA [DBLJ14], have shown explicit exponential convergence. However, these results, once applied to machine learning where the function F_i is the loss function associated with the i-th observation of a finite training data set of size k, say nothing about the loss on unseen data (test loss). The rates we present in this paper are on *unseen* data.

Second, assuming that at the optimum all stochastic gradients are equal to zero, then for strongly-convex problems (e.g., linear predictions with low-correlated features), linear convergence rates can be obtained for test losses [Sol98, SL13]. However, for supervised machine learning, this has limited relevance as having zero gradients for all stochastic gradients at the optimum essentially implies prediction problems with no uncertainty (that is, the output is a deterministic function of the input). Moreover, we can only get an exponential rate for strongly-convex problems and thus this imposes a parametric noiseless problem, which limits the applicability (even if the problem was noiseless, this can only reasonably be in a non-parametric way with neural networks or positive definite kernels). Our rates are on noisy problems and on infinite-dimensional problems where we can hope that we approach the optimal prediction function with large numbers of observations. For prediction functions described by a reproducing kernel Hilbert space, and for the square loss, the excess testing loss (equal to testing loss minus the minimal testing loss over all measurable prediction functions) is known to converge to zero at a subexponential rate typically greater than $O(1/n)$ [DB16, DFB17], these rates being optimal for the estimation of testing losses.

Going back to the origins of supervised machine learning with binary labels, we will not consider getting to the optimal testing *loss* (using a convex surrogate such as logistic, hinge or least-squares) but the testing *error* (number of mistakes in predictions), also referred to as the 0-1 loss.

It is known that the excess testing error (testing error minus the minimal testing error over all measurable prediction functions) is upper bounded by a function of the excess testing loss [Zha04, BJM06], but always with a loss in the convergence rate (e.g., no difference or taking square roots). Thus a slow rate in $O(1/n)$ or $O(1/\sqrt{n})$ on the excess loss leads to a slow(er) rate on the excess testing error.

Such general relationships between excess loss and excess error have been refined with the use of *margin conditions*, which characterize how hard the prediction problems are [MT99]. Simplest input points

are points where the label is deterministic (i.e., conditional probabilities of the label are equal to zero or one), while hardest points are the ones where the conditional probabilities are equal to $1/2$. Margin conditions quantify the mass of input points which are hardest to predict, and lead to improved transfer functions from testing losses to testing errors, but still no exponential convergence rates [BJM06].

In this paper, we consider the strongest margin condition, that is conditional probabilities are bounded away from $1/2$, but not necessarily equal to 0 or 1. This assumption on the learning problem has been used in the past to show that regularized empirical (convex) risk minimization leads to exponential convergence rates [AT07, KB05]. Our main contribution is to show that stochastic gradient descent also achieves similar rates (see an empirical illustration in Figure 10 in the Appendix A.1). This requires several side contributions that are interesting on their own, that is, a new and simple formalization of the learning problem that allows exponential rates of estimation (regardless of the algorithms used to find the estimator) and a new concentration result for averaged stochastic gradient descent (SGD) applied to least-squares, which is finer than existing work [BM13].

The paper is organized as follows: in Section 1.2, we present the learning set-up, namely binary classification with positive definite kernels, with a particular focus on the relationship between errors and losses. Our main results rely on a generic condition for which we give concrete examples in Section 1.3. In Section 1.4, we present our version of stochastic gradient descent, with the use of tail averaging [JKK+16], and provide new deviation inequalities, which we apply in Section 1.5 to our learning problem, leading to exponential convergence rates for the testing errors. We conclude in Section 1.6 by providing several avenues for future work. Finally, synthetic experiments illustrating our results can be found in Section A.1 of the Appendix.

Main contributions of the paper. We would like to underline that our main contributions are in the two following results; (a) we show in Theorem 4 the exponential convergence of stochastic gradient descent on the testing error, and (b) this result strongly rests on a new deviation inequality stated in Corollary 1 for stochastic gradient descent for least-squares problems. This last result is interesting on its own and gives an improved high-probability result which does not depend on the dimension of the problem and has a tighter dependence on the strongly convex parameter –through the effective dimension of the problem, see [CDV07, DB16].

1.2. Problem Set-up

In this section, we present the general machine learning set-up, from generic assumptions to more specific assumptions.

1.2.1. Generic assumptions

We consider a measurable set $\mathcal{X}$ and a probability distribution ρ on data $(x, y) \in \mathcal{X} \times \{-1, 1\}$; we denote by $\rho_{\mathcal{X}}$ the marginal probability on x, and by $\rho(\pm 1|x)$ the conditional probability that $y = \pm 1$ given x. We have $\mathbb{E}(y|x) = \rho(1|x) - \rho(-1|x)$. Our main margin condition is the following (and independent of the learning framework):

(A1) $|\mathbb{E}(y|x)| \geqslant \delta$ *almost surely for some* $\delta \in (0, 1]$.

This margin condition (often referred to as a low-noise condition) is commonly used in the theoretical study of binary classification [MT99, AT07, KB05], and usually takes the following form: $\forall \delta > 0$, $\mathbb{P}(|\mathbb{E}(y|x)| < \delta) = O(\delta^{\alpha})$ for $\alpha > 0$. Here, however, δ is a fixed constant. Our stronger margin condition **(A1)** is necessary to show exponential convergence rates but we give also explicit rates in the case of the latter low-noise condition. This extension is derived in Appendix A.10 and more precisely in Corollary 4. Note that the smaller the α, the larger the mass of inputs with hard-to-predict labels. Our condition corresponds

to $\alpha = +\infty$, and simply states that for all inputs, the problem is never totally ambiguous, and the degree of non-ambiguity is bounded from below by δ. When $\delta = 1$, then the label $y \in \{-1, 1\}$ is a deterministic function of x, but our results apply for all $\delta \in (0, 1]$ and thus to noisy problems (with low noise). Note that problems like image classification or object recognition are well characterized by **(A1)**. Indeed, the noise in classifying an image between two disparate classes (cars/pedestrians, bikes/airplanes) is usually way smaller that $1/2$.

We will consider learning functions in a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ with kernel function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and dot-product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. We make the following standard assumptions on $\mathcal{H}$:

(A2) *$\mathcal{H}$ is a separable Hilbert space and there exists $R > 0$, such that for all $x \in \mathcal{X}$, $K(x, x) \leqslant R^2$.*

For $x \in \mathcal{X}$, we consider the function $K_x : \mathcal{X} \to \mathbb{R}$ defined as $K_x(x') = K(x, x')$. We have the classical reproducing property for $g \in \mathcal{H}$, $g(x) = \langle g, K_x \rangle_{\mathcal{H}}$ [STC04, SS02]. We will consider other norms, beyond the RKHS norm $\|g\|_{\mathcal{H}}$, that is the L_2-norm (always with respect to $\rho_{\mathcal{X}}$), defined as $\|g\|_{L_2}^2 = \int_{\mathcal{X}} g(x)^2 d\rho_{\mathcal{X}}(x)$, as well as the L_∞-norm $\| \cdot \|_{L_\infty}$ on the support of $\rho_{\mathcal{X}}$. A key property is that **(A2)** implies $\|g\|_{L_\infty} \leqslant R \|g\|_{\mathcal{H}}$.

Finally, we will consider observations with standard assumptions:

(A3) *The observations $(x_n, y_n) \in \mathcal{X} \times \{-1, 1\}, n \in \mathbb{N}^*$ are independent and identically distributed with respect to the distribution ρ.*

1.2.2. Ridge regression

In this paper, we focus primarily on least-squares estimation to obtain estimators. We define g_* as the minimizer over L_2 of

$$\mathbb{E}(y - g(x))^2 = \int_{\mathcal{X} \times \{-1,1\}} (y - g(x))^2 d\rho(x, y).$$

We always have $g_*(x) = \mathbb{E}(y|x) = \rho(1|x) - \rho(-1|x)$, but we *do not require* $g_* \in \mathcal{H}$. We also consider the ridge regression problem [CDV07] and denote by g_λ the unique (when $\lambda > 0$) minimizer in $\mathcal{H}$ of

$$\mathbb{E}(y - g(x))^2 + \lambda \|g\|_{\mathcal{H}}^2.$$

The function g_λ always exists for $\lambda > 0$ and is always an element of $\mathcal{H}$. When $\mathcal{H}$ is dense in L_2 our results depend on the L_∞-error $\|g_\lambda - g_*\|_\infty$, which is weaker than $\|g_\lambda - g_*\|_{\mathcal{H}}$ which itself only exists when $g_* \in \mathcal{H}$ (which we do not assume). When $\mathcal{H}$ is not dense we simply define $\tilde{g}_*$ as the orthonormal projector for the L_2 norm on $\mathcal{H}$ of $g_* = \mathbb{E}(y|x)$ so that our bound will the depend on $\|g_\lambda - \tilde{g}_*\|_\infty$. Note that $\tilde{g}_*$ is the minimizer of $\mathbb{E}(y - g(x))^2$ with respect to g in the closure of $\mathcal{H}$ in L_2.

Moreover our main technical assumption is:

(A4) *There exists $\lambda > 0$ such that almost surely,* $\operatorname{sign}(\mathbb{E}(y|x)) g_\lambda(x) \geqslant \dfrac{\delta}{2}$.

In the assumption above, we could replace $\delta/2$ by any multiplicative constants in $(0, 1)$ times δ (instead of $1/2$). Note that with **(A4)**, λ depends on δ and on the probability measure ρ, which are both fixed (respectively by **(A1)** and the problem), so that λ is fixed too. It implies that for any estimator $\hat{g}$ such that $\|g_\lambda - \hat{g}\|_{L_\infty} < \delta/2$, the predictions from $\hat{g}$ (obtained by taking the sign of $\hat{g}(x)$ for any x), are the same as the sign of the optimal prediction $\operatorname{sign}(\mathbb{E}(y|x))$. Note that a sufficient condition is $\|g_\lambda - \hat{g}\|_{\mathcal{H}} < \delta/(2R)$ (which does not assume that $g_* \in \mathcal{H}$), see next subsection.

Note that more generally, for all problems for which **(A1)** is true and ridge regression (in the population case) is so that $\|g_\lambda - g_*\|_{L_\infty}$ tends to zero as λ tends to zero then **(A4)** is satisfied, since $\|g_\lambda - g_*\|_{L_\infty} \leqslant \delta/2$ for λ small enough, together with **(A1)** then implies **(A4)**.

In Section 1.3, we provide concrete examples where **(A4)** is satisfied and we then present the SGD algorithm and our convergence results. Before we relate excess testing losses to excess testing errors.

1.2.3. From testing losses to testing error

Here we provide some results that will be useful to prove exponential rates for classification with squared loss and stochastic gradient descent. First we define the 0-1 loss defining the classification error:

$$\mathcal{R}(g) = \rho(\{(x, y) : \operatorname{sign}(g(x)) \neq y\}),$$

where $\operatorname{sign} u = +1$ for $u \geq 0$ and -1 for $u < 0$. In particular denote by $\mathcal{R}^*$ the so-called *Bayes risk* $\mathcal{R}^* = \mathcal{R}(E[y|x])$ which is the minimum achievable classification error [DGL13].

A well known approach to bound the testing errors by testing losses is *via transfer functions*. In particular we recall the following result [DGL13, BJM06], let $g_*(x)$ be equal to $E[y|x]$ a.e., then

$$\mathcal{R}(g) - \mathcal{R}^* \leq \phi(\|g - g_*\|_{L^2}^2), \qquad \forall g \in L^2(d\rho_X),$$

with $\phi(u) = \sqrt{u}$ (or $\phi(u) = u^\beta$, with $\beta \in [1/2, 1]$, depending on some properties of ρ [BJM06]. While this result does not require **(A1)** or **(A4)**, it does not readily lead to exponential rates since the squared loss excess risk has minimax lower bounds that are polynomial in n [CDV07].

Here we follow a different approach, requiring via **(A4)** the existence of g_λ having the same sign as g_* and with absolute value uniformly bounded from below. Then we can bound the 0-1 error with respect to the distance in $\mathcal{H}$ of the estimator $\widehat{g}$ from g_λ as shown in the next lemma (proof in Appendix A.3). This will lead to exponential rates when the distribution satisfies a margin condition **(A1)** as we prove in the next section and in Section 1.5. Note also that for the sake of completeness we recalled in Appendix A.4 that exponential rates could be achieved for kernel ridge regression.

Lemma 1 (*From approximately correct sign to 0-1 error*)

*Let $q \in (0, 1)$. Under **(A1)**, **(A2)**, **(A4)**, $\widehat{g} \in \mathcal{H}$ a random function such that $\|\widehat{g} - g_\lambda\|_{\mathcal{H}} < \frac{\delta}{2R}$, with probability at least $1 - q$. Then*

$$\mathcal{R}(\widehat{g}) = \mathcal{R}^*, \text{ with probability at least } 1 - q, \text{ and in particular } \mathbb{E}\mathcal{R}(\widehat{g}) - \mathcal{R}^* \leq q.$$

In the next section we provide sufficient conditions and explicit settings naturally satisfying **(A4)**.

1.3. Concrete Examples and Related Work

In this section we illustrate specific settings that naturally satisfy **(A4)**. We start by the following simple result showing that the existence of $g_* \in \mathcal{H}$ such that $g_*(x) = E[y|x]$ a.e. on the support of ρ_X, is sufficient to have **(A4)** (proof in Appendix A.5.1).

Proposition 3

*Under **(A1)**, assume that there exists $g_* \in \mathcal{H}$ such that $g_*(x) := E[y|x]$ on the support of ρ_X, then for any δ, there exists $\lambda > 0$ satisfying **(A4)**, that is, $\operatorname{sign}(\mathbb{E}(y|x))g_\lambda(x) \geqslant \frac{\delta}{2}$.*

We are going to use the proposition above to derive more specific settings. In particular we consider the case where the positive and negative classes are separated by a margin that is strictly positive. Let $\mathcal{X} \subseteq \mathbb{R}^d$ and denote by $\mathcal{S}$ the support of the probability ρ_X and by $\mathcal{S}_+ = \{x \in \mathcal{X} : g_*(x) > 0\}$ the part associated to the positive class, and by $\mathcal{S}_-$ the one associated with the negative class. Consider the following assumption:

(A5) *There exists $\mu > 0$ such that $\min_{x \in \mathcal{S}_+, x' \in \mathcal{S}_-} \|x - x'\| \geq \mu$.*

Denote by $W^{s,2}$ the Sobolev space of order s defined with respect to the L^2 norm, on $\mathbb{R}^d$ (see [AF03] and Appendix A.5.2). We also introduce the following assumption:

(A6) *$\mathcal{X} \subseteq \mathbb{R}^d$ and the kernel is such that $W^{s,2} \subseteq \mathcal{H}$, with $s > d/2$.*

An example of kernel such that $\mathcal{H} = W^{s,2}$, with $s > d/2$ is the Abel kernel $K(x, x') = e^{-\frac{1}{\sigma}\|x-x'\|}$, for $\sigma > 0$. In the following proposition we show that if there exist two functions in $\mathcal{H}$, one matching $E[y|x]$ on $\mathcal{S}_+$ and the second matching $E[y|x]$ on $\mathcal{S}_-$ and if the kernel satisfies **(A6)**, then **(A4)** is satisfied.

Proposition 4

*Under **(A1)**, **(A5)**, **(A6)**, if there exist two functions $g_+^*, g_-^* \in W^{s,2}$ such that $g_+^*(x) = E[y|x]$ on $\mathcal{S}_+$ and $g_-^*(x) = E[y|x]$ on $\mathcal{S}_-$, then **(A4)** is satisfied.*

Finally we are able to introduce another setting where **(A4)** is naturally satisfied (the proof of the proposition above and the example below are given in Appendix A.5.2).

Example 1 (*Independent noise on the labels*)

*Let $\rho_{\mathcal{X}}$ be a probability distribution on $\mathcal{X} \subseteq \mathbb{R}^d$ and let $\mathcal{S}_+, \mathcal{S}_- \subseteq \mathcal{X}$ be a partition of the support of $\rho_{\mathcal{X}}$ satisfying $\rho_{\mathcal{X}}(S_+), \rho_{\mathcal{X}}(S_-) > 0$ and **(A5)**. Let $n \in \mathbb{N}^*$. For $1 \leq i \leq n$, x_i independently sampled from $\rho_{\mathcal{X}}$ and the label y_i defined by the law*

$$y_i = \begin{cases} \zeta_i & \text{if } x_i \in S_+ \\ -\zeta_i & \text{if } x_i \in S_-, \end{cases}$$

*with ζ_i independently distributed as $\zeta_i = -1$ with probability $p \in [0, 1/2)$ and $\zeta_i = 1$ with probability $1 - p$. Then **(A1)** is satisfied with $\delta = 1 - 2p$ and **(A4)** is satisfied as soon as **(A2)** and **(A6)** are, that is, the kernel is bounded and $\mathcal{H}$ is rich enough (see an example in Appendix A.5 Figure 12).*

Finally note that the results of this section can be easily generalized from $\mathcal{X} = \mathbb{R}^d$ to any Polish space, by using a *separating* kernel [DVRT14, RCDVR14] instead of **(A6)**.

1.4. Stochastic Gradient descent

We now consider the stochastic gradient algorithm to solve the ridge regression problem with a fixed strictly positive regularization parameter λ. We consider solving the regularized problem with regularization $\|g - g_0\|_{\mathcal{H}}^2$ through stochastic approximation starting from a function $g_0 \in \mathcal{H}$ (typically 0).[1] Denote by $F : \mathcal{H} \to \mathbb{R}$, the functional

$$F(g) = \mathbb{E}(Y - g(X))^2 = \mathbb{E}(Y - \langle K_X, g\rangle)^2,$$

where the last identity is due to the reproducing property of the RKHS $\mathcal{H}$. Note that F has the following gradient $\nabla F(g) = -2\mathbb{E}\left[(Y - \langle K_X, g\rangle)K_X\right]$. We consider also $F_\lambda = F + \lambda\|\cdot - g_0\|_{\mathcal{H}}^2$, for which $\nabla F_\lambda(g) = \nabla F(g) + 2\lambda(g - g_0)$, and we have for each pair of observation (x_n, y_n) that $F_\lambda(g) = \mathbb{E}\big[F_{n,\lambda}(g)\big] = \mathbb{E}(\langle g, K_{x_n}\rangle - y_n)^2 + \lambda\|g - g_0\|_{\mathcal{H}}^2$, with $F_{n,\lambda}(g) = (\langle g, K_{x_n}\rangle - y_n)^2 + \lambda\|g - g_0\|_{\mathcal{H}}^2$.

Denoting $\Sigma = \mathbb{E}\big[K_{x_n} \otimes K_{x_n}\big]$ the covariance operator defined as a linear operator from $\mathcal{H}$ to $\mathcal{H}$ (see [FBJ04] and references therein), we have the optimality conditions for g_λ and $\tilde{g}_*$:

$$\Sigma g_\lambda - \mathbb{E}\left(y_n K_{x_n}\right) + \lambda(g_\lambda - g_0) = 0, \qquad \mathbb{E}\left[(y_n - \tilde{g}_*(x_n))\, K_{x_n}\right] = 0,$$

see [CDV07] or Appendix A.6.1 for the proof of the last identity. Let $(\gamma_n)_{n\geqslant 1}$ be a positive sequence; we consider the stochastic gradient recursion[2] in $\mathcal{H}$ started at g_0:

$$g_n = g_{n-1} - \frac{\gamma_n}{2}\nabla F_{n,\lambda}(g_{n-1}) = g_{n-1} - \gamma_n\left[(\langle K_{x_n}, g_{n-1}\rangle - y_n)K_{x_n} + \lambda(g_{n-1} - g_0)\right]. \tag{55}$$

[1] Note that g_0 is the initialization of the recursion, and is not the limit of g_λ when λ tends to zero (this limit being $\tilde{g}_*$).

[2] The complexity of n steps of the recursion is $O(n^2)$ if using kernel functions or $O(\tau n)$ when using explicit feature representations, with τ the complexity of computing dot-products and adding feature vectors.

We are going to consider Polyak-Ruppert averaging [PJ92], that is $\bar{g}_n = \frac{1}{n+1}\sum_{i=0}^n g_i$, as well as the tail-averaging estimate $\bar{g}_n^{\text{tail}} = \frac{1}{\lfloor n/2 \rfloor}\sum_{i=\lfloor n/2 \rfloor}^n g_i$, studied by [JKK+16]. For the sake of clarity, all the results in the main text are for the tail averaged estimate but note that all of them have been also proved for the full average in Appendix A.9.

As explained earlier (see Lemma 1), we need to show the convergence of g_n to g_λ in $\mathcal{H}$-norm. We are going to consider two cases: (1) for the non-averaged recursion (γ_n) is a decreasing sequence, with the important particular case $\gamma_n = \gamma/n^\alpha$, for $\alpha \in [0,1]$; (2) for the averaged or tail-averaged functions (γ_n) is a constant sequence equal to γ. For all the proofs of this section see Appendix A.7. In the next subsection we reformulate the recursion in Eq. (55) as a least-squares recursion converging to g_λ.

1.4.1. Reformulation as noisy recursion

We can first reformulate the SGD recursion equation in Eq. (55) as a regular least-squares SGD recursion with noise, with the notation $\xi_n = y_n - \tilde{g}_*(x_n)$, which satisfies $\mathbb{E}\big[\xi_n K_{x_n}\big] = 0$. This is the object of the following lemma (for the proof see Appendix A.6.2.):

Lemma 2

The SGD recursion can be rewritten as follows:

$$g_n - g_\lambda = \big[I - \gamma_n(K_{x_n}\otimes K_{x_n} + \lambda I)\big](g_{n-1} - g_\lambda) + \gamma_n \varepsilon_n, \tag{56}$$

with the noise term $\varepsilon_k = \xi_k K_{x_k} + (\tilde{g}_*(x_k) - g_\lambda(x_k))K_{x_k} - \mathbb{E}\left[(\tilde{g}_*(x_k) - g_\lambda(x_k))K_{x_k}\right] \in \mathcal{H}$.

We are thus in presence of a least-squares problem in the Hilbert space $\mathcal{H}$, to estimate a function $g_\lambda \in \mathcal{H}$ with a specific noise ε_n in the gradient and feature vector K_x. In the next section, we will consider the generic recursion above, which will require some bounds on the noise. In our setting, we have the following almost sure bounds and the noise (see Lemma 9 of Appendix A.7):

$$\begin{aligned}\|\varepsilon_n\|_{\mathcal{H}} &\leqslant R(1 + 2\|\tilde{g}_* - g_\lambda\|_{L_\infty}) \\ \mathbb{E}\big[\varepsilon_n \otimes \varepsilon_n\big] &\preccurlyeq 2\left(1 + \|\tilde{g}_* - g_\lambda\|_\infty^2\right)\Sigma,\end{aligned}$$

where $\Sigma = \mathbb{E}\big[K_{x_n}\otimes K_{x_n}\big]$ is the covariance operator.

1.4.2. SGD for general Least-Square problems

We now consider results on (averaged) SGD for least-squares that are interesting on their own. As said before, we show results in two different settings depending on the step-size sequence. First, we consider (γ_n) as a decreasing sequence, second we take (γ_n) constant but prove the convergence of the (tail-)averaged iterates.

Since the results we need could be of interest (even for finite-dimensional models), in this section, we study the following general recursion:

$$\eta_n = (I - \gamma H_n)\eta_{n-1} + \gamma_n \varepsilon_n, \tag{57}$$

We make the following assumptions:

(H1) *We start at some* $\eta_0 \in \mathcal{H}$.

(H2) $(H_n, \varepsilon_n)_{n\geqslant 1}$ *are i.i.d. and* H_n *is a positive self-adjoint operator so that almost surely* $H_n \succcurlyeq \lambda I$, *and* $H := \mathbb{E}H_n$.

(H3) *Noise:* $\mathbb{E}\varepsilon_n = 0$, $\|\varepsilon_n\|_{\mathcal{H}} \leqslant c^{1/2}$ *almost surely and* $\mathbb{E}(\varepsilon_n \otimes \varepsilon_n) \preccurlyeq C$, *with* C *commuting with* H. *Note that one consequence of this assumption is* $\mathbb{E}\|\varepsilon_n\|_{\mathcal{H}}^2 \leqslant \operatorname{tr} C$.

(H4) *For all* $n \geqslant 1$, $\mathbb{E}\left[H_n C H^{-1} H_n\right] \preccurlyeq \gamma_0^{-1} C$ *and* $\gamma \leqslant \gamma_0$.

(H5) A *is a positive self-adjoint operator which commutes with* H.

Note that we will later apply the results of this section to $H_n = K_{x_n} \otimes K_{x_n} + \lambda I$, $H = \Sigma + \lambda I$, $C = \Sigma$ and $A \in \{I, \Sigma\}$. We first consider the non-averaged SGD recursion, then the (tail-)averaged recursion. The key difference with existing bounds is the need for precise probabilistic deviation results.

For least-squares, one can always separate the impact of the initial condition η_0 and of the noise terms ε_k, namely $\eta_n = \eta_n^{\text{bias}} + \eta_n^{\text{variance}}$, where η_n^{bias} is the recursion with no noise ($\varepsilon_k = 0$), and η_n^{variance} is the recursion started at $\eta_0 = 0$. The final performance will be bounded by the sum of the two separate performances (see, e.g.,[DB15]). Hence all of our bounds will depend on these two. See more details in Appendix A.7.

1.4.3. Non-averaged SGD

In this section, we prove results for the recursion defined by Eq. (57) in the case where for $\alpha \in [0, 1]$, $\gamma_n = \gamma / n^\alpha$. These results extend the ones of [BM11] by providing deviation inequalities, but are limited to least-squares. For general loss functions and in the strongly-convex case, see also [KT09].

Theorem 1 (*SGD, decreasing step size:* $\gamma_n = \gamma / n^\alpha$)

Assume ***(H1)***, ***(H2)***, ***(H3)***, $\gamma_n = \gamma / n^\alpha$, $\gamma\lambda < 1$ *and denote by* $\eta_n \in \mathcal{H}$ *the n-th iterate of the recursion in Eq.* (57). *We have for* $t > 0, n \geqslant 1$ *and* $\alpha \in (0, 1)$,

$$\|g_n - g_\lambda\|_{\mathcal{H}} \leqslant \exp\left(-\frac{\gamma\lambda}{1-\alpha}\left((n+1)^{1-\alpha} - 1\right)\right)\|g_0 - g_\lambda\|_{\mathcal{H}} + V_n,$$

almost surely for n *large enough* [3], *with* $\mathbb{P}\left(V_n \geqslant t\right) \leqslant 2\exp\left(-\frac{t^2}{8\gamma \mathrm{tr} C/\lambda + \gamma c^{1/2} t} \cdot n^\alpha\right)$.

We can make the following observations:

- The proof technique (see Appendix A.7.1 for the detailed proof) relies on the following scheme: we notice that η_n can be decomposed in two terms, (a) the bias: obtained from a product of n contractant operators, and (b) the variance: a sum of increments of a martingale. We treat separately the two terms. For the second one, we prove almost sure bounds on the increments and on the variance that lead to a Bernstein-type concentration result on the tail $\mathbb{P}(V_n \geqslant t)$. Following this proof technique, the coefficient in the latter exponential is composed of the variance bound plus the almost sure bound of the increments of martingale times t.
- Note that we only presented in Theorem 1 the case where $\alpha \in (0, 1)$. Indeed, we only focused on the case where we had exponential convergence (see the whole result in the Appendix: Proposition 8). Actually, that there are three different regimes. For $\alpha = 0$ (constant step-size), the algorithm is not converging, as the tail probability bound on $\mathbb{P}\left(V_n \geqslant t\right)$ is not dependent on n. For $\alpha = 1$, confirming results from [BM11], there is no exponential forgetting of initial conditions. And for $\alpha \in (0, 1)$, the forgetting of initial conditions and the tail probability are converging to zero exponentially fast, respectively, as $\exp(-Cn^{1-\alpha})$ and $\exp(-Cn^\alpha)$, for a constant C, hence the natural choice of $\alpha = 1/2$ in our experiments.

1.4.4. Averaged and Tail-averaged SGD with constant step-size

In the subsection, we take: $\forall n \geqslant 1,\ \gamma_n = \gamma$. We first start with a result on the variance term, whose proof extends the work of [DFB17] to deviation inequalities which are sharper than the ones from [BM13].

Theorem 2 (*Convergence of the variance term in averaged SGD*)

Assume (H1), (H2), (H3), (H4), (H5) and consider the average of the $n+1$ first iterates of the sequence defined in Eq. (57): $\bar{\eta}_n = \frac{1}{n+1}\sum_{i=0}^{n}\eta_i$. Assume $\eta_0 = 0$. We have for $t > 0, n \geqslant 1$:

$$\mathbb{P}\left(\left\|A^{1/2}\bar{\eta}_n\right\|_{\mathcal{H}} \geqslant t\right) \leqslant 2\exp\left[-\frac{(n+1)t^2}{E_t}\right], \tag{58}$$

where E_t is defined with respect to the constants introduced in the assumptions:

$$E_t = 4\mathrm{tr}(AH^{-2}C) + \frac{2c^{1/2}\|A^{1/2}\|_{op}}{3\lambda}\cdot t. \tag{59}$$

The work that remains to be done is to bound the bias term of the recursion $\bar{\eta}_n^{\text{bias}}$. We have done it for the full averaged sequence (see Appendix A.9.1 Theorem 6) but as it is quite technical and could lower a bit the clarity of the reasoning, we have decided to leave it in the Appendix. We present here another approach and consider the tail-averaged recursion, $\bar{\eta}_n^{\text{tail}} = \frac{1}{\lfloor n/2\rfloor}\sum_{i=\lfloor n/2\rfloor}^{n}\eta_i$ (as proposed by [JKK+16, Sha11]). For this, we use the simple almost sure bound $\|\eta_i^{\text{bias}}\|_{\mathcal{H}} \leqslant (1-\lambda\gamma)^i\|\eta_0\|_{\mathcal{H}}$, such that $\|\bar{\eta}_n^{\text{tail, bias}}\|_{\mathcal{H}} \leqslant (1-\lambda\gamma)^{n/2}\|\eta_0\|_{\mathcal{H}}$. For the variance term, we can simply use the result above for n and $n/2$, as $\bar{\eta}_n^{\text{tail}} = 2\bar{\eta}_n - \bar{\eta}_{n/2}$. This leads to:

Corollary 1 (*Convergence of tail-averaged SGD*)

Assume (H1), (H2), (H3), (H4), (H5) and consider the tail-average of the sequence defined in Eq. (57): $\bar{\eta}_n^{tail} = \frac{1}{\lfloor n/2\rfloor}\sum_{i=\lfloor n/2\rfloor}^{n}\eta_i$. We have for $t > 0, n \geqslant 1$:

$$\left\|A^{1/2}\bar{\eta}_n^{tail}\right\|_{\mathcal{H}} \leqslant (1-\gamma\lambda)^{n/2}\|A^{1/2}\|_{op}\|\eta_0\|_{\mathcal{H}} + L_n \quad , \text{ with} \tag{60}$$

$$\mathbb{P}(L_n \geqslant t) \leqslant 4\exp\left(-(n+1)t^2/(4E_t)\right), \tag{61}$$

where L_n is defined in the proof (see Appendix A.7.3) and is the variance term of the tail-averaged recursion.

We can make the following observations on the two previous results:

- The proof technique (see Appendix A.7.2 and A.7.3 for the detailed proofs) relies on concentration inequality of Bernstein type. Indeed, we notice that (in the setting of Theorem 2) $\bar{\eta}_n$ is a sum of increments of a martingale. We prove almost sure bounds on the increments and on the variance (following the proof technique of [DFB17]) that lead to a Bernstein type concentration result on the tail $\mathbb{P}(V_n \geqslant t)$. Following the proof technique summed-up before, we see that E_t is composed of the variance bound plus the almost sure bound times t.
- Remark that classically, A and C are proportional to H for excess risk predictions. In the finite d-dimensional setting this leads us to the usual variance bound proportional to the dimension d: $\mathrm{tr}(AH^{-2}C) \cong \mathrm{tr}I = d$. The result is general in the sense that we can apply it for all matrices A commuting with H (this can be used to prove results in L_2 or in $\mathcal{H}$).
- Finally, note that we improved the variance bound with respect to the strong convexity parameter λ which is usually of the order $1/\lambda^2$ (see [Sha11]), and is here $\mathrm{tr}(AH^{-2}C)$. Indeed, in our setting, we will apply it for $A = C = \Sigma$ and $H = \Sigma + \lambda I$, so that $\mathrm{tr}(AH^{-2}C)$ is upper bounded by the effective dimension $\mathrm{tr}(\Sigma(\Sigma+\lambda I)^{-1})$ which can be way smaller than $1/\lambda^2$ (see [CDV07, DB16]).
- The complete proof for the full average is written in Appendix A.9.1 and more precisely in Theorem 6. In this case the initial conditions are not forgotten exponentially fast though.

1.5. Exponentially Convergent SGD for Classification error

In this section we want to show our main results, on the error made (on unseen data) by the n-th iterate of the regularized SGD algorithm. Hence, we go back to the original SGD recursion defined in Eq. (56). Let us recall it:

$$g_n - g_\lambda = \left[I - \gamma_n(K_{x_n} \otimes K_{x_n} + \lambda I)\right](g_{n-1} - g_\lambda) + \gamma_n \varepsilon_n,$$

with the noise term $\varepsilon_k = \xi_k K_{x_k} + (\tilde{g}_*(x_k) - g_\lambda(x_k))K_{x_k} - \mathbb{E}\left[(\tilde{g}_*(x_k) - g_\lambda(x_k))K_{x_k}\right] \in \mathcal{H}$. Like in the previous section we are going to state two results in two different settings, the first one for SGD with decreasing step-size ($\gamma_n = \gamma/n^\alpha$) and the second one for tail averaged SGD with constant step-size. For all the proofs of this section see the Appendix (section A.8).

1.5.1. SGD with decreasing step-size

In this section, we focus on decreasing step-sizes $\gamma_n = \gamma/n^\alpha$ for $\alpha \in (0,1)$, which lead to exponential convergence rates. Results for $\alpha = 1$ and $\alpha = 0$ can be derived in a similar way (but do not lead to exponential rates).

Theorem 3

Assume (A1), (A2), (A3), (A4) and $\gamma_n = \gamma/n^\alpha$, $\alpha \in (0,1)$ for any n and $\gamma\lambda < 1$. Let g_n be the n-th iterate of the recursion defined in Eq. (56), as soon as n satisfies $\exp\left(-\frac{\gamma\lambda}{1-\alpha}\left((n+1)^{1-\alpha} - 1\right)\right) \leqslant \delta/(5R\|g_0 - g_\lambda\|_{\mathcal{H}})$, then

$$\mathcal{R}(g_n) = \mathcal{R}^*, \text{ with probability at least } 1 - 2\exp\left(-\frac{\delta^2}{C_R}\cdot n^\alpha\right),$$

with $C_R = 2^{\alpha+7}\gamma R^2 \mathrm{tr}\Sigma\left(1 + \|\tilde{g}_* - g_\lambda\|_\infty^2\right)/\lambda + 8\gamma R^2\delta(1 + 2\|\tilde{g}_* - g_\lambda\|_\infty)/3$, and in particular

$$\mathbb{E}\mathcal{R}(g_n) - \mathcal{R}^* \leqslant 2\exp\left(-\frac{\delta^2}{C_R}\cdot n^\alpha\right).$$

Note that Theorem 3 shows that with probability at least $1 - 2\exp\left(-\frac{\delta^2}{C_R}\cdot n^\alpha\right)$, the predictions of g_n are perfect. We can also make the following observations:

- The idea of the proof (see Appendix A.8.1 for the detailed proof) is the following: we know that as soon as $\|g_n - g_\lambda\|_{\mathcal{H}} \leqslant \delta/(2R)$, the predictions of g_n are perfect (Lemma 1). We just have to apply Theorem 1 for to the original SGD recursion and make sure to bound each term by $\delta/(4R)$. Similar results for non-averaged SGD could be derived beyond least-squares (e.g., hinge or logistic loss) using results from [KT09].
- Also note that the larger the α, the smaller the bound. However, it is only valid for n larger that a certain quantity depending of $\lambda\gamma$. A good trade-off is $\alpha = 1/2$, for which we get an excess error of $2\exp\left(-\frac{\delta^2}{C_R}n^{1/2}\right)$, which is valid as soon as $n \geqslant \log(10R\|g_0 - g_\lambda\|_{\mathcal{H}}/\delta)/(4\lambda^2\gamma^2)$. Notice also that we should go for large $\gamma\lambda$ to increase the factor in the exponential and make the condition happen as soon as possible.
- If we want to emphasize the dependence of the bound on the important parameters, we can write that: $\mathbb{E}\mathcal{R}(g_n) - \mathcal{R}^* \lesssim 2\exp\left(-\lambda\delta^2 n^\alpha/R^2\right)$.

- When the condition on n is not met, then we still have the usual bound obtained by taking directly the excess loss [BJM06] but we lose exponential convergence.

1.5.2. Tail averaged SGD with constant step-size

We now consider the tail-averaged recursion[4], with the following result:

Theorem 4

Assume (A1), (A2), (A3), (A4) and $\gamma_n = \gamma$ for any n, $\gamma\lambda < 1$ and $\gamma \leqslant \gamma_0 = (R^2 + 2\lambda)^{-1}$. Let g_n be the n-th iterate of the recursion defined in Eq. (56), and $\bar{g}_n^{tail} = \frac{1}{\lfloor n/2 \rfloor} \sum_{i=\lfloor n/2 \rfloor}^{n} g_i$, as soon as $n \geqslant 2/(\gamma\lambda) \ln(5R\|g_0 - g_\lambda\|_{\mathcal{H}}/\delta)$, then

$$\mathcal{R}(\bar{g}_n^{tail}) = \mathcal{R}^*, \text{ with probability at least } 1 - 4\exp\left(-\delta^2 K_R(n+1)\right),$$

with $K_R^{-1} = 2^9 R^2 \left(1 + \|\tilde{g}_* - g_\lambda\|_\infty^2\right) \mathrm{tr}(\Sigma(\Sigma + \lambda I)^{-2}) + 32\delta R^2(1 + 2\|\tilde{g}_* - g_\lambda\|_\infty)/(3\lambda)$, and in particular

$$\mathbb{E}\mathcal{R}(\bar{g}_n^{tail}) - \mathcal{R}^* \leqslant 4\exp\left(-\delta^2 K_R(n+1)\right).$$

Theorem 4 shows that with probability at least $1 - 4\exp\left(-\delta^2 K_R(n+1)\right)$, the predictions of $\bar{g}_n^{\text{tail}}$ are perfect. We can also make the following observations:

- The idea of the proof (see Appendix A.8.2 for the detailed proof) is the following: we know that as soon as $\|\bar{g}_n^{\text{tail}} - g_\lambda\|_{\mathcal{H}} \leqslant \delta/(2R)$, the predictions of $\bar{g}_n^{\text{tail}}$ are perfect (Lemma 1). We just have to apply Corollary 1 to the original SGD recursion, and make sure to bound each term by $\delta/(4R)$.
- If we want to emphasize the dependence of the bound on the important parameters, we can write that: $\mathbb{E}\mathcal{R}(g_n) - \mathcal{R}^* \lesssim 2\exp\left(-\lambda^2\delta^2 n/R^4\right)$. Note that the λ^2 could be made much smaller with assumptions on the decrease of eigenvalues of Σ (it has been shown [CDV07] that if the decay happens at speed $1/n^\beta$: $\mathrm{tr}\Sigma(\Sigma + \lambda I)^{-2} \leqslant \lambda^{-1}\mathrm{tr}\Sigma(\Sigma + \lambda I)^{-1} \leqslant R^2/\lambda^{1+1/\beta}$).
- We want to take $\gamma\lambda$ as big as possible to satisfy quickly the condition. In comparison to the convergence rate in the case of decreasing step-sizes, the dependence on n is improved as the convergence is really an exponential of n (and not of some power of n as in the previous result).
- Finally, the complete proof for the full average is contained in Appendix A.9.2 and more precisely in Theorem 7.

1.6. Conclusion

In this paper, we have shown that stochastic gradient could be exponentially convergent, once some margin conditions are assumed; and even if a weaker margin condition is assumed, fast rates can be achieved (see Appendix A.10). This is obtained by running averaged stochastic gradient on a least-squares problem, and proving new deviation inequalities.

Our work could be extended in several natural ways: (a) our work relies on new concentration results for the least-mean-squares algorithm (i.e., SGD for square loss), it is natural to extend it to other losses, such as the logistic or hinge loss; (b) going beyond binary classification is also natural with the square loss [CRR16, OBLJ17] or without [TCKG05]; (c) in our experiments, we use regularization, but we have experimented with unregularized recursions, which do exhibit fast convergence, but for which proofs

[4] The full averaging result corresponding to Theorem 4 is proved in Appendix A.9.2, Theorem 7.

are usually harder [DB16]; finally, (d) in order to avoid the $O(n^2)$ complexity, extending the results of [RCR17, RR17] would lead to a subquadratic complexity.

⋆

⋆ ⋆

A. Appendix of Exponential convergence of testing error for stochastic gradient descent

A.1. *Experiments*
where the experiments and their settings are explained.

A.2. *Probabilistic lemmas*
where concentration inequalities in Hilbert spaces used in section A.7 are recalled.

A.3. *From $\mathcal{H}$ to 0-1 loss*
where, from high probability bound for $\|\cdot\|_{\mathcal{H}}$, we derived bound for the 0-1 error.

A.4. *Proofs of Exponential rates for Kernel Ridge Regression*
where exponential rates for Kernel Ridge Regression are proven (Theorem 5).

A.5. *Proofs and additional results about concrete examples*
where additional results and croncrete examples to satisfy **(A4)** are given.

A.6. *Preliminaries for Stochastic Gradient Descent*
where the SGD recursion is derived.

A.7. *Proof of stochastic gradient descent results*
where high probability bounds for the general SGD recursion are shown (Theorems 1 and 2).

A.8. *Exponentially convergent SGD for classification error*
where exponential convergence of test error are shown (Theorems 3 and 4).

A.9. *Extension for the full averaged case*
where previous results are extended for full averaged SGD (instead of tail-averaged).

A.10. *Convergence under weaker margin assumption*
where previous results are extended in the case of a weaker margin assumption.

A.1. Experiments

To illustrate our results, we consider one-dimensional synthetic examples ($\mathcal{X} = [0, 1]$) for which our assumptions are easily satisfied. Indeed, we consider the following set-up that fulfils our assumptions:

- **(A1)**, **(A3)** We consider here $X \sim U([0, (1-\varepsilon)/2] \cup [(1+\varepsilon)/2, 1])$ and with the notations of Example 1, we take $\mathcal{S}_+ = [0, (1-\varepsilon)/2]$ and $\mathcal{S}_- = [(1+\varepsilon)/2, 1]$. For $1 \leq i \leq n$, x_i independently sampled from $\rho_{\mathcal{X}}$ we define $y_i = 1$ if $x_i \in \mathcal{S}_+$ and $y_i = -1$ if $x_i \in \mathcal{S}_-$.
- **(A2)** We take the kernel to be the exponential kernel $K(x, x') = \exp(-|x - x'|)$ for which the RKHS is a Sobolev space $\mathcal{H} = W^{s,2}$, with $s > d/2$, which is dense in L_2 [AF03].
- **(A4)** With this setting we could find a closed form for g_λ and checked that it verified **(A4)**. Indeed we could solve the optimality equation satisfied by g_λ :

$$\forall z \in [0, 1], \ \int_0^1 K(x, z) g_\lambda(x) d\rho_X(x) + \lambda g_\lambda(z) = \int_0^1 K(x, z) g_\rho(x) d\rho_X(x),$$

the solution being a linear combination of exponentials in each set : $[0, (1-\varepsilon)/2]$, $[(1-\varepsilon)/2, (1+\varepsilon)/2]$ and $[(1+\varepsilon)/2, 1]$.

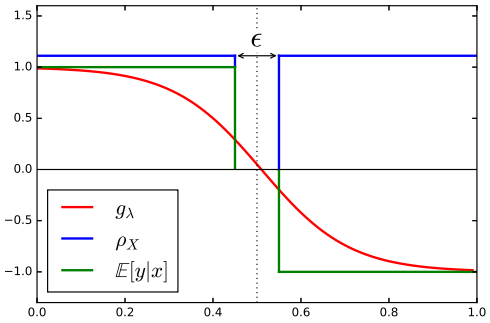

Figure 9: Representing the $\rho_{\mathcal{X}}$ density (uniform with ε-margin), the best estimator, i.e., $\mathbb{E}(x|y)$ and g_λ used for the simulations ($\lambda = 0.01$).

In the case of SGD with decreasing step size, we computed only the test error $\mathbb{E}(\mathcal{R}(g_n) - \mathcal{R}^*)$. For tail averaged SGD with constant step size, we computed the test error as well as the training error, the test loss (which corresponds to the L_2 loss : $\int_0^1 (g_n(x) - g_\lambda(x))^2 d\rho(x)$) and the training loss. In all cases we computed the errors of the n-th iterate with respect to the calculated g_λ, taking $g_0 = 0$. For any $n \geqslant 1$,

$$g_n = g_{n-1} - \gamma_n \big[(g_{n-1}(x_n) - y_n)K_{x_n} + \lambda g_{n-1}\big].$$

We can use representants to find the recursion on the coefficients. Indeed, if $g_n = \sum_{i=1}^n a_i^n K_{x_i}$, then the following recursion for the (a_i^n) reads :

$$\begin{aligned} \text{for } i \leqslant n-1,\ a_i^n &= (1 - \gamma_n \lambda) a_i^{n-1} \\ a_n^n &= -\gamma_n \Big(\sum_{i=1}^{n-1} a_i^{n-1} K(x_n, x_i) - y_n\Big). \end{aligned}$$

From (a_i^n), we can also compute the coefficients of $\bar{g}_n$ and $\bar{g}_n^{\text{tail}}$ that we note $\bar{a}_i^n$ and $\bar{a}_i^{n,\text{tail}}$ respectively: $\bar{a}_i^n = \sum_{k=i}^n \frac{a_i^k}{n+1}$ and $\bar{a}_i^{n,\text{tail}} = \frac{1}{\lfloor n/2 \rfloor} \sum_{k=\lfloor n/2 \rfloor}^n a_i^k$. To show our theoretical results we have decided to present the following figures:

- For the exponential convergence of the averaged and tail averaged cases, we plotted the error $\log_{10} \mathbb{E}(\mathcal{R}(g_n) - \mathcal{R}^*))$ as a function of n. With this scale and following our results it goes as a line after a certain n (Figures 10 and 11 right).
- We recover the results of [DFB17] that show convergence at speed $1/n$ for the loss (Figure 10 left). We adapted the scale to compare with the error plot.
- For Figure 11 left, we plotted $-\log(-\log(\mathbb{E}(\mathcal{R}(g_n) - \mathcal{R}^*)))$ of the excess error with respect to the $\log$ of n to show a line of slope $-1/2$. It meets our theoretical bound of the form $\exp(-K\sqrt{n})$,

Note that for the plots where we plotted the expected excess errors, i.e., $\mathbb{E}(\mathcal{R}(g_n) - \mathcal{R}^*)$, we plotted the mean of the errors over 1000 replications until $n = 200$, whereas for the plots where we plotted the losses, i.e., a function of $\|g_n - g_*\|_2$, we plotted the mean of the loss over 100 replications until $n = 2000$.

We can make the following observations:

First remark that between plots of losses and errors (Figure 10 left and right resp.), there is a factor 10 between the numbers of samples (200 for errors and 2000 for losses) and another factor 10 between errors

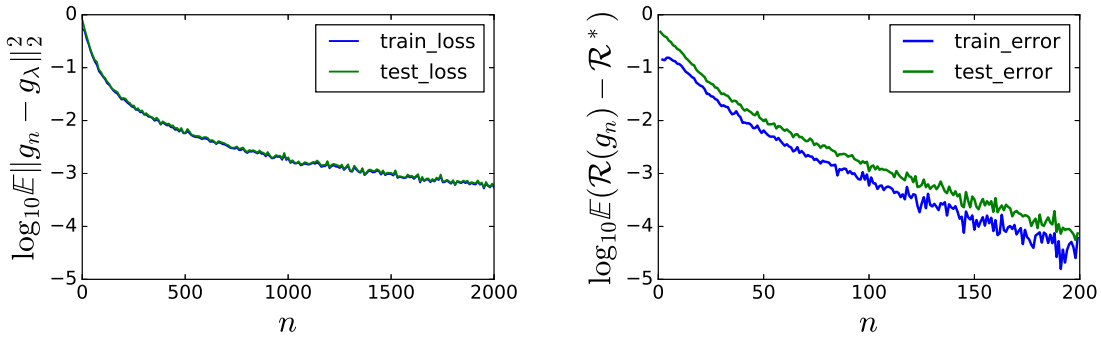

Figure 10: Showing linear convergence for the L^{01} errors in the case of margin of width ε. **Left** figure corresponds to the test and training loss in the averaged case whereas the **right** one corresponds to the error in the same setting. Note that the y-axis is the same while the x-axis is different of a factor 10. The fact that the error plot is a line after a certain n matches our theoretical results. We took the following parameters : $\varepsilon = 0.05$, $\gamma = 0.25$, $\lambda = 0.01$.

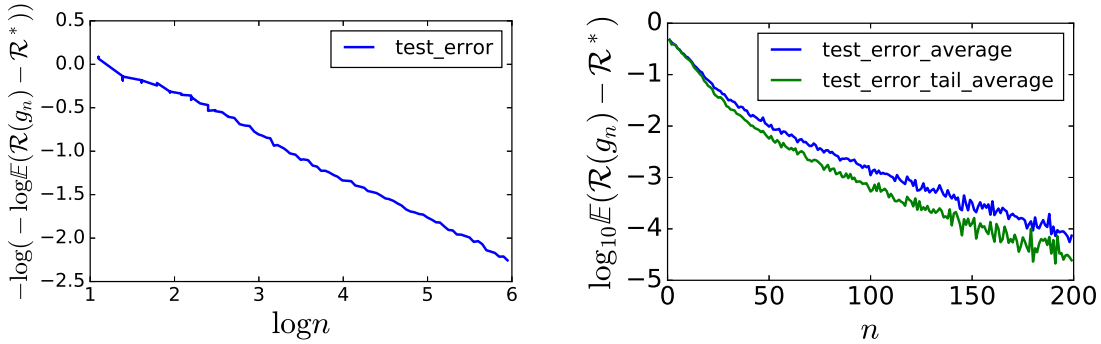

Figure 11: **Left** plot shows the error in the non-averaged case for $\gamma_n = \gamma/\sqrt{n}$ and **right** compares the test error between averaged and tail averaged case. We took the following parameters : $\varepsilon = 0.05$, $\gamma = 0.25$, $\lambda = 0.01$.

and losses (10^{-4} for errors and 10^{-3} for losses). That underlines well our theoretical result which is the difference between exponential rates of convergence of the excess error and $1/n$ rate of convergence of the loss.

Moreover, we see that even if the excess error with tail averaging seems a bit faster, we have linear rates too for the convergence of the excess error in the averaged case. Finally, we remark that the error on the train set is always below the one for a unknown test set (of what seems to be close to a factor 2).

A.2. Probabilistic lemmas

In this section we recall two fundamental results for concentration inequalities in Hilbert spaces shown in [Pin94].

Proposition 5

Let $(X_k)_{k\in\mathbb{N}}$ be a sequence of vectors of $\mathcal{H}$ adapted to a non decreasing sequence of σ-fields $(\mathcal{F}_k)$ such that $\mathbb{E}[X_k|\mathcal{F}_{k-1}] = 0$, $\sup_{k\leqslant n}\|X_k\| \leqslant a_n$ and $\sum_{k=1}^n \mathbb{E}\left[\|X_k\|^2|\mathcal{F}_{k-1}\right] \leqslant b_n^2$ for some sequences $(a_n), (b_n) \in \left(\mathbb{R}_+^*\right)^{\mathbb{N}}$. Then, for all $t \geqslant 0$, $n \geqslant 1$,

$$\mathbb{P}\left(\left\|\sum_{k=1}^n X_k\right\| \geqslant t\right) \leqslant 2\exp\left(\frac{t}{a_n} - \left(\frac{t}{a_n} + \frac{b_n^2}{a_n^2}\right)\ln\left(1 + \frac{ta_n}{b_n}\right)\right). \tag{62}$$

Proof : As $\mathbb{E}\left[X_k|\mathcal{F}_{k-1}\right]=0$, the $\mathcal{F}_j$-adapted sequence (f_j) defined by $f_j=\sum_{k=1}^{j}X_k$ is a martingale and so is the stopped-martingale $(f_{j\wedge n})$. By applying Theorem 3.4 of [Pin94] to the martingale $(f_{j\wedge n})$, we have the result. ■

Corollary 2

Let $(X_k)_{k\in\mathbb{N}}$ be a sequence of vectors of $\mathcal{H}$ adapted to a non decreasing sequence of σ-fields $(\mathcal{F}_k)$ such that $\mathbb{E}\left[X_k|\mathcal{F}_{k-1}\right]=0$, $\sup_{k\leqslant n}\|X_k\|\leqslant a_n$ and $\sum_{k=1}^{n}\mathbb{E}\left[\|X_k\|^2|\mathcal{F}_{k-1}\right]\leqslant b_n^2$ for some sequences $(a_n),(b_n)\in\left(\mathbb{R}_+^*\right)^{\mathbb{N}}$. Then, for all $t\geqslant 0$, $n\geqslant 1$,

$$\mathbb{P}\left(\left\|\sum_{k=1}^{n}X_k\right\|\geqslant t\right)\leqslant 2\exp\left(-\frac{t^2}{2\left(b_n^2+a_nt/3\right)}\right). \tag{63}$$

Proof : We apply 5 and simply notice that

$$\begin{aligned}\frac{t}{a_n}-\left(\frac{t}{a_n}+\frac{b_n^2}{a_n^2}\right)\ln\left(1+\frac{ta_n}{b_n}\right) &= -\frac{b_n^2}{a_n^2}\left(\left(1+\frac{a_nt}{b_n^2}\right)\ln\left(1+\frac{a_nt}{b_n^2}\right)-\frac{a_nt}{b_n^2}\right)\\ &= -\frac{b_n^2}{a_n^2}\phi\left(\frac{a_nt}{b_n^2}\right),\end{aligned}$$

where $\phi(u)=(1+u)\ln(1+u)-u$ for $u>0$. Moreover $\phi(u)\geqslant\dfrac{u^2}{2\left(1+u/3\right)}$, so that:

$$\frac{t}{a_n}-\left(\frac{t}{a_n}+\frac{b_n^2}{a_n^2}\right)\ln\left(1+\frac{ta_n}{b_n}\right)\leqslant-\frac{b_n^2}{a_n^2}\frac{(a_nt/b_n^2)^2}{2\left(1+a_nt/3b_n^2\right)}=-\frac{t^2}{2\left(b_n^2+a_nt/3\right)}.$$

■

A.3. From $\mathcal{H}$ to 0-1 loss

In this section we prove Lemma 1. Note that **(A4)** requires the existence of g_λ having the same sign of g_* almost everywhere on the support of $\rho_{\mathcal{X}}$ and with absolute value uniformly bounded from below. In Lemma 1 we prove that we can bound the 0-1 error with respect to the distance in $\mathcal{H}$ of the estimator $\widehat{g}$ form g_λ.

Proof of Lemma 1 : Denote by W the event such that $\left\|\widehat{g}-g_\lambda\right\|_{\mathcal{H}}<\delta/(2R)$. Note that for any $f\in\mathcal{H}$,

$$f(x)=\langle f,K_x\rangle_{\mathcal{H}}\leq\left\|K_x\right\|_{\mathcal{H}}\left\|f\right\|_{\mathcal{H}}\leq R\left\|f\right\|_{\mathcal{H}},$$

for any $x\in\mathcal{X}$. So for $\widehat{g}\in W$, we have

$$|\widehat{g}(x)-g_\lambda(x)|\leq R\left\|\widehat{g}-g_\lambda\right\|_{\mathcal{H}}<\delta/2\quad\forall x\in\mathcal{X}.$$

Let x be in the support of $\rho_{\mathcal{X}}$. By **(A4)** $|g_\lambda(x)|\geq\delta/2$ a.e.. Let $\widehat{g}\in W$ and $x\in\mathcal{X}$ such that $g_\lambda(x)>0$, we have

$$\widehat{g}(x)=g_\lambda(x)-(g_\lambda(x)-\widehat{g}(x))\geq g_\lambda(x)-|g_\lambda(x)-\widehat{g}(x)|>0,$$

so $\operatorname{sign}(\widehat{g}(x))=\operatorname{sign}(g_\lambda(x))=+1$. Similarly let $\widehat{g}\in W$ and $x\in\mathcal{X}$ such that $g_\lambda(x)<0$, we have

$$\widehat{g}(x)=g_\lambda(x)+(\widehat{g}(x)-g_\lambda(x))\leq g_\lambda(x)+|g_\lambda(x)-\widehat{g}(x)|<0,$$

so $\operatorname{sign}(\widehat{g}(x))=\operatorname{sign}(g_\lambda(x))=-1$. Finally note that for any $\widehat{g}\in\mathcal{H}$, by **(A4)**, either $g_\lambda(x)>0$ or $g_\lambda(x)<0$ a.e., so $\operatorname{sign}(\widehat{g}(x))=\operatorname{sign}(g_\lambda(x))$ a.e.

Now note that by **(A1)**, **(A4)** we have that $\operatorname{sign}(g_*(x))=\operatorname{sign}(g_\lambda(x))$ a.e., where $g_*(x):=\mathbb{E}[y|x]$. So when $\widehat{g}\in W$, we have that $\operatorname{sign}(\widehat{g}(x))=\operatorname{sign}(g_\lambda(x))=\operatorname{sign}(g_*(x))$ a.e., so

$$\mathcal{R}(\widehat{g})=\rho(\{(x,y):\operatorname{sign}(\widehat{g}(x))\neq y\})=\rho(\{(x,y):\operatorname{sign}(g_*(x))\neq y\})=\mathcal{R}^*.$$

Finally note that

$$\mathbb{E}\mathcal{R}(\widehat{g}) = \mathbb{E}\mathcal{R}(\widehat{g})\mathbf{1}_W + \mathbb{E}\mathcal{R}(\widehat{g})\mathbf{1}_{W^c},$$

where $\mathbf{1}_W$ is 1 on the set W and 0 outside, W^c is the complement set of W. So, when $\widehat{g} \in W$, we have

$$\mathbb{E}\mathcal{R}(\widehat{g})\mathbf{1}_W = \mathcal{R}^*\mathbb{E}\mathbf{1}_W \leq \mathcal{R}^*,$$

while

$$\mathbb{E}\mathcal{R}(\widehat{g})\mathbf{1}_{W^c} \leq \mathbb{E}\mathbf{1}_{W^c} \leq q.$$

■

A.4. Exponential rates for Kernel Ridge Regression

A.4.1. Results

In this section, we first specialize some results already known in literature about the consistency of kernel ridge least-squares regression (KRLS) in $\mathcal{H}$-norm [CDV07] and then we derive exponential classification learning rates. Let $(x_i, y_i)_{i=1}^n$ be n examples independently and identically distributed according to ρ, that is Assumption **(A3)**. Denote by $\Sigma, \widehat{\Sigma}$ the linear operators on $\mathcal{H}$ defined by

$$\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^n K_{x_i} \otimes K_{x_i}, \quad \Sigma = \int_{\mathcal{X}} (K_x \otimes K_x) d\rho_{\mathcal{X}}(x),$$

referred to as the covariance and empirical (non-centered) covariance operators (see [FBJ04] and references therein). We recall that the KRLS estimator $\widehat{g}_\lambda \in \mathcal{H}$, which minimizes the regularized empirical risk, is defined as follows in terms of $\widehat{\Sigma}$,

$$\widehat{g}_\lambda = (\widehat{\Sigma} + \lambda I)^{-1} \left(\frac{1}{n}\sum_{i=1}^n y_i K_{x_i} \right).$$

Moreover we recall that the population regularized estimator g_λ is characterized by see ([CDV07])

$$g_\lambda = (\Sigma + \lambda I)^{-1} (\mathbb{E} y K_x).$$

The following lemma bounds the empirical regularized estimator with respect to the population one in terms of λ, n and is essentially contained in the work of [CDV07]; here we rederive it in a subcase (see below for the proof).

Lemma 3

*Under assumption **(A2)**, **(A3)** for any $\lambda > 0$, note $u_n = \|\frac{1}{n}\sum_{i=1}^n y_i K_{x_i} - \mathbb{E} y K_x\|_{\mathcal{H}}$ and $v_n = \|\Sigma - \widehat{\Sigma}\|_{op}$, we have:*

$$\|\widehat{g}_\lambda - g_\lambda\|_{\mathcal{H}} \leq \frac{u_n}{\lambda} + \frac{R v_n}{\lambda^2}.$$

By using deviation inequalities for u_n, v_n in Lemma 3 and then applying Lemma 1, we obtain the following exponential bound for kernel ridge regression (see complete proof below):

Theorem 5

*Under **(A1)**, **(A2)**, **(A3)**, **(A4)** we have that for any $n \in \mathbb{N}$,*

$$\mathcal{R}(\widehat{g}_\lambda) - \mathcal{R}^* = 0 \text{ with probability at least } 1 - 4\exp\left(-\frac{C_0 \lambda^4 \delta^2}{R^8} n\right).$$

Moreover, $\mathbb{E}\mathcal{R}(\widehat{g}_\lambda) - \mathcal{R}^* \leq 4\exp\left(-C_0\lambda^4\delta^2 n/R^8\right)$, with $C_0^{-1} := 72(1+\lambda R^2)^2$.

The result above is a refinement of Thm. 2.6 from [YRC07]. We improved the dependency in n and removed the requirements that $g^* \in \mathcal{H}$ or $g^* = \Sigma^r w$ for a $w \in L^2(d\rho_X)$ and $r > 1/2$. Similar results exist for losses that are usually considered more suitable for classification, like the hinge or logistic loss and more generally losses that are non-decreasing [KB05]. With respect to this latter work, our analysis uses the explicit characterization of the kernel ridge regression estimator in terms of linear operators on $\mathcal{H}$ [CDV07]. This, together with **(A4)**, allows us to use analytic tools specific to reproducing kernel Hilbert spaces, leading to proofs that are comparatively simpler, with explicit constants and a clearer problem setting (consisting essentially in **(A1)**, **(A4)** and no assumptions on $\mathbb{E}[y|x]$).

Finally note that the exponent of λ could be reduced by using a refined analysis under additional regularity assumption of ρ_X and $\mathbb{E}[y|x]$ (as *source condition* and *intrinsic dimension* from [CDV07]), but it is beyond the scope of this paper.

A.4.2. Proofs

Here we prove that Kernel Ridge Regression achieves exponential classification rates under assumptions **(A1)**, **(A4)**. In particular by Lemma 3 we bound $\|\widehat{g}_\lambda - g_\lambda\|_{\mathcal{H}}$ in high probability and then we use Lemma 1 that gives exponential classfication rates when $\|\widehat{g}_\lambda - g_\lambda\|_{\mathcal{H}}$ is small enough in high probability.

Proof of Lemma 3 : Denote by $\widehat{\Sigma}_\lambda$ the operator $\widehat{\Sigma} + \lambda I$ and with Σ_λ the operator $\Sigma + \lambda I$. We have

$$\begin{aligned}\widehat{g}_\lambda - g_\lambda &= \widehat{\Sigma}_\lambda^{-1}\left(\frac{1}{n}\sum_{i=1}^n y_i K_{x_i}\right) - \Sigma_\lambda^{-1}(\mathbb{E}yK_x) \\ &= \widehat{\Sigma}_\lambda^{-1}\left(\frac{1}{n}\sum_{i=1}^n y_i K_{x_i} - \mathbb{E}yK_x\right) + (\widehat{\Sigma}_\lambda^{-1} - \Sigma_\lambda^{-1})\mathbb{E}yK_x.\end{aligned}$$

For the first term, since $\|\widehat{\Sigma}_\lambda^{-1}\|_{\text{op}} \leq \lambda^{-1}$, we have

$$\begin{aligned}\|\widehat{\Sigma}_\lambda^{-1}\left(\frac{1}{n}\sum_{i=1}^n y_i K_{x_i} - \mathbb{E}yK_x\right)\|_{\mathcal{H}} &\leq \|\widehat{\Sigma}_\lambda^{-1}\|_{\text{op}}\|\frac{1}{n}\sum_{i=1}^n y_i K_{x_i} - \mathbb{E}yK_x\|_{\mathcal{H}} \\ &\leq \frac{1}{\lambda}\|\frac{1}{n}\sum_{i=1}^n y_i K_{x_i} - \mathbb{E}yK_x\|_{\mathcal{H}}.\end{aligned}$$

For the second term, since $\|\Sigma_\lambda^{-1}\|_{\text{op}} \leq \lambda^{-1}$ and $\|\mathbb{E}yK_x\| \leq \mathbb{E}\|yK_x\| \leq R$, we have

$$\begin{aligned}\|(\widehat{\Sigma}_\lambda^{-1} - \Sigma_\lambda^{-1})\mathbb{E}yK_x\|_{\mathcal{H}} &= \|\widehat{\Sigma}_\lambda^{-1}(\Sigma - \widehat{\Sigma})\Sigma_\lambda^{-1}\mathbb{E}yK_x\|_{\mathcal{H}} \\ &\leq \|\widehat{\Sigma}_\lambda^{-1}\|_{\text{op}}\|\Sigma - \widehat{\Sigma}\|_{\text{op}}\|\Sigma_\lambda^{-1}\|_{\text{op}}\|\mathbb{E}yK_x\|_{\mathcal{H}} \leq \frac{R}{\lambda^2}\|\Sigma - \widehat{\Sigma}\|_{\text{op}}.\end{aligned}$$

■

Proof of Theorem 5 : Let $\tau > 0$. By Lemma 2 we know that

$$\|\widehat{g}_\lambda - g_\lambda\|_{\mathcal{H}} \leq \frac{u_n}{\lambda} + \frac{Rv_n}{\lambda^2},$$

with $u_n = \|\frac{1}{n}\sum_{i=1}^n (y_i K_{x_i} - \mathbb{E}yK_x)\|_{\mathcal{H}}$ and $v_n = \|\Sigma - \widehat{\Sigma}\|_{\text{op}}$. For u_n we can apply Pinelis inequality (Thm. 3.5, [Pin94]), since $(x_i, y_i)_{i=1}^n$ are sampled independently according to the probability ρ and that $y_i K_{x_i} - \mathbb{E}yK_x$ is zero mean. Since

$$\|\frac{1}{n}(y_i K_{x_i} - \mathbb{E}yK_x)\|_{\mathcal{H}} \leq \frac{2R}{n}$$

a.e. and $\mathcal{H}$ is a Hilbert space, then we apply Pinelis inequality with $b_*^2 = \frac{4R^2}{n}$ and $D = 1$, obtaining

$$u_n \leq \sqrt{\frac{8R^2\tau}{n}},$$

with probability at least $1-2e^{-\tau}$. Now, denote by $\|\cdot\|_{HS}$ the Hilbert-Schmidt norm and recall that $\|\cdot\| \leq \|\cdot\|_{HS}$. To bound v_n we apply again the Pinelis inequality [RBV10] considering that the space of Hilbert-Schmidt operators is again a Hilbert space and that $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} K_{x_i} \otimes K_{x_i}$, that $(x_i)_{i=1}^n$ are independently sampled from ρ_X and that $\mathbb{E}K_{x_i} \otimes K_{x_i} = \Sigma$. In particular we apply it with $D=1$ and $b_*^2 = \frac{4R^4}{n}$, so

$$v_n = \|\Sigma - \widehat{\Sigma}\| \leq \|\Sigma - \widehat{\Sigma}\|_{HS} \leq \sqrt{\frac{8R^4\tau}{n}},$$

with probability $1-2e^{-\tau}$. Finally we take the intersection bound of the two events obtaining, with probability at least $1-4e^{-\tau}$,

$$\|\widehat{g}_\lambda - g_\lambda\|_{\mathcal{H}} \leq \sqrt{\frac{8R^2\tau}{\lambda^2 n}} + \sqrt{\frac{8R^6\tau}{\lambda^4 n}}.$$

By selecting $\tau = \frac{\delta^2}{9R^2(\sqrt{\frac{8R^2}{\lambda^2 n}}+\sqrt{\frac{8R^6}{\lambda^4 n}})^2}$, we obtain $\|\widehat{g}_\lambda - g_\lambda\|_{\mathcal{H}} \leq \frac{\delta}{3R}$, with probability $1-4e^{-\tau}$. Now we can apply Lemma 1 to have the exponential bound for the classification error. ∎

A.5. Proofs and additional results about concrete examples

In the next subsection we prove that $g_* \in \mathcal{H}$ is sufficient to satisfy **(A4)**, while in subsection A.5.2 we prove that specific settings naturally satisfy **(A4)**.

A.5.1. From $g_* \in \mathcal{H}$ to (A4)

Here we assume that there exists $g_* \in \mathcal{H}$ such that $g_*(x) = \mathbb{E}[y|x]$ a.e. on the support of ρ_X. First we introduce $A(\lambda)$, that is a quantity related to the approximation error of g_λ with respect to g_* and we study its behavior when $\lambda \to 0$. Then we express $\|g_\lambda - g_*\|_{\mathcal{H}}$ in terms of $A(\lambda)$. Finally we prove that for any δ given by **(A1)**, there exists λ such that **(A4)** is satisfied.

Let $(\sigma_t, u_t)_{t\in\mathbb{N}}$ be an eigenbasis of Σ with $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$, and let $\alpha_j = \langle g_*, u_j\rangle$ we introduce the following quantity

$$A(\lambda) = \sum_{t:\sigma_t \leq \lambda} \alpha_t^2.$$

Lemma 4

*Under **(A2)**, $A(\lambda)$ is decreasing for any $\lambda > 0$ and*

$$\lim_{\lambda\to 0} A(\lambda) = 0.$$

Proof: Under **(A2)** and the linearity of trace, we have that

$$\sum_{j\in N} \sigma_j = \mathrm{tr}(\Sigma) = \int \mathrm{tr}\left(K_x \otimes K_x\right) d\rho_X(x) = \int \langle K_x, K_x\rangle_{\mathcal{H}}\, d\rho_X(x) = \int K(x,x) d\rho_X(x) \leq R^2.$$

Denote by $t_\lambda \in \mathbb{N}$, the number $\min\{t \in \mathbb{N} \mid \sigma_t \leq \lambda\}$. Since the $(\sigma_j)_{j\in\mathbb{N}}$ is a non-decreasing summable sequence, then it converges to 0, then

$$\lim_{\lambda\to 0} t_\lambda = \infty.$$

Finally, since $(\alpha_j^2)_{j\in\mathbb{N}}$ is a summable sequence we have that

$$\lim_{\lambda\to 0} A(\lambda) = \lim_{\lambda\to 0} \sum_{t:\sigma_t\leq\lambda} \alpha_t^2 = \lim_{\lambda\to 0} \sum_{j=t_\lambda} \alpha_j^2 = \lim_{t\to\infty} \sum_{j=t}^{\infty} \alpha_j^2 = 0.$$

■

Here we express $\|g_\lambda - g_*\|_{\mathcal{H}}$ in terms of $\|g_*\|_{\mathcal{H}}$ and of $A(\sqrt{\lambda})$.

Lemma 5

Under (A2), for any $\lambda > 0$ we have

$$\|g_\lambda - g_*\|_{\mathcal{H}} \leq \sqrt{\sqrt{\lambda}\|g_*\|_{\mathcal{H}}^2 + A(\sqrt{\lambda})}.$$

Proof : Denote by Σ_λ the operator $\Sigma + \lambda I$. Note that since $g_* \in \mathcal{H}$, then

$$\mathbb{E}yK_x = \mathbb{E}g_*(x)K_x = \mathbb{E}(K_x \otimes K_x)g_* = \mathbb{E}K_x \otimes K_x g_* = \Sigma g_*,$$

then $g_\lambda = \Sigma_\lambda^{-1}\mathbb{E}yK_x = \Sigma_\lambda^{-1}\Sigma g_*$. So we have

$$\|g_\lambda - g_*\|_{\mathcal{H}} = \|\Sigma_\lambda^{-1}\Sigma g_* - g_*\|_{\mathcal{H}} = \|(\Sigma_\lambda^{-1}\Sigma - I)g_*\|_{\mathcal{H}} = \lambda\|\Sigma_\lambda^{-1}g_*\|_{\mathcal{H}}.$$

Moreover

$$\lambda\|(\Sigma+\lambda I)^{-1}g_*\|_{\mathcal{H}} \leq \sqrt{\lambda}\|(\Sigma+\lambda I)^{-1/2}\|\,\sqrt{\lambda}\|(\Sigma+\lambda I)^{-1/2}g_*\|_{\mathcal{H}} \leq \sqrt{\lambda}\|(\Sigma+\lambda I)^{-1/2}g_*\|_{\mathcal{H}}.$$

Now we express $\sqrt{\lambda}\|(\Sigma+\lambda I)^{-1/2}g_*\|_{\mathcal{H}}$ in terms of $A(\lambda)$. We have that

$$\begin{aligned}\lambda\|(\Sigma+\lambda I)^{-1/2}g_*\|_{\mathcal{H}}^2 = \lambda\langle g_*, (\Sigma+\lambda I)^{-1}g_*\rangle &= \lambda\left\langle g_*, \left(\sum_{j\in\mathbb{N}}(\sigma_j+\lambda)^{-1}u_j\otimes u_j\right)g_*\right\rangle \\ &= \sum_{j\in\mathbb{N}}\frac{\lambda\alpha_j^2}{\sigma_j+\lambda}.\end{aligned}$$

Now divide the series in two parts

$$\sum_{j\in\mathbb{N}}\frac{\lambda\alpha_j^2}{\sigma_j+\lambda} = S_1(\lambda)+S_2(\lambda), \quad S_1(\lambda) = \sum_{j:\sigma_j\geq\sqrt{\lambda}}\frac{\lambda\alpha_j^2}{\sigma_j+\lambda}, \quad S_2(\lambda) = \sum_{j:\sigma_j<\sqrt{\lambda}}\frac{\lambda\alpha_j^2}{\sigma_j+\lambda}.$$

For each term in S_1, since j is selected such that $\sigma_j \geq \sqrt{\lambda}$ we have that $\lambda(\sigma_j+\lambda)^{-1} \leq \lambda(\sqrt{\lambda}+\lambda)^{-1} \leq \lambda/\sqrt{\lambda} \leq \sqrt{\lambda}$, so

$$S_1(\lambda) \leq \sqrt{\lambda}\sum_{j:\sigma_j\geq\sqrt{\lambda}}\alpha_j^2 \leq \sqrt{\lambda}\sum_{j\in\mathbb{N}}\alpha_j^2 = \sqrt{\lambda}\|g_*\|^2.$$

For S_2, we have that $\lambda(\sigma_j+\lambda)^{-1} \leq 1$, so

$$S_2(\lambda) \leq \sum_{j:\sigma_j<\sqrt{\lambda}}\alpha_j^2 = A(\sqrt{\lambda}).$$

■

Proof of Proposition 3 : By Lemma 5 we have that

$$\|g_\lambda - g_*\|_{\mathcal{H}} \leq \sqrt{\sqrt{\lambda}\|g_*\|_{\mathcal{H}}^2 + A(\sqrt{\lambda})}.$$

Now note that the r.h.s. is non-decreasing in λ, and is 0 when $\lambda \to 0$, due to Lemma 4. Then there exists λ such that $\|g_\lambda - g_*\|_{\mathcal{H}} < \frac{\delta}{2R}$.

Since $|f(x)| \leq R\|f\|_{\mathcal{H}}$ for any $f \in \mathcal{H}$ when the kernel satisfies (A2) and moreover (A1) holds, we have that for any $x \in \mathcal{X}$ such that $g_*(x) > 0$ we have

$$g_\lambda(x) = g_*(x) - (g_*(x) - g_\lambda(x)) \geq g_*(x) - |g_*(x) - g_\lambda(x)| \geq \delta - R\|g_\lambda - g_*\| \geq \delta/2,$$

so $\operatorname{sign}(g_*(x)) = \operatorname{sign}(g_\lambda(x)) = +1$ and $\operatorname{sign}(g_*(x))g_\lambda(x) \geq \delta/2$. Analogously for any $x \in \mathcal{X}$ such that $g_*(x) < 0$ we have

$$g_\lambda(x) = g_*(x) + (g_\lambda(x) - g_*(x)) \leq g_*(x) + |g_*(x) - g_\lambda(x)| \leq -\delta + R\|g_\lambda - g_*\| \leq -\delta/2,$$

so $\operatorname{sign}(g_*(x)) = \operatorname{sign}(g_\lambda(x)) = -1$ and $\operatorname{sign}(g_*(x))g_\lambda(x) \geq \delta/2$. Note finally that $g_*(x) = 0$ on a zero measure set by (A4). ■

A.5.2. Examples

In this subsection we first introduce some notation and basic results about Sobolev spaces, then we prove Prop. 4 and Example 1.

In what follows denote by A_t the t-fattening of a set $A \subseteq \mathbb{R}^d$, that is $A_t = \bigcup_{x\in P} B_t(x)$ where $B_t(x)$ is the open ball of ray t centered in x. We denote by $W^{s,2}(\mathbb{R}^d)$ the Sobolev space endowed with norm

$$\left\|f\right\|_{W^{s,2}} = \left\{ f \in \mathcal{L}^1(\mathbb{R}^d) \cap \mathcal{L}^2(\mathbb{R}^d) \;\middle|\; \int_{\mathbb{R}^d} \mathcal{F}(f)(\omega)^2 (1+\|\omega\|^2)^{s/2} d\omega < \infty \right\}.$$

Finally we define the function $\phi_{s,t} : \mathcal{X} \to \mathbb{R}$, that will be used in the proofs as follows

$$\phi_{s,t}(x) \;=\; q_{d,\delta}\, t^{-d}\, 1_{\{0\}_t}(x)\, (1 - \|x/t\|^2)^{s-d/2},$$

with $q_{d,s} = \pi^{-d/2}\Gamma(1+s)/\Gamma(1+s-d/2)$ and $t > 0, s \geq d/2$. Note that $\phi_{s,t}(x)$ is supported on $\{0\}_{\epsilon/2}$, satisfies

$$\int_{\mathbb{R}^d} \phi_{s,t}(y) dy = 1$$

and it is continuous and belongs to $W^{s,2}(\mathbb{R}^d)$.

Proposition 6

Let P, N two compact subsets of $\mathbb{R}^d$ with Hausdorff distance at least $\epsilon > 0$. There exists $g_{P,N} \in W^{s,2}$ such that

$$g_{P,N}(x) = 1, \;\; \forall\, x \in P, \qquad q_{P,N}(x) = 0, \;\; \forall\, x \in N.$$

In particular $g_{P,N} = 1_{P_{\epsilon/2}} * \phi_{s,\epsilon/2}$.

Proof : Denote by $v_{\epsilon,s}$ the function $(1 - \|2x/\epsilon\|^2)^{s-d/2}$. We have

$$\begin{aligned} g_{P,N}(x) \;&=\; q_{d,s}(\epsilon/2)^{-d} \int_{\mathbb{R}^d} 1_{P_{\epsilon/2}}(x-y)\, 1_{\{0\}_{\epsilon/2}}(y)\, v_{\epsilon,s}(y)\, dy \\ &=\; q_{d,s}(\epsilon/2)^{-d} \int_{\{0\}_{\epsilon/2}} 1_{P_{\epsilon/2}}(x-y)\, v_{\epsilon,s}(y)\, dy \\ &=\; q_{d,s}(\epsilon/2)^{-d} \int_{\{x\}_{\epsilon/2}} 1_{P_{\epsilon/2}}(y)\, v_{\epsilon,s}(y-x)\, dy \end{aligned}$$

Now when $x \in P$, then $\{x\}_{\epsilon/2} \subseteq P_{\epsilon/2}$, so

$$\begin{aligned} g_{P,N}(x) \;&=\; q_{d,s}(\epsilon/2)^{-d} \int_{\{x\}_{\epsilon/2}} 1_{P_{\epsilon/2}}(y)\, v_{\epsilon,s}(y-x)\, dy \\ &=\; q_{d,s}(\epsilon/2)^{-d} \int_{\{x\}_{\epsilon/2}} v_{\epsilon,s}(y-x) dy \;=\; q_{d,s}\epsilon^{-d} \int_{\{0\}_{\epsilon/2}} v_{\epsilon,s}(y) dy \\ &=\; q_{d,s}(\epsilon/2)^{-d} \int_{\mathbb{R}^d} 1_{\{0\}_{\epsilon/2}}(y) v_{\epsilon,s}(y) dy \;=\; \int_{\mathbb{R}^d} \phi_{s,\epsilon/2}(y) dy \;=\; 1. \end{aligned}$$

Conversely, when $x \in N$, then $\{x\}_{\epsilon/2} \cap P_{\epsilon/2} = \emptyset$, so

$$g_{P,N}(x) \;=\; q_{d,s}(\epsilon/2)^{-d} \int_{\{x\}_{\epsilon/2}} 1_{P_{\epsilon/2}}(y)\, v_{\epsilon,s}(y-x)\, dy \;=\; 0.$$

Now we prove that $g_{P,N} \in W^{s,2}$. First note that $P_{\epsilon/2}$ is compact whenever P is compact. This implies that $1_{P_{\epsilon/2}}$ is in $L^2(\mathbb{R}^d)$. Since g_δ is the convolution of an $L^2(\mathbb{R}^d)$ function and a $W^{s,2}$, then it belongs to $W^{s,2}$. ■

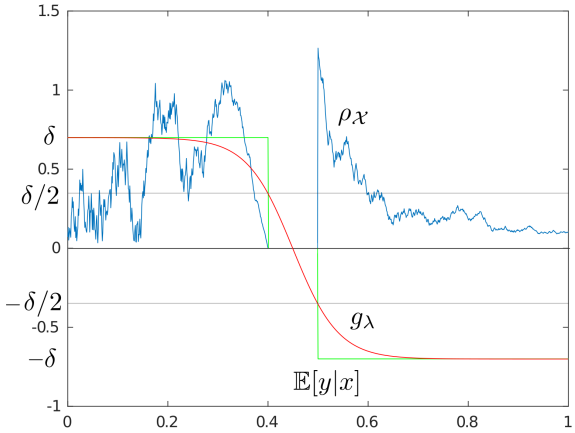

Figure 12: Pictorial representation of a model in 1D satisfying Example 1, ($p = 0.15$). Blue: $\rho_{\mathcal{X}}$, green: $\mathbb{E}[y|x]$, red: g_λ.

Proof of Proposition 4 : Since we are under **(A5)**, we can apply Prop. 6 that prove the existence two functions $q_{\mathcal{S}_+,\mathcal{S}_-}, q_{\mathcal{S}_-,\mathcal{S}_+} \in W^{s,2}$ with the property to be respectively equal to 1 on $\mathcal{S}_+$, 0 on $\mathcal{S}_-$, and 1 on $\mathcal{S}_-$, 0 on $\mathcal{S}_+$. Since $W^{s,2}$ is a Banach algebra [AF03], then $gh \in W^{s,2}$ for any $g, h \in W^{s,2}$. So in particular

$$g_* \;=\; g_+^* q_{\mathcal{S}_+,\mathcal{S}_-} \;-\; g_-^* q_{\mathcal{S}_-,\mathcal{S}_+},$$

belongs to $W^{s,2}$ (and so to $\mathcal{H}$) and is equal to $\mathbb{E}[y|x]$ a.e. on the support of $\rho_{\mathcal{X}}$ by definition. Finally, **(A4)** is satisfied, by Prop. 3. ■

Proof of Example 1 : By definition of y, we have that

$$\mathbb{E}[y|x] = (1-2p)g(x), \quad g(x) = \mathbf{1}_{\mathcal{S}_+} - \mathbf{1}_{\mathcal{S}_-}.$$

In particular note that **(A1)** is satisfied with $\delta = 1 - 2p > 0$ since $p \in [0, 1/2)$. Moreover note that $\mathbb{E}[y|x]$ is constant δ on $\mathcal{S}_+$ and $-\delta$ on $\mathcal{S}_-$. Note now that there exists two functions in $W^{s,2} \subseteq \mathcal{H}$ (due to **(A6)**) that are, respectively δ on $\mathcal{S}_+$ and $-\delta$ on $\mathcal{S}_-$. They are exactly $g_+^* := \delta q_{\mathcal{S}_+,\mathcal{S}_-}$ and $g_-^* = -\delta q_{\mathcal{S}_-,\mathcal{S}_+}$, from Prop. 6. So we can apply Prop. 4, that given g_+^*, g_-^* guarantees that **(A4)** is satisfied. See an example in Figure 12. ■

A.6. Preliminaries for Stochastic Gradient Descent

In this section we show two preliminary results on stochastic gradient descent.

A.6.1. Proof of the optimality condition on g_*

In this subsection we prove the optimality condition on g_*:

$$\mathbb{E}\left[(y_n - \tilde{g}_*(x_n))\, K_{x_n}\right] = 0.$$

Let us recall that as $\mathcal{H}$ is not necessarily dense in L_2, we have defined $\tilde{g}_*$ as the orthonormal projector for the L_2 norm on $\mathcal{H}$ of $g_* = \mathbb{E}(y|x)$ which is the minimizer over all $g \in L_2$ of $\mathbb{E}(y - g(x))^2$. Let $\mathcal{F}$ be the linear space $\overline{\mathcal{H}}^{L_2}$ equipped with the L_2 norm, remark that $\tilde{g}_*$ verifies $\tilde{g}_* = \underset{g\in\mathcal{F}}{\operatorname{argmin}} \|g - g_*\|_{L_2}^2$ and that $g_* - \tilde{g}_* = \mathcal{P}_{\mathcal{H}^\perp}(g_*) \in \mathcal{F}^\perp$.

$$\begin{aligned}\mathbb{E}\left[(y_n - \tilde{g}_*(x_n))\, K_{x_n}\right] &= \mathbb{E}\left[(y_n - \mathbb{E}(y_n|x_n) + \mathbb{E}(y_n|x_n) - \tilde{g}_*(x_n))\, K_{x_n}\right] \\ &= \mathbb{E}\left[(y_n - \mathbb{E}(y_n|x_n))\, K_{x_n}\right] + \mathbb{E}\left[(g_*(x_n) - \tilde{g}_*(x_n))\, K_{x_n}\right] \\ &= \mathbb{E}\left[\mathcal{P}_{\mathcal{H}^\perp}(g_*)(x_n) K_{x_n}\right] \\ &= 0,\end{aligned}$$

where the last equality is true because we have $< \mathcal{P}_{\mathcal{H}^\perp}(g_*), K(\cdot, z) >_{L_2} = 0$ and,

$$\begin{aligned}\|\mathbb{E}\left[\mathcal{P}_{\mathcal{H}^\perp}(g_*)(x_n) K_{x_n}\right]\|_{\mathcal{H}}^2 &= \left\| \int_x \mathcal{P}_{\mathcal{H}^\perp}(g_*)(x) K_x d\rho(x) \right\|_{\mathcal{H}}^2 \\ &= \int_z \mathcal{P}_{\mathcal{H}^\perp}(g_*)(z) \left(\underbrace{\int_x \mathcal{P}_{\mathcal{H}^\perp}(g_*)(x) K(x,z) d\rho(x)}_{=0} \right) d\rho(z) = 0.\end{aligned}$$

A.6.2. Proof of Lemma 2: reformulation of SGD as noisy recursion

Let $n \geqslant 1$ and $g_0 \in \mathcal{H}$, we start form the SGD recursion defined by (55):

$$\begin{aligned}g_n &= g_{n-1} - \gamma_n \big[(\langle K_{x_n}, g_{n-1}\rangle - y_n) K_{x_n} + \lambda(g_{n-1} - g_0)\big] \\ &= g_{n-1} - \gamma_n \big[K_{x_n} \otimes K_{x_n} g_{n-1} - y_n K_{x_n} + \lambda(g_{n-1} - g_0)\big] \\ &= g_{n-1} - \gamma_n \big[K_{x_n} \otimes K_{x_n} g_{n-1} - \tilde{g}_*(x_n) K_{x_n} - \xi_n K_{x_n} + \lambda(g_{n-1} - g_0)\big],\end{aligned}$$

leading to (using the optimality conditions for g_λ and g_*):

$$\begin{aligned}g_n - g_\lambda &= g_{n-1} - g_\lambda - \gamma_n \big[K_{x_n} \otimes K_{x_n}(g_{n-1} - g_\lambda) + \lambda(g_{n-1} - g_0) \\ &\qquad + (K_{x_n} \otimes K_{x_n}) g_\lambda - \tilde{g}_*(x_n) K_{x_n}\big] + \gamma_n \xi_n K_{x_n} \\ &= g_{n-1} - g_\lambda - \gamma_n \big[K_{x_n} \otimes K_{x_n}(g_{n-1} - g_\lambda) + \lambda(g_{n-1} - g_0) \\ &\qquad + (K_{x_n} \otimes K_{x_n} - \Sigma) g_\lambda + \Sigma g_\lambda - \tilde{g}_*(x_n) K_{x_n}\big] + \gamma_n \xi_n K_{x_n} \\ &= g_{n-1} - g_\lambda - \gamma_n \big[K_{x_n} \otimes K_{x_n}(g_{n-1} - g_\lambda) + \lambda g_{n-1} + (K_{x_n} \otimes K_{x_n} - \Sigma) g_\lambda \\ &\qquad - \lambda g_\lambda + \mathbb{E}\left[\tilde{g}_*(x_n) K_{x_n}\right] - \tilde{g}_*(x_n) K_{x_n}\big] + \gamma_n \xi_n K_{x_n} \\ &= g_{n-1} - g_\lambda - \gamma_n \big[(K_{x_n} \otimes K_{x_n} + \lambda I)(g_{n-1} - g_\lambda) + (K_{x_n} \otimes K_{x_n} - \Sigma) g_\lambda \\ &\qquad + \mathbb{E}\left[\tilde{g}_*(x_n) K_{x_n}\right] - \tilde{g}_*(x_n) K_{x_n}\big] + \gamma_n \xi_n K_{x_n} \\ &= \big[I - \gamma_n (K_{x_n} \otimes K_{x_n} + \lambda I)\big](g_{n-1} - g_\lambda) \\ &\qquad + \gamma_n \left[\xi_n K_{x_n} + (\Sigma - K_{x_n} \otimes K_{x_n}) g_\lambda + \tilde{g}_*(x_n) K_{x_n} - \mathbb{E}\left[\tilde{g}_*(x_n) K_{x_n}\right]\right] \\ &= \big[I - \gamma_n (K_{x_n} \otimes K_{x_n} + \lambda I)\big](g_{n-1} - g_\lambda) \\ &\qquad + \gamma_n \left[\xi_n K_{x_n} - (K_{x_n} \otimes K_{x_n}) g_\lambda + \tilde{g}_*(x_n) K_{x_n} + \Sigma g_\lambda - \mathbb{E}\left[\tilde{g}_*(x_n) K_{x_n}\right]\right] \\ &= \big[I - \gamma_n (K_{x_n} \otimes K_{x_n} + \lambda I)\big](g_{n-1} - g_\lambda) \\ &\qquad + \gamma_n \left[\xi_n K_{x_n} + (\tilde{g}_*(x_n) - g_\lambda(x_n)) K_{x_n} - \mathbb{E}\left[(\tilde{g}_*(x_n) - g_\lambda(x_n)) K_{x_n}\right]\right].\end{aligned}$$

A.7. Proof of stochastic gradient descent results

Let us recall for the Appendix the SGD recursion defined in Eq. (57):

$$\eta_n = (I - \gamma H_n)\eta_{n-1} + \gamma_n \varepsilon_n,$$

for which we assume **(H1)**, **(H2)**, **(H3)**, **(H4)**, **(H5)**.

Notations. We define the following notations, which will be useful during all the proofs of the section:

- the following contractant operators: for $i \geqslant k$,

$$M(i,k) = (I - \gamma H_i) \cdots (I - \gamma H_k), \text{ and } M(i, i+1) = I,$$

- the following sequences $Z_k = M(n, k+1)\varepsilon_k$ and $W_n = \sum_{k=1}^n \gamma_k Z_k$.

then,

$$\eta_n = M(n,n)\eta_{n-1} + \gamma_n \varepsilon_n \tag{64}$$

$$\eta_n = M(n,1)\eta_0 + \sum_{k=1}^n \gamma_k M(n,k+1)\varepsilon_k, \tag{65}$$

Note that in all this section, when there is no ambiguity, we will use $\|\cdot\|$ instead of $\|\cdot\|_{\mathcal{H}}$.

A.7.1. Non-averaged SGD - Proof of Theorem 1

In this section, we define the three following sequences: $\alpha_n = \prod_{i=1}^n (1-\gamma_i\lambda)$, $\beta_n = \sum_{k=1}^n \gamma_k^2 \prod_{i=k+1}^n (1-\gamma_i\lambda)^2$ and $\zeta_n = \sup_{k \leqslant n} \gamma_k \prod_{i=k+1}^n (1-\gamma_i\lambda)$.

We can decompose η_n in two terms:

$$\eta_n = \underbrace{M(n,1)\eta_0}_{\text{Biais term}} + \underbrace{W_n}_{\text{Noise term}}, \tag{66}$$

- The biais term represents the speed at which we forget initial conditions. It is the product of n contracting operators

$$\|M(n,1)\eta_0\| \leqslant \prod_{i=1}^n (1-\gamma_i\lambda)\|\eta_0\| = \alpha_n\|\eta_0\|.$$

- The noise term W_n which is a martingale. We are going to show by using a concentration inequality that the probability of the event $\{\|W_n\| \geq t\}$ goes to zero exponentially fast.

General result for all (γ_n). As $W_n = \sum_{k=1}^n \gamma_k Z_k$, we want to apply Corollary 2 of section A.2 to $(\gamma_k Z_k)_{k\in\mathbb{N}}$ that is why we need the following lemma:

Lemma 6

We have the following bounds:

$$\sup_{k\leqslant n} \|\gamma_k Z_k\| \leqslant c^{1/2}\zeta_n, \text{ and} \tag{67}$$

$$\sum_{k=1}^n \mathbb{E}\left[\|\gamma_k Z_k\|^2 | \mathcal{F}_{k-1}\right] \leqslant \mathrm{tr} C\beta_n, \tag{68}$$

*where c and C are defined by **(H3)**.*

Proof : First, $\|\gamma_k Z_k\| = \gamma_k \|M(n,k+1)\varepsilon_k\| \leq \gamma_k \|M(n,k+1)\|_{\text{op}} \|\varepsilon_k\| \leq \gamma_k \frac{\alpha_n}{\alpha_k} \|\varepsilon_k\| \leqslant \zeta_n c^{1/2}$.

Second,

$$\begin{aligned}\sum_{k=1}^{n} \mathbb{E}\left[\|\gamma_k Z_k\|^2 | \mathcal{F}_{k-1}\right] &\leqslant \sum_{k=1}^{n} \frac{\alpha_n^2}{\alpha_k^2} \gamma_k^2 \, \mathbb{E}\|\varepsilon_k\|^2 \\ &\leqslant \sum_{k=1}^{n} \frac{\alpha_n^2}{\alpha_k^2} \gamma_k^2 \, \mathrm{tr}C.\end{aligned}$$

Hence,

$$\begin{aligned}\sum_{k=1}^{n} \mathbb{E}\left[\|\gamma_k Z_k\|^2 | \mathcal{F}_{k-1}\right] &\leqslant \sum_{k=1}^{n} \gamma_k^2 \prod_{i=k+1}^{n} (1-\gamma_i \lambda)^2 \mathrm{tr}C \\ &= \mathrm{tr}C\beta_n.\end{aligned}$$

■

Proposition 7

We have the following inequality: for $t > 0, n \geqslant 1$,

$$\|\eta_n\| \leqslant \alpha_n \|\eta_0\| + V_n, \quad \textit{with} \tag{69}$$

$$\mathbb{P}\left(V_n \geqslant t\right) \leqslant 2\exp\left(-\frac{t^2}{2(\mathrm{tr}C\beta_n + c^{1/2}\zeta_n t/3)}\right). \tag{70}$$

Proof : We just need to apply Lemma 6 and Corollary 2 to the martingale W_n and $V_n = \|W_n\|$ for all n. ■

Result for $\gamma_n = \gamma/n^\alpha$. We now derive estimates of α_n, β_n and ζ_n to have explicit bound for the previous result in the case where $\gamma_n = \dfrac{\gamma}{n^\alpha}$ for $\alpha \in [0,1]$. Some of the estimations are taken from [BM11].

Lemma 7

In the interesting particular case where $\gamma_n = \dfrac{\gamma}{n^\alpha}$ for $\alpha \in [0,1]$:

- *for $\alpha = 1$, $i.e\ \gamma_n = \dfrac{\gamma}{n}$, then $\zeta_n = \dfrac{\gamma}{1-\gamma\lambda}\alpha_n$, and we have the following estimations for $\gamma\lambda < 1/2$:*
  (i) $\alpha_n \leqslant \dfrac{1}{n^{\gamma\lambda}}$, *(ii)* $\beta_n \leqslant \dfrac{2(1-\gamma\lambda)}{1-2\gamma\lambda}\dfrac{4^{\gamma\lambda}\gamma^2}{n^{2\gamma\lambda}}$, *(iii)* $\zeta_n \leqslant \dfrac{\gamma}{(1-\lambda\gamma)n^{\gamma\lambda}}$.
- *for $\alpha = 0$, $i.e\ \gamma_n = \gamma$, then $\zeta_n = \gamma$, and we have the following:*
  (i) $\alpha_n = (1-\gamma\lambda)^n$, *(ii)* $\beta_n \leqslant \dfrac{\gamma}{\lambda}$, *(iii)* $\zeta_n = \gamma$.
- *for $\alpha \in]0,1[$, $\zeta_n = \max\left\{\gamma_n, \dfrac{\gamma}{1-\gamma\lambda}\alpha_n\right\}$, and we have the following estimations:*
  (i) $\alpha_n \leqslant \exp\left(-\dfrac{\gamma\lambda}{1-\alpha}\left((n+1)^{1-\alpha}-1\right)\right)$,
  (ii) Denoting $L_\alpha = \frac{2\lambda\gamma}{1-\alpha}2^{1-\alpha}\left(1-\left(\frac{3}{4}\right)^{1-\alpha}\right)$, *we distinguish three cases:*
  - $\alpha > 1/2$, $\beta_n \leqslant \gamma^2 \frac{2\alpha}{2\alpha-1}\exp\left(-L_\alpha n^{1-\alpha}\right) + \frac{2^\alpha\gamma}{\lambda n^\alpha}$,
  - $\alpha = 1/2$, $\beta_n \leqslant \gamma^2 \ln(3n)\exp\left(-L_\alpha n^{1-\alpha}\right) + \frac{2^\alpha\gamma}{\lambda n^\alpha}$,
  - $\alpha < 1/2$, $\beta_n \leqslant \gamma^2 \frac{n^{1-2\alpha}}{1-2\alpha}\exp\left(-L_\alpha n^{1-\alpha}\right) + \frac{2^\alpha\gamma}{\lambda n^\alpha}$.

(iii) $\zeta_n \leqslant \max\left\{\frac{\gamma}{1-\gamma\lambda}\exp\left(-\frac{\gamma\lambda}{1-\alpha}\left((n+1)^{1-\alpha}-1\right)\right), \frac{\gamma}{n^\alpha}\right\}$.

Note that in this case for n large enough we have the following estimations:
(i) $\alpha_n \leqslant \exp\left(-\frac{\gamma\lambda}{2^{1-\alpha}(1-\alpha)}n^{1-\alpha}\right)$, *(ii)* $\beta_n \leqslant \frac{2^{\alpha+1}\gamma}{\lambda n^\alpha}$, *(iii)* $\zeta_n \leqslant \frac{\gamma}{n^\alpha}$.

Proof : First we show for $\alpha \in [0,1]$ the equality for ζ_n. Denote $a_k = \gamma_k \prod_{i=k+1}^n (1-\gamma_i\lambda)$, we want to find $\zeta_n = \sup_{k\leqslant n} a_k$. We show for $\gamma_n = \frac{\gamma}{n^\alpha}$ that $(a_k)_{k\geqslant 1}$ decreases then increases so that $\zeta_n = \max\{a_1, a_n\}$. Let $k \leqslant n-1$,

$$\begin{aligned}\frac{a_{k+1}}{a_k} &= \frac{\gamma_{k+1}}{\gamma_k}\frac{1}{(1-\gamma_{k+1}\lambda)} \\ &= \frac{1}{\frac{\gamma_k}{\gamma_{k+1}}-\gamma_k\lambda}\end{aligned}$$

Hence, $\frac{a_k}{a_{k+1}} - 1 = \frac{\gamma_k}{\gamma_{k+1}} - \gamma_k\lambda - 1$. Take $\alpha \in]0,1[$, in this case where $\gamma_n = \frac{\gamma}{n^\alpha}$,

$$\frac{a_k}{a_{k+1}} - 1 = \left(1+\frac{1}{k}\right)^\alpha - \frac{\gamma\lambda}{k^\alpha} - 1.$$

A rapid study of the function $f_\alpha(x) = \left(1+\frac{1}{x}\right)^\alpha - \frac{\gamma\lambda}{x^\alpha} - 1$ in $\mathbb{R}_+^\star$ shows that it decreases until $x_\star = (\gamma\lambda)^{\frac{1}{(\alpha-1)}} - 1$ then increases. This concludes the proof for $\alpha \in]0,1[$. By a direct calculation for $\alpha = 1$, $\frac{a_k}{a_{k+1}} - 1 = \frac{1-\gamma\lambda}{k} \geqslant 0$ thus a_k is non increasing and $\zeta_n = a_1 = \frac{\gamma}{1-\gamma\lambda}\alpha_n$. Similarly, for $\alpha = 0$, $\frac{a_k}{a_{k+1}} - 1 = \gamma\lambda < 0$ thus a_k is increasing and $\zeta_n = a_n = \gamma_n$.

We show now the different estimations we have for α_n, β_n and ζ_n for the three cases above.

- for $\alpha = 1$,

$$\begin{aligned}\ln\alpha_n &= \sum_{i=1}^n \ln\left(1-\frac{\gamma\lambda}{i}\right) \leqslant -\gamma\lambda\sum_{i=1}^n\frac{1}{i} \leqslant -\gamma\lambda\ln n \\ \alpha_n &\leqslant \frac{1}{n^{\gamma\lambda}}.\end{aligned}$$

Then,

$$\begin{aligned}\beta_n &= \gamma^2\sum_{k=1}^n\frac{1}{k^2}\prod_{i=k+1}^n\left(1-\frac{\gamma\lambda}{i}\right)^2 \\ \beta_n &\leqslant \gamma^2\sum_{k=1}^n\frac{1}{k^2}\exp\left(-2\gamma\lambda\sum_{i=k+1}^n\frac{1}{i}\right) \\ &\leqslant \gamma^2\sum_{k=1}^n\frac{1}{k^2}\exp\left(-2\gamma\lambda\ln\left(\frac{n+1}{k+1}\right)\right) \\ &\leqslant \gamma^2\sum_{k=1}^n\frac{1}{k^2}\left(\frac{k+1}{n+1}\right)^{2\gamma\lambda} \\ &\leqslant 4^{\gamma\lambda}\gamma^2\sum_{k=1}^n\frac{1}{k^2}\left(\frac{k}{n}\right)^{2\gamma\lambda} \\ &\leqslant \frac{4^{\gamma\lambda}\gamma^2}{n^{2\gamma\lambda}}\sum_{k=1}^n k^{2\gamma\lambda-2},\end{aligned}$$

Moreover for $\gamma\lambda < \frac{1}{2}$, $\sum_{k=1}^n k^{2\gamma\lambda-2} \leqslant 1 - \frac{1}{2\gamma\lambda-1} = \frac{2(1-\gamma\lambda)}{1-2\gamma\lambda}$, hence,

$$\beta_n \leqslant \frac{2(1-\gamma\lambda)}{1-2\gamma\lambda}\frac{4^{\gamma\lambda}\gamma^2}{n^{2\gamma\lambda}}$$

Finally,

$$\zeta_n = \frac{\gamma}{1-\gamma\lambda}\alpha_n \leqslant \frac{\gamma}{1-\gamma\lambda}\frac{1}{n^{\gamma\lambda}}.$$

- for $\alpha = 0$,

$$\alpha_n = \prod_{i=1}^{n}(1-\gamma\lambda) = (1-\gamma\lambda)^n.$$

Then,

$$\beta_n = \gamma^2\sum_{k=1}^{n}\prod_{i=k+1}^{n}(1-\gamma\lambda)^2 = \gamma^2\sum_{k=1}^{n}(1-\gamma\lambda)^{2(n-k)} \leqslant \frac{1}{1-(1-\lambda\gamma)^2} \leqslant \frac{\gamma}{\lambda}.$$

Finally,

$$\zeta_n = \gamma_n = \gamma.$$

- for $\alpha \in]0,1[$,

$$\begin{aligned}\ln\alpha_n &= \sum_{i=1}^{n}\ln\left(1-\frac{\gamma\lambda}{i^\alpha}\right) \leqslant -\gamma\lambda\sum_{i=1}^{n}\frac{1}{i^\alpha} \leqslant -\gamma\lambda\frac{(n+1)^{1-\alpha}-1}{1-\alpha}\\ \alpha_n &\leqslant \exp\left(-\frac{\gamma\lambda}{1-\alpha}\left((n+1)^{1-\alpha}-1\right)\right).\end{aligned}$$

To have an estimation on β_n, we are going to split it into two sums. Let $m \in [\![1,n]\!]$,

$$\begin{aligned}\beta_n &= \sum_{k=1}^{n}\gamma_k^2\prod_{i=k+1}^{n}(1-\gamma_i\lambda)^2 = \sum_{k=1}^{m}\gamma_k^2\prod_{i=k+1}^{n}(1-\gamma_i\lambda)^2 + \sum_{k=m+1}^{n}\gamma_k^2\prod_{i=k+1}^{n}(1-\gamma_i\lambda)^2\\
\beta_n &\leqslant \sum_{k=1}^{m}\gamma_k^2\exp\left(-2\lambda\sum_{i=m+1}^{n}\gamma_i\right) + \frac{\gamma_m}{\lambda}\sum_{k=m+1}^{n}\prod_{i=k+1}^{n}(1-\gamma_i\lambda)^2\lambda\gamma_k\\
&\leqslant \sum_{k=1}^{n}\gamma_k^2\exp\left(-2\lambda\sum_{i=m+1}^{n}\gamma_i\right)\\
&\qquad + \frac{\gamma_m}{\lambda}\sum_{k=m+1}^{n}\left[\prod_{i=k+1}^{n}(1-\gamma_i\lambda)^2 - \prod_{i=k+1}^{n}(1-\gamma_i\lambda)^2(1-\gamma_k\lambda)\right]\\
&\leqslant \sum_{k=1}^{n}\gamma_k^2\exp\left(-2\lambda\sum_{i=m+1}^{n}\gamma_i\right) + \frac{\gamma_m}{\lambda}\sum_{k=m+1}^{n}\left[\prod_{i=k+1}^{n}(1-\gamma_i\lambda)^2 - \prod_{i=k}^{n}(1-\gamma_i\lambda)^2\right]\\
&\leqslant \sum_{k=1}^{n}\gamma_k^2\exp\left(-2\lambda\sum_{i=m+1}^{n}\gamma_i\right) + \frac{\gamma_m}{\lambda}\left(1-\prod_{i=m+1}^{n}(1-\gamma_i\lambda)^2\right)\\
&\leqslant \sum_{k=1}^{n}\gamma_k^2\exp\left(-2\lambda\sum_{i=m+1}^{n}\gamma_i\right) + \frac{\gamma_m}{\lambda}.\end{aligned}$$

By taking $\gamma_n = \frac{\gamma}{n^\alpha}$ and $m = \lfloor \frac{n}{2} \rfloor$, we get:

$$\begin{aligned}
\beta_n &\leqslant \gamma^2 \sum_{k=1}^{n} \frac{1}{k^{2\alpha}} \exp\left(-2\lambda\gamma \sum_{i=\lfloor \frac{n}{2} \rfloor + 1}^{n} \frac{1}{i^\alpha} \right) + \frac{2^\alpha \gamma}{\lambda n^\alpha} \\
&\leqslant \gamma^2 \sum_{k=1}^{n} \frac{1}{k^{2\alpha}} \exp\left(-\frac{2\lambda\gamma}{1-\alpha} \left((n+1)^{1-\alpha} - \left(\frac{n}{2}+1\right)^{1-\alpha} \right) \right) + \frac{2^\alpha \gamma}{\lambda n^\alpha} \\
&\leqslant \gamma^2 \sum_{k=1}^{n} \frac{1}{k^{2\alpha}} \exp\left(-\frac{2\lambda\gamma}{1-\alpha} n^{1-\alpha} \left(\left(1+\frac{1}{n}\right)^{1-\alpha} - \left(\frac{1}{2}+\frac{1}{n}\right)^{1-\alpha} \right) \right) + \frac{2^\alpha \gamma}{\lambda n^\alpha} \\
&\leqslant \gamma^2 \sum_{k=1}^{n} \frac{1}{k^{2\alpha}} \exp\left(-\frac{2\lambda\gamma}{1-\alpha} n^{1-\alpha} 2^{1-\alpha} \left(1 - \left(\frac{3}{4}\right)^{1-\alpha} \right) \right) + \frac{2^\alpha \gamma}{\lambda n^\alpha}.
\end{aligned}$$

Calling $S_n^\alpha = \sum_{k=1}^{n} \frac{1}{k^{2\alpha}}$ and noting that: for $\alpha > 1/2$, $S_n^\alpha \leqslant \frac{2\alpha}{2\alpha-1}$, $\alpha = 1/2$, $S_n^\alpha \leqslant \ln(3n)$ and $\alpha < 1/2$, $S_n^\alpha \leqslant \frac{n^{1-2\alpha}}{1-2\alpha}$ we have the expected result.
Finally,

$$\zeta_n \leqslant \max\left\{ \frac{\gamma}{1-\gamma\lambda} \exp\left(-\frac{\gamma\lambda}{1-\alpha} \left((n+1)^{1-\alpha} - 1 \right) \right), \frac{\gamma}{n^\alpha} \right\}.$$

■

With this estimations we can easily show the Theorem 1. In the following we recall the main result of this Theorem and give an extension for $\alpha = 0$ and $\alpha = 1$ that cannot be found in the main text.

Proposition 8 (*SGD, decreasing step size: $\gamma_n = \gamma/n^\alpha$*)

Assume (H1), (H2), (H3), $\gamma_n = \gamma/n^\alpha$, $\gamma\lambda < 1$ and denote by $\eta_n \in \mathcal{H}$ the n-th iterate of the recursion in Eq. (57). We have for $t > 0, n \geqslant 1$,

- *for $\alpha = 1$ and $\gamma\lambda < 1/2$, $\|g_n - g_\lambda\|_{\mathcal{H}} \leqslant \frac{\|g_0 - g_\lambda\|_{\mathcal{H}}}{n^{\gamma\lambda}} + V_n$, almost surely, with*

$$\mathbb{P}\left(V_n \geqslant t\right) \leqslant 2\exp\left(-\frac{t^2}{4^{3/2}(\mathrm{tr}C)\gamma^2/((1-2\gamma\lambda)n^{\gamma\lambda}) + 4tc^{1/2}\gamma/3} \cdot n^{\gamma\lambda} \right);$$

- *for $\alpha = 0$, $\|g_n - g_\lambda\|_{\mathcal{H}} \leqslant (1-\gamma\lambda)^n \|g_0 - g_\lambda\|_{\mathcal{H}} + V_n$, almost surely, with*

$$\mathbb{P}\left(V_n \geqslant t\right) \leqslant 2\exp\left(-\frac{t^2}{2\gamma(\mathrm{tr}C/\lambda + tc^{1/2}/3)} \right);$$

- *for $\alpha \in (0,1)$, $\|g_n - g_\lambda\|_{\mathcal{H}} \leqslant \exp\left(-\frac{\gamma\lambda}{1-\alpha} \left((n+1)^{1-\alpha} - 1 \right) \right) \|g_0 - g_\lambda\|_{\mathcal{H}} + V_n$, almost surely for n large enough [5], with*

$$\mathbb{P}\left(V_n \geqslant t\right) \leqslant 2\exp\left(-\frac{t^2}{\gamma(2^{\alpha+2}\mathrm{tr}C/\lambda + 2c^{1/2}t/3)} \cdot n^\alpha \right).$$

Proof of Theorem 1 : We apply Proposition 7, and the bound found on α_n, β_n and ζ_n in Lemma 7 to get the results. ■

A.7.2. Averaged SGD for the variance term ($\eta_0 = 0$) - Proof of Theorem 2

We consider the same recursion but with $\gamma_n = \gamma$:

$$\eta_n = (I - \gamma H_n)\eta_{n-1} + \gamma\varepsilon_n,$$

started at $\eta_0 = 0$ and with assumptions **(H1)**, **(H2)**, **(H3)**,**(H4)**, **(H5)**.

However, in this section, we consider the averaged:

$$\bar{\eta}_n = \frac{1}{n+1}\sum_{i=0}^{n}\eta_i.$$

Thus, we get

$$\bar{\eta}_n = \frac{1}{n+1}\sum_{i=0}^{n}\gamma\sum_{k=1}^{i}M(i,k+1)\varepsilon_k = \frac{\gamma}{n+1}\sum_{k=1}^{n}\Big(\sum_{i=k}^{n}M(i,k+1)\Big)\varepsilon_k = \frac{\gamma}{n+1}\sum_{k=1}^{n}\bar{Z}_k.$$

Our the goal is to bound $\mathbb{P}\left(\|\bar{\eta}_n\| \geqslant t\right)$ using Propostion 5 that is going to lead us to some Bernstein concentration inquality. Calling, as above, $\bar{Z}_k = \sum_{i=k}^{n} M(i,k+1)\varepsilon_k$, and as $\mathbb{E}\left[\bar{Z}_k|\mathcal{F}_{k-1}\right] = 0$ we just need to bound, $\sup_{k\leqslant n}\|\bar{Z}_k\|$ and $\sum_{k=1}^{n}\mathbb{E}\left[\|\bar{Z}_k\|^2|\mathcal{F}_{k-1}\right]$. For a more general result, we consider in the following lemma $(A^{1/2}\bar{Z}_k)_k$.

Lemma 8

*Assuming **(H1)**, **(H2)**, **(H3)**,**(H4)**, **(H5)**, we have the following bounds for $\bar{Z}_k = \sum_{i=k}^{n} M(i,k+1)\varepsilon_k$:*

$$\sup_{k\leqslant n}\|A^{1/2}\bar{Z}_k\| \leqslant \frac{c^{1/2}\|A\|_{op}^{1/2}}{\gamma\lambda} \tag{71}$$

$$\sum_{k=1}^{n}\mathbb{E}\left[\|A^{1/2}\bar{Z}_k\|^2|\mathcal{F}_{k-1}\right] \leqslant n\frac{1}{\gamma^2}\frac{1}{1-\gamma/2\gamma_0}\mathrm{tr}\left(AH^{-2}\cdot C\right). \tag{72}$$

Proof : First $\|A^{1/2}\bar{Z}_k\| \leqslant \|A\|_{op}^{1/2}\|\bar{Z}_k\|$ and we have, almost surely, $\|\varepsilon_k\| \leqslant c^{1/2}$ and $H_n \succcurlyeq \lambda I$, thus for all k, as $\gamma\lambda \leqslant 1$, $I - \gamma H_k \preccurlyeq (1-\gamma\lambda)I$. Hence, $\|M(i,k+1)\|_{\mathrm{op}} \leqslant (1-\gamma\lambda)^{i-k}$ and,

$$\|\bar{Z}_k\| \leqslant \|\varepsilon_k\|\sum_{i=k}^{n}\|M(i,k+1)\|_{\mathrm{op}} \leqslant c^{1/2}\sum_{i=k}^{n}(1-\gamma\lambda)^{i-k} \leqslant \frac{c^{1/2}}{\gamma\lambda}$$

Second, we need an upper bound on $\mathbb{E}\left[\|A^{1/2}\bar{Z}_k\|^2|\mathcal{F}_{k-1}\right]$, we are going to find it in two steps:

- **Step 1:** we first show that the upper bound depends of the trace of some operator involving H^{-1}.

$$\mathbb{E}\left[\|A^{1/2}\bar{Z}_k\|^2|\mathcal{F}_{k-1}\right] \qu\leqslant \quad 2\sum_{i=k}^{n}\mathrm{tr}\left(A\left(\gamma H\right)^{-1}\mathbb{E}\left[M(i,k+1)CM(i,k+1)^*\right]\right),$$

- **Step 2:** we then upperbound this sum to a telescopic one involving H^{-2} to finally show:

$$\mathbb{E}\left[\|A^{1/2}\bar{Z}_k\|^2|\mathcal{F}_{k-1}\right] \quad\leqslant \quad \frac{1}{\gamma^2}\frac{1}{1-\gamma/2\gamma_0}\mathrm{tr}\left(AH^{-2}C\right).$$

Step 1: We write,

$$\begin{aligned}
\mathbb{E}\left[\|A^{1/2}\bar{Z}_k\|^2|\mathcal{F}_{k-1}\right] &= \mathbb{E}\left[\sum_{k\leqslant i,j\leqslant n}\left\langle A^{1/2}M(i,k+1)\varepsilon_k, A^{1/2}M(j,k+1)\varepsilon_k\right\rangle|\mathcal{F}_{k-1}\right] \\
&= \mathbb{E}\left[\sum_{k\leqslant i,j\leqslant n}\langle M(i,k+1)\varepsilon_k, AM(j,k+1)\varepsilon_k\rangle|\mathcal{F}_{k-1}\right] \\
&= \sum_{k\leqslant i,j\leqslant n}\mathbb{E}\left[\operatorname{tr}\left(M(i,k+1)^*AM(j,k+1)\cdot\varepsilon_k\otimes\varepsilon_k\right)\right] \\
&= \sum_{k\leqslant i,j\leqslant n}\operatorname{tr}\left(\mathbb{E}\left[M(i,k+1)^*AM(j,k+1)\right]\cdot\mathbb{E}\left[\varepsilon_k\otimes\varepsilon_k\right]\right).
\end{aligned}$$

We have $\mathbb{E}\left[\varepsilon_k\otimes\varepsilon_k\right]\preccurlyeq C$ so that as every operators are positive semi-definite,

$$\mathbb{E}\left[\|A^{1/2}\bar{Z}_k\|^2|\mathcal{F}_{k-1}\right] \leqslant \sum_{k\leqslant i,j\leqslant n}\operatorname{tr}\left(\mathbb{E}\left[M(i,k+1)^*AM(j,k+1)\right]\cdot C\right).$$

We now bound the last expression by dividing it into two terms, noting $M(i,k)=M_k^i$ for more compact notations (only until the end of the proof),

$$\begin{aligned}
\sum_{k\leqslant i,j\leqslant n}\operatorname{tr}\left(\mathbb{E}\left[{M_{k+1}^i}^*AM_{k+1}^j\right]\cdot C\right) &= \sum_{i=k}^n\operatorname{tr}\left(\mathbb{E}\left[{M_{k+1}^i}^*AM_{k+1}^i\right]\cdot C\right) \\
&\quad + 2\sum_{k\leqslant i<j\leqslant n}\operatorname{tr}\left(\mathbb{E}\left[{M_{k+1}^i}^*AM_{k+1}^j\right]\cdot C\right).
\end{aligned}$$

Moreover,

$$\begin{aligned}
&\sum_{k\leqslant i<j\leqslant n}\operatorname{tr}\left(\mathbb{E}\left[{M_{k+1}^i}^*AM_{k+1}^j\right]\cdot C\right) \\
=\ &\sum_{k\leqslant i<j\leqslant n}\operatorname{tr}\left(\mathbb{E}\left[{M_{k+1}^i}^*A\left(I-\gamma H\right)^{j-i}M_{k+1}^i\right]\cdot C\right) \\
=\ &\sum_{i=k}^n\operatorname{tr}\left(\mathbb{E}\left[{M_{k+1}^i}^*A\sum_{j=i+1}^n\left(I-\gamma H\right)^{j-i}M_{k+1}^i\right]\cdot C\right) \\
=\ &\sum_{i=k}^n\operatorname{tr}\left(\mathbb{E}\left[{M_{k+1}^i}^*A\left[\left(I-\gamma H\right)\left(I-\left(I-\gamma H\right)^{n-i}\right)(\gamma H)^{-1}\right]M_{k+1}^i\right]\cdot C\right) \\
\leqslant\ &\sum_{i=k}^n\operatorname{tr}\left(\mathbb{E}\left[{M_{k+1}^i}^*A\left[(\gamma H)^{-1}-I\right]M_{k+1}^i\right]\cdot C\right) \\
\leqslant\ &\sum_{i=k}^n\operatorname{tr}\left(\mathbb{E}\left[{M_{k+1}^i}^*A\left(\gamma H\right)^{-1}M_{k+1}^i\right]\cdot C\right)-\sum_{i=k}^n\operatorname{tr}\left(\mathbb{E}\left[{M_{k+1}^i}^*AM_{k+1}^i\right]\cdot C\right).
\end{aligned}$$

Hence,

$$\begin{aligned}
&\sum_{k\leqslant i,j\leqslant n}\operatorname{tr}\left(\mathbb{E}\left[{M_{k+1}^i}^*AM_{k+1}^j\right]\cdot C\right) \\
&\quad= \sum_{i=k}^n\operatorname{tr}\left(\mathbb{E}\left[{M_{k+1}^i}^*AM_{k+1}^i\right]\cdot C\right)+2\sum_{k\leqslant i<j\leqslant n}\operatorname{tr}\left(\mathbb{E}\left[{M_{k+1}^i}^*AM_{k+1}^j\right]\cdot C\right) \\
&\quad\leqslant 2\sum_{i=k}^n\operatorname{tr}\left(\mathbb{E}\left[{M_{k+1}^i}^*A\left(\gamma H\right)^{-1}M_{k+1}^i\right]\cdot C\right)-\sum_{i=k}^n\operatorname{tr}\left(\mathbb{E}\left[{M_{k+1}^i}^*AM_{k+1}^i\right]\cdot C\right) \\
&\quad\leqslant 2\sum_{i=k}^n\operatorname{tr}\left(\mathbb{E}\left[{M_{k+1}^i}^*A\left(\gamma H\right)^{-1}M_{k+1}^i\right]\cdot C\right) \\
&\quad\leqslant 2\sum_{i=k}^n\operatorname{tr}\left(A\left(\gamma H\right)^{-1}\mathbb{E}\left[M_{k+1}^iC{M_{k+1}^i}^*\right]\right)
\end{aligned}$$

This concludes step 1.

Step 2: Let us now try to bound $\sum_{i=k}^{n} \mathrm{tr}\left(A\left(\gamma H\right)^{-1}\mathbb{E}\left[M_{k+1}^{i} C {M_{k+1}^{i}}^{*}\right]\right)$. We will do so by bounding it by a telescopic sum. Indeed,

$$\begin{aligned}
&\mathbb{E}\left[M_{k+1}^{i+1} C\left(\gamma H\right)^{-1} {M_{k+1}^{i+1}}^{*}\right] = \mathbb{E}\left[M_{k+1}^{i}\left(I-\gamma H_{i+1}\right) C\left(\gamma H\right)^{-1}\left(I-\gamma H_{i+1}\right) {M_{k+1}^{i}}^{*}\right] \\
&= \mathbb{E}\left[M_{k+1}^{i}\mathbb{E}\left[C\left(\gamma H\right)^{-1} - CH^{-1}H_{i+1} - H_{i+1}CH^{-1} + \gamma H_{i+1}CH^{-1}H_{i+1}\right] {M_{k+1}^{i}}^{*}\right] \\
&= \mathbb{E}\left[M_{k+1}^{i} C\left(\gamma H\right)^{-1} {M_{k+1}^{i}}^{*}\right] - 2\mathbb{E}\left[M_{k+1}^{i} C {M_{k+1}^{i}}^{*}\right] + \gamma\mathbb{E}\left[M_{k+1}^{i}\mathbb{E}\left[H_{i+1}CH^{-1}H_{i+1}\right] {M_{k+1}^{i}}^{*}\right],
\end{aligned}$$

such that, by multiplying the previous equality by $A\left(\gamma H\right)^{-1}$ and taking the trace we have,

$$\begin{aligned}
\mathrm{tr}\left(A\left(\gamma H\right)^{-1}\mathbb{E}\left[M_{k+1}^{i+1} C\left(\gamma H\right)^{-1} {M_{k+1}^{i+1}}^{*}\right]\right) &= \mathrm{tr}\left(A\left(\gamma H\right)^{-1}\mathbb{E}\left[M_{k+1}^{i} C\left(\gamma H\right)^{-1} {M_{k+1}^{i}}^{*}\right]\right) \\
&\quad - 2\mathrm{tr}\left(A\left(\gamma H\right)^{-1}\mathbb{E}\left[M_{k+1}^{i} C {M_{k+1}^{i}}^{*}\right]\right) \\
&\quad + \gamma\mathrm{tr}\left(A\left(\gamma H\right)^{-1}\mathbb{E}\left[M_{k+1}^{i}\mathbb{E}\left[H_{i+1}CH^{-1}H_{i+1}\right] {M_{k+1}^{i}}^{*}\right]\right),
\end{aligned}$$

And as $\mathbb{E}\left[H_k CH^{-1}H_k\right] \preccurlyeq \gamma_0^{-1} C$ we have,

$$\gamma\mathrm{tr}\left(A\left(\gamma H\right)^{-1}\mathbb{E}\left[M_{k+1}^{i}\mathbb{E}\left[H_{i+1}CH^{-1}H_{i+1}\right] {M_{k+1}^{i}}^{*}\right]\right) \leqslant \gamma/\gamma_0\mathrm{tr}\left(A\left(\gamma H\right)^{-1}\mathbb{E}\left[M_{k+1}^{i} C {M_{k+1}^{i}}^{*}\right]\right),$$

thus,

$$\begin{aligned}
\mathrm{tr}\left(A\left(\gamma H\right)^{-1}\mathbb{E}\left[M_{k+1}^{i+1} C\left(\gamma H\right)^{-1} {M_{k+1}^{i+1}}^{*}\right]\right) &\leqslant \mathrm{tr}\left(A\left(\gamma H\right)^{-1}\mathbb{E}\left[M_{k+1}^{i} C\left(\gamma H\right)^{-1} {M_{k+1}^{i}}^{*}\right]\right) \\
&\quad - 2\mathrm{tr}\left(A\left(\gamma H\right)^{-1}\mathbb{E}\left[M_{k+1}^{i} C {M_{k+1}^{i}}^{*}\right]\right) \\
&\quad + \gamma/\gamma_0\mathrm{tr}\left(A\left(\gamma H\right)^{-1}\mathbb{E}\left[M_{k+1}^{i} C {M_{k+1}^{i}}^{*}\right]\right)
\end{aligned}$$

$$\begin{aligned}
&\mathrm{tr}\left(A\left(\gamma H\right)^{-1}\mathbb{E}\left[M_{k+1}^{i} C {M_{k+1}^{i}}^{*}\right]\right) \\
&\leqslant \frac{1}{2-\frac{\gamma}{\gamma_0}}\left(\mathrm{tr}\left(A\left(\gamma H\right)^{-1}\mathbb{E}\left[M_{k+1}^{i} C\left(\gamma H\right)^{-1} {M_{k+1}^{i}}^{*}\right]\right) - \mathrm{tr}\left(A\left(\gamma H\right)^{-1}\mathbb{E}\left[M_{k+1}^{i+1} C\left(\gamma H\right)^{-1} {M_{k+1}^{i+1}}^{*}\right]\right)\right).
\end{aligned}$$

If we take all the calculations from the beginning,

$$\begin{aligned}
\mathbb{E}\left[\|A^{1/2}\bar{Z}_k\|^2|\mathcal{F}_{k-1}\right] &\leqslant \sum_{k\leqslant i,j\leqslant n}\mathrm{tr}\left(\mathbb{E}\left[{M_{k+1}^{i}}^{*} A M_{k+1}^{j}\right]\cdot C\right) \\
&\leqslant 2\sum_{i=k}^{n}\mathrm{tr}\left(A\left(\gamma H\right)^{-1}\mathbb{E}\left[M_{k+1}^{i} C {M_{k+1}^{i}}^{*}\right]\right) \\
&\leqslant \frac{2}{2-\gamma/\gamma_0}\sum_{i=k}^{n}\mathrm{tr}\left(A\left(\gamma H\right)^{-1}\mathbb{E}\left[M_{k+1}^{i} C\left(\gamma H\right)^{-1} {M_{k+1}^{i}}^{*}\right]\right) \\
&\qquad -\mathrm{tr}\left(A\left(\gamma H\right)^{-1}\mathbb{E}\left[M_{k+1}^{i+1} C\left(\gamma H\right)^{-1} {M_{k+1}^{i+1}}^{*}\right]\right) \\
&\leqslant \frac{2}{2-\gamma/\gamma_0}\mathrm{tr}\left(A\left(\gamma H\right)^{-1}\mathbb{E}\left[M_{k+1}^{k} C\left(\gamma H\right)^{-1} {M_{k+1}^{k}}^{*}\right]\right) \\
&\leqslant \frac{1}{\gamma^2}\frac{1}{1-\gamma/2\gamma_0}\mathrm{tr}\left(AH^{-2}\cdot C\right),
\end{aligned}$$

which concludes the proof if we sum this inequality from 1 to n. ■

We can now prove Theorem 2:

Proof of Theorem 2 : We apply Corollary 2 to the sequence $\left(\frac{\gamma}{n+1}A^{1/2}Z_k\right)_{k\leqslant n}$ thanks to Lemma 8. We have:

$$\begin{aligned}\sup_{k\leqslant n}\|\frac{\gamma}{n+1}A^{1/2}Z_k\| &\leqslant \frac{c^{1/2}\|A^{1/2}\|}{(n+1)\lambda}\\ \sum_{k=1}^{n}\mathbb{E}\left[\|\frac{\gamma}{n+1}A^{1/2}Z_k\|^2|\mathcal{F}_{k-1}\right] &\leqslant \frac{1}{n+1}\frac{1}{1-\gamma/2\gamma_0}\mathrm{tr}\left(AH^{-2}\cdot C\right),\end{aligned}$$

so that,

$$\mathbb{P}\left(\left\|A^{1/2}\bar{\eta}_n\right\|\geqslant t\right)=\mathbb{P}\left(\left\|\sum_{k=1}^{n}\frac{\gamma}{n+1}A^{1/2}Z_k\right\|\geqslant t\right)\leqslant 2\exp\left(-\frac{t^2}{2\left(\frac{\mathrm{tr}\left(AH^{-2}\cdot C\right)}{(n+1)(1-\gamma/2\gamma_0)}+\frac{c^{1/2}\|A^{1/2}\|t}{3\lambda(n+1)}\right)}\right)$$

$$\mathbb{P}\left(\left\|A^{1/2}\bar{\eta}_n\right\|\geqslant t\right)\leqslant 2\exp\left(-\frac{(n+1)t^2}{\frac{2\mathrm{tr}\left(AH^{-2}\cdot C\right)}{(1-\gamma/2\gamma_0)}+\frac{2\|A^{1/2}\|c^{1/2}t}{3\lambda}}\right).$$

■

A.7.3. Tail-averaged SGD - Proof of Corollary 1

We now prove the result for tail-averaging that allow us to relax the assumption that $\eta_0=0$. The proof relies on the fact that the bias term can easily be bounded as $\|\bar{\eta}_n^{\text{tail, bias}}\|_{\mathcal{H}}\leqslant(1-\lambda\gamma)^{n/2}\|\eta_0\|_{\mathcal{H}}$. For the variance term, we can simply use the Theorem 2 for n and $n/2$, as $\bar{\eta}_n^{\text{tail}}=2\bar{\eta}_n-\bar{\eta}_{n/2}$.

Proof Proof of Corollary 1 : Let $n\geqslant 1$ and n an even number for the sake of clarity (the case where n is an odd number can be solved similarly),

$$\begin{aligned}A^{1/2}\bar{\eta}_n^{\text{tail}} &= \frac{1}{n/2}\sum_{k=n/2}^{n}A^{1/2}\eta_k\\ &= \frac{1}{n/2}\sum_{k=n/2}^{n}A^{1/2}M(k,1)\eta_0+\frac{1}{n/2}\sum_{k=n/2}^{n}A^{1/2}W_k\\ &= \frac{1}{n/2}\sum_{k=n/2}^{n}A^{1/2}M(k,1)\eta_0+2A^{1/2}\overline{W}_n-A^{1/2}\overline{W}_{n/2}.\end{aligned}$$

Hence,

$$\begin{aligned}\left\|A^{1/2}\bar{\eta}_n^{\text{tail}}\right\| &\leqslant \left\|\frac{1}{n/2}\sum_{k=n/2}^{n}A^{1/2}M(k,1)\eta_0\right\|+2\left\|A^{1/2}\overline{W}_n\right\|+\left\|A^{1/2}\overline{W}_{n/2}\right\|\\ &\leqslant \frac{1}{n/2}\sum_{k=n/2}^{n}\left\|A^{1/2}M(k,1)\right\|_{op}\|\eta_0\|+2\left\|A^{1/2}\overline{W}_n\right\|+\left\|A^{1/2}\overline{W}_{n/2}\right\|,\end{aligned}$$

Let $L_n=2\left\|A^{1/2}\overline{W}_n\right\|+\left\|A^{1/2}\overline{W}_{n/2}\right\|$,

$$\begin{aligned}\left\|A^{1/2}\bar{\eta}_n^{\text{tail}}\right\| &\leqslant \frac{1}{n/2}\sum_{k=n/2}^{n}\|A^{1/2}\|_{op}(1-\gamma\lambda)^k\|\eta_0\|+L_n\\ \left\|A^{1/2}\bar{\eta}_n^{\text{tail}}\right\| &\leqslant (1-\gamma\lambda)^{n/2}\|A^{1/2}\|_{op}\|\eta_0\|+L_n,\end{aligned}$$

And finally for $t \geqslant 0$,

$$\begin{aligned}\mathbb{P}(L_n \geqslant t) &= \mathbb{P}(2\left\|A^{1/2}\overline{W}_n\right\| + \left\|A^{1/2}\overline{W}_{n/2}\right\| \geqslant t) \\ &\leqslant \mathbb{P}\left(2\left\|A^{1/2}\overline{W}_n\right\| \geqslant t\right) + \mathbb{P}\left(\left\|A^{1/2}\overline{W}_{n/2}\right\| \geqslant t\right) \\ &\leqslant 2\left[\exp\left(-\frac{(n+1)(t/2)^2}{E_{t/2}}\right) + \exp\left(-\frac{(n/2+1)t^2}{E_t}\right)\right].\end{aligned}$$

Let us remark that $E_{t/2} \leqslant E_t$. Hence,

$$\begin{aligned}\mathbb{P}(L_n \geqslant t) &\leqslant 2\left[\exp\left(-\frac{(n+1)t^2}{4E_t}\right) + \exp\left(-\frac{(n+1)t^2}{2E_t}\right)\right] \\ &\leqslant 4\exp\left(-\frac{(n+1)t^2}{4E_t}\right).\end{aligned}$$

■

A.8. Exponentially convergent SGD for classification error

In this section we prove the results for the error in the case of SGD. Let us recall the recursion:

$$g_n - g_\lambda = \left[I - \gamma_n(K_{x_n} \otimes K_{x_n} + \lambda I)\right](g_{n-1} - g_\lambda) + \gamma_n \varepsilon_n,$$

with the noise term $\varepsilon_k = \xi_k K_{x_k} + (\tilde{g}_*(x_k) - g_\lambda(x_k))K_{x_k} - \mathbb{E}\left[(\tilde{g}_*(x_k) - g_\lambda(x_k))K_{x_k}\right] \in \mathcal{H}$. This is the same recursion as in Eq (57):

$$\eta_n = (I - \gamma H_n)\eta_{n-1} + \gamma_n \varepsilon_n,$$

with $H_n = K_{x_n} \otimes K_{x_n} + \lambda I$ and $\eta_n = g_n - g_\lambda$. First we begin by showing that for this recursion and assuming **(A2)**, **(A3)**, we can show **(H1)**, **(H2)**, **(H3)**,**(H4)**.

Lemma 9 (*Showing (H1), (H2), (H3),(H4) for SGD recursion.*)

Let us assume (A2), (A3),

- ***(H1)*** *We start at some $g_0 - g_\lambda \in \mathcal{H}$.*
- ***(H2)*** *(H_n, ε_n) i.i.d. and H_n is a positive self-adjoint operator so that almost surely $H_n \succcurlyeq \lambda I$, with $H = \mathbb{E}H_n = \Sigma + \lambda I$.*
- ***(H3)*** *We have the two following bounds on the noise:*

$$\begin{aligned}\|\varepsilon_n\| &\leqslant R(1 + 2\|\tilde{g}_* - g_\lambda\|_{L_\infty}) = c^{1/2} \\ \mathbb{E}\varepsilon_n \otimes \varepsilon_n &\preccurlyeq 2\left(1 + \|\tilde{g}_* - g_\lambda\|_\infty^2\right)\Sigma = C \\ \mathbb{E}\|\varepsilon_n\|^2 &\leqslant 2\left(1 + \|\tilde{g}_* - g_\lambda\|_\infty^2\right)\mathrm{tr}\Sigma = \mathrm{tr}C.\end{aligned}$$

- ***(H4)*** *We have:*

$$\mathbb{E}\left[H_k C H^{-1} H_k\right] \preccurlyeq \left(R^2 + 2\lambda\right)C = \gamma_0^{-1} C\,.$$

Proof : (H1), (H2) are obviously satisfied.

Let us show (H3):

$$\begin{aligned}\|\varepsilon_n\| &= \|\xi_n K_{x_n} + (\tilde{g}_*(x_n) - g_\lambda(x_n))K_{x_n} - \mathbb{E}\left[(\tilde{g}_*(x_n) - g_\lambda(x_n))K_{x_n}\right]\| \\ &\leqslant (|\xi_n| + |\tilde{g}_*(x_n) - g_\lambda(x_n)|)\|K_{x_n}\| + \mathbb{E}\left[|\tilde{g}_*(x_n) - g_\lambda(x_n)|\|K_{x_n}\|\right] \\ &\leqslant (1 + \|\tilde{g}_* - g_\lambda\|_\infty)R + \|\tilde{g}_* - g_\lambda\|_\infty R \\ &= R(1 + 2\|\tilde{g}_* - g_\lambda\|_\infty)\end{aligned}$$

We have [6]:

$$\begin{aligned}\varepsilon_n \otimes \varepsilon_n \preccurlyeq\ & 2\xi_n K_{x_n} \otimes \xi_n K_{x_n} + 2\left((\tilde{g}_*(x_n) - g_\lambda(x_n))K_{x_n} - \mathbb{E}\left[(\tilde{g}_*(x_n) - g_\lambda(x_n))K_{x_n}\right]\right) \\ & \otimes \left((\tilde{g}_*(x_n) - g_\lambda(x_n))K_{x_n} - \mathbb{E}\left[(\tilde{g}_*(x_n) - g_\lambda(x_n))K_{x_n}\right]\right)\end{aligned}$$

Moreover, $\mathbb{E}[\xi_n K_{x_n} \otimes \xi_n K_{x_n}] = \mathbb{E}[\xi_n^2 K_{x_n} \otimes K_{x_n}] \preccurlyeq \Sigma$, And,

$$\begin{aligned}&\mathbb{E}[((\tilde{g}_*(x_n) - g_\lambda(x_n))K_{x_n} - \mathbb{E}\left[(\tilde{g}_*(x_n) - g_\lambda(x_n)K_{x_n}\right]) \\ &\qquad \otimes \left((\tilde{g}_*(x_n) - g_\lambda(x_n))K_{x_n} - \mathbb{E}\left[(\tilde{g}_*(x_n) - g_\lambda(x_n))K_{x_n}\right]\right)] \\ &= \mathbb{E}\left[(\tilde{g}_*(x_n) - g_\lambda(x_n))^2(x_n)K_{x_n} \otimes K_{x_n}\right] - \mathbb{E}\left[(\tilde{g}_*(x_n) - g_\lambda(x_n))K_{x_n}\right] \\ &\qquad \otimes \mathbb{E}\left[(\tilde{g}_*(x_n) - g_\lambda(x_n))K_{x_n}\right] \\ &\preccurlyeq \mathbb{E}\left[(\tilde{g}_*(x_n) - g_\lambda(x_n))^2(x_n)K_{x_n} \otimes K_{x_n}\right] \\ &\preccurlyeq \|\tilde{g}_* - g_\lambda\|_\infty^2 \Sigma.\end{aligned}$$

So that,

$$\mathbb{E}\varepsilon_n \otimes \varepsilon_n \preccurlyeq 2\left(1 + \|\tilde{g}_* - g_\lambda\|_\infty^2\right)\Sigma$$

Finally $\mathbb{E}\varepsilon_n \otimes \varepsilon_n \preccurlyeq 2\left(1 + \|\tilde{g}_* - g_\lambda\|_\infty^2\right)\Sigma$, we have $\mathrm{tr}\mathbb{E}\varepsilon_n \otimes \varepsilon_n \leqslant 2\left(1 + \|\tilde{g}_* - g_\lambda\|_\infty^2\right)\mathrm{tr}\Sigma$, thus

$$\mathrm{tr}\mathbb{E}\varepsilon_n \otimes \varepsilon_n = \mathbb{E}\mathrm{tr}\varepsilon_n \otimes \varepsilon_n = \mathbb{E}\|\varepsilon_n\|^2 \leqslant 2\left(1 + \|\tilde{g}_* - g_\lambda\|_\infty^2\right)\mathrm{tr}\Sigma.$$

To conclude the proof of this lemma, let us show (H4). We have:

$$\begin{aligned}\mathbb{E}\Big[(K_{x_k} \otimes K_{x_k} + \lambda I)\Sigma(\Sigma + \lambda I)^{-1}(K_{x_k} \otimes K_{x_k} + \lambda I)\Big] &= \mathbb{E}\Big[K_{x_k} \otimes K_{x_k}\Sigma(\Sigma + \lambda I)^{-1}K_{x_k} \otimes K_{x_k}\Big] \\ &\quad + \lambda\Sigma\Sigma(\Sigma + \lambda I)^{-1} + \lambda\Sigma\end{aligned}$$

Moreover, $\lambda\Sigma\Sigma(\Sigma+\lambda I)^{-1} = \lambda\Sigma(\Sigma+\lambda I-\lambda I)(\Sigma+\lambda I)^{-1} = \lambda\Sigma - \lambda^2\Sigma(\Sigma+\lambda I)^{-1} \preccurlyeq \lambda\Sigma$, and similarly, $\mathbb{E}\Big[K_{x_k}\otimes K_{x_k}\Sigma(\Sigma+\lambda I)^{-1}K_{x_k}\otimes K_{x_k}\Big] = \mathbb{E}\Big[(K_{x_k}\otimes K_{x_k})^2\Big] - \lambda\mathbb{E}\Big[K_{x_k}\otimes K_{x_k}(\Sigma+\lambda I)^{-1}K_{x_k}\otimes K_{x_k}\Big] \preccurlyeq R^2\Sigma$.

Finally we obtain $\mathbb{E}\Big[(K_{x_k} \otimes K_{x_k} + \lambda I)\Sigma(\Sigma + \lambda I)^{-1}(K_{x_k} \otimes K_{x_k} + \lambda I)\Big] \preccurlyeq R^2\Sigma + \lambda\Sigma + \lambda\Sigma = (R^2 + 2\lambda)\Sigma$. ∎

A.8.1. SGD with decreasing step-size: proof of Theorem 3

Proof of Theorem 3 : Let us apply Theorem 1 to $g_n - g_\lambda$. We assume (A2), (A3) and $A = I$, such that (A2), (A3), we can show that (H1), (H2), (H3),(H4), (H5) are verified (Lemma 9). Let δ correspond to the one of (A4). We have for $t = \delta/(4R), n \geqslant 1$:

$$\|g_n - g_\lambda\|_{\mathcal{H}} \leqslant \exp\left(-\frac{\gamma\lambda}{1-\alpha}\left((n+1)^{1-\alpha} - 1\right)\right)\|g_0 - g_\lambda\|_{\mathcal{H}} + \|W_n\|_{\mathcal{H}}, \text{ a.s, with}$$

$$\mathbb{P}\left(\|W_n\|_{\mathcal{H}} \geqslant \delta/(4R)\right) \leqslant 2\exp\left(-\frac{\delta^2}{C_R}n^\alpha\right), \quad C_R = \gamma(2^{\alpha+6}R^2\mathrm{tr}C/\lambda + 8Rc^{1/2}\delta/3).$$

[6] We use the following inequality: for all a and $b \in \mathcal{H}$, $(a+b)\otimes(a+b) \preccurlyeq 2a\otimes a + 2b\otimes b$. Indeed, for all $x \in \mathcal{H}$, $\langle x, (a+b)\otimes(a+b)x\rangle = (\langle a+b, x\rangle)^2 = (\langle a, x\rangle + \langle b, x\rangle)^2 \leqslant 2\langle a, x\rangle^2 + 2\langle b, x\rangle^2 = 2\langle x, (a\otimes a)x\rangle + 2\langle x, (b\otimes b)x\rangle$.

Then if n is such that $\exp\left(-\frac{\gamma\lambda}{1-\alpha}\left((n+1)^{1-\alpha}-1\right)\right) \leqslant \frac{\delta}{5R\|g_0-g_\lambda\|_{\mathcal{H}}}$,

$$\begin{aligned}\|g_n-g_\lambda\|_{\mathcal{H}} &\leqslant \frac{\delta}{5R}+\frac{\delta}{4R}, \text{ with probability } 1-2\exp\left(-\frac{\delta^2}{C_R}n^\alpha\right),\\ \|g_n-g_\lambda\|_{\mathcal{H}} &< \frac{\delta}{2R}, \text{ with probability } 1-2\exp\left(-\frac{\delta^2}{C_R}n^\alpha\right).\end{aligned}$$

Now assume (A1), (A4), we simply apply Lemma 1 to g_n with $q = 2\exp\left(-\frac{\delta^2}{C_R}n^\alpha\right)$ And

$$\begin{aligned}C_R &= \gamma(2^{\alpha+6}R^2\mathrm{tr}C/\lambda + 8Rc^{1/2}\delta/3)\\ C_R &== \gamma\left(\frac{2^{\alpha+7}R^2\mathrm{tr}\Sigma\left(1+\|\tilde{g}_*-g_\lambda\|_\infty^2\right)}{\lambda}+\frac{8R^2\delta(1+2\|\tilde{g}_*-g_\lambda\|_\infty)}{3}\right).\end{aligned}$$

■

A.8.2. Tail averaged SGD with constant step-size: proof of Theorem 4

Proof of Theorem 4 : Let us apply Corollary 1 to $g_n - g_\lambda$. We assume (A2), (A3) and $A = I$, such that (H1), (H2), (H3),(H4), (H5) are verified (Lemma 9). Let δ correspond to the one of (A4). We have for $t = \delta/(4R), n \geqslant 1$:

$$\begin{aligned}\left\|\bar{g}_n^{\text{tail}}-g_\lambda\right\|_{\mathcal{H}} &\leqslant (1-\gamma\lambda)^{n/2}\|g_0-g_\lambda\|_{\mathcal{H}}+L_n \quad \text{, with}\\ \mathbb{P}(L_n\geqslant t) &\leqslant 4\exp\left(-(n+1)t^2/(4E_t)\right).\end{aligned}$$

Then as soon as $(1-\gamma\lambda)^{n/2} \leqslant \frac{\delta}{5R\|g_0-g_\lambda\|_{\mathcal{H}}}$,

$$\begin{aligned}\left\|\bar{g}_n^{\text{tail}}-g_\lambda\right\|_{\mathcal{H}} &\leqslant \frac{\delta}{5R}+\frac{\delta}{4R}, \text{ with probability } 1-4\exp\left(-(n+1)\delta^2/(64R^2E_{\delta/(4R)})\right),\\ \left\|\bar{g}_n^{\text{tail}}-g_\lambda\right\|_{\mathcal{H}} &< \frac{\delta}{2R}, \text{ with probability } 1-4\exp\left(-(n+1)\delta^2/(64R^2E_{\delta/(4R)})\right).\end{aligned}$$

Now assume (A1), (A4), we simply apply Lemma 1 to $\bar{g}_n^{\text{tail}}$ with $q = 4\exp\left(-(n+1)\delta^2/K_R\right)$. And

$$\begin{aligned}K_R = 64R^2E_{\delta/(4R)} &= 64R^2\left(4\mathrm{tr}(H^{-2}C)+\frac{2c^{1/2}}{3\lambda}\cdot\frac{\delta}{4R}\right)\\ &= 512R^2\left(1+\|\tilde{g}_*-g_\lambda\|_\infty^2\right)\mathrm{tr}((\Sigma+\lambda I)^{-2}\Sigma)+\frac{32\delta R^2(1+2\|\tilde{g}_*-g_\lambda\|_\infty)}{3\lambda}.\end{aligned}$$

■

A.9. Extension of Corollary 1 and Theorem 4 for the full averaged case.

A.9.1. Extension of Corollary 1 for the full averaged case.

Let us recall the SGD abstract recursion defined in Eq. (57) that we are going to further apply with $\eta_n = g_n - g_\lambda$, $H_n = K_{x_n}\otimes K_{x_n}+\lambda I$ and $H = \Sigma+\lambda I$:

$$\begin{aligned}\eta_n &= (I-\gamma H_n)\eta_{n-1}+\gamma_n\varepsilon_n,\\ \eta_n &= \underbrace{M(n,1)\eta_0}_{\eta_n^{\text{bias}}}+\underbrace{\sum_{k=1}^{n}\gamma_k M(n,k+1)\varepsilon_k}_{\eta_n^{\text{variance}}}.\end{aligned}$$

Notations. The second term, η_n^{variance}, is treated by Theorem 2 of the article. Now consider that $\eta_0 \neq 0$ and let us bound the initial condition term i.e., $\eta_n^{\text{bias}} = M(n,1)\eta_0$. Let us define also an auxiliary sequence (u_n) that follows the same recursion as η_n^{bias} but with H:

$$\begin{aligned} \eta_n^{\text{bias}} &= (I - \gamma H_n)\eta_{n-1}^{\text{bias}} \\ u_n &= (I - \gamma H)u_{n-1}, \qquad u_0 = \eta_0^{\text{bias}} = \eta_0. \end{aligned}$$

We define $w_n = \eta_n^{\text{bias}} - u_n$ and as always we consider the first n average of each of these sequences that we are going to denote $\bar{w}_n$, $\bar{\eta}_n^{\text{bias}}$ and $\bar{u}_n$ respectively.

Note $\tilde{\varepsilon}_n = (H - H_n)\eta_{n-1}^{\text{bias}}$ and $\tilde{H}_n = H$, then w_n follows the recursion : $w_0 = 0$, and

$$w_n = (I - \gamma \tilde{H}_n)w_{n-1} + \gamma \tilde{\varepsilon}_n. \tag{73}$$

Thus, w_n follows the same recursion as Eq.(57) with $(\tilde{H}_n, \tilde{\varepsilon}_n)$. We thus have the following corollary:

Corollary 3

Assume that the sequence (w_n) defined in Eq. (73) verifies (H1), (H2), (H3), (H4) and (H5) with $(\tilde{H}_n, \tilde{\varepsilon}_n)$, then for $t > 0, n \geqslant 1$:

$$\mathbb{P}\left(\left\|A^{1/2}\bar{w}_n\right\|_{\mathcal{H}} \geqslant t\right) \leqslant 2\exp\left[-\frac{(n+1)t^2}{\tilde{E}_t}\right],$$

where $\tilde{E}_t$ is defined with respect to the constants introduced in the assumptions (with a tilde):

$$\tilde{E}_t = 4\mathrm{tr}(AH^{-2}\tilde{C}) + \frac{2\tilde{c}^{1/2}\|A^{1/2}\|_{op}}{3\lambda} \cdot t.$$

Proof : Apply Theorem 2 to the sequence (w_n) defined in Eq. (73). ■

Now, we can decompose η_n in three terms: $\eta_n = \eta_n^{\text{bias}} + \eta_n^{\text{variance}} = w_n + u_n + \eta_n^{\text{variance}}$. We can thus state the following general result:

Theorem 6

Assume (H1), (H2), (H3), (H4), (H5) for both (H_n, ε_n) and $(\tilde{H}_n, \tilde{\varepsilon}_n)$, and consider the average of the sequence defined in Eq. (57). We have for $t > 0, n \geqslant 1$:

$$\left\|A^{1/2}\bar{\eta}_n\right\|_{\mathcal{H}} \leqslant \frac{\left\|A^{1/2}\right\|\|\eta_0\|_{\mathcal{H}}}{(n+1)\gamma\lambda} + L_n \quad \text{, with} \tag{74}$$

$$\mathbb{P}(L_n \geqslant t) \leqslant 4\exp\left(-\frac{(n+1)t^2}{\max(E_t, \tilde{E}_t)}\right). \tag{75}$$

Proof of Theorem 6 : As $\bar{\eta}_n = \bar{\eta}_n^{\text{bias}} + \bar{\eta}_n^{\text{variance}} = \bar{w}_n + \bar{u}_n + \bar{\eta}_n^{\text{variance}}$, we are going to bound $\bar{u}_n$, then the sum $\bar{w}_n + \bar{\eta}_n^{\text{variance}}$.

First, $\|\bar{u}_n\| = \left\|\frac{1}{n+1}\sum_{k=0}^{n} u_k\right\| \leqslant \frac{1}{n+1}\sum_{k=0}^{n}\|u_k\| \leqslant \frac{1}{n+1}\sum_{k=0}^{n}(1-\gamma\lambda)^k\|\eta_0\| \leqslant \frac{\|\eta_0\|}{(n+1)\gamma\lambda}$.

Thus, we have:

$$\left\|A^{1/2}\bar{\eta}_n\right\| \leqslant \frac{\left\|A^{1/2}\right\|\|\eta_0\|}{(n+1)\gamma\lambda} + \left\|A^{1/2}\bar{w}_n\right\| + \left\|A^{1/2}\bar{\eta}_n^{\text{variance}}\right\|,$$

Let $L_n = \left\|A^{1/2}\bar{w}_n\right\| + \left\|A^{1/2}\bar{\eta}_n^{\text{variance}}\right\|$, for $t \geqslant 0$,

$$\begin{aligned}\mathbb{P}(L_n \geqslant t) &= \mathbb{P}(\left\|A^{1/2}\bar{w}_n\right\| + \left\|A^{1/2}\bar{\eta}_n^{\text{variance}}\right\| \geqslant t) \\ &\leqslant \mathbb{P}\left(\left\|A^{1/2}\bar{w}_n\right\| \geqslant t\right) + \mathbb{P}\left(\left\|A^{1/2}\bar{\eta}_n^{\text{variance}}\right\| \geqslant t\right) \\ &\leqslant 2\left[\exp\left[-\frac{(n+1)t^2}{\tilde{E}_t}\right] + \exp\left[-\frac{(n+1)t^2}{E_t}\right]\right].\end{aligned}$$

Hence,

$$\mathbb{P}(L_n \geqslant t) \leqslant 4\exp\left(-\frac{(n+1)t^2}{\max(E_t, \tilde{E}_t)}\right). \qquad \blacksquare$$

A.9.2. Extension of Theorem 4 for the full averaged case.

Same situation here, we want to apply full averaged SGD instead of the tail-averaged technique.

Theorem 7

Assume (A1), (A2), (A3), (A4) and $\gamma_n = \gamma$ for any n, $\gamma\lambda < 1$ and $\gamma \leqslant \gamma_0 = (R^2+\lambda)^{-1}$. Let $\bar{g}_n$ be the average of the first n iterate of the SGD recursion defined in Eq. (56), as soon as: $n \geqslant \dfrac{5R\|g_0 - g_\lambda\|_{\mathcal{H}}}{\lambda\gamma\delta}$, then

$$\mathcal{R}(\bar{g}_n^{tail}) = \mathcal{R}^*, \text{ with probability at least } 1 - 4\exp\left(-\delta^2 K_R(n+1)\right),$$

and in particular

$$\mathbb{E}\mathcal{R}(\bar{g}_n^{tail}) - \mathcal{R}^* \leqslant 4\exp\left(-\delta^2 K_R(n+1)\right),$$

with

$$K_R^{-1} = \max\left\{\begin{array}{l} 128R^2\left(1 + \|\tilde{g}_* - g_\lambda\|_\infty^2\right)\operatorname{tr}((\Sigma+\lambda I)^{-2}\Sigma) + \dfrac{8R^2(1+2\|\tilde{g}_* - g_\lambda\|_\infty)}{3\lambda} \\ 64R^4\|g_0 - g_\lambda\|_{\mathcal{H}}\operatorname{tr}((\Sigma+\lambda I)^{-2}\Sigma) + \dfrac{16R^4\|g_0 - g_\lambda\|_{\mathcal{H}}}{3\lambda}. \end{array}\right\}$$

Proof of Theorem 7 : We want to apply Theorem 6 to the SGD recursion. We thus want to check that assumptions **(H1)**, **(H2)**, **(H3)**, **(H4)**, **(H5)** are verified for both (H_n, ε_n) and $(\tilde{H}_n, \tilde{\varepsilon}_n)$. For the recursion involving (H_n, ε_n), this corresponds to Lemma 9. For the recursion involving $(\tilde{H}_n = H, \tilde{\varepsilon}_n = (H - H_n)M(n-1,1)(g_0 - g_\lambda)$, this corresponds to the following lemma:

Lemma 10 (*Showing (H1), (H2), (H3), (H4) for the auxiliary recursion.*)

Let us assume (A2), (A3),

- ***(H1)*** *We start at some $g_0 - g_\lambda \in \mathcal{H}$.*
- ***(H2)*** *$(\tilde{H}_n, \tilde{\varepsilon}_n)$ i.i.d. and $\tilde{H}_n$ is a positive self-adjoint operator so that almost surely $\tilde{H}_n \succcurlyeq \lambda I$, with $H = \mathbb{E}\tilde{H}_n = \Sigma + \lambda I$.*
- ***(H3)*** *We have the two following bounds on the noise:*

$$\begin{aligned}\|\tilde{\varepsilon}_n\| &\leqslant 2R^2\|g_0 - g_\lambda\|_{\mathcal{H}} = \tilde{c}^{1/2} \\ \mathbb{E}\tilde{\varepsilon}_n \otimes \tilde{\varepsilon}_n &\preccurlyeq R^2\|g_0 - g_\lambda\|_{\mathcal{H}}\Sigma = \tilde{C} \\ \mathbb{E}\|\tilde{\varepsilon}_n\|^2 &\leqslant R^2\|g_0 - g_\lambda\|_{\mathcal{H}}\operatorname{tr}\Sigma = \operatorname{tr}\tilde{C}.\end{aligned}$$

- ***(H4)*** *We have:*

$$\mathbb{E}\left[\tilde{H}_k\tilde{C}H^{-1}\tilde{H}_k\right] \preccurlyeq \left(R^2+\lambda\right)\tilde{C} = \tilde{\gamma_0}^{-1}\tilde{C}.$$

Proof : (H1), (H2) are obviously satisfied.

Let us show (H3): For the first one:

$$\begin{aligned}\|\tilde{\varepsilon}_n\| &= \|(H - H_n)M(n-1,1)(g_0 - g_\lambda)\| \\ &\leqslant \|(\Sigma - K_{x_n} \otimes K_{x_n})\| \, \|M(n-1,1)\| \, \|g_0 - g_\lambda\| \\ &\leqslant 2R^2 \|g_0 - g_\lambda\|_{\mathcal{H}}.\end{aligned}$$

$$\begin{aligned}\|\tilde{\varepsilon}_n\| &= \|(H - H_n)M(n-1,1)(g_0 - g_\lambda)\| \\ &\leqslant \|(\Sigma - K_{x_n} \otimes K_{x_n})\| \, \|M(n-1,1)\| \, \|g_0 - g_\lambda\| \\ &\leqslant 2R^2 \|g_0 - g_\lambda\|_{\mathcal{H}}.\end{aligned}$$

And for the second inquality:

$$\begin{aligned}\mathbb{E}\left[\tilde{\varepsilon}_n \otimes \tilde{\varepsilon}_n | \mathcal{F}_{n-1}\right] &= \mathbb{E}\left[(\Sigma - K_{x_n} \otimes K_{x_n}) \, \eta_n^{\text{bias}} \otimes \eta_n^{\text{bias}} \, (\Sigma - K_{x_n} \otimes K_{x_n}) \, | \mathcal{F}_{n-1}\right] \\ &= \Sigma \eta_n^{\text{bias}} \otimes \eta_n^{\text{bias}} \Sigma - 2\Sigma \eta_n^{\text{bias}} \otimes \eta_n^{\text{bias}} \Sigma + \mathbb{E}\left[K_{x_n} \otimes K_{x_n} \eta_n^{\text{bias}} \otimes \eta_n^{\text{bias}} K_{x_n} \otimes K_{x_n}\right] \\ &= -\Sigma \eta_n^{\text{bias}} \otimes \eta_n^{\text{bias}} \Sigma + \mathbb{E}\left[\langle K_{x_n}, \eta_n^{\text{bias}} \rangle^2 K_{x_n} \otimes K_{x_n}\right] \\ &\preccurlyeq R^2 \|g_0 - g_\lambda\|_{\mathcal{H}} \Sigma.\end{aligned}$$

Finally, we have for (H4) :

$$\begin{aligned}\mathbb{E}\left[\tilde{H}_k \tilde{C} H^{-1} \tilde{H}_k\right] = H\tilde{C} = R^2 \|g_0 - g_\lambda\|_{\mathcal{H}} (\Sigma^2 + \lambda\Sigma) &\preccurlyeq R^2 \|g_0 - g_\lambda\|_{\mathcal{H}} (\|\Sigma\|_{\text{op}} + \lambda)\Sigma \\ &\preccurlyeq \left(R^2 + \lambda\right) \tilde{C} = \tilde{\gamma_0}^{-1} \tilde{C}.\end{aligned}$$

■

Let us apply now Theorem 6 to $g_n - g_\lambda$. We assume (A2), (A3) and $A = I$, such that (H1), (H2), (H3), (H4), (H5) are verified for both problems $((H_n, \varepsilon_n)$ and $(\tilde{H}_n, \tilde{\varepsilon}_n))$ (Lemma 9,10). Let δ correspond to the one of Assumption (A4). We have for $t = \delta/(4R), n \geqslant 1$:

$$\begin{aligned}\|\bar{g}_n - g_\lambda\|_{\mathcal{H}} &\leqslant \frac{\|g_0 - g_\lambda\|_{\mathcal{H}}}{(n+1)\gamma\lambda} + L_n \quad \text{, with} \\ \mathbb{P}(L_n \geqslant t) &\leqslant 4 \exp\left(-\frac{(n+1)t^2}{\max(E_t, \tilde{E}_t)}\right).\end{aligned}$$

Then as soon as $\dfrac{1}{(n+1)\lambda\gamma} \leqslant \dfrac{\delta}{5R\|g_0 - g_\lambda\|_{\mathcal{H}}}$,

$$\begin{aligned}\|\bar{g}_n - g_\lambda\|_{\mathcal{H}} &\leqslant \frac{\delta}{5R} + \frac{\delta}{4R}, \text{ with probability } 1 - 4\exp\left(-\frac{(n+1)\delta^2}{16R^2 \max(E_{\delta/4R}, \tilde{E}_{\delta/4R})}\right), \\ \|\bar{g}_n - g_\lambda\|_{\mathcal{H}} &< \frac{\delta}{2R}, \text{ with probability } 1 - 4\exp\left(-\frac{(n+1)\delta^2}{16R^2 \max(E_{\delta/4R}, \tilde{E}_{\delta/4R})}\right).\end{aligned}$$

Now assume (A1), (A4), we now only have to apply Lemma 1 to the estimator $\bar{g}_n$ with the probability $q = 4\exp\left(-\dfrac{(n+1)\delta^2}{16R^2 \max(E_{\delta/4R}, \tilde{E}_{\delta/4R})}\right)$. And,

$$\begin{aligned}K_R^{-1} &= 16R^2 \max(E_{\delta/4R}, \tilde{E}_{\delta/4R}) \\ &= \max\begin{cases} 128R^2 \left(1 + \|\tilde{g}_* - g_\lambda\|_\infty^2\right) \text{tr}((\Sigma + \lambda I)^{-2}\Sigma) + \dfrac{8R^2(1 + 2\|\tilde{g}_* - g_\lambda\|_\infty)}{3\lambda} \\ 64R^4 \|g_0 - g_\lambda\|_{\mathcal{H}} \text{tr}((\Sigma + \lambda I)^{-2}\Sigma) + \dfrac{16R^4 \|g_0 - g_\lambda\|_{\mathcal{H}}}{3\lambda}. \end{cases}\end{aligned}$$

■

A.10. Convergence rate under weaker margin assumption

We make the following assumptions:

(A7) $\forall \delta > 0,\ \mathbb{P}\left(|g_*| \leqslant 2\delta\right) \leqslant \delta^{\alpha}$.

(A8) *There exists* [7] $\gamma > 0$ *such that* $\forall \lambda > 0,\ \|g_* - g_\lambda\|_\infty \leqslant \lambda^{\gamma}$.

(A9) *The eigenvalues of* Σ *decrease as* $1/n^{\beta}$ *for* $\beta > 1$.

Note that **(A7)** is weaker than **(A1)** and to balance this we need a stronger condition on g_λ than **(A4)** which is **(A8)**. **(A9)** is just a technical assumption needed to give explicit rate. The following Corollary corresponds to Theorem 4 with the new assumptions. Note that it could also be shown for the full average sequence $\bar{g}_n$.

Corollary 4 (*Explicit onvergence rate under weaker margin condition*)

Assume **(A2)**, **(A3)**, **(A7)**, **(A8)** *and* **(A9)**. *Let* $\gamma_n = \gamma$ *for any n,* $\gamma\lambda < 1$ *and* $\gamma \leqslant \gamma_0 = (R^2 + 2\lambda)^{-1}$. *Let* $\bar{g}_n^{tail}$ *be the n-th iterate of the recursion defined in Eq. (56), and* $\bar{g}_n^{tail} = \frac{1}{\lfloor n/2 \rfloor} \sum_{i=\lfloor n/2 \rfloor}^{n} g_i$, *as soon as* $n \geqslant \dfrac{2}{\gamma\lambda} \ln\left(\dfrac{5R\|g_0 - g_\lambda\|_{\mathcal{H}}}{\delta}\right)$, *then*

$$\mathbb{E}\left[R(\bar{g}_n^{tail}) - R^*\right] \leqslant \frac{C_{\alpha,\beta}}{n^{\alpha \cdot q_{\gamma,\beta}}}.$$

Proof : The proof technique follows the one of [AT07].

Let $\delta, \lambda > 0$, such that $\|g_* - g_\lambda\|_\infty \leqslant \delta$. Remark that $\forall j \in \mathbb{N}$,
$\mathbb{P}\left(\text{sign}(g_*(X))g_\lambda(X) \leqslant 2^j\delta\right) \leqslant \mathbb{P}\left(|g_\lambda(X)| \leqslant 2^j\delta\right) \leqslant \mathbb{P}\left(|g_*(X)| \leqslant 2^{j+1}\delta\right) \leqslant 2^{\alpha j}\delta^{\alpha}$.
Note $A_0 = \{x \in \mathcal{X} |\ \text{sign}(g_*)g_\lambda \leqslant \delta\}$ and for $j \geqslant 1$, $A_j = \{x \in \mathcal{X} |\ 2^{j-1}\delta < \text{sign}(g_*)g_\lambda \leqslant 2^j\delta\}$. Then,

$$\begin{aligned}
\mathbb{E}\left[R(\bar{g}_n^{\text{tail}}) - R^*\right] &= \sum_{j\in\mathbb{N}} \mathbb{E}\left[\left(R(\bar{g}_n^{\text{tail}}) - R^*\right)\mathbf{1}_{A_j}\right] \\
&= \mathbb{E}\left[\left(R(\bar{g}_n^{\text{tail}}) - R^*\right)\mathbf{1}_{\text{sign}(g_*)g_\lambda \leqslant \delta}\right] + \sum_{j\geqslant 1} \mathbb{E}\left[\left(R(\bar{g}_n^{\text{tail}}) - R^*\right)\mathbf{1}_{A_j}\right] \\
&\leqslant \mathbb{P}\left(\text{sign}(g_*(X))g_\lambda(X) \leqslant \delta\right) + \sum_{j\geqslant 1} \mathbb{E}\left[\left(R(\bar{g}_n^{\text{tail}}) - R^*\right)\mathbf{1}_{2^{j-1}\delta < \text{sign}(g_*(X))g_\lambda(X) \leqslant 2^j\delta}\right] \\
&\leqslant \delta^{\alpha} + \sum_{j\geqslant 1} \mathbb{E}_X\left[\mathbb{E}_{x_1,\dots,x_n}\left[\underbrace{\left(R(\bar{g}_n^{\text{tail}}) - R^*\right)\mathbf{1}_{2^{j-1}\delta < \text{sign}(g_*(X))g_\lambda(X)}}_{\text{Theorem 4}} | x_1,\dots,x_n\right]\right. \\
&\qquad \left. \cdot \mathbf{1}_{\text{sign}(g_*(X))g_\lambda(X) \leqslant 2^j\delta}\right] \\
&\leqslant \delta^{\alpha} + 4\sum_{j\geqslant 1} \mathbb{P}\left(\text{sign}(g_*(X))g_\lambda(X) \leqslant 2^j\delta\right) \exp\left(-(2^j\delta)^2 K_R(\delta)(n+1)\right) \\
&\leqslant \delta^{\alpha} + 4\delta^{\alpha}\sum_{j\geqslant 1} 2^{\alpha j} \exp\left(-(2^j\delta)^2 K_R(\delta)(n+1)\right),
\end{aligned}$$

and $K_R(\delta)^{-1} = 2^9R^2\left(1 + \|\tilde{g}_* - g_\lambda\|_\infty^2\right)\text{tr}(\Sigma(\Sigma + \lambda I)^{-2}) + \dfrac{32\delta R^2(1 + 2\|\tilde{g}_* - g_\lambda\|_\infty)}{3\lambda}$. Let us now choose δ as a function of n to cancel the dependence on n in the exponential term. In the following, as we

[7] This assumption is verified for the following source condition $\exists g \in \mathcal{H}, r > 0$ s.t. $\mathbb{P}_{\mathcal{H}}(g) = \Sigma^r g_*$. If the additionnal assumption **(A9)** is verified then **(A8)** is verified with $\gamma = \frac{r-1/2}{2r+1/\beta}$ [CDV07].

assumed **(A8)**, we chose $\lambda = \delta^{1/\gamma}$ such that $\|g_* - g_\lambda\|_\infty \leqslant \lambda^\gamma = \delta$. Second, **(A9)** implies (see [CDV07]) that $\operatorname{tr}(\Sigma(\Sigma + \lambda I)^{-2}) \leqslant \dfrac{\beta}{(\beta-1)\lambda^{1+1/\beta}}$. For δ small enough, we have:

$$K_R(\delta)^{-1} \leqslant 2^{10} \frac{\beta R^2}{(\beta-1)\delta^{\frac{1+1/\beta}{\gamma}}} + 32\delta^{(\gamma-1)/\gamma} R^2$$
$$K_R(\delta)^{-1} \leqslant 2^{11} \frac{\beta R^2}{(\beta-1)} \cdot \delta^{-(\beta+1)/\beta\gamma}$$

Hence, if we take $\delta^2 \delta^{(\beta+1)/\beta\gamma} = 1/n$, i.e., $\delta = n^{-\gamma/(2\gamma+1+1/\beta)}$, we have:

$$\mathbb{E}\left[R(\bar{g}_n^{\text{tail}}) - R^*\right] \leqslant \frac{1 + \sum_{j\geqslant 1} 2^{\alpha j+2} \exp\left(-4^j(\beta-1)/(2^{11}\beta R^2)\right)}{n^{\alpha\gamma/(2\gamma+1+1/\beta)}}.$$

As the sum converges, we have proved the result. ■

⋆

⋆ ⋆

2. Statistical Optimality of SGD on Hard Learning Problems through Multiple Passes

2.1. Introduction

Stochastic gradient descent (SGD) and its multiple variants —averaged [PJ92], accelerated [Lan12], variance-reduced [RSB12, JZ13, DBLJ14]— are the workhorses of large-scale machine learning, because (a) these methods looks at only a few observations before updating the corresponding model, and (b) they are known in theory and in practice to generalize well to unseen data [BCN18].

Beyond the choice of step-size (often referred to as the learning rate), the number of passes to make on the data remains an important practical and theoretical issue. In the context of finite-dimensional models (least-squares regression or logistic regression), the theoretical answer has been known for many years: a single passes suffices for the optimal statistical performance [PJ92, NY83]. Worse, most of the theoretical work only apply to single pass algorithms, with some exceptions leading to analyses of multiple passes when the step-size is taken smaller than the best known setting [HRS16, LR17].

However, in practice, multiple passes are always performed as they empirically lead to better generalization (e.g., loss on unseen test data) [BCN18]. But no analysis so far has been able to show that, given the appropriate step-size, multiple pass SGD was theoretically better than single pass SGD.

The main contribution of this paper is to show that for least-squares regression, while single pass averaged SGD is optimal for a certain class of "easy" problems, multiple passes are needed to reach optimal prediction performance on another class of "hard" problems.

In order to define and characterize these classes of problems, we need to use tools from infinite-dimensional models which are common in the analysis of kernel methods. De facto, our analysis will be done in infinite-dimensional feature spaces, and for finite-dimensional problems where the dimension far exceeds the number of samples, using these tools are the only way to obtain non-vacuous dimension-independent bounds. Thus, overall, our analysis applies both to finite-dimensional models with explicit features (parametric estimation), and to kernel methods (non-parametric estimation).

The two important quantities in the analysis are:

(a) The decay of eigenvalues of the covariance matrix Σ of the input features, so that the ordered eigenvalues λ_m decay as $O(m^{-\alpha})$; the parameter $\alpha \geqslant 1$ characterizes the size of the feature space, $\alpha = 1$ corresponding to the largest feature spaces and $\alpha = +\infty$ to finite-dimensional spaces. The decay will be measured through $\mathrm{tr}\Sigma^{1/\alpha} = \sum_m \lambda_m^{1/\alpha}$, which is small when the decay of eigenvalues is faster than $O(m^{-\alpha})$.

(b) The complexity of the optimal predictor θ_* as measured through the covariance matrix Σ, that is with coefficients $\langle e_m, \theta_* \rangle$ in the eigenbasis $(e_m)_m$ of the covariance matrix that decay so that $\langle \theta_*, \Sigma^{1-2r}\theta_* \rangle$ is small. The parameter $r \geqslant 0$ characterizes the difficulty of the learning problem: $r = 1/2$ corresponds to characterizing the complexity of the predictor through the squared norm $\|\theta_*\|^2$, and thus r close to zero corresponds to the hardest problems while r larger, and in particular $r \geqslant 1/2$, corresponds to simpler problems.

Dealing with non-parametric estimation provides a simple way to evaluate the optimality of learning procedures. Indeed, given problems with parameters r and α, the best prediction performance (averaged square loss on unseen data) is well known [FS17] and decay as $O(n^{\frac{-2r\alpha}{2r\alpha+1}})$, with $\alpha = +\infty$ leading to the usual parametric rate $O(n^{-1})$. For *easy problems*, that is for which $r \geqslant \frac{\alpha-1}{2\alpha}$, then it is known that most iterative algorithms achieve this optimal rate of convergence (but with various running-time complexities),

such as exact regularized risk minimization [CDV07], gradient descent on the empirical risk [YRC07], or averaged stochastic gradient descent [DB16].

We show that for *hard problems*, that is for which $r \leqslant \frac{\alpha-1}{2\alpha}$ (see Example 2 for a typical hard problem), then multiple passes are superior to a single pass. More precisely, under additional assumptions detailed in Section 2.2 that will lead to a subset of the hard problems, with $\Theta(n^{(\alpha-1-2r\alpha)/(1+2r\alpha)})$ passes, we achieve the optimal statistical performance $O(n^{\frac{-2r\alpha}{2r\alpha+1}})$, while for all other hard problems, a single pass only achieves $O(n^{-2r})$. This is illustrated in Figure 13.

We thus get a number of passes that grows with the number of observations n and depends precisely on the quantities r and α. In synthetic experiments with kernel methods where α and r are known, these scalings are precisely observed. In experiments on parametric models with large dimensions, we also exhibit an increasing number of required passes when the number of observations increases.

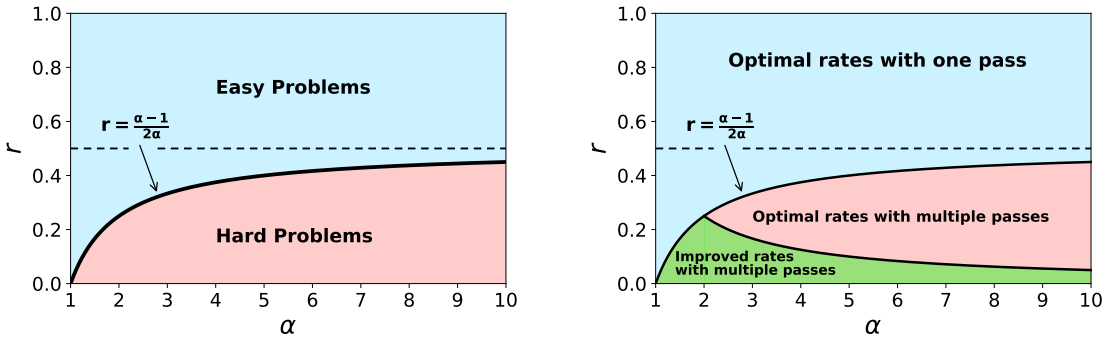

Figure 13: (Left) easy and hard problems in the (α, r)-plane. (Right) different regions for which multiple passes improved known previous bounds (green region) or reaches optimality (red region).

2.2. Least-squares regression in finite dimension

We consider a joint distribution ρ on pairs of input/output $(x, y) \in \mathcal{X} \times \mathbb{R}$, where $\mathcal{X}$ is any input space, and we consider a feature map Φ from the input space $\mathcal{X}$ to a feature space $\mathcal{H}$, which we assume Euclidean in this section, so that all quantities are well-defined. In Section 2.4, we will extend all the notions to Hilbert spaces.

2.2.1. Main assumptions

We are considering predicting y as a linear function $f_\theta(x) = \langle \theta, \Phi(x) \rangle_{\mathcal{H}}$ of $\Phi(x)$, that is estimating $\theta \in \mathcal{H}$ such that $F(\theta) = \frac{1}{2}\mathbb{E}(y - \langle \theta, \Phi(x) \rangle_{\mathcal{H}})^2$ is as small as possible. Estimators will depend on n observations, with standard sampling assumptions:

(A6) *The n observations $(x_i, y_i) \in \mathcal{X} \times \mathbb{R}$, $i = 1, \dots, n$, are independent and identically distributed from the distribution ρ.*

Since $\mathcal{H}$ is finite-dimensional, F always has a (potentially non-unique) minimizer in $\mathcal{H}$ which we denote θ_*. We make the following standard boundedness assumptions:

(A7) *$\|\Phi(x)\| \leqslant R$ almost surely, $|y - \langle \theta_*, \Phi(x) \rangle_{\mathcal{H}}|$ is almost surely bounded by σ and $|y|$ is almost surely bounded by M.*

In order to obtain improved rates with multiple passes, and motivated by the equivalent previously used condition in reproducing kernel Hilbert spaces presented in Section 2.4, we make the following extra assumption (we denote by $\Sigma = \mathbb{E}[\Phi(x) \otimes_{\mathcal{H}} \Phi(x)]$ the (non-centered) covariance matrix).

(A8) *For $\mu \in [0,1]$, there exists $\kappa_\mu \geqslant 0$ such that, almost surely, $\Phi(x) \otimes_{\mathcal{H}} \Phi(x) \preccurlyeq_{\mathcal{H}} \kappa_\mu^2 R^{2\mu} \Sigma^{1-\mu}$. Note that it can also be written as $\|\Sigma^{\mu/2-1/2}\Phi(x)\|_{\mathcal{H}} \leqslant \kappa_\mu R^\mu$.*

Assumption (A8) is always satisfied with any $\mu \in [0,1]$, and has particular values for $\mu = 1$, with $\kappa_1 = 1$, and $\mu = 0$, where κ_0 has to be larger than the dimension of the space $\mathcal{H}$.

We will also introduce a parameter α that characterizes the decay of eigenvalues of Σ through the quantity $\mathrm{tr}\Sigma^{1/\alpha}$, as well as the difficulty of the learning problem through $\|\Sigma^{1/2-r}\theta_*\|_{\mathcal{H}}$, for $r \in [0,1]$. In the finite-dimensional case, these quantities can always be defined and most often finite, but may be very large compared to sample size. In the following assumptions the quantities are assumed to be finite and small compared to n.

(A9) *There exists $\alpha > 1$ such that* $\mathrm{tr}\ \Sigma^{1/\alpha} < \infty$.

Assumption (A9) is often called the "capacity condition". First note that this assumption implies that the decreasing sequence of the eigenvalues of Σ, $(\lambda_m)_{m \geqslant 1}$, satisfies $\lambda_m = o(1/m^\alpha)$. Note that $\mathrm{tr}\Sigma^\mu \leqslant \kappa_\mu^2 R^{2\mu}$ and thus often we have $\mu \geqslant 1/\alpha$, and in the most favorable cases in Section 2.4, this bound will be achieved. We also assume:

(A10) *There exists $r \geqslant 0$, such that $\|\Sigma^{1/2-r}\theta_*\|_{\mathcal{H}} < \infty$.*

Assumption (A10) is often called the "source condition". Note also that for $r = 1/2$, this simply says that the optimal predictor has a small norm.

In the subsequent sections, we essentially assume that α, μ and r are chosen (by the theoretical analysis, not by the algorithm) so that all quantities R_μ, $\|\Sigma^{1/2-r}\theta_*\|_{\mathcal{H}}$ and $\mathrm{tr}\Sigma^{1/\alpha}$ are finite and small. As recalled in the introduction, these parameters are often used in the non-parametric literature to quantify the hardness of the learning problem (Figure 13).

We will use result with $O(\cdot)$ and $\Theta(\cdot)$ notations, which will all be independent of n and t (number of observations and number of iterations) but can depend on other finite constants. Explicit dependence on all parameters of the problem is given in proofs. More precisely, we will use the usual $O(\cdot)$ and $\Theta(\cdot)$ notations for sequences b_{nt} and a_{nt} that can depend on n and t, as $a_{nt} = O(b_{nt})$ if and only if, there exists $M > 0$ such that for all n, t, $a_{nt} \leqslant M b_{nt}$, and $a_{nt} = \Theta(b_{nt})$ if and only if, there exist $M, M' > 0$ such that for all n, t, $M' b_{nt} \leqslant a_{nt} \leqslant M b_{nt}$.

2.2.2. Related work

Given our assumptions above, several algorithms have been developed for obtaining low values of the expected excess risk $\mathbb{E}\big[F(\theta)\big] - F(\theta_*)$.

Regularized empirical risk minimization. Forming the empirical risk $\hat{F}(\theta)$, it minimizes $\hat{F}(\theta) + \lambda\|\theta\|_{\mathcal{H}}^2$, for appropriate values of λ. It is known that for easy problems where $r \geqslant \frac{\alpha-1}{2\alpha}$, it achieves the optimal rate of convergence $O(n^{\frac{-2r\alpha}{2r\alpha+1}})$ [CDV07]. However, algorithmically, this requires to solve a linear system of size n times the dimension of $\mathcal{H}$. One could also use fast variance-reduced stochastic gradient algorithms such as SAG [RSB12], SVRG [JZ13] or SAGA [DBLJ14], with a complexity proportional to the dimension of $\mathcal{H}$ times $n + R^2/\lambda$.

Early-stopped gradient descent on the empirical risk. Instead of solving the linear system directly, one can use gradient descent with early stopping [YRC07, LRRC18]. Similarly to the regularized empirical risk minimization case, a rate of $O(n^{-\frac{2r\alpha}{2r\alpha+1}})$ is achieved for the easy problems, where $r \geqslant \frac{\alpha-1}{2\alpha}$. Different

iterative regularization techniques beyond batch gradient descent with early stopping have been considered, with computational complexities ranging from $O(n^{1+\frac{\alpha}{2r\alpha+1}})$ to $O(n^{1+\frac{\alpha}{4r\alpha+2}})$ times the dimension of $\mathcal{H}$ (or n in the kernel case in Section 2.4) for optimal predictions [YRC07, GRO+08, RV15, BK16, LRRC18].

Stochastic gradient. The usual stochastic gradient recursion is iterating from $i = 1$ to n,

$$\theta_i = \theta_{i-1} + \gamma\big(y_i - \langle \theta_{i-1}, \Phi(x_i)\rangle_{\mathcal{H}}\big)\Phi(x_i),$$

with the averaged iterate $\bar{\theta}_n = \frac{1}{n}\sum_{i=1}^n \theta_i$. Starting from $\theta_0 = 0$, [BM13] shows that the expected excess performance $\mathbb{E}[F(\bar{\theta}_n)] - F(\theta_*)$ decomposes into a *variance* term that depends on the noise σ^2 in the prediction problem, and a *bias term*, that depends on the deviation $\theta_* - \theta_0 = \theta_*$ between the initialization and the optimal predictor. Their bound is, up to universal constants, $\frac{\sigma^2 \dim(\mathcal{H})}{n} + \frac{\|\theta_*\|^2_{\mathcal{H}}}{\gamma n}$.

Further, [DB16] considered the quantities α and r above to get the bound, up to constant factors:

$$\frac{\sigma^2 \mathrm{tr}\Sigma^{1/\alpha}(\gamma n)^{1/\alpha}}{n} + \frac{\|\Sigma^{1/2-r}\theta_*\|^2}{\gamma^{2r} n^{2r}}.$$

We recover the finite-dimensional bound for $\alpha = +\infty$ and $r = 1/2$. The bounds above are valid for all $\alpha \geqslant 1$ and all $r \in [0, 1]$, and the step-size γ is such that $\gamma R^2 \leqslant 1/4$, and thus we see a natural trade-off appearing for the step-size γ, between bias and variance.

When $r \geqslant \frac{\alpha-1}{2\alpha}$, then the optimal step-size minimizing the bound above is $\gamma \propto n^{\frac{-2\alpha\min\{r,1\}-1+\alpha}{2\alpha\min\{r,1\}+1}}$, and the obtained rate is optimal. Thus a single pass is optimal. However, when $r \leqslant \frac{\alpha-1}{2\alpha}$, the best step-size does not depend on n, and one can only achieve $O(n^{-2r})$.

Finally, in the same multiple pass set-up as ours, [LR17] has shown that for easy problems where $r \geqslant \frac{\alpha-1}{2\alpha}$ (and single-pass averaged SGD is already optimal) that multiple-pass non-averaged SGD is becoming optimal after a correct number of passes (while single-pass is not). Our proof principle of comparing to batch gradient is taken from [LR17], but we apply it to harder problems where $r \leqslant \frac{\alpha-1}{2\alpha}$. Moreover we consider the multi-pass averaged-SGD algorithm, instead of non-averaged SGD, and take explicitly into account the effect of Assumption (A8).

2.3. Averaged SGD with multiple passes

We consider the following algorithm, which is stochastic gradient descent with sampling with replacement with multiple passes over the data (we experiment in Section B.5 of the Appendix with cycling over the data, with or without reshuffling between each pass).

- **Initialization**: $\theta_0 = \bar{\theta}_0 = 0$, t = maximal number of iterations, $\gamma = 1/(4R^2)$ = step-size
- **Iteration**: for $u = 1$ to t, sample $i(u)$ uniformly from $\{1, \dots, n\}$ and make the step

$$\theta_u = \theta_{u-1} + \gamma\big(y_{i(u)} - \langle \theta_{t-1}, \Phi(x_{i(u)})\rangle_{\mathcal{H}}\big)\Phi(x_{i(u)}) \quad \text{and} \quad \bar{\theta}_u = (1 - \tfrac{1}{u})\bar{\theta}_{u-1} + \tfrac{1}{u}\theta_u.$$

In this paper, following [BM13, DB16], but as opposed to [DFB17], we consider unregularized recursions. This removes a unnecessary regularization parameter (at the expense of harder proofs).

Convergence rate and optimal number of passes. Our main result is the following (see full proof in Appendix):

Theorem 8

Let $n \in \mathbb{N}^*$ and $t \geqslant n$, under Assumptions (A6), (A7), (A8), (A9), (A10), (A11), with $\gamma = 1/(4R^2)$.

- *For $\mu\alpha < 2r\alpha + 1 < \alpha$, if we take $t = \Theta(n^{\alpha/(2r\alpha+1)})$, we obtain the following rate:*

$$\mathbb{E}F(\bar{\theta}_t) - F(\theta_*) = O(n^{-2r\alpha/(2r\alpha+1)}).$$

- *For $\mu\alpha \geqslant 2r\alpha + 1$, if we take $t = \Theta(n^{1/\mu} (\log n)^{\frac{1}{\mu}})$, we obtain the following rate:*

$$\mathbb{E}F(\bar{\theta}_t) - F(\theta_*) \leqslant O(n^{-2r/\mu}).$$

Sketch of proof. The main difficulty in extending proofs from the single pass case [BM13, DB16] is that as soon as an observation is processed twice, then statistical dependences are introduced and the proof does not go through. In a similar context, some authors have considered stability results [HRS16], but the large step-sizes that we consider do not allow this technique. Rather, we follow [RV15, LR17] and compare our multi-pass stochastic recursion θ_t to the batch gradient descent iterate η_t defined as $\eta_t = \eta_{t-1} + \frac{\gamma}{n} \sum_{i=1}^n \big(y_i - \langle \eta_{t-1}, \Phi(x_i)\rangle_{\mathcal{H}}\big)\Phi(x_i)$ with its averaged iterate $\bar{\eta}_t$. We thus need to study the predictive performance of $\bar{\eta}_t$ and the deviation $\bar{\theta}_t - \bar{\eta}_t$. It turns out that, given the data, the deviation $\theta_t - \eta_t$ satisfies an SGD recursion (with the respect to the randomness of the sampling with replacement). For a more detailed summary of the proof technique see Section B.2.

The novelty compared to [RV15, LR17] is (a) to use refined results on averaged SGD for least-squares, in particular convergence in various norms for the deviation $\bar{\theta}_t - \bar{\eta}_t$ (see Section B.1), that can use our new Assumption (A8). Moreover, (b) we need to extend the convergence results for the batch gradient descent recursion from [LRRC18], also to take into account the new assumption (see Section B.4). These two results are interesting on their own.

Improved rates with multiple passes. We can draw the following conclusions:

- If $2\alpha r + 1 \geqslant \alpha$, that is, easy problems, it has been shown by [DB16] that a single pass with a smaller step-size than the one we propose here is optimal, and our result does not apply.
- If $\mu\alpha < 2r\alpha + 1 < \alpha$, then our proposed number of iterations is $t = \Theta(n^{\alpha/(2\alpha r+1)})$, which is now greater than n; the convergence rate is then $O(n^{\frac{-2r\alpha}{2r\alpha+1}})$, and, as we will see in Section 2.4.2, the predictive performance is then optimal when $\mu \leqslant 2r$.
- If $\mu\alpha \geqslant 2r\alpha + 1$, then with a number of iterations is $t = \Theta(n^{1/\mu})$, which is greater than n (thus several passes), with a convergence rate equal to $O(n^{-2r/\mu})$, which improves upon the best known rates of $O(n^{-2r})$. As we will see in Section 2.4.2, this is not optimal.

Note that these rates are theoretically only bounds on the optimal number of passes over the data, and one should be cautious when drawing conclusions; however our simulations on synthetic data, see Figure 14 in Section 2.5, confirm that our proposed scalings for the number of passes is observed in practice.

2.4. Application to kernel methods

In the section above, we have assumed that $\mathcal{H}$ was finite-dimensional, so that the optimal predictor $\theta_* \in \mathcal{H}$ was always defined. Note however, that our bounds that depends on α, r and μ are *independent of the dimension*, and hence, intuitively, following [DFB17], should apply immediately to infinite-dimensional spaces.

We now first show in Section 2.4.1 how this intuition can be formalized and how using kernel methods provides a particularly interesting example. Moreover, this interpretation allows to characterize the statistical optimality of our results in Section 2.4.2.

2.4.1. Extension to Hilbert spaces, kernel methods and non-parametric estimation

Our main result in Theorem 8 extends directly to the case where $\mathcal{H}$ is an infinite-dimensional Hilbert space. In particular, given a feature map $\Phi : \mathcal{X} \to \mathcal{H}$, any vector $\theta \in \mathcal{H}$ is naturally associated to a function defined as $f_\theta(x) = \langle \theta, \Phi(x) \rangle_{\mathcal{H}}$. Algorithms can then be run with infinite-dimensional objects if the *kernel* $K(x', x) = \langle \Phi(x'), \Phi(x) \rangle_{\mathcal{H}}$ can be computed efficiently. This identification of elements θ of $\mathcal{H}$ with functions f_θ endows the various quantities we have introduced in the previous sections, with natural interpretations in terms of functions. The stochastic gradient descent described in Section 2.3 adapts instantly to this new framework as the iterates $(\theta_u)_{u \leqslant t}$ are linear combinations of feature vectors $\Phi(x_i)$, $i = 1, \dots, n$, and the algorithms can classically be "kernelized" [YP08, DB16], with an overall running time complexity of $O(nt)$.

First note that Assumption (A8) is equivalent to, for all $x \in \mathcal{X}$ and $\theta \in \mathcal{H}$, $|f_\theta(x)|^2 \leqslant \kappa_\mu^2 R^{2\mu} \langle f_\theta, \Sigma^{1-\mu} f_\theta \rangle_{\mathcal{H}}$, that is, $\|g\|_{L_\infty}^2 \leqslant \kappa_\mu^2 R^{2\mu} \|\Sigma^{1/2-\mu/2} g\|_{\mathcal{H}}^2$ for any $g \in \mathcal{H}$ and also implies[8] $\|g\|_{L_\infty} \leqslant \kappa_\mu R^\mu \|g\|_{\mathcal{H}}^\mu \|g\|_{L_2}^{1-\mu}$, which are common assumptions in the context of kernel methods [SHS09], essentially controlling in a more refined way the regularity of the whole space of functions associated to $\mathcal{H}$, with respect to the L^∞-norm, compared to the too crude inequality $\|g\|_{L^\infty} = \sup_x |\langle \Phi(x), g \rangle_{\mathcal{H}}| \leqslant \sup_x \|\Phi(x)\|_{\mathcal{H}} \|g\|_{\mathcal{H}} \leqslant R \|g\|_{\mathcal{H}}$.

The natural relation with functions allows to analyze effects that are crucial in the context of learning, but difficult to grasp in the finite-dimensional setting. Consider the following prototypical example of a hard learning problem,

Example 2 (*Prototypical hard problem on simple Sobolev space*)

Let $\mathcal{X} = [0, 1]$, with x sampled uniformly on X and

$$y = \operatorname{sign}(x - 1/2) + \epsilon, \quad \Phi(x) = \{|k|^{-1} e^{2ik\pi x}\}_{k \in \mathbb{Z}^*}.$$

This corresponds to the kernel $K(x, y) = \sum_{k \in \mathbb{Z}^*} |k|^{-2} e^{2ik\pi(x-y)}$, which is well defined (and lead to the simplest Sobolev space). Note that for any $\theta \in \mathcal{H}$, which is here identified as the space of square-summable sequences $\ell^2(\mathbb{Z})$, we have $f_\theta(x) = \langle \theta, \Phi(x) \rangle_{\ell^2(\mathbb{Z})} = \sum_{k \in \mathbb{Z}^*} \frac{\theta_k}{|k|} e^{2ik\pi x}$. This means that for any estimator $\hat{\theta}$ given by the algorithm, $f_{\hat{\theta}}$ is at least once continuously differentiable, while the target function $\operatorname{sign}(\cdot - 1/2)$ is not even continuous. Hence, we are in a situation where θ_*, the minimizer of the excess risk, does not belong to $\mathcal{H}$. Indeed let represent $\operatorname{sign}(\cdot - 1/2)$ in $\mathcal{H}$, for almost all $x \in [0, 1]$, by its Fourier series $\operatorname{sign}(x - 1/2) = \sum_{k \in \mathbb{Z}_*} \alpha_k e^{2ik\pi x}$, with $|\alpha_k| \sim 1/k$, an informal reasoning would lead to $(\theta_*)_k = \alpha_k |k| \sim 1$, which is not square-summable and thus $\theta_* \notin \mathcal{H}$. For more details, see [AF03, Wah90].

This setting generalizes important properties that are valid for Sobolev spaces, as shown in the following example, where α, r, μ are characterized in terms of the smoothness of the functions in $\mathcal{H}$, the smoothness of f^* and the dimensionality of the input space $\mathcal{X}$.

[8] Indeed, for any $g \in \mathcal{H}$, $\|\Sigma^{1/2-\mu/2} g\|_{\mathcal{H}} = \|\Sigma^{-\mu/2} g\|_{L_2} \leqslant \|\Sigma^{-1/2} g\|_{L_2}^\mu \|g\|_{L_2}^{1-\mu} = \|g\|_{\mathcal{H}}^\mu \|g\|_{L_2}^{1-\mu}$, where we used that for any $g \in \mathcal{H}$, any bounded operator A, $s \in [0, 1]$: $\|A^s g\|_{L_2} \leqslant \|Ag\|_{L_2}^s \|g\|_{L_2}^{1-s}$ (see [RR17]).

Example 3 (*Sobolev Spaces [Wen04, SHS09, Bac17, FS17]*)

Let $\mathcal{X} \subseteq \mathbb{R}^d$, $d \in \mathbb{N}$, with $\rho_{\mathcal{X}}$ supported on $\mathcal{X}$, absolutely continous with the uniform distribution and such that $\rho_{\mathcal{X}}(x) \geqslant a > 0$ almost everywhere, for a given a. Assume that $f^*(x) = \mathbb{E}[y|x]$ is s-times differentiable, with $s > 0$. Choose a kernel, inducing Sobolev spaces of smoothness m with $m > d/2$, as the Matérn kernel

$$K(x', x) = \|x' - x\|^{m-d/2} \mathcal{K}_{d/2-m}(\|x' - x\|),$$

where $\mathcal{K}_{d/2-m}$ is the modified Bessel function of the second kind. Then the assumptions are satisfied for any $\epsilon > 0$, with $\alpha = \frac{2m}{d}$, $\quad \mu = \frac{d}{2m} + \epsilon, \quad r = \frac{s}{2m}$.

In the following subsection we compare the rates obtained in Thm. 8, with known lower bounds under the same assumptions.

2.4.2. Minimax lower bounds

In this section we recall known lower bounds on the rates for classes of learning problems satisfying the conditions in Sect. 2.2.1. Interestingly, the comparison below shows that our results in Theorem 8 are optimal in the setting $2r \geqslant \mu$. While the optimality of SGD was known for the regime $\{2r\alpha + 1 \geqslant \alpha \cap 2r \geqslant \mu\}$, here we extend the optimality to the new regime $\alpha \geqslant 2r\alpha + 1 \geqslant \mu\alpha$, covering essentially all the region $2r \geqslant \mu$, as it is possible to observe in Figure 13, where for clarity we plotted the best possible value for μ that is $\mu = 1/\alpha$ [FS17] (which is true for Sobolev spaces).

When $r \in (0, 1]$ is fixed, but there are no assumptions on α or μ, then the optimal minimax rate of convergence is $O(n^{-2r/(2r+1)})$, attained by regularized empirical risk minimization [CDV07] and other spectral filters on the empirical covariance operator [BM17].

When $r \in (0, 1]$ and $\alpha \geqslant 1$ are fixed (but there are no constraints on μ), the optimal minimax rate of convergence $O(n^{\frac{-2r\alpha}{2r\alpha+1}})$ is attained when $r \geqslant \frac{\alpha-1}{2\alpha}$, with empirical risk minimization [LRRC18] or stochastic gradient descent [DB16].

When $r \geqslant \frac{\alpha-1}{2\alpha}$, the rate of convergence $O(n^{\frac{-2r\alpha}{2r\alpha+1}})$ is known to be a lower bound on the optimal minimax rate, but the best upper-bound so far is $O(n^{-2r})$ and is achieved by empirical risk minimization [LRRC18] or stochastic gradient descent [DB16], and the optimal rate is not known.

When $r \in (0, 1]$, $\alpha \geqslant 1$ and $\mu \in [1/\alpha, 1]$ are fixed, then the rate of convergence $O(n^{\frac{-\max\{\mu,2r\}\alpha}{2\max\{\mu,2r\}\alpha+1}})$ is known to be a lower bound on the optimal minimax rate [FS17]. This is attained by regularized empirical risk minimization when $2r \geqslant \mu$ [FS17], and *now by SGD with multiple passes*, and it is thus the optimal rate in this situation. When $2r < \mu$, the only known upper bound is $O(n^{-2\alpha r/(\mu\alpha+1)})$, and the optimal rate is not known.

2.5. Experiments

In our experiments, the main goal is to show that with more that one pass over the data, we can improve the accuracy of SGD when the problem is hard. We also want to highlight our dependence of the optimal number of passes (that is t/n) with respect to the number of observations n.

Synthetic experiments. Our main experiments are performed on artificial data following the setting in [RR17]. For this purpose, we take kernels K corresponding to splines of order q (see [Wah90]) that fulfill Assumptions (A6) (A7) (A8) (A9) (A10) (A11). Indeed, let us consider the following function

$$\Lambda_q(x, z) = \sum_{k \in \mathbb{Z}} \frac{e^{2i\pi k(x-z)}}{|k|^q},$$

defined almost everywhere on $[0,1]$, with $q \in \mathbb{R}$, and for which we have the interesting relationship: $\langle \Lambda_q(x,\cdot), \Lambda_{q'}(z,\cdot)\rangle_{L_2(d\rho_\mathcal{X})} = \Lambda_{q+q'}(x,z)$ for any $q, q' \in \mathbb{R}$. Our setting is the following:

- **Input distribution:** $\mathcal{X} = [0,1]$ and $\rho_\mathcal{X}$ is the uniform distribution.
- **Kernel:** $\forall (x,z) \in [0,1],\ K(x,z) = \Lambda_\alpha(x,z)$.
- **Target function:** $\forall x \in [0,1],\ \theta_* = \Lambda_{r\alpha+\frac{1}{2}}(x,0)$.
- **Output distribution :** $\rho(y|x)$ is a Gaussian with variance σ^2 and mean θ_*.

For this setting we can show that the learning problem satisfies Assumptions (A6) (A7) (A8) (A9) (A10) (A11) with $r,\ \alpha,$ and$\mu = 1/\alpha$. We take different values of these parameters to encounter all the different regimes of the problems shown in Figure 13.

For each n from 100 to 1000, we found the optimal number of steps $t_*(n)$ that minimizes the test error $F(\bar{\theta}_t) - F(\theta_*)$. Note that because of overfitting the test error increases for $t > t_*(n)$. In Figure 14, we show $t_*(n)$ with respect to n in log scale. As expected, for the easy problems (where $r \geqslant \frac{\alpha-1}{2\alpha}$, see top left and right plots), the slope of the plot is 1 as one pass over the data is enough: $t_*(n) = \Theta(n)$. But we see that for hard problems (where $r \leqslant \frac{\alpha-1}{2\alpha}$, see bottom left and right plots), we need more than one pass to achieve optimality as the optimal number of iterations is very close to $t_*(n) = \Theta\left(n^{\frac{\alpha}{2r\alpha+1}}\right)$. That matches the theoretical predictions of Theorem 8. We also notice in the plots that, the bigger $\frac{\alpha}{2r\alpha+1}$ the harder the problem is and the bigger the number of epochs we have to take. Note, that to reduce the noise on the estimation of $t_*(n)$, plots show an average over 100 replications.

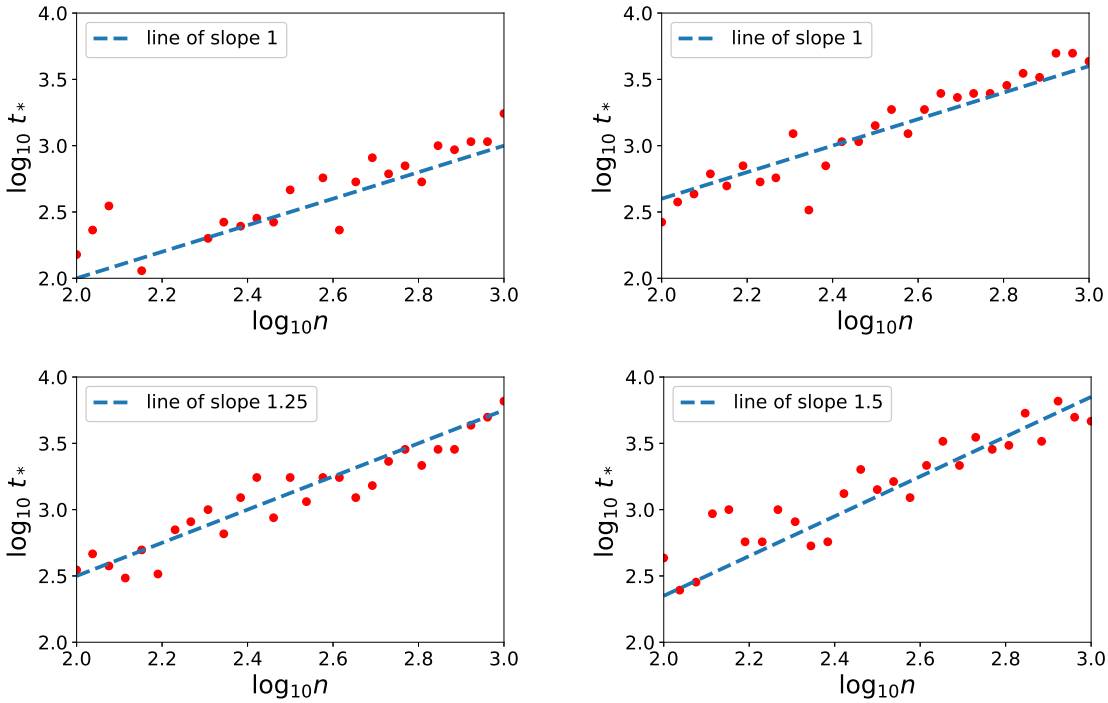

Figure 14: The four plots represent each a different configuration on the (α, r) plan represented in Figure 13, for $r = 1/(2\alpha)$. **Top left** ($\alpha = 1.5$) and **right** ($\alpha = 2$) are two easy problems (Top right is the limiting case where $r = \frac{\alpha-1}{2\alpha}$) for which one pass over the data is optimal. **Bottom left** ($\alpha = 2.5$) and **right** ($\alpha = 3$) are two hard problems for which an increasing number of passes is required. The blue dotted line are the slopes predicted by the theoretical result in Theorem 8.

To conclude, the experiments presented in the section correspond exactly to the theoretical setting of the article (sampling with replacement), however we present in Figures 16 and 17 of Section B.5 of the

Appendix results on the same datasets for two different ways of sampling the data: (a)*without replacement*: for which we select randomly the data points but never use twice the same point in one epoch, (b) *cycles*: for which we pick successively the data points in the same order. The obtained scalings relating number of iterations or passes to number of observations are the same.

Linear model. To illustrate our result with some real data, we show how the optimal number of passes over the data increases with the number of samples. In Figure 15, we simply performed linear least-squares regression on the MNIST dataset and plotted the optimal number of passes over the data that leads to the smallest error on the test set. Evaluating α and r from Assumptions (A9) and (A10), we found $\alpha = 1.7$ and $r = 0.18$. As $r = 0.18 \leqslant \frac{\alpha-1}{2\alpha} \sim 0.2$, Theorem 8 indicates that this corresponds to a situation where only one pass on the data is not enough, confirming the behavior of Figure 15. This suggests that learning MNIST with linear regression is a *hard problem*.

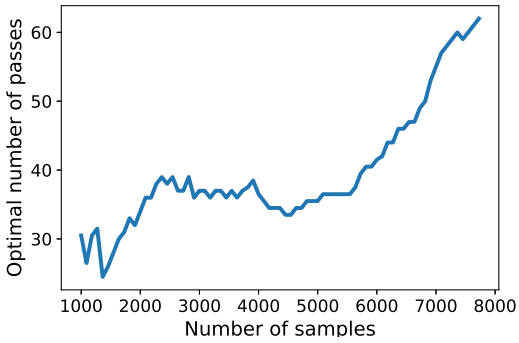

Figure 15: For the MNIST data set, we show the optimal number of passes over the data with respect to the number of samples in the case of the linear regression.

2.6. Conclusion

In this paper, we have shown that for least-squares regression, in hard problems where single-pass SGD is not statistically optimal ($r < \frac{\alpha-1}{2\alpha}$), then multiple passes lead to statistical optimality with a number of passes that somewhat surprisingly needs to grow with sample size, with a convergence rate which is superior to previous analyses of stochastic gradient. Using a non-parametric estimation, we show that under certain conditions ($2r \geqslant \mu$), we attain statistical optimality.

Our work could be extended in several ways: (a) our experiments suggest that cycling over the data and cycling with random reshuffling perform similarly to sampling with replacement, it would be interesting to combine our theoretical analysis with work aiming at analyzing other sampling schemes [Sha16, GOP15]. (b) Mini-batches could be also considered with a potentially interesting effects compared to the streaming setting. Also, (c) our analysis focuses on least-squares regression, an extension to all smooth loss functions would widen its applicability. Moreover, (d) providing optimal efficient algorithms for the situation $2r < \mu$ is a clear open problem (for which the optimal rate is not known, even for non-efficient algorithms). Additionally, (e) in the context of classification, we could combine our analysis with [PVRB18] to study the potential discrepancies between training and testing losses and errors when considering high-dimensional models [ZBH+16]. More generally, (f) we could explore the effect of our analysis for methods based on the least-squares estimator in the context of structured prediction [CRR16, OBLJ17, CBR18] and (non-linear) multitask learning [CRRP17]. Finally, (g) to reduce the computational complexity of the algorithm, while retaining the (optimal) statistical guarantees, we could combine multi-pass stochastic gradient descent,

with approximation techniques like *random features* [RR08], extending the analysis of [CRR18] to the more general setting considered in this paper.

Acknowledgements

We acknowledge support from the European Research Council (grant SEQUOIA 724063). We also thank Raphaël Berthier and Yann Labbé for their enlightening advices on this project.

⋆

⋆ ⋆

B. Appendix of Statistical Optimality of SGD on Hard Learning Problems through Multiple Passes

The appendix in constructed as follows:

- We first present in Section B.1 a new result for stochastic gradient recursions which generalizes the work of [BM13] and [DB16] to more general norms. This result could be used in other contexts.
- The proof technique for Theorem 8 is presented in Section B.2.
- In Section B.3 we give a proof of the various lemmas needed in the first part of the proof of Theorem 8 (deviation between SGD and batch gradient descent).
- In Section B.4 we provide new results for the analysis of batch gradient descent, which are adapted to our new (A8), and instrumental in proving Theorem 8 in Section B.2.
- Finally, in Section B.5 we present experiments for different sampling techniques.

B.1. A general result for the SGD variance term

Independently of the problem studied in this paper, we consider i.i.d. observations $(z_t, \xi_t) \in \mathcal{H} \times \mathcal{H}$ a Hilbert space, and the recursion started from $\mu_0 = 0$.

$$\mu_t = (I - \gamma z_t \otimes z_t)\, \mu_{t-1} + \gamma \xi_t \tag{76}$$

(this will applied with $z_t = \Phi(x_{i(t)})$). This corresponds to the variance term of SGD. We denote by $\bar{\mu}_t$ the averaged iterate $\bar{\mu}_t = \frac{1}{t} \sum_{i=1}^{t} \mu_i$.

The goal of the proposition below is to provide a bound on $\mathbb{E}\left[\left\|H^{u/2} \bar{\mu}_t\right\|^2\right]$ for $u \in [0, \frac{1}{\alpha} + 1]$, where $H = \mathbb{E}\left[z_t \otimes z_t\right]$ is such that $\operatorname{tr} H^{1/\alpha}$ is finite. Existing results only cover the case $u = 1$.

Proposition 9 (*A general result for the SGD variance term*)

Let us consider the recursion in (76) started at $\mu_0 = 0$. Denote $\mathbb{E}\left[z_t \otimes z_t\right] = H$, assume that $\operatorname{tr} H^{1/\alpha}$ is finite, $\mathbb{E}\left[\xi_t\right] = 0$, $\mathbb{E}\left[(z_t \otimes z_t)^2\right] \preccurlyeq R^2 H$, $\mathbb{E}\left[\xi_t \otimes \xi_t\right] \preccurlyeq \sigma^2 H$ and $\gamma R^2 \leqslant 1/4$, then for $u \in [0, \frac{1}{\alpha} + 1]$:

$$\mathbb{E}\left[\left\|H^{u/2} \bar{\mu}_t\right\|^2\right] \leqslant 4\sigma^2 \gamma^{1-u}\, \frac{\gamma^{1/\alpha} \operatorname{tr} H^{1/\alpha}}{t^{u-1/\alpha}}. \tag{77}$$

B.1.1. Proof principle

We follow closely the proof technique of [BM13], and prove Proposition 9 by showing it first for a "semi-stochastic" recursion, where $z_t \otimes z_t$ is replaced by its expectation (see Lemma 11). We will then compare our general recursion to the semi-stochastic one.

B.1.2. Semi-stochastic recursion

Lemma 11 (*Semi-stochastic SGD*)

Let us consider the following recursion $\mu_t = (I - \gamma H)\,\mu_{t-1} + \gamma\xi_t$ *started at* $\mu_0 = 0$. *Assume that* $\mathrm{tr}H^{1/\alpha}$ *is finite,* $\mathbb{E}[\xi_t] = 0$, $\mathbb{E}[\xi_t \otimes \xi_t] \preccurlyeq \sigma^2 H$ *and* $\gamma H \preccurlyeq I$, *then for* $u \in [0, \frac{1}{\alpha} + 1]$:

$$\mathbb{E}\left[\left\|H^{u/2}\bar{\mu}_t\right\|^2\right] \leqslant \sigma^2\gamma^{1-u}\,\gamma^{1/\alpha}\mathrm{tr}H^{1/\alpha}t^{1/\alpha-u}. \tag{78}$$

Proof : For $t \geqslant 1$ and $u \in [0, \frac{1}{\alpha} + 1]$, using an explicit formula for μ_t and $\bar{\mu}_t$ (see [BM13] for details), we get:

$$\begin{aligned}
\mu_t &= (I - \gamma H)\,\mu_{t-1} + \gamma\xi_t = (I - \gamma H)^t\,\mu_0 + \gamma\sum_{k=1}^{t}(I - \gamma H)^{t-k}\,\xi_k \\
\bar{\mu}_t &= \frac{1}{t}\sum_{u=1}^{t}\mu_u = \frac{\gamma}{t}\sum_{u=1}^{t}\sum_{k=1}^{u}(I - \gamma H)^{u-k}\,\xi_k = \frac{1}{t}\sum_{k=1}^{t}H^{-1}\left(I - (I - \gamma H)^{t-k+1}\right)\xi_k \\
\mathbb{E}\left[\left\|H^{u/2}\bar{\mu}_t\right\|^2\right] &= \frac{1}{t^2}\mathbb{E}\sum_{k=1}^{t}\mathrm{tr}\left[\left(I - (I - \gamma H)^{t-k+1}\right)^2 H^{u-2}\xi_k \otimes \xi_k\right] \\
&\leqslant \frac{\sigma^2}{t^2}\sum_{k=1}^{t}\mathrm{tr}\left[\left(I - (I - \gamma H)^{k}\right)^2 H^{u-1}\right] \text{ using } \mathbb{E}[\xi_t \otimes \xi_t] \preccurlyeq \sigma^2 H.
\end{aligned}$$

Now, let $(\lambda_i)_{i\in\mathbb{N}^*}$ be the non-increasing sequence of eigenvalues of the operator H. We obtain:

$$\mathbb{E}\left[\left\|H^{u/2}\bar{\mu}_t\right\|^2\right] \leqslant \frac{\sigma^2}{t^2}\sum_{k=1}^{t}\sum_{i=1}^{\infty}\left(I - (I - \gamma\lambda_i)^k\right)^2\lambda_i^{u-1}.$$

We can now use a simple result[9] that for any $\rho \in [0,1]$, $k \geqslant 1$ and $u \in [0, \frac{1}{\alpha} + 1]$, we have : $(1 - (1-\rho)^k)^2 \leqslant (k\rho)^{1-u+1/\alpha}$, applied to $\rho = \gamma\lambda_i$. We get, by comparing sums to integrals:

$$\begin{aligned}
\mathbb{E}\left[\left\|H^{u/2}\bar{\mu}_t\right\|^2\right] &\leqslant \frac{\sigma^2}{t^2}\sum_{k=1}^{t}\sum_{i=1}^{\infty}\left(I - (I - \gamma\lambda_i)^k\right)^2\lambda_i^{u-1} \\
&\leqslant \frac{\sigma^2}{t^2}\sum_{k=1}^{t}\sum_{i=1}^{\infty}(k\gamma\lambda_i)^{1-u+1/\alpha}\lambda_i^{u-1} \\
&\leqslant \frac{\sigma^2}{t^2}\gamma^{1-u+1/\alpha}\mathrm{tr}H^{1/\alpha}\sum_{k=1}^{t}k^{1-u+1/\alpha} \\
&\leqslant \frac{\sigma^2}{t^2}\gamma^{1-u+1/\alpha}\mathrm{tr}H^{1/\alpha}\int_1^t y^{1-u+1/\alpha}dy \\
&\leqslant \frac{\sigma^2}{t^2}\gamma^{1-u}\,\gamma^{1/\alpha}\mathrm{tr}H^{1/\alpha}\frac{t^{2-u+1/\alpha}}{2-u+1/\alpha} \\
&\leqslant \sigma^2\gamma^{1-u}\,\gamma^{1/\alpha}\mathrm{tr}H^{1/\alpha}t^{1/\alpha-u},
\end{aligned}$$

which shows the desired result. ■

[9] Indeed, adapting a similar result from [BM13], on the one hand, $1 - (1-\rho)^k \leqslant 1$ implying that $(1 - (1-\rho)^k)^{1-1/\alpha+u} \leqslant 1$. On the other hand, $1 - (1 - \gamma x)^k \leqslant \gamma k x$ implying that $(1 - (1-\rho)^k)^{1+1/\alpha-u} \leqslant (k\rho)^{1+1/\alpha-u}$. Thus by multiplying the two we get $(1 - (1-\rho)^k)^2 \leqslant (k\rho)^{1-u+1/\alpha}$.

B.1.3. Relating the semi-stochastic recursion to the main recursion

Then, to relate the semi-stochastic recursion with the true one, we use an expansion in the powers of γ using recursively the perturbation idea from [AMP00].

For $r \geqslant 0$, we define the sequence $(\mu_t^r)_{t\in\mathbb{N}}$, for $t \geqslant 1$,

$$\mu_t^r = (I - \gamma H)\mu_{t-1}^r + \gamma \Xi_t^r, \text{ with } \Xi_t^r = \begin{cases} (H - z_t \otimes z_t)\mu_{t-1}^{r-1} \text{ if } r \geqslant 1 \\ \Xi_t^0 = \xi_t \end{cases}. \tag{79}$$

We will show that $\mu_t \simeq \sum_{i=0}^{\infty} \mu_t^i$. To do so, notice that for $r \geqslant 0$, $\mu_t - \sum_{i=0}^{r} \mu_t^i$ follows the recursion:

$$\mu_t - \sum_{i=0}^{r} \mu_t^i = (I - z_t \otimes z_t)\left(\mu_{t-1} - \sum_{i=0}^{r} \mu_{t-1}^i\right) + \gamma \Xi_t^{r+1}, \tag{80}$$

so that by bounding the covariance operator we can apply a classical SGD result. This is the purpose of the following lemma.

Lemma 12 (*Bound on covariance operator*)

For any $r \geqslant 0$, we have the following inequalities:

$$\mathbb{E}\left[\Xi_t^r \otimes \Xi_t^r\right] \preccurlyeq \gamma^r R^{2r}\sigma^2 H \quad \textit{and} \quad \mathbb{E}\left[\mu_t^r \otimes \mu_t^r\right] \preccurlyeq \gamma^{r+1} R^{2r}\sigma^2 I. \tag{81}$$

Proof : We propose a proof by induction on r. For $r = 0$, and $t \geqslant 0$, $\mathbb{E}\left[\Xi_t^0 \otimes \Xi_t^0\right] = \mathbb{E}\left[\xi_t \otimes \xi_t\right] \preccurlyeq \sigma^2 H$ by assumption. Moreover,

$$\mathbb{E}\left[\mu_t^0 \otimes \mu_t^0\right] = \gamma^2 \sum_{k=1}^{t-1} (I - \gamma H)^{t-k} \mathbb{E}\left[\Xi_t^0 \otimes \Xi_t^0\right] (I - \gamma H)^{t-k} \preccurlyeq \gamma^2\sigma^2 \sum_{k=1}^{t-1} (I - \gamma H)^{2(t-k)} H \preccurlyeq \gamma\sigma^2 I.$$

Then, for $r \geqslant 1$,

$$\begin{aligned} \mathbb{E}\left[\Xi_t^{r+1} \otimes \Xi_t^{r+1}\right] &\preccurlyeq \mathbb{E}[(H - z_t \otimes z_t)\mu_{t-1}^r \otimes \mu_{t-1}^r (H - z_t \otimes z_t)] \\ &= \mathbb{E}[(H - z_t \otimes z_t)\mathbb{E}[\mu_{t-1}^r \otimes \mu_{t-1}^r](H - z_t \otimes z_t)] \\ &\preccurlyeq \gamma^{r+1} R^{2r}\sigma^2 \mathbb{E}[(H - z_t \otimes z_t)^2] \\ &\preccurlyeq \gamma^{r+1} R^{2r+2}\sigma^2 H. \end{aligned}$$

And,

$$\begin{aligned} \mathbb{E}\left[\mu_t^{r+1} \otimes \mu_t^{r+1}\right] &= \gamma^2 \sum_{k=1}^{t-1} (I - \gamma H)^{t-k} \mathbb{E}\left[\Xi_t^{r+1} \otimes \Xi_t^{r+1}\right] (I - \gamma H)^{t-k} \\ &\preccurlyeq \gamma^{r+3} R^{2r+2}\sigma^2 \sum_{k=1}^{t-1} (I - \gamma H)^{2(t-k)} H \preccurlyeq \gamma^{r+2} R^{2r+2}\sigma^2 I, \end{aligned}$$

which thus shows the lemma by induction. ■

To bound $\mu_t - \sum_{i=0}^{r} \mu_t^i$, we prove a very loose result for the average iterate, that will be sufficient for our purpose.

Lemma 13 (*Bounding SGD recursion*)

Let us consider the following recursion $\mu_t = (I - \gamma z_t \otimes z_t)\,\mu_{t-1} + \gamma\xi_t$ starting at $\mu_0 = 0$. Assume that

$\mathbb{E}[z_t \otimes z_t] = H$, $\mathbb{E}[\xi_t] = 0$, $\|x_t\|^2 \leqslant R^2$, $\mathbb{E}[\xi_t \otimes \xi_t] \preccurlyeq \sigma^2 H$ *and* $\gamma R^2 < I$, *then for* $u \in [0, \frac{1}{\alpha} + 1]$:

$$\mathbb{E}\left[\left\|H^{u/2}\bar{\mu}_t\right\|^2\right] \leqslant \sigma^2\gamma^2 R^u \mathrm{tr} H \; t. \tag{82}$$

Proof : Let us define the operators for $j \leqslant i : M_j^i = (I - \gamma z_{i(i)} \otimes z_{i(i)}) \cdots (I - \gamma z_{i(j)} \otimes z_{i(j)})$ and $M_{i+1}^i = I$. Since $\mu_0 = 0$, note that we have we have, $\mu_i = \gamma \sum_{k=1}^i M_{k+1}^i \xi_k$. Hence, for $i \geqslant 1$,

$$\begin{aligned}
\mathbb{E}\left\|H^{u/2}\mu_i\right\|^2 &= \gamma^2 \mathbb{E} \sum_{k,j} \langle M_{j+1}^i \xi_j, H^u M_{k+1}^i \xi_k \rangle \\
&= \gamma^2 \mathbb{E} \sum_{k=1}^i \langle M_{k+1}^i \xi_k, H^u M_{k+1}^i \xi_k \rangle \\
&= \gamma^2 \mathrm{tr}\left(\mathbb{E}\left[\sum_{k=1}^i {M_{k+1}^i}^* H^u M_{k+1}^i \xi_k \otimes \xi_k\right]\right) \leqslant \sigma^2\gamma^2 \mathbb{E}\left[\sum_{k=1}^i \mathrm{tr}\left({M_{k+1}^i}^* H^u M_{k+1}^i H\right)\right] \\
&\leqslant \sigma^2\gamma^2 R^u i \; \mathrm{tr} H,
\end{aligned}$$

because $\mathrm{tr}\left({M_{k+1}^i}^* H^u M_{k+1}^i H\right) \leqslant R^u \mathrm{tr} H$. Then,

$$\begin{aligned}
\mathbb{E}\left\|H^{u/2}\bar{\mu}_t\right\|^2 &= \frac{1}{t^2} \sum_{i,j} \langle H^{u/2}\mu_i, H^{u/2}\mu_j \rangle \\
&\leqslant \frac{1}{t^2} \mathbb{E}\left(\sum_{i=1}^t \left\|H^{u/2}\mu_i\right\|\right)^2 \leqslant \frac{1}{t} \sum_{i=1}^t \mathbb{E}\left\|H^{u/2}\mu_i\right\|^2 \leqslant \sigma^2\gamma^2 R^u \mathrm{tr} H \; t,
\end{aligned}$$

which finishes the proof of Lemma 13. ■

B.1.4. Final steps of the proof

We have now all the material to conclude. Indeed by the triangular inequality:

$$\left(\mathbb{E}\left\|H^{u/2}\bar{\mu}_t\right\|^2\right)^{1/2} \leqslant \sum_{i=1}^r \left(\underbrace{\mathbb{E}\left\|H^{u/2}\bar{\mu}_t^i\right\|^2}_{\text{Lemma 11}}\right)^{1/2} + \left(\underbrace{\mathbb{E}\left\|H^{u/2}\left(\bar{\mu}_t - \sum_{i=1}^r \bar{\mu}_t^i\right)\right\|^2}_{\text{Lemma 13}}\right)^{1/2}.$$

With Lemma 12, we have all the bounds on the covariance of the noise, so that:

$$\begin{aligned}
\left(\mathbb{E}\left\|H^{u/2}\bar{\mu}_t\right\|^2\right)^{1/2} &\leqslant \sum_{i=1}^r \left(\gamma^i R^{2i} \sigma^2 \gamma^{1-u} \; \gamma^{1/\alpha} \mathrm{tr} H^{1/\alpha} t^{1/\alpha - u}\right)^{1/2} + \left(\gamma^{r+2} R^{2r+u} \mathrm{tr} H \; t\right)^{1/2} \\
&\leqslant \left(\sigma^2 \gamma^{1-u} \; \gamma^{1/\alpha} \mathrm{tr} H^{1/\alpha} t^{1/\alpha - u}\right)^{1/2} \sum_{i=1}^r \left(\gamma R^2\right)^{i/2} + \left(\gamma^{r+2} R^{2r+u} \mathrm{tr} H \; t\right)^{1/2}.
\end{aligned}$$

Now we make r go to infinity and we obtain:

$$\left(\mathbb{E}\left\|H^{u/2}\bar{\mu}_t\right\|^2\right)^{1/2} \leqslant \left(\sigma^2 \gamma^{1-u} \; \gamma^{1/\alpha} \mathrm{tr} H^{1/\alpha} t^{1/\alpha - u}\right)^{1/2} \frac{1}{1 - \sqrt{\gamma R^2}} + \underbrace{\left(\gamma^{r+2} R^{2r+u} \mathrm{tr} H \; t\right)^{1/2}}_{\underset{r \to \infty}{\longrightarrow} 0}$$

Hence with $\gamma R^2 \leqslant 1/4$,

$$\mathbb{E}\left\|H^{u/2}\bar{\mu}_t\right\|^2 \leqslant 4\sigma^2\gamma^{1-u}\,\gamma^{1/\alpha}\mathrm{tr}H^{1/\alpha}t^{1/\alpha-u},$$

which finishes to prove Proposition 9.

B.2. Proof sketch for Theorem 8

We consider the batch gradient descent recursion, started from $\eta_0 = 0$, with the same step-size:

$$\eta_t = \eta_{t-1} + \frac{\gamma}{n}\sum_{i=1}^n \big(y_i - \langle \eta_{t-1}, \Phi(x_i)\rangle_{\mathcal{H}}\big)\Phi(x_i),$$

as well as its averaged version $\bar{\eta}_t = \frac{1}{t}\sum_{i=0}^t \eta_i$. We obtain a recursion for $\theta_t - \eta_t$, with the initialization $\theta_0 - \eta_0 = 0$, as follows:

$$\theta_t - \eta_t = \big[I - \Phi(x_{i(u)}) \otimes_{\mathcal{H}} \Phi(x_{i(u)})\big](\theta_{t-1} - \eta_{t-1}) + \gamma\xi_t^1 + \gamma\xi_t^2,$$

with $\xi_t^1 = y_{i(u)}\Phi(x_{i(u)}) - \frac{1}{n}\sum_{i=1}^n y_i\Phi(x_i)$ and $\xi_t^2 = \big[\Phi(x_{i(u)}) \otimes_{\mathcal{H}} \Phi(x_{i(u)}) - \frac{1}{n}\sum_{i=1}^n \Phi(x_i) \otimes_{\mathcal{H}} \Phi(x_i)\big]\eta_{t-1}$. We decompose the performance $F(\theta_t)$ in two parts, one analyzing the performance of batch gradient descent, one analyzing the deviation $\theta_t - \eta_t$, using

$$\mathbb{E}F(\bar{\theta}_t) - F(\theta_*) \leqslant 2\mathbb{E}\big[\|\Sigma^{1/2}(\theta_t - \eta_t)\|_{\mathcal{H}}^2\big] + 2\big[\mathbb{E}F(\bar{\eta}_t) - F(\theta_*)\big].$$

We denote by $\hat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^n \Phi(x_i) \otimes \Phi(x_i)$ the empirical second-order moment.

Deviation $\theta_t - \eta_t$. Denoting by $\mathcal{G}$ the σ-field generated by the data and by $\mathcal{F}_t$ the σ-field generated by $i(1), \dots, i(t)$, then, we have $\mathbb{E}(\xi_t^1|\mathcal{G}, \mathcal{F}_{t-1}) = \mathbb{E}(\xi_t^2|\mathcal{G}, \mathcal{F}_{t-1}) = 0$, thus we can apply results for averaged SGD (see Proposition 9 of the Appendix) to get the following lemma.

Lemma 14

For any $t \geqslant 1$, if $\mathbb{E}\big[(\xi_t^1+\xi_t^2)\otimes_{\mathcal{H}}(\xi_t^1+\xi_t^2)|\mathcal{G}\big] \preccurlyeq \tau^2\hat{\Sigma}_n$, and $4\gamma R^2 = 1$, under Assumptions (A6), (A7), (A9),

$$\mathbb{E}\big[\|\hat{\Sigma}_n^{1/2}(\bar{\theta}_t - \bar{\eta}_t)\|_{\mathcal{H}}^2|\mathcal{G}\big] \leqslant \frac{8\tau^2\gamma^{1/\alpha}\mathrm{tr}\,\hat{\Sigma}_n^{1/\alpha}}{t^{1-1/\alpha}}. \tag{83}$$

In order to obtain the bound, we need to bound τ^2 (which is dependent on $\mathcal{G}$) and go from a bound with the empirical covariance matrix $\hat{\Sigma}_n$ to bounds with the population covariance matrix Σ.

We have

$$\mathbb{E}\big[\xi_t^1 \otimes_{\mathcal{H}} \xi_t^1|\mathcal{G}\big] \preccurlyeq_{\mathcal{H}} \mathbb{E}\big[y_{i(u)}^2\Phi(x_{i(u)}) \otimes_{\mathcal{H}} \Phi(x_{i(u)})|\mathcal{G}\big] \preccurlyeq_{\mathcal{H}} \|y\|_\infty^2\hat{\Sigma}_n \preccurlyeq_{\mathcal{H}} (\sigma + \sup_{x\in\mathcal{X}}\langle\theta_*, \Phi(x)\rangle_{\mathcal{H}})^2\hat{\Sigma}_n$$

$$\mathbb{E}\big[\xi_t^2 \otimes \xi_t^2|\mathcal{G}\big] \preccurlyeq_{\mathcal{H}} \mathbb{E}\big[\langle\eta_{t-1}, \Phi(x_{i(u)})\rangle^2\Phi(x_{i(u)}) \otimes_{\mathcal{H}} \Phi(x_{i(u)})|\mathcal{G}\big] \preccurlyeq_{\mathcal{H}} \sup_{t\in\{0,\dots,T-1\}}\sup_{x\in\mathcal{X}}\langle\eta_t, \Phi(x)\rangle_{\mathcal{H}}^2\hat{\Sigma}_n$$

Therefore $\tau^2 = 2M^2 + 2\sup_{t\in\{0,\dots,T-1\}}\sup_{x\in\mathcal{X}}\langle\eta_t, \Phi(x)\rangle_{\mathcal{H}}^2$ or using Assumption (A8) $\tau^2 = 2M^2 + 2\sup_{t\in\{0,\dots,T-1\}} R^{2\mu}\kappa_\mu^2\|\Sigma^{1/2-\mu/2}\eta_t\|_{\mathcal{H}}^2$.

In the proof, we rely on an event (that depend on $\mathcal{G}$) where $\hat{\Sigma}_n$ is close to Σ. This leads to the the following Lemma that bounds the deviation $\bar{\theta}_t - \bar{\eta}_t$.

Lemma 15

For any $t \geqslant 1$, $4\gamma R^2 = 1$, under Assumptions (A6), (A7), (A9),

$$\mathbb{E}\left[\|\Sigma^{1/2}(\bar{\theta}_t - \bar{\eta}_t)\|_{\mathfrak{H}}^2\right] \leqslant 16\tau_\infty^2 \left[R^{-2/\alpha}\mathrm{tr}\,\Sigma^{1/\alpha} t^{1/\alpha}\left(\frac{1}{t} + \left(\frac{4}{\mu}\frac{\log n}{n}\right)^{1/\mu}\right) + 1\right]. \tag{84}$$

We make the following remark on the bound.

Remark 5

Note that as defined in the proof τ_∞ may diverge in some cases as

$$\tau_\infty^2 = \begin{cases} O(1) & \text{when } \mu \leqslant 2r, \\ O(n^{\mu-2r}) & \text{when } 2r \leqslant \mu \leqslant 2r + 1/\alpha, \\ O(n^{1-2r/\mu}) & \text{when } \mu \geqslant 2r + 1/\alpha, \end{cases}$$

with $O(\cdot)$ are defined explicitly in the proof.

Convergence of batch gradient descent. The main result is summed up in the following lemma, with $t = O(n^{1/\mu})$ and $t \geqslant n$.

Lemma 16

Let $t > 1$, under Assumptions (A6), (A7), (A8), (A9), (A10), (A11), when, with $4\gamma R^2 = 1$,

$$t = \begin{cases} \Theta(n^{\alpha/(2r\alpha+1)}) & 2r\alpha + 1 > \mu\alpha \\ \Theta(n^{1/\mu}\,(\log n)^{\frac{1}{\mu}}) & 2r\alpha + 1 \leqslant \mu\alpha. \end{cases} \tag{85}$$

then,

$$\mathbb{E}F(\bar{\eta}_t) - F(\theta_*) \leqslant \begin{cases} O(n^{-2r\alpha/(2r\alpha+1)}) & 2r\alpha + 1 > \mu\alpha \\ O(n^{-2r/\mu}) & 2r\alpha + 1 \leqslant \mu\alpha \end{cases} \tag{86}$$

with $O(\cdot)$ are defined explicitly in the proof.

Remark 6

In all cases, we can notice that the speed of convergence of Lemma 16 are slower that the ones in Lemma 15, hence, the convergence of the gradient descent controls the rates of convergence of the algorithm.

B.3. Bounding the deviation between SGD and batch gradient descent

In this section, following the proof sketch from Section B.2, we provide a bound on the deviation $\theta_t - \eta_t$. In all the following let us denote $\mu_t = \theta_t - \eta_t$ that deviation between the stochastic gradient descent recursion and the batch gradient descent recursion.

B.3.1. Proof of Lemma 15

We need to (a) go from $\hat{\Sigma}_n$ to Σ in the result of Lemma 14 and (b) to have a bound on τ. To prove this result we are going to need the two following lemmas:

Lemma 17

Let $\lambda > 0$, $\delta \in (0,1]$. Under Assumption (A8), when $n \geqslant 11(1+\kappa_\mu^2 R^{2\mu}\gamma^\mu t^\mu)\log\frac{8R^2}{\lambda\delta}$, the following holds with probability $1-\delta$,

$$\left\|(\Sigma+\lambda I)^{1/2}(\hat{\Sigma}_n+\lambda I)^{-1/2}\right\|^2 \leqslant 2. \tag{87}$$

Proof : This Lemma is proven and stated lately in Lemma 24 in Section B.4.3. We recalled it here for the sake of clarity. ■

Lemma 18

Let $\lambda > 0$, $\delta \in (0,1]$. Under Assumption (A8), for $t = O\left(\frac{1}{n^{1/\mu}}\right)$ then the following holds with probability $1-\delta$,

$$\tau^2 \leqslant \tau_\infty^2 \quad \text{and} \quad \tau_\infty^2 = \begin{cases} O(1), & \text{when } \mu \leqslant 2r, \\ O\left(n^{\mu-2r}\right), & \text{when } 2r \leqslant \mu \leqslant 2r+1/\alpha, \\ O\left(n^{1-2r/\mu}\right) & \text{when } \mu \geqslant 2r+1/\alpha, \end{cases} \tag{88}$$

where the $O(\cdot)$-notation depend only on the parameters of the problem (and is independent of n and t).

Proof : This Lemma is a direct implication of Corollary 6 in Section B.4.3. We recalled it here for the sake of clarity. ■

Note that we can take $\lambda_n^\delta = \left(\frac{\log\frac{n}{\delta}}{n}\right)^{1/\mu}$ so that Lemma 17 result holds. Now we are ready to prove Lemma 15.

Proof of Lemma 15 : Let A_{δ_a} be the set for which inequality (87) holds and let B_{δ_b} be the set for which inequality (88) holds. Note that $\mathbb{P}(A_{\delta_a}^c) = \delta_a$ and $\mathbb{P}(B_{\delta_b}^c) = \delta_b$. We use the following decomposition:

$$\mathbb{E}\left\|\Sigma^{1/2}\bar{\mu}_t\right\|^2 \leqslant \mathbb{E}\left[\left\|\Sigma^{1/2}\bar{\mu}_t\right\|^2 \mathbf{1}_{A_{\delta_a}\cap B_{\delta_b}}\right] + \mathbb{E}\left[\left\|\Sigma^{1/2}\bar{\mu}_t\right\|^2 \mathbf{1}_{A_{\delta_a}^c}\right] + \mathbb{E}\left[\left\|\Sigma^{1/2}\bar{\mu}_t\right\|^2 \mathbf{1}_{B_{\delta_b}^c}\right].$$

First, let us bound roughly $\|\bar{\mu}_t\|^2$.

First, for $i \geqslant 1$, $\|\mu_i\|^2 \leqslant \gamma^2\left(\sum_{i=1}^t \|\xi_i^1\| + \|\xi_i^2\|\right)^2 \leqslant 16R^2\gamma^2\tau^2t^2$, so that $\|\bar{\mu}_t\|^2 \leqslant \frac{1}{t}\sum_{i=1}^t\|\mu_i\|^2 \leqslant 16R^2\gamma^2\tau^2t^2$. We can bound similarly $\tau^2 \leqslant 4M^2\gamma^2R^4t^2$, so that $\|\bar{\mu}_t\|^2 \leqslant 64R^2M^2\gamma^4t^4$. Thus, for the second term:

$$\mathbb{E}\left[\left\|\Sigma^{1/2}\bar{\mu}_t\right\|^2 \mathbf{1}_{A_{\delta_a}^c}\right] \leqslant 64R^8M^2\gamma^4t^4\mathbb{E}\mathbf{1}_{A_{\delta_a}^c} \leqslant 64R^8M^2\gamma^4t^4\delta_a,$$

and for the third term:

$$\mathbb{E}\left[\left\|\Sigma^{1/2}\bar{\mu}_t\right\|^2 \mathbf{1}_{B_{\delta_b}^c}\right] \leqslant 64R^8M^2\gamma^4t^4\mathbb{E}\mathbf{1}_{B_{\delta_b}^c} \leqslant 64R^8M^2\gamma^4t^4\delta_b.$$

And on for the first term,

$$\begin{aligned}
\mathbb{E}\left[\left\|\Sigma^{1/2}\bar{\mu}_t\right\|^2 \mathbf{1}_{A_{\delta_a}\cap B_{\delta_b}}\right] &\leqslant \mathbb{E}\left[\left\|\Sigma^{1/2}(\Sigma+\lambda_n^\delta I)^{-1/2}\right\|^2 \left\|(\Sigma+\lambda_n^\delta I)^{1/2}(\hat{\Sigma}_n+\lambda_n^\delta I)^{-1/2}\right\|^2 \right.\\
&\qquad\qquad \left. \left\|(\hat{\Sigma}_n+\lambda_n^\delta I)^{1/2}\bar{\mu}_t\right\|^2 \mathbf{1}_{A_{\delta_a}\cap B_{\delta_b}} \mid \mathcal{G}\right]\\
&\leqslant 2\mathbb{E}\left[\left\|(\hat{\Sigma}_n+\lambda_n^\delta I)^{1/2}\bar{\mu}_t\right\|^2 \mid \mathcal{G}\right]\\
&= 2\mathbb{E}\left[\left\|\hat{\Sigma}_n^{1/2}\bar{\mu}_t\right\|^2 \mid \mathcal{G}\right] + 2\lambda_n^\delta \mathbb{E}\left[\|\bar{\mu}_t\|^2 \mid \mathcal{G}\right]\\
&\leqslant 16\tau_\infty^2 \frac{\gamma^{1/\alpha}\mathbb{E}\left[\operatorname{tr}\hat{\Sigma}_n^{1/\alpha}\right]}{t^{1-1/\alpha}} + 8\lambda_n^\delta\tau_\infty^2\,\gamma^{1/\alpha}\mathbb{E}\left[\operatorname{tr}\hat{\Sigma}_n^{1/\alpha}\right] t^{1/\alpha},
\end{aligned}$$

using Proposition 9 twice with $u=1$ for the left term and $u=1$ for the right one.

As $x\to x^{1/\alpha}$ is a concave function, we can apply Jensen's inequality to have :

$$\mathbb{E}\left[\operatorname{tr}(\hat{\Sigma}_n^{1/\alpha})\right] \leqslant \operatorname{tr}\Sigma^{1/\alpha},$$

so that:

$$\begin{aligned}
\mathbb{E}\left[\left\|\Sigma^{1/2}\bar{\mu}_t\right\|^2 \mathbf{1}_{A_{\delta_a}\cap B_{\delta_b}}\right] &\leqslant 16\tau_\infty^2 \frac{\gamma^{1/\alpha}\operatorname{tr}\Sigma^{1/\alpha}}{t^{1-1/\alpha}} + 8\lambda_n^\delta\tau_\infty^2\gamma\,\gamma^{1/\alpha}\operatorname{tr}\Sigma^{1/\alpha}t^{1/\alpha}\\
&\leqslant 16\tau_\infty^2\gamma^{1/\alpha}\operatorname{tr}\Sigma^{1/\alpha}t^{1/\alpha}\left(\frac{1}{t}+\lambda_n^\delta\right).
\end{aligned}$$

Now, we take $\delta_a=\delta_b=\frac{\tau_\infty^2}{4M^2R^8\gamma^4t^4}$ and this concludes the proof of Lemma 15, with the bound:

$$\mathbb{E}\left\|\Sigma^{1/2}\bar{\mu}_t\right\|^2 \leqslant 16\tau_\infty^2\gamma^{1/\alpha}\operatorname{tr}\Sigma^{1/\alpha}t^{1/\alpha}\left(\frac{1}{t}+\left(\frac{2+2\log M+4\log(\gamma R^2)+4\log t}{n}\right)^{1/\mu}\right). \qquad \blacksquare$$

B.4. Convergence of batch gradient descent

In this section we prove the convergence of averaged batch gradient descent to the target function. In particular, since the proof technique is valid for the wider class of algorithms known as spectral filters [GRO+08, LRRC18], we will do the proof for a generic spectral filter (in Lemma 19, Sect. B.4.1 we prove that averaged batch gradient descent is a spectral filter).

In Section B.4.1 we provide the required notation and additional definitions. In Section B.4.2, in particular in Theorem B.4.2 we perform an analytical decomposition of the excess risk of the averaged batch gradient descent, in terms of basic quantities that will be controlled in expectation (or probability) in the next sections. In Section B.4.3 the various quantites obtained by the analytical decomposition are controlled, in particular, Corollary 6 controls the L^∞ norm of the averaged batch gradient descent algorithm. Finally in Section B.4.4, the main result, Theorem 10 controlling in expectation of the excess risk of the averaged batch gradient descent estimator is provided. In Corollary 7, a version of the result of Theorem 10 is given, with explicit rates for the regularization parameters and of the excess risk.

B.4.1. Notations

In this subsection, we study the convergence of batch gradient descent. For the sake of clarity we consider the RKHS framework (which includes the finite-dimensional case). We will thus consider elements of $\mathcal{H}$ that are naturally embedded in $L_2(d\rho_X)$ by the operator S from $\mathcal{H}$ to $L_2(d\rho_X)$ and such

that: $(Sg)(x) = \langle g, K_x \rangle$, where we have $\Phi(x) = K_x = K(\cdot, x)$ where $K : \mathcal{X} \to \mathcal{X} \to \mathbb{R}$ is the kernel. We recall the recursion for η_t in the case of an RKHS feature space with kernel K:

$$\eta_t = \eta_{t-1} + \frac{\gamma}{n} \sum_{i=1}^{n} \big(y_i - \langle \eta_{t-1}, K_{x_i} \rangle_{\mathcal{H}}\big) K_{x_i},$$

Let us begin with some notations. In the following we will often use the letter g to denote vectors of $\mathcal{H}$, hence, Sg will denote functions of $L_2(d\rho_{\mathcal{X}})$. We also define the following operators (we may also use their adjoints, denoted with a $*$):

- The operator $\widehat{\mathsf{S}}_n$ from $\mathcal{H}$ to $\mathbb{R}^n$, $\widehat{\mathsf{S}}_n g = \frac{1}{\sqrt{n}}(g(x_1), \dots g(x_n))$.
- The operators from $\mathcal{H}$ to $\mathcal{H}$, Σ and $\widehat{\Sigma}_n$, defined respectively as $\Sigma = \mathbb{E}\left[K_x \otimes K_x\right] = \int_{\mathcal{X}} K_x \otimes K_x d\rho_{\mathcal{X}}$ and $\widehat{\Sigma}_n = \frac{1}{n} \sum_{i=1}^{n} K_{x_i} \otimes K_{x_i}$. Note that Σ is the covariance operator.
- The operator $\mathcal{L} : L_2(d\rho_{\mathcal{X}}) \to L_2(d\rho_{\mathcal{X}})$ is defined by

  $$(\mathcal{L}f)(x) = \int_{\mathcal{X}} K(x, z) f(z) d\rho_{\mathcal{X}}(x), \quad \forall f \in L_2(d\rho_{\mathcal{X}}).$$

  Moreover denote by $\mathcal{N}(\lambda)$ the so called *effective dimension* of the learning problem, that is defined as

  $$\mathcal{N}(\lambda) = \mathrm{tr}(\mathcal{L}(\mathcal{L} + \lambda I)^{-1}),$$

  for $\lambda > 0$. Recall that by Assumption (A9), there exists $\alpha \geqslant 1$ and $Q > 0$ such that

  $$\mathcal{N}(\lambda) \leqslant Q \lambda^{-1/\alpha}, \quad \forall \lambda > 0.$$

  We can take $Q = \mathrm{tr}\Sigma^{1/\alpha}$.
- $P : L_2(d\rho_{\mathcal{X}}) \to L_2(d\rho_{\mathcal{X}})$ projection operator on $\mathcal{H}$ for the $L_2(d\rho_{\mathcal{X}})$ norm s.t. $\mathrm{ran}P = \mathrm{ran}S$.

Denote by f_ρ the function so that $f_\rho(x) = \mathbb{E}[y|x] \ \in L_2(d\rho_{\mathcal{X}})$ the minimizer of the expected risk, defined by $F(f) = \int_{X \times \mathbb{R}} (f(x) - y)^2 d\rho(x, y)$.

Remark 7 (*On Assumption (A10)*)

With the notation above, we express assumption (A10), more formally, w.r.t. Hilbert spaces with infinite dimensions, as follows. There exists $r \in [0, 1]$ and $\phi \in L_2(d\rho_{\mathcal{X}})$, such that

$$Pf_\rho = \mathcal{L}^r \phi.$$

(A11) *Let $q \in [1, \infty]$ be such that $\|f_\rho - Pf_\rho\|_{L^{2q}(\mathcal{X}, \rho_{\mathcal{X}})} < \infty$.*

The assumption above is always true for $q = 1$, moreover when the kernel is universal it is true even for $q = \infty$. Moreover if $r \geqslant 1/2$ then it is true for $q = \infty$. Note that we make the calculation in this Appendix for a general $q \in [1, \infty]$, but we presented the results for $q = \infty$ in the main paper. The following proposition relates the excess risk to a certain norm.

Proposition 10

When $\widehat{g} \in \mathcal{H}$,

$$F(\widehat{g}) - \inf_{g \in \mathcal{H}} F(g) = \|S\widehat{g} - Pf_\rho\|^2_{L_2(d\rho_{\mathcal{X}})}.$$

We introduce the following function $g_\lambda \in \mathcal{H}$ that will be useful in the rest of the paper $g_\lambda = (\Sigma + \lambda I)^{-1} S^* f_\rho$.

We introduce the estimators of the form, for $\lambda > 0$,

$$\widehat{g}_\lambda = q_\lambda(\widehat{\Sigma}_n)\widehat{\mathsf{S}}_n^* \widehat{y},$$

where $q_\lambda : \mathbb{R}_+ \to \mathbb{R}_+$ is a function called *filter*, that essentially approximates x^{-1} with the approximation controlled by λ. Denote moreover with r_λ the function $r_\lambda(x) = 1 - x q_\lambda(x)$. The following definition precises the form of the filters we want to analyze. We then prove in Lemma 19 that our estimator corresponds to such a filter.

Definition 2 (*Spectral filters*)

Let $q_\lambda : \mathbb{R}_+ \to \mathbb{R}_+$ be a function parametrized by $\lambda > 0$. q_λ is called a filter *when there exists $c_q > 0$ for which*

$$\lambda q_\lambda(x) \leqslant c_q, \quad r_\lambda(x) x^u \leqslant c_q \lambda^u, \quad \forall x > 0, \lambda > 0, u \in [0, 1].$$

We now justify that we study estimators of the form $\widehat{g}_\lambda = q_\lambda(\widehat{\Sigma}_n)\widehat{\mathsf{S}}_n^* \widehat{y}$ with the following lemma. Indeed, we show that the average of batch gradient descent can be represented as a filter estimator, $\widehat{g}_\lambda$, for $\lambda = 1/(\gamma t)$.

Lemma 19

For $t > 1$, $\lambda = 1/(\gamma t)$, $\bar{\eta}_t = \widehat{g}_\lambda$, with respect to the filter, $q^\eta(x) = \left(1 - \frac{1-(1-\gamma x)^t}{\gamma t x}\right)\frac{1}{x}$.

Proof : Indeed, for $t > 1$,

$$\begin{aligned}
\eta_t &= \eta_{t-1} + \frac{\gamma}{n}\sum_{i=1}^n \left(y_i - \langle \eta_{t-1}, K_{x_i}\rangle_{\mathcal{H}}\right) K_{x_i} \\
&= \eta_{t-1} + \gamma(\widehat{\mathsf{S}}_n^* \widehat{y} - \widehat{\Sigma}_n \eta_{t-1}) \\
&= (I - \gamma\widehat{\Sigma}_n)\eta_{t-1} + \gamma \widehat{\mathsf{S}}_n^* \widehat{y} \\
&= \gamma \sum_{k=0}^{t-1} (I - \gamma\widehat{\Sigma}_n)^k \widehat{\mathsf{S}}_n^* \widehat{y} = \left[I - (I - \gamma\widehat{\Sigma}_n)^t\right] \widehat{\Sigma}_n^{-1} \widehat{\mathsf{S}}_n^* \widehat{y},
\end{aligned}$$

leading to

$$\bar{\eta}_t = \frac{1}{t}\sum_{i=0}^t \eta_i = q^\eta\left(\widehat{\Sigma}_n\right)\widehat{\mathsf{S}}_n^* \widehat{y}.$$

Now, we prove that q has the properties of a filter. First, for $t > 1$, $\frac{1}{\gamma t} q^\eta(x) = \left(1 - \frac{1-(1-\gamma x)^t}{\gamma t x}\right)\frac{1}{\gamma t x}$ is a decreasing function so that $\frac{1}{\gamma t} q^\eta(x) \leqslant \frac{1}{\gamma t} q^\eta(0) \leqslant 1$. Second for $u \in [0, 1]$, $x^u(1 - x q^\eta(x)) = \frac{1-(1-\gamma x)^t}{\gamma t x} x^u$. As used in Section B.1.2, $1 - (1 - \gamma x)^t \leqslant (\gamma t x)^{1-u}$, so that, $r^\eta(x) x^u \leqslant \frac{(\gamma t x)^{1-u}}{\gamma t x} x^u = \frac{1}{(\gamma t)^u}$, this concludes the proof that q^η is indeed a filter. ■

B.4.2. Analytical decomposition

Lemma 20

Let $\lambda > 0$ and $s \in (0, 1/2]$. Under Assumption (A10) (see Rem. 7), the following holds

$$\|\mathcal{L}^{-s} S(\widehat{g}_\lambda - g_\lambda)\|_{L_2(d\rho_X)} \leqslant 2\lambda^{-s}\beta^2 c_q \|\Sigma_\lambda^{-1/2}(\widehat{\mathsf{S}}_n^* \widehat{y} - \widehat{\Sigma}_n g_\lambda)\|_{\mathcal{H}} + 2\beta c_q \|\phi\|_{L_2(d\rho_X)} \lambda^{r-s},$$

where $\beta := \|\Sigma_\lambda^{1/2} \widehat{\Sigma}_{n\lambda}^{-1/2}\|$.

Proof : By Prop. 10, we can characterize the excess risk of $\widehat{g}_\lambda$ in terms of the $L_2(d\rho_{\mathcal{X}})$ squared norm of $S\widehat{g}_\lambda - Pf_\rho$. In this paper, simplifying the analysis of [LRRC18], we perform the following decomposition

$$\begin{aligned}\mathcal{L}^{-s}S(\widehat{g}_\lambda - g_\lambda) = \mathcal{L}^{-s}S\widehat{g}_\lambda - \mathcal{L}^{-s}Sq_\lambda(\widehat{\Sigma}_n)\widehat{\Sigma}_n g_\lambda \\ + \ \mathcal{L}^{-s}Sq_\lambda(\widehat{\Sigma}_n)\widehat{\Sigma}_n g_\lambda - \mathcal{L}^{-s}Sg_\lambda.\end{aligned}$$

Upper bound for the first term. By using the definition of $\widehat{g}_\lambda$ and multiplying and dividing by $\Sigma_\lambda^{1/2}$, we have that

$$\begin{aligned}\mathcal{L}^{-s}S\widehat{g}_\lambda - \mathcal{L}^{-s}Sq_\lambda(\widehat{\Sigma}_n)\widehat{\Sigma}_n g_\lambda &= \mathcal{L}^{-s}Sq_\lambda(\widehat{\Sigma}_n)(\widehat{\mathsf{S}}_n^*\widehat{y} - \widehat{\Sigma}_n g_\lambda) \\ &= \mathcal{L}^{-s}Sq_\lambda(\widehat{\Sigma}_n)\Sigma_\lambda^{1/2}\,\Sigma_\lambda^{-1/2}(\widehat{\mathsf{S}}_n^*\widehat{y} - \widehat{\Sigma}_n g_\lambda),\end{aligned}$$

from which

$$\|\mathcal{L}^{-s}S(\widehat{g}_\lambda - q_\lambda(\widehat{\Sigma}_n)\widehat{\Sigma}_n g_\lambda)\|_{L_2(d\rho_{\mathcal{X}})} \leqslant \|\mathcal{L}^{-s}Sq_\lambda(\widehat{\Sigma}_n)\Sigma_\lambda^{1/2}\|\,\|\Sigma_\lambda^{-1/2}(\widehat{\mathsf{S}}_n^*\widehat{y} - \widehat{\Sigma}_n g_\lambda)\|_{\mathcal{H}}.$$

Upper bound for the second term. By definition of $r_\lambda(x) = 1 - xq_\lambda(x)$ and $g_\lambda = \Sigma_\lambda^{-1}S^*f_\rho$,

$$\begin{aligned}\mathcal{L}^{-s}Sq_\lambda(\widehat{\Sigma}_n)\widehat{\Sigma}_n g_\lambda - \mathcal{L}^{-s}Sg_\lambda &= \mathcal{L}^{-s}S(q_\lambda(\widehat{\Sigma}_n)\widehat{\Sigma}_n - I)g_\lambda \\ &= -\mathcal{L}^{-s}Sr_\lambda(\widehat{\Sigma}_n)\,\Sigma_\lambda^{-(1/2-r)}\,\Sigma_\lambda^{-1/2-r}S^*\mathcal{L}^r\,\phi,\end{aligned}$$

where in the last step we used the fact that $S^*f_\rho = S^*Pf_\rho = S^*\mathcal{L}^r\phi$, by Asm. (A10) (see Rem. 7). Then

$$\begin{aligned}\|\mathcal{L}^{-s}S(q_\lambda(\widehat{\Sigma}_n)\widehat{\Sigma}_n - I)g_\lambda)\|_{L_2(d\rho_{\mathcal{X}})} &\leqslant \|\mathcal{L}^{-s}Sr_\lambda(\widehat{\Sigma}_n)\|\|\Sigma_\lambda^{-(1/2-r)}\|\|\Sigma_\lambda^{-1/2-r}S^*\mathcal{L}^r\|\|\phi\|_{L_2(d\rho_{\mathcal{X}})} \\ &\leqslant \lambda^{-(1/2-r)}\|\mathcal{L}^{-s}Sr_\lambda(\widehat{\Sigma}_n)\|\|\phi\|_{L_2(d\rho_{\mathcal{X}})},\end{aligned}$$

where the last step is due to the fact that $\|\Sigma_\lambda^{-(1/2-r)}\| \leqslant \lambda^{-(1/2-r)}$ and that $S^*\mathcal{L}^{2r}S = S^*(SS^*)^{2r}S = (S^*S)^{2r}S^*S = \Sigma^{1+2r}$ from which

$$\|\Sigma_\lambda^{-1/2-r}S^*\mathcal{L}^r\|^2 = \|\Sigma_\lambda^{-1/2-r}S^*\mathcal{L}^{2r}S\Sigma_\lambda^{-1/2-r}\| = \|\Sigma_\lambda^{-1/2-r}\Sigma^{1+2r}\Sigma_\lambda^{-1/2-r}\| \leqslant 1. \tag{89}$$

Additional decompositions. We further bound $\|\mathcal{L}^{-s}Sr_\lambda(\widehat{\Sigma}_n)\|$ and $\|\mathcal{L}^{-s}Sq_\lambda(\widehat{\Sigma}_n)\Sigma_\lambda^{1/2}\|$. For the first, by the identity $\mathcal{L}^{-s}Sr_\lambda(\widehat{\Sigma}_n) = \mathcal{L}^{-s}S\widehat{\Sigma}_{n\lambda}^{-1/2}\widehat{\Sigma}_{n\lambda}^{1/2}r_\lambda(\widehat{\Sigma}_n)$, we have

$$\|\mathcal{L}^{-s}Sr_\lambda(\widehat{\Sigma}_n)\| = \|\mathcal{L}^{-s}S\widehat{\Sigma}_{n\lambda}^{-1/2}\|\|\widehat{\Sigma}_{n\lambda}^{1/2}r_\lambda(\widehat{\Sigma}_n)\|,$$

where

$$\|\widehat{\Sigma}_{n\lambda}^{1/2}r_\lambda(\widehat{\Sigma}_n)\| = \sup_{\sigma\in\sigma(\widehat{\Sigma}_n)}(\sigma+\lambda)^{1/2}r_\lambda(\sigma) \leqslant \sup_{\sigma\geqslant 0}(\sigma+\lambda)^{1/2}r_\lambda(\sigma) \leqslant 2c_q\lambda^{1/2}.$$

Similarly, by using the identity

$$\mathcal{L}^{-s}Sq_\lambda(\widehat{\Sigma}_n)\Sigma_\lambda^{1/2} = \mathcal{L}^{-s}S\widehat{\Sigma}_{n\lambda}^{-1/2}\,\widehat{\Sigma}_{n\lambda}^{1/2}q_\lambda(\widehat{\Sigma}_n)\widehat{\Sigma}_{n\lambda}^{1/2}\,\widehat{\Sigma}_{n\lambda}^{-1/2}\Sigma_\lambda^{1/2},$$

we have

$$\|\mathcal{L}^{-s}Sq_\lambda(\widehat{\Sigma}_n)\Sigma_\lambda^{1/2}\| = \|\mathcal{L}^{-s}S\widehat{\Sigma}_{n\lambda}^{-1/2}\|\,\|\widehat{\Sigma}_{n\lambda}^{1/2}q_\lambda(\widehat{\Sigma}_n)\widehat{\Sigma}_{n\lambda}^{1/2}\|\,\|\widehat{\Sigma}_{n\lambda}^{-1/2}\Sigma_\lambda^{1/2}\|.$$

Finally note that

$$\|\mathcal{L}^{-s}S\widehat{\Sigma}_{n\lambda}^{-1/2}\| \leqslant \|\mathcal{L}^{-s}S\Sigma_\lambda^{-1/2+s}\|\|\Sigma_\lambda^{-s}\|\|\Sigma_\lambda^{1/2}\widehat{\Sigma}_{n\lambda}^{-1/2}\|,$$

and $\|\mathcal{L}^{-s}S\Sigma_\lambda^{-1/2+s}\| \leqslant 1$, $\|\Sigma_\lambda^{-s}\| \leqslant \lambda^{-s}$, and moreover

$$\|\widehat{\Sigma}_{n\lambda}^{1/2}q_\lambda(\widehat{\Sigma}_n)\widehat{\Sigma}_{n\lambda}^{1/2}\| = \sup_{\sigma\in\sigma(\widehat{\Sigma}_n)}(\sigma+\lambda)q_\lambda(\sigma) \leqslant \sup_{\sigma\geqslant 0}(\sigma+\lambda)q_\lambda(\sigma) \leqslant 2c_q,$$

so, in conclusion

$$\|\mathcal{L}^{-s}Sr_\lambda(\widehat{\Sigma}_n)\| \leqslant 2c_q\lambda^{1/2-s}\beta, \quad \|\mathcal{L}^{-s}Sq_\lambda(\widehat{\Sigma}_n)\Sigma_\lambda^{1/2}\| \leqslant 2c_q\lambda^{-s}\beta^2.$$

The final result is obtained by gathering the upper bounds for the three terms above and the additional terms of this last section. ∎

Lemma 21

Let $\lambda > 0$ and $s \in (0, \min(r, 1/2)]$. Under Assumption (A10) (see Rem. 7), the following holds

$$\|\mathcal{L}^{-s}(S\widehat{g}_\lambda - Pf_\rho)\|_{L_2(d\rho_X)} \leqslant \lambda^{r-s}\|\phi\|_{L_2(d\rho_X)}.$$

Proof: Since $S\Sigma_\lambda^{-1}S^* = \mathcal{L}\mathcal{L}_\lambda^{-1} = I - \lambda\mathcal{L}_\lambda^{-1}$, we have

$$\begin{aligned}\mathcal{L}^{-s}(Sg_\lambda - Pf_\rho) &= \mathcal{L}^{-s}(S\Sigma_\lambda^{-1}S^*f_\rho - Pf_\rho) = \mathcal{L}^{-s}(S\Sigma_\lambda^{-1}S^*Pf_\rho - Pf_\rho)\\ &= \mathcal{L}^{-s}(S\Sigma_\lambda^{-1}S^* - I)Pf_\rho = \mathcal{L}^{-s}(S\Sigma_\lambda^{-1}S^* - I)\mathcal{L}^r\,\phi\\ &= -\lambda\mathcal{L}^{-s}\mathcal{L}_\lambda^{-1}L^r\phi = -\lambda^{r-s}\,\lambda^{1-r+s}\mathcal{L}_\lambda^{-(1-r+s)}\,\mathcal{L}_\lambda^{-(r-s)}\mathcal{L}^{r-s}\,\phi,\end{aligned}$$

from which

$$\begin{aligned}\|\mathcal{L}^{-s}(Sg_\lambda - Pf_\rho)\|_{L_2(d\rho_X)} &\leqslant \lambda^{r-s}\|\lambda^{1-r+s}\mathcal{L}_\lambda^{-(1-r+s)}\|\|\mathcal{L}_\lambda^{-(r-s)}\mathcal{L}^{r-s}\|\,\|\phi\|_{L_2(d\rho_X)}\\ &\leqslant \lambda^{r-s}\|\phi\|_{L_2(d\rho_X)}.\end{aligned}$$

■

Theorem 9

Let $\lambda > 0$ and $s \in (0, \min(r, 1/2)]$. Under Assumption (A10) (see Rem. 7), the following holds

$$\|\mathcal{L}^{-s}(S\widehat{g}_\lambda - Pf_\rho)\|_{L_2(d\rho_X)} \leqslant 2\lambda^{-s}\beta^2 c_q\|\Sigma_\lambda^{-1/2}(\widehat{\mathsf{S}}_n^*\widehat{y} - \widehat{\Sigma}_n g_\lambda)\|_{\mathcal{H}} + \left(1 + \beta 2c_q\|\phi\|_{L_2(d\rho_X)}\right)\lambda^{r-s}$$

where $\beta := \|\Sigma_\lambda^{1/2}\widehat{\Sigma}_{n\lambda}^{-1/2}\|$.

Proof: By Prop. 10, we can characterize the excess risk of $\widehat{g}_\lambda$ in terms of the $L_2(d\rho_X)$ squared norm of $S\widehat{g}_\lambda - Pf_\rho$. In this paper, simplifying the analysis of [LRRC18], we perform the following decomposition

$$\begin{aligned}\mathcal{L}^{-s}(S\widehat{g}_\lambda - Pf_\rho) &= \mathcal{L}^{-s}S\widehat{g}_\lambda - \mathcal{L}^{-s}Sg_\lambda\\ &\quad + \mathcal{L}^{-s}(Sg_\lambda - Pf_\rho).\end{aligned}$$

The first term is bounded by Lemma 20, the second is bounded by Lemma 21. ■

B.4.3. Probabilistic bounds

In this section denote by $\mathcal{N}_\infty(\lambda)$, the quantity

$$\mathcal{N}_\infty(\lambda) = \sup_{x\in S}\|\Sigma_\lambda^{-1/2}K_x\|_{\mathcal{H}}^2,$$

where $S \subseteq \mathcal{X}$ is the support of the probability measure ρ_X.

Lemma 22

Under Asm. (A8), we have that for any $g \in \mathcal{H}$

$$\sup_{x\in supp(\rho_X)}|g(x)| \leqslant \kappa_\mu R^\mu\|\Sigma^{1/2(1-\mu)}g\|_{\mathcal{H}} = \kappa_\mu R^\mu\|\mathcal{L}^{-\mu/2}Sg\|_{L_2(d\rho_X)}.$$

Proof: Note that, Asm. (A8) is equivalent to

$$\|\Sigma^{-1/2(1-\mu)}K_x\| \leqslant \kappa_\mu R^\mu,$$

for all x in the support of $\rho_{\mathcal{X}}$. Then we have, for any x in the support of $\rho_{\mathcal{X}}$,

$$|g(x)| = \langle g, K_x \rangle_{\mathcal{H}} = \left\langle \Sigma^{1/2(1-\mu)} g, \Sigma^{-1/2(1-\mu)} K_x \right\rangle_{\mathcal{H}}$$
$$\leqslant \|\Sigma^{1/2(1-\mu)} g\|_{\mathcal{H}} \|\Sigma^{-1/2(1-\mu)} K_x\| \leqslant \kappa_\mu R^\mu \|\Sigma^{1/2(1-\mu)} g\|_{\mathcal{H}}.$$

Now note that, since $\Sigma^{1-\mu} = S^* \mathcal{L}^{-\mu} S$, we have

$$\|\Sigma^{1/2(1-\mu)} g\|_{\mathcal{H}}^2 = \langle g, \Sigma^{1-\mu} g \rangle_{\mathcal{H}} = \left\langle \mathcal{L}^{-\mu/2} Sg, \mathcal{L}^{-\mu/2} Sg \right\rangle_{L_2(d\rho_{\mathcal{X}})}.$$

■

Lemma 23

Under Assumption (A8), we have

$$\mathcal{N}_\infty(\lambda) \leqslant \kappa_\mu^2 R^{2\mu} \lambda^{-\mu}.$$

Proof: First denote with $f_{\lambda,u} \in \mathcal{H}$ the function $\Sigma_\lambda^{-1/2} u$ for any $u \in \mathcal{H}$ and $\lambda > 0$. Note that

$$\|f_{\lambda,u}\|_{\mathcal{H}} = \|\Sigma_\lambda^{-1/2} u\|_{\mathcal{H}} \leqslant \|\Sigma_\lambda^{-1/2}\| \|u\|_{\mathcal{H}} \leqslant \lambda^{-1/2} \|u\|_{\mathcal{H}}.$$

Moreover, since for any $g \in \mathcal{H}$ the identity $\|g\|_{L_2(d\rho_{\mathcal{X}})} = \|Sg\|_{\mathcal{H}}$, we have

$$\|f_{\lambda,u}\|_{L_2(d\rho_{\mathcal{X}})} = \|S\Sigma_\lambda^{-1/2} u\|_{\mathcal{H}} \leqslant \|S\Sigma_\lambda^{-1/2}\| \|u\|_{\mathcal{H}} \leqslant \|u\|_{\mathcal{H}}.$$

Now denote with $B(\mathcal{H})$ the unit ball in $\mathcal{H}$, by applying Asm. (A8) to $f_{\lambda,u}$ we have that

$$\mathcal{N}_\infty(\lambda) = \sup_{x \in S} \|\Sigma_\lambda^{-1/2} K_x\|^2 = \sup_{x \in S, u \in B(\mathcal{H})} \left\langle u, \Sigma_\lambda^{-1/2} K_x \right\rangle_{\mathcal{H}}^2$$
$$= \sup_{x \in S, u \in B(\mathcal{H})} \langle f_{\lambda,u}, K_x \rangle_{\mathcal{H}}^2 = \sup_{u \in B(\mathcal{H})} \sup_{x \in S} |f_{\lambda,u}(x)|^2$$
$$\leqslant \kappa_\mu^2 R^{2\mu} \sup_{u \in B(\mathcal{H})} \|f_{\lambda,u}\|_{\mathcal{H}}^{2\mu} \|f_{\lambda,u}\|_{L_2(d\rho_{\mathcal{X}})}^{2-2\mu}$$
$$\leqslant \kappa_\mu^2 R^{2\mu} \lambda^{-\mu} \sup_{u \in B(\mathcal{H})} \|u\|_{\mathcal{H}}^2 \leqslant \kappa_\mu^2 R^{2\mu} \lambda^{-\mu}.$$

■

Lemma 24

Let $\lambda > 0$, $\delta \in (0, 1]$ and $n \in \mathbb{N}$. Under Assumption (A8), we have that, when

$$n \geqslant 11(1 + \kappa_\mu^2 R^{2\mu} \lambda^{-\mu}) \log \frac{8R^2}{\lambda\delta},$$

then the following holds with probability $1 - \delta$,

$$\|\Sigma_\lambda^{1/2} \widehat{\Sigma}_{n\lambda}^{-1/2}\|^2 \leqslant 2.$$

Proof: This result is a refinement of the one in [RCR13] and is based on non-commutative Bernstein inequalities for random matrices [Tro12a]. By Prop. 8 in [RR17], we have that

$$\|\Sigma_\lambda^{1/2} \widehat{\Sigma}_{n\lambda}^{-1/2}\|^2 \leqslant (1-t)^{-1}, \quad t := \|\Sigma_\lambda^{-1/2} (\Sigma - \widehat{\Sigma}_n) \Sigma_\lambda^{-1/2}\|.$$

When $0 < \lambda \leqslant \|\Sigma\|$, by Prop. 6 of [RR17] (see also [RCR17] Lemma 9 for more refined constants), we have that the following holds with probability at least $1 - \delta$,

$$t \leqslant \frac{2\eta(1 + \mathcal{N}_\infty(\lambda))}{3n} + \sqrt{\frac{2\eta \mathcal{N}_\infty(\lambda)}{n}},$$

with $\eta = \log \frac{8R^2}{\lambda\delta}$. Finally, by selecting $n \geqslant 11(1 + \kappa_\mu^2 R^{2\mu} \lambda^{-\mu})\eta$, we have that $t \leqslant 1/2$ and so $\|\Sigma_\lambda^{1/2} \widehat{\Sigma}_{n\lambda}^{-1/2}\|^2 \leqslant (1-t)^{-1} \leqslant 2$, with probability $1 - \delta$.

To conclude note that when $\lambda \geqslant \|\Sigma\|$, we have

$$\|\Sigma_\lambda^{1/2}\widehat{\Sigma}_{n\lambda}^{-1/2}\|^2 \leqslant \|\Sigma + \lambda I\|\|(\widehat{\Sigma}_n + \lambda I)^{-1}\| \leqslant \frac{\|\Sigma\| + \lambda}{\lambda} = 1 + \frac{\|\Sigma\|}{\lambda} \leqslant 2.$$

■

Lemma 25

Under Assumption (A8), (A9), (A10) (see Rem. 7), (A11) we have

1. *Let $\lambda > 0$, $n \in \mathbb{N}$, the following holds*

$$\mathbb{E}[\|\Sigma_\lambda^{-1/2}(\widehat{\mathsf{S}}_n^*\widehat{y} - \widehat{\Sigma}_n g_\lambda)\|_{\mathcal{H}}^2] \leqslant \|\phi\|_{L_2(d\rho_{\mathcal{X}})}^2 \lambda^{2r} + \frac{2\kappa_\mu^2 R^{2\mu}\lambda^{-(\mu-2r)}}{n} + \frac{4\kappa_\mu^2 R^{2\mu} AQ\lambda^{-\frac{q+\mu\alpha}{q\alpha+\alpha}}}{n},$$

   where $A := \|f_\rho - Pf_\rho\|_{L^{2q}(\mathcal{X},\rho_{\mathcal{X}})}^{2-2/(q+1)}$.

2. *Let $\delta \in (0,1]$, under the same assumptions, the following holds with probability at least $1-\delta$*

$$\|\Sigma_\lambda^{-1/2}(\widehat{\mathsf{S}}_n^*\widehat{y} - \widehat{\Sigma}_n g_\lambda)\|_{\mathcal{H}} \leqslant c_0\lambda^r + \frac{4(c_1\lambda^{-\frac{\mu}{2}} + c_2\lambda^{-r-\mu})\log\frac{2}{\delta}}{n} + \sqrt{\frac{16\kappa_\mu^2 R^{2\mu}(\lambda^{-(\mu-2r)} + 2AQ\lambda^{-\frac{q+\mu\alpha}{q\alpha+\alpha}})\log\frac{2}{\delta}}{n}},$$

   with $c_0 = \|\phi\|_{L_2(d\rho_{\mathcal{X}})}$, $c_1 = \kappa_\mu R^\mu M + \kappa_\mu^2 R^{2\mu}(2R)^{2r-\mu}\|\phi\|_{L_2(d\rho_{\mathcal{X}})}$, $c_2 = \kappa_\mu^2 R^{2\mu}\|\phi\|_{L_2(d\rho_{\mathcal{X}})}$

Proof : First denote with ζ_i the random variable

$$\zeta_i = (y_i - g_\lambda(x_i))\Sigma_\lambda^{-1/2}K_{x_i}.$$

In particular note that, by using the definitions of $\widehat{\mathsf{S}}_n$, $\widehat{y}$ and $\widehat{\Sigma}_n$, we have

$$\Sigma_\lambda^{-1/2}(\widehat{\mathsf{S}}_n^*\widehat{y} - \widehat{\Sigma}_n g_\lambda) = \Sigma_\lambda^{-1/2}(\frac{1}{n}\sum_{i=1}^n K_{x_i}y_i - \frac{1}{n}(K_{x_i}\otimes K_{x_i})g_\lambda) = \frac{1}{n}\sum_{i=1}^n \zeta_i.$$

I So, by noting that ζ_i are independent and identically distributed, we have

$$\begin{aligned}\mathbb{E}[\|\Sigma_\lambda^{-1/2}(\widehat{\mathsf{S}}_n^*\widehat{y} - \widehat{\Sigma}_n g_\lambda)\|_{\mathcal{H}}^2] &= \mathbb{E}[\|\frac{1}{n}\sum_{i=1}^n \zeta_i\|_{\mathcal{H}}^2] = \frac{1}{n^2}\sum_{i,j=1}^n \mathbb{E}[\langle\zeta_i,\zeta_j\rangle_{\mathcal{H}}] \\ &= \frac{1}{n}\mathbb{E}[\|\zeta_1\|_{\mathcal{H}}^2] + \frac{n-1}{n}\|\mathbb{E}[\zeta_1]\|_{\mathcal{H}}^2.\end{aligned}$$

Now note that

$$\mathbb{E}[\zeta_1] = \Sigma_\lambda^{-1/2}(\mathbb{E}[K_{x_1}y_1] - \mathbb{E}[K_{x_1}\otimes K_{x_1}]g_\lambda) = \Sigma_\lambda^{-1/2}(S^*f_\rho - \Sigma g_\lambda).$$

In particular, by the fact that $S^*f_\rho = Pf_\rho$, $Pf_\rho = \mathcal{L}^r\phi$ and $\Sigma g_\lambda = \Sigma\Sigma_\lambda^{-1}S^*f_\rho$ and $\Sigma\Sigma_\lambda^{-1} = I - \lambda\Sigma_\lambda^{-1}$, we have

$$\Sigma_\lambda^{-1/2}(S^*f_\rho - \Sigma g_\lambda) = \lambda\Sigma_\lambda^{-3/2}S^*f_\rho = \lambda^r\ \lambda^{1-r}\Sigma_\lambda^{-(1-r)}\ \Sigma_\lambda^{-1/2-r}S^*\mathcal{L}^r\ \phi.$$

So, since $\|\Sigma_\lambda^{-1/2-r}S^*\mathcal{L}^r\| \leqslant 1$, as proven in Eq. 89, then

$$\|\mathbb{E}[\zeta_1]\|_{\mathcal{H}} \leqslant \lambda^r\|\lambda^{1-r}\Sigma_\lambda^{-(1-r)}\|\ \|\Sigma_\lambda^{-1/2-r}S^*\mathcal{L}^r\|\ \|\phi\|_{L_2(d\rho_{\mathcal{X}})} \leqslant \lambda^r\|\phi\|_{L_2(d\rho_{\mathcal{X}})} := Z.$$

Morever

$$\begin{aligned}\mathbb{E}[\|\zeta_1\|_{\mathcal{H}}^2] &= \mathbb{E}[\|\Sigma_\lambda^{-1/2}K_{x_1}\|_{\mathcal{H}}^2(y_1 - g_\lambda(x_1))^2] = \mathbb{E}_{x_1}\mathbb{E}_{y_1|x_1}[\|\Sigma_\lambda^{-1/2}K_{x_1}\|_{\mathcal{H}}^2(y_1 - g_\lambda(x_1))^2] \\ &= \mathbb{E}_{x_1}[\|\Sigma_\lambda^{-1/2}K_{x_1}\|_{\mathcal{H}}^2(f_\rho(x_1) - g_\lambda(x_1))^2].\end{aligned}$$

Moreover we have

$$\begin{aligned}\mathbb{E}[\|\zeta_1\|_{\mathcal{H}}^2] &= \mathbb{E}_x[\|\Sigma_\lambda^{-1/2}K_x\|_{\mathcal{H}}^2(f_\rho(x)-g_\lambda(x))^2] \\ &= \mathbb{E}_x[\|\Sigma_\lambda^{-1/2}K_x\|_{\mathcal{H}}^2((f_\rho(x)-(Pf_\rho)(x))+((Pf_\rho)(x)-g_\lambda(x)))^2] \\ &\leqslant 2\mathbb{E}_x[\|\Sigma_\lambda^{-1/2}K_x\|_{\mathcal{H}}^2(f_\rho(x)-(Pf_\rho)(x))^2]+2\mathbb{E}_x[\|\Sigma_\lambda^{-1/2}K_x\|_{\mathcal{H}}^2((Pf_\rho)(x)-g_\lambda(x))^2].\end{aligned}$$

Now since $\mathbb{E}[AB] \leqslant (\operatorname{ess\,sup} A)\mathbb{E}[B]$, for any two random variables A, B, we have

$$\begin{aligned}\mathbb{E}_x[\|\Sigma_\lambda^{-1/2}K_x\|_{\mathcal{H}}^2((Pf_\rho)(x)-g_\lambda(x))^2] &\leqslant \mathcal{N}_\infty(\lambda)\mathbb{E}_x[((Pf_\rho)(x)-g_\lambda(x))^2] \\ &= \mathcal{N}_\infty(\lambda)\|Pf_\rho - Sg_\lambda\|_{L_2(d\rho_{\mathcal{X}})}^2 \\ &\leqslant \kappa_\mu^2 R^{2\mu}\lambda^{-(\mu-2r)},\end{aligned}$$

where in the last step we bounded $\mathcal{N}_\infty(\lambda)$ via Lemma 23 and $\|Pf_\rho - Sg_\lambda\|_{L_2(d\rho_{\mathcal{X}})}^2$, via Lemma. 21 applied with $s=0$. Finally, denoting by $a(x) = \|\Sigma_\lambda^{-1/2}K_x\|_{\mathcal{H}}^2$ and $b(x) = (f_\rho(x)-(Pf_\rho)(x))^2$ and noting that by Markov inequality we have $\mathbb{E}_x[\mathbf{1}_{\{b(x)>t\}}] = \rho_{\mathcal{X}}(\{b(x)>t\}) = \rho_{\mathcal{X}}(\{b(x)^q > t^q\}) \leqslant \mathbb{E}_x[b(x)^q]t^{-q}$, for any $t>0$. Then for any $t>0$ the following holds

$$\begin{aligned}\mathbb{E}_x[a(x)b(x)] &= \mathbb{E}_x[a(x)b(x)\mathbf{1}_{\{b(x)\leqslant t\}}]+\mathbb{E}_x[a(x)b(x)\mathbf{1}_{\{b(x)>t\}}] \\ &\leqslant t\mathbb{E}_x[a(x)]+N_\infty(\lambda)\mathbb{E}_x[b(x)\mathbf{1}_{\{b(x)>t\}}] \\ &\leqslant tN(\lambda)+N_\infty(\lambda)\mathbb{E}_x[b(x)^q]t^{-q}.\end{aligned}$$

By minimizing the quantity above in t, we obtain

$$\begin{aligned}\mathbb{E}_x[\|\Sigma_\lambda^{-1/2}K_x\|_{\mathcal{H}}^2(f_\rho(x)-(Pf_\rho)(x))^2] &\leqslant 2\|f_\rho - Pf_\rho\|_{L^q(X,\rho_{\mathcal{X}})}^{\frac{q}{q+1}}\mathcal{N}(\lambda)^{\frac{q}{q+1}}\mathcal{N}_\infty(\lambda)^{\frac{1}{q+1}} \\ &\leqslant 2\kappa_\mu^2 R^{2\mu}AQ\lambda^{-\frac{q+\mu\alpha}{q\alpha+\alpha}}.\end{aligned}$$

So finally

$$\mathbb{E}[\|\zeta_1\|_{\mathcal{H}}^2] \leqslant 2\kappa_\mu^2 R^{2\mu}\lambda^{-(\mu-2r)} + 4\kappa_\mu^2 R^{2\mu}AQ\lambda^{-\frac{q+\mu\alpha}{q\alpha+\alpha}} := W^2.$$

To conclude the proof, let us obtain the bound in high probability. We need to bound the higher moments of ζ_1. First note that

$$\mathbb{E}[\|\zeta_1 - \mathbb{E}[\zeta_1]\|_{\mathcal{H}}^p] \leqslant \mathbb{E}[\|\zeta_1-\zeta_2\|_{\mathcal{H}}^p] \leqslant 2^{p-1}\mathbb{E}[\|\zeta_1|_{\mathcal{H}}^p + \|\zeta_2\|_{\mathcal{H}}^p] \leqslant 2^p\mathbb{E}[\|\zeta_1|_{\mathcal{H}}^p].$$

Moreover, denoting by $S \subseteq \mathcal{X}$ the support of $\rho_{\mathcal{X}}$ and recalling that y is bounded in $[-M, M]$, the following bound holds almost surely

$$\begin{aligned}\|\zeta_1\| &\leqslant \sup_{x\in S}\|\Sigma_\lambda^{-1/2}K_x\|(M+|g_\lambda(x)|) \leqslant (\sup_{x\in S}\|\Sigma_\lambda^{-1/2}K_x\|)(M+\sup_{x\in S}|g_\lambda(x)|) \\ &\leqslant \kappa_\mu R^\mu\lambda^{-\mu/2}(M+\kappa_\mu R^\mu\|\Sigma^{1/2(1-\mu)}g_\lambda\|_{\mathcal{H}}).\end{aligned}$$

where in the last step we applied Lemma 23 and Lemma 22. In particular, by definition of g_λ, the fact that $S^*f_\rho = S^*Pf_\rho$, that $Pf_\rho = \mathcal{L}^r\phi$ and that $\|\Sigma_\lambda^{-(1/2+r)}S^*\mathcal{L}^r\| \leqslant 1$ as proven in Eq. 89, we have

$$\begin{aligned}\|\Sigma^{1/2(1-\mu)}g_\lambda\|_{\mathcal{H}} &= \|\Sigma^{1/2(1-\mu)}\Sigma_\lambda^{-1}S^*\mathcal{L}^r\phi\|_{\mathcal{H}} \\ &\leqslant \|\Sigma^{1/2(1-\mu)}\Sigma^{-1/2(1-\mu)}\|\|\Sigma_\lambda^{-(\mu/2-r)}\|\|\Sigma_\lambda^{-(1/2+r)}S^*\mathcal{L}^r\|\|\phi\|_{L_2(d\rho_{\mathcal{X}})} \\ &\leqslant \|\Sigma_\lambda^{r-\mu/2}\|\|\phi\|_{L_2(d\rho_{\mathcal{X}})}.\end{aligned}$$

Finally note that if $r \leqslant \mu/2$ then $\|\Sigma_\lambda^{r-\mu/2}\| \leqslant \lambda^{-(\mu/2-r)}$, if $r \geqslant \mu/2$ then

$$\|\Sigma_\lambda^{r-\mu/2}\| = (\|C\|+\lambda)^{r-\mu/2} \leqslant (2\|C\|)^{r-\mu/2} \leqslant (2R)^{2r-\mu}.$$

So in particular

$$\|\Sigma_\lambda^{r-\mu/2}\| \leqslant (2R)^{2r-\mu} + \lambda^{-(\mu/2-r)}.$$

Then the following holds almost surely

$$\|\zeta_1\| \leqslant (\kappa_\mu R^\mu M + \kappa_\mu^2 R^{2\mu}(2R)^{2r-\mu}\|\phi\|_{L_2(d\rho_\mathcal{X})})\lambda^{-\mu/2} + \kappa_\mu^2 R^{2\mu}\|\phi\|_{L_2(d\rho_\mathcal{X})}\lambda^{r-\mu} := V.$$

So finally

$$\mathbb{E}[\|\zeta_1 - \mathbb{E}[\zeta_1]\|_\mathcal{H}^p] \leqslant 2^p\mathbb{E}[\|\zeta_1\|_\mathcal{H}^p] \leqslant \frac{p!}{2}(2V)^{p-2}(4W^2).$$

By applying Pinelis inequality, the following holds with probability $1-\delta$

$$\|\frac{1}{n}\sum_{i=1}^n(\zeta_i - \mathbb{E}[\zeta_i])\|_\mathcal{H} \leqslant \frac{4V\log\frac{2}{\delta}}{n} + \sqrt{\frac{8W\log\frac{2}{\delta}}{n}}.$$

So with the same probability

$$\|\frac{1}{n}\sum_{i=1}^n\zeta_i\|_\mathcal{H} \leqslant \|\frac{1}{n}\sum_{i=1}^n(\zeta_i - \mathbb{E}[\zeta_i])\|_\mathcal{H} + \|\mathbb{E}[\zeta_1]\|_\mathcal{H} \leqslant Z + \frac{4V\log\frac{2}{\delta}}{n} + \sqrt{\frac{8W\log\frac{2}{\delta}}{n}}.$$

■

Lemma 26

Let $\lambda > 0$, $n \in \mathbb{N}$ and $s \in (0, 1/2]$. Let $\delta \in (0,1]$. Under Assumption (A8), (A9), (A10) (see Rem. 7), (A11), when

$$n \geqslant 11(1 + \kappa_\mu^2 R^{2\mu}\lambda^{-\mu})\log\frac{16R^2}{\lambda\delta},$$

then the following holds with probability $1-\delta$,

$$\|\mathcal{L}^{-s}S(\widehat{g}_\lambda - g_\lambda)\|_{L_2(d\rho_\mathcal{X})} \leq c_0\lambda^{r-s} + \frac{(c_1\lambda^{-\frac{\mu}{2}-s} + c_2\lambda^{r-\mu-s})\log\frac{4}{\delta}}{n} + \sqrt{\frac{(c_3\lambda^{-(\mu+2s-2r)} + c_4\lambda^{-\frac{q+\mu\alpha}{q\alpha+\alpha}-2s})\log\frac{4}{\delta}}{n}}.$$

with $c_0 = 7c_q\|\phi\|_{L_2(d\rho_\mathcal{X})}$, $c_1 = 16c_q(\kappa_\mu R^\mu M + \kappa_\mu^2 R^{2\mu}(2R)^{2r-\mu}\|\phi\|_{L_2(d\rho_\mathcal{X})})$, $c_2 = 16c_q\kappa_\mu^2 R^{2\mu}\|\phi\|_{L_2(d\rho_\mathcal{X})}$, $c_3 = 64\kappa_\mu^2 R^{2\mu}c_q^2$, $c_4 = 128\kappa_\mu^2 R^{2\mu}AQc_q^2$.

Proof : Let $\tau = \delta/2$, the result is obtained by combining Lemma 20, with Lemma 25 with probability τ, and Lemma 24, with probability τ and then taking the intersection bound of the two events. ■

Corollary 5

Let $\lambda > 0$, $n \in \mathbb{N}$ and $s \in (0, 1/2]$. Let $\delta \in (0,1]$. Under the assumptions of Lemma 26, when

$$n \geqslant 11(1 + \kappa_\mu^2 R^{2\mu}\lambda^{-\mu})\log\frac{16R^2}{\lambda\delta},$$

then the following holds with probability $1-\delta$,

$$\|\mathcal{L}^{-s}S\widehat{g}_\lambda\|_{L_2(d\rho_\mathcal{X})} \leq R^{2r-2s} + (1+c_0)\lambda^{r-s} + \frac{(c_1\lambda^{-\frac{\mu}{2}-s} + c_2\lambda^{r-\mu-s})\log\frac{4}{\delta}}{n} + \sqrt{\frac{(c_3\lambda^{-(\mu+2s-2r)} + c_4\lambda^{-\frac{q+\mu\alpha}{q\alpha+\alpha}-2s})\log\frac{4}{\delta}}{n}} +$$

with the same constants $c_0, \ldots, c_4$ as in Lemma 26.

Proof : First note that

$$\|\mathcal{L}^{-s}S\widehat{g}_\lambda\|_{L_2(d\rho_\mathcal{X})} \leqslant \|\mathcal{L}^{-s}S(\widehat{g}_\lambda - g_\lambda)\|_{L_2(d\rho_\mathcal{X})} + \|\mathcal{L}^{-s}Sg_\lambda\|_{L_2(d\rho_\mathcal{X})}.$$

The first term on the right hand side is controlled by Lemma 26, for the second, by using the definition of g_λ and Asm. (A10) (see Rem. 7), we have

$$\begin{aligned}\|\mathcal{L}^{-s} S g_\lambda\|_{L_2(d\rho_\mathcal{X})} &\leqslant \|\mathcal{L}^{-s} S \Sigma_\lambda^{-1/2+s}\| \|\Sigma_\lambda^{-(s-r)}\| \|\Sigma_\lambda^{-1/2-r} S^* \mathcal{L}^r\| \|\phi\|_{L_2(d\rho_\mathcal{X})} \\ &\leqslant \|\Sigma_\lambda^{r-s}\| \|\phi\|_{L_2(d\rho_\mathcal{X})},\end{aligned}$$

where $\|\Sigma_\lambda^{-1/2-r} S^* \mathcal{L}^r\| \leqslant 1$ by Eq. 89 and analogously $\|\mathcal{L}^{-s} S \Sigma_\lambda^{-1/2+s}\| \leqslant 1$. Note that if $s \geqslant r$ then $\|\Sigma_\lambda^{r-s}\| \leqslant \lambda^{-(s-r)}$. If $s < r$, we have

$$\|\Sigma_\lambda^{r-s}\| = (\|\Sigma\| + \lambda)^{r-s} \leqslant \|C\|^{r-s} + \lambda^{r-s} \leqslant R^{2r-2s} + \lambda^{r-s}.$$

So finally $\|\Sigma_\lambda^{r-s}\| \leqslant R^{2r-2s} + \lambda^{r-s}$. ■

Corollary 6

Let $\lambda > 0$, $n \in \mathbb{N}$ and $s \in (0, 1/2]$. Let $\delta \in (0, 1]$. Under Assumption (A8), (A9), (A10) (see Rem. 7), (A11), when

$$n \geqslant 11(1 + \kappa_\mu^2 R^{2\mu} \lambda^{-\mu}) \log \frac{16 R^2}{\lambda \delta},$$

then the following holds with probability $1 - \delta$,

$$\begin{aligned}\sup_{x \in \mathcal{X}} |\widehat{g}_\lambda(x)| \;\leq\; & \kappa_\mu R^\mu R^{2r-2s} \;+\; \kappa_\mu R^\mu (1 + c_0) \lambda^{r-\mu/2} \;+\; \kappa_\mu R^\mu \frac{(c_1 \lambda^{-\mu} + c_2 \lambda^{r-3/2\mu}) \log \frac{4}{\delta}}{n} \\ & + \kappa_\mu R^\mu \sqrt{\frac{(c_3 \lambda^{-(2\mu-2r)} + \kappa_\mu R^\mu c_4 \lambda^{-\frac{q+\mu\alpha}{q\alpha+\alpha} - \mu}) \log \frac{4}{\delta}}{n}}.\end{aligned}$$

with the same constants $c_0, \ldots, c_4$ in Lemma 26.

Proof : The proof is obtained by applying Lemma 22 on $\widehat{g}_\lambda$ and then Corollary 5. ■

B.4.4. Main Result

Theorem 10

Let $\lambda > 0$, $n \in \mathbb{N}$ and $s \in (0, \min(r, 1/2)]$. Under Assumption (A8), (A9), (A10) (see Rem. 7), (A11), when

$$n \geqslant 11(1 + \kappa_\mu^2 R^{2\mu} \lambda^{-\mu}) \log \frac{c_0}{\lambda^{3+4r-4s}},$$

then

$$\mathbb{E}[\|\mathcal{L}^{-s}(S\widehat{g}_\lambda - P f_\rho)\|^2_{L_2(d\rho_\mathcal{X})}] \leqslant c_1 \frac{\lambda^{-(\mu+2s-2r)}}{n} + c_2 \frac{\lambda^{-\frac{q+\mu\alpha}{q\alpha+\alpha} - 2s}}{n} + c_3 \lambda^{2r-2s},$$

where $m_4 = M^4$, $c_0 = 32R^{4-4s} m_4 + 32R^{8-8r-8s} \|\phi\|^4_{L_2(d\rho_\mathcal{X})}$, $c_1 = 16 c_q^2 \kappa_\mu^2 R^{2\mu}$, $c_2 = 32 c_q^2 \kappa_\mu^2 R^{2\mu} AQ$, $c_3 = 3 + 8 c_q^2 \|\phi\|^2_{L_2(d\rho_\mathcal{X})}$.

Proof : Denote by $R(\widehat{g}_\lambda)$, the expected risk $R(\widehat{g}_\lambda) = \mathcal{E}(\widehat{g}_\lambda) - \inf_{g \in \mathcal{H}} \mathcal{E}(g)$. First, note that by Prop. 10, we have

$$R_s(\widehat{g}_\lambda) = \|\mathcal{L}^{-s}(S\widehat{g}_\lambda - P f_\rho)\|^2_{L_2(d\rho_\mathcal{X})}.$$

Denote by E the event such that β as defined in Thm. 9, satisfies $\beta \leqslant 2$. Then we have

$$\mathbb{E}[R_s(\widehat{g}_\lambda)] = \mathbb{E}[R_s(\widehat{g}_\lambda) \mathbf{1}_E] + \mathbb{E}[R(\widehat{g}_\lambda) \mathbf{1}_{E^c}].$$

For the first term, by Thm. 9 and Lemma 25, we have

$$\begin{aligned}\mathbb{E}[R_s(\widehat{g}_\lambda)\mathbf{1}_E] &\leqslant \mathbb{E}[\Big(2\lambda^{-2s}\beta^4 c_q^2\|\Sigma_\lambda^{-1/2}(\widehat{\mathsf{S}}_n^*\widehat{y}-\widehat{\Sigma}_n g_\lambda)\|_{\mathcal{H}}^2 \\ &\qquad + 2\left(1+\beta^2 2c_q^2\|\phi\|_{L_2(d\rho_{\mathcal{X}})}^2\right)\lambda^{2r-2s}\Big)\mathbf{1}_E] \\ &\leqslant 8\lambda^{-2s}c_q^2\mathbb{E}[\|\Sigma_\lambda^{-1/2}(\widehat{\mathsf{S}}_n^*\widehat{y}-\widehat{\Sigma}_n g_\lambda)\|_{\mathcal{H}}^2] + 2\left(1+4c_q^2\|\phi\|_{L_2(d\rho_{\mathcal{X}})}^2\right)\lambda^{2r-2s} \\ &\leqslant \frac{16c_q^2\kappa_\mu^2 R^{2\mu}\lambda^{-\mu+2r-2s}}{n} + \frac{32c_q^2\kappa_\mu^2 R^{2\mu}AQ\lambda^{-\frac{q+\mu\alpha}{q\alpha+\alpha}-2s}}{n} + \left(2+8c_q^2\|\phi\|_{L_2(d\rho_{\mathcal{X}})}^2\right)\lambda^{2r-2s}.\end{aligned}$$

For the second term, since $\widehat{\Sigma}_{n\lambda}^{1/2}q_\lambda(\widehat{\Sigma}_n)\widehat{\Sigma}_{n\lambda}^{1/2} = \widehat{\Sigma}_{n\lambda}q_\lambda(\widehat{\Sigma}_n) \leqslant \sup_{\sigma>0}(\sigma+\lambda)q_\lambda(\sigma) \leqslant c_q$ by definition of filters, and that $Pf_\rho = L^r\phi$, we have

$$\begin{aligned}R_s(\widehat{g}_\lambda)^{1/2} &\leqslant \|\mathcal{L}^{-s}S\widehat{g}_\lambda\|_{L_2(d\rho_{\mathcal{X}})} + \|\mathcal{L}^{-s}Pf_\rho\|_{L_2(d\rho_{\mathcal{X}})} \\ &\leqslant \|\mathcal{L}^{-s}S\|\|\widehat{\Sigma}_{n\lambda}^{-1/2}\|\|\widehat{\Sigma}_{n\lambda}^{1/2}q_\lambda(\widehat{\Sigma}_n)\widehat{\Sigma}_{n\lambda}^{1/2}\|\|\widehat{\Sigma}_{n\lambda}^{-1/2}\widehat{\mathsf{S}}_n^*\|\|\widehat{y}\| + \|\mathcal{L}^{-s}\mathcal{L}^r\|\|\phi\|_{L_2(d\rho_{\mathcal{X}})} \\ &\leqslant R^{1/2-s}\lambda^{-1/2}\|\widehat{y}\| + R^{2r-2s}\|\phi\|_{L_2(d\rho_{\mathcal{X}})} \\ &\leqslant \lambda^{-1/2}(R^{1/2-s}(n^{-1}\sum_{i=1}^n y_i) + R^{1+2r-2s}\|\phi\|_{L_2(d\rho_{\mathcal{X}})}),\end{aligned}$$

where the last step is due to the fact that $1 \leqslant \lambda^{-1/2}\|\mathcal{L}\|^{1/2}$ since λ satisfies $0 < \lambda \leqslant \|\Sigma\| = \|\mathcal{L}\| \leqslant R^2$. Denote with δ the quantity $\delta = \lambda^{2+4r-4s}/c_0$. Since $\mathbb{E}[\mathbf{1}_E^c]$ corresponds to the probability of the event E^c, and, by Lemma 24, we have that E^c holds with probability at most δ since $n \geqslant 11(1+\kappa_\mu^2 R^{2\mu}\lambda^{-\mu})\log\frac{8R^2}{\lambda\delta}$, then we have that

$$\begin{aligned}\mathbb{E}[R(\widehat{g}_\lambda)\mathbf{1}_{E^c}] &\leqslant \mathbb{E}[\|S\widehat{g}_\lambda\|_{L_2(d\rho_{\mathcal{X}})}^2\mathbf{1}_{E^c}] \leqslant \sqrt{\mathbb{E}[\|S\widehat{g}_\lambda\|_{L_2(d\rho_{\mathcal{X}})}^4]}\sqrt{\mathbb{E}[\mathbf{1}_{E^c}]} \\ &\leqslant \sqrt{\frac{4R^{2-4s}n^{-2}(\sum_{i,j=1}^n\mathbb{E}[y_i^2y_j^2]) + 4R^{4-8r-8s}\|\phi\|_{L_2(d\rho_{\mathcal{X}})}^4}{\lambda^2}}\sqrt{\delta} \\ &\leqslant \frac{\sqrt{\delta}}{\lambda}\sqrt{4R^{2-4s}m_4 + 4R^{4-8r-8s}\|\phi\|_{L_2(d\rho_{\mathcal{X}})}^4} \\ &= \frac{\sqrt{\delta c_0/(8R^2)}}{\lambda} \leqslant \lambda^{2r-2s}.\end{aligned}$$

■

Corollary 7

Let $\lambda > 0$ and $n \in \mathbb{N}$ and $s = 0$. Under Assumption (A8), (A9), (A10) (see Rem. 7), (A11), when

$$\lambda \;=\; B_1\begin{cases} n^{-\alpha/\left(2r\alpha+1+\frac{\mu\alpha-1}{q+1}\right)} & 2r\alpha+1+\frac{\mu\alpha-1}{q+1} > \mu\alpha \\ n^{-1/\mu}\,(\log B_2 n)^{\frac{1}{\mu}} & 2r\alpha+1+\frac{\mu\alpha-1}{q+1} \leqslant \mu\alpha. \end{cases} \tag{90}$$

then,

$$\mathbb{E}\,\mathcal{E}(\widehat{g}_\lambda) - \inf_{g\in\mathcal{H}}\mathcal{E}(g) \leqslant B_3\begin{cases} n^{-2r\alpha/\left(2r\alpha+1+\frac{\mu\alpha-1}{q+1}\right)} & 2r\alpha+1+\frac{\mu\alpha-1}{q+1} > \mu\alpha \\ n^{-2r/\mu} & 2r\alpha+1+\frac{\mu\alpha-1}{q+1} \leqslant \mu\alpha \end{cases} \tag{91}$$

where $B_2 = 3 \vee (32R^6m_4)^{\frac{\mu}{3+4r}}B_1^{-\mu}$ and B_1 defined explicitly in the proof.

Proof : The proof of this corollary is a direct application of Thm. 10. In the rest of the proof we find the constants to guarantee that the condition relating n, λ in the theorem is always satisfied. Indeed to guarantee the applicability of Thm. 10, we need to be sure that $n \geqslant 11(1+\kappa_\mu^2R^{2\mu}\lambda^{-\mu})\log\frac{32R^6m_4}{\lambda^{3+4r}}$. This is satisfied when both the following conditions hold $n \geqslant 22\log\frac{32R^6m_4}{\lambda^{3+4r}}$ and $n \geqslant 2\kappa_\mu^2R^{2\mu}\lambda^{-\mu}\log\frac{32R^6m_4}{\lambda^{3+4r}}$. To study the last two conditions, we recall that for $A, B, s, q > 0$ we have that $An^{-s}\log(Bn^q)$ satisfy

$$An^{-s}\log(Bn^q) = \frac{qAB^{s/q}}{s}\frac{\log B^{s/q}n^s}{B^{s/q}n^s} \leqslant \frac{qAB^{s/q}}{es},$$

for any $n > 0$, since $\frac{\log x}{x} \leqslant \frac{1}{e}$ for any $x > 0$. Now we define explicitly B_1, let $\tau = \alpha / \left(2r\alpha + 1 + \frac{\mu\alpha - 1}{q+1}\right)$, we have

$$B_1 = \left(\frac{22(3+4r)}{e\mu}(32R^6 m_4)^{\frac{\mu}{3+4r}}\right)^{\frac{1}{\mu}} \vee \tag{92}$$

$$\vee \begin{cases} \left(\frac{2M(3+4r)}{e(1/\tau-\mu)}(32R^6 m_4)^{\frac{1/\tau-\mu}{3+4r}}\right)^{\tau} & 2r\alpha + 1 + \frac{\mu\alpha-1}{q+1} > \mu\alpha \\ \left(\frac{2M(3+4r)}{\mu}\right)^{\frac{1}{\mu}} & 2r\alpha + 1 + \frac{\mu\alpha-1}{q+1} \leqslant \mu\alpha \end{cases}. \tag{93}$$

For the first condition, we use the fact that λ is always larger than $B_1 n^{-1/\mu}$, so we have

$$\frac{22}{n}\log\frac{32R^6 m_4}{\lambda^{3+4r}} \leqslant \frac{22}{n}\log\frac{32R^6 m_4 n^{(3+4r)/\mu}}{B_1^{3+4r}} \leqslant \frac{22(3+4r)(32R^6 m_4)^{\mu/(3+4r)}}{e\mu B_1^{\mu}} \leqslant 1.$$

For the second inequality, when $2r\alpha + 1 + \frac{\mu\alpha-1}{q+1} \geqslant \mu\alpha$, we have $\lambda = B_1 n^{-\tau}$, so

$$\begin{aligned}\frac{2\kappa_\mu^2 R^{2\mu}}{n}\lambda^{-\mu}\log\frac{32R^6 m_4}{\lambda^{3+4r}} &\leqslant \frac{2\kappa_\mu^2 R^{2\mu}}{B_1^{\mu} n^{1-\mu\tau}}\log\frac{32R^6 m_4 n^{(3+4r)\tau}}{B_1^{3+4r}} \\ &\leqslant \frac{2\kappa_\mu^2 R^{2\mu}(3+4r)\tau}{e(1-\mu\tau)}\frac{(32R^6 m_4)^{\frac{1/\tau-\mu}{3+4r}}}{B_1^{1/\tau}} \leqslant 1.\end{aligned}$$

Finally, when $2r\alpha + 1 + \frac{\mu\alpha-1}{q+1} \geqslant \mu\alpha$, we have $\lambda = B_1 n^{-1/\mu}(\log B_2 n)^{1/\mu}$. So since $\log(B_2 n) > 1$, we have

$$\frac{2\kappa_\mu^2 R^{2\mu}}{n}\log\frac{32R^6 m_4}{\lambda^{3+4r}} \leqslant \frac{2\kappa_\mu^2 R^{2\mu}}{B_1^{\mu}}\frac{\log\frac{32R^6 m_4 n^{(3+4r)/\mu}}{B_1^{3+4r}}}{\log(B_2 n)} = \frac{2\kappa_\mu^2 R^{2\mu}(3+4r)}{\mu B_1^{\mu}}\frac{\log\frac{(32R^6 m_4)^{\mu/(3+4r)} n}{B_1^{\mu}}}{\log(B_2 n)} \leqslant 1.$$

So by selecting λ as in Eq. 90, we guarantee that the condition required by Thm. 10 is satisfied.

Finally the constant B_3 is obtained by

$$B_3 = c_1 \max(1, w)^{-(\mu+2s-2r)} + c_2 \max(1, w)^{-\frac{q+\mu\alpha}{q\alpha+\alpha}-2s} + c_3 \max(1, w)^{2r-2s},$$

with $w = B_1 \log(1 + B_2)$ and c_1, c_2, c_3 as in Thm. 10. ■

B.5. Experiments with different sampling

We present here the results for two different types of sampling, which seem to be more stable, perform better and are widely used in practice :

Without replacement (Figure 16): for which we select randomly the data points but never use two times over the same point in one epoch.

Cycles (Figure 17): for which we pick successively the data points in the same order.

⋆

⋆ ⋆

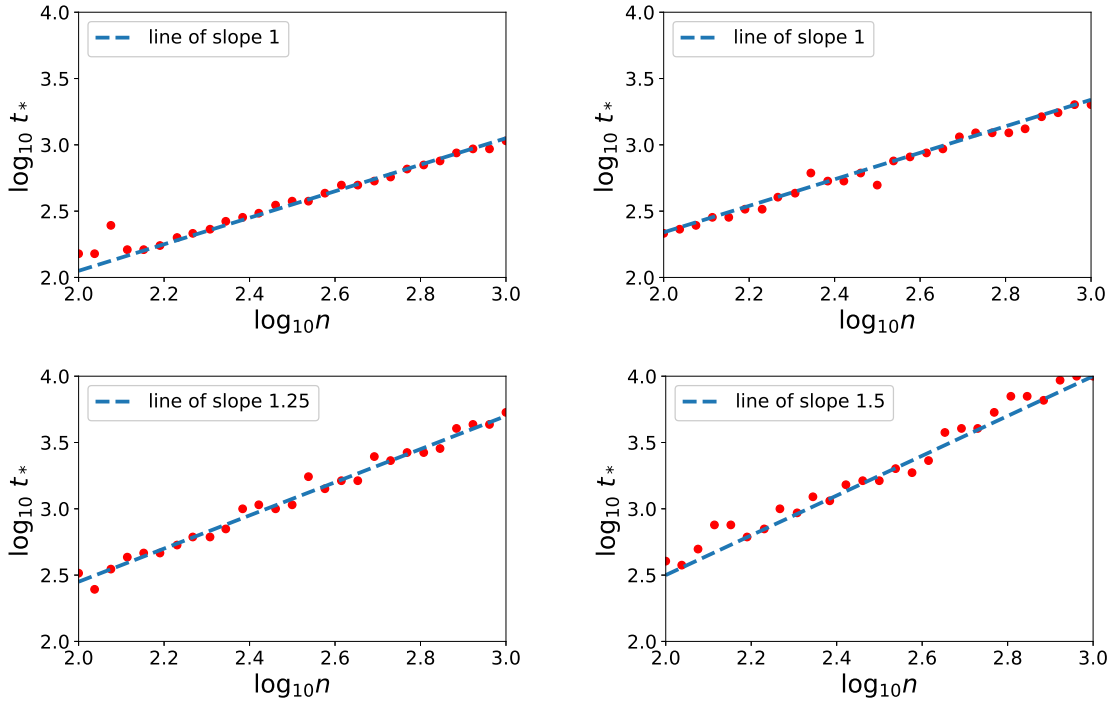

Figure 16: The sampling is performed by **cycling over the data** The four plots represent each a different configuration on the (α, r) plan represented in Figure 13, for $r = 1/(2\alpha)$. **Top left** ($\alpha = 1.5$) and **right** ($\alpha = 2$) are two easy problems (Top right is the limiting case where $r = \frac{\alpha-1}{2\alpha}$) for which one pass over the data is optimal. **Bottom left** ($\alpha = 2.5$) and **right** ($\alpha = 3$) are two hard problems for which an increasing number of passes is recquired. The blue dotted line are the slopes predicted by the theoretical result in Theorem 8.

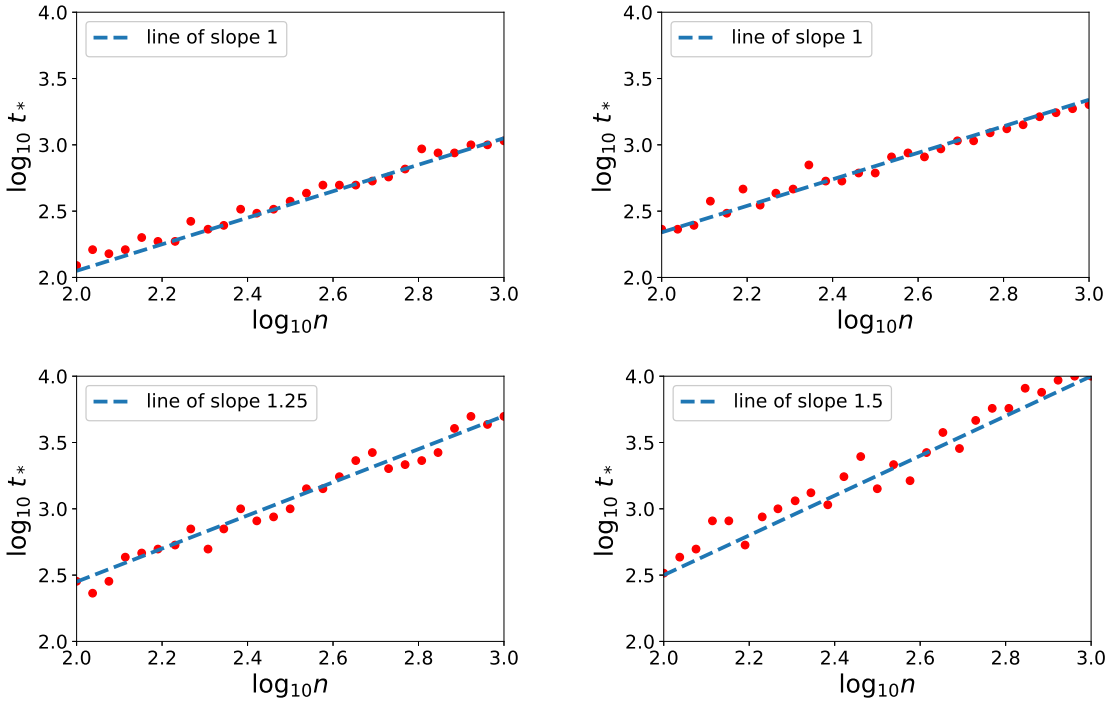

Figure 17: The sampling is performed **without replacement**. The four plots represent each a different configuration on the (α, r) plan represented in Figure 13, for $r = 1/(2\alpha)$. **Top left** ($\alpha = 1.5$) and **right** ($\alpha = 2$) are two easy problems (Top right is the limiting case where $r = \frac{\alpha-1}{2\alpha}$) for which one pass over the data is optimal. **Bottom left** ($\alpha = 2.5$) and **right** ($\alpha = 3$) are two hard problems for which an increasing number of passes is recquired. The blue dotted line are the slopes predicted by the theoretical result in Theorem 8.

Part III

Statistical estimation of Laplacian

We divide this part into two contributions for the statistical estimation of Laplacian.

The Section 1 together with its Appendix 1.6 shows how it is possible to estimate the Poincaré constant of a distribution through samples of it. Then, we explain how to use this estimation to design an algorithm looking for *reaction coordinates* associated to the overdamped Langevin dynamics associated with the measure. As explain in the introduction these *reaction coordinates* are the cornerstone of accelerating dynamics in this context. This section is based on our work, **Statistical Estimation of the Poincaré constant and Application to Sampling Multimodal Distributions**, L. Pillaud-Vivien, F. Bach, T. Lelievre, A. Rudi, G. Stoltz, published in the *International Conference on Artificial Intelligence and Statistics* in 2020.

The following Section 2 is the natural continuation of the work presented in the previous Section 1. However, besides being more mature, the focus of this work is quite different from the previous one: while previously we leveraged the estimation of the first eigenvalue to find reduced order models in physical systems, we will focus in this work on the estimation of the whole spectrum of the diffusion operator and try to be more precise on its convergence properties. Finally, note that even if the story behind it is almost complete, this work is still unfinished. An explicit discussion at the end, in subsection 2.4.3, details where exactly it stands.

Contents

1 Statistical estimation of the Poincaré constant and application to sampling multimodal distributions **134**
1.1 Introduction . . . 134
1.2 Poincaré Inequalities . . . 135
1.3 Statistical Estimation of the Poincaré Constant . . . 137
1.4 Learning a Reaction Coordinate . . . 140
1.5 Numerical experiments . . . 142
1.6 Conclusion and Perspectives . . . 144

C Appendix of Statistical estimation of the Poincaré constant and application to sampling multimodal distributions **145**
C.1 Proofs of Proposition 11 and 12 . . . 145
C.2 Analysis of the bias: convergence of the regularized Poincaré constant to the true one . . . 146
C.3 Technical inequalities . . . 150
C.4 Calculation of the bias in the Gaussian case . . . 154

2 Statistical estimation of Laplacian and application to dimensionality reduction **164**
2.1 Introduction . . . 164
2.2 Diffusion operator . . . 165
2.3 Approximation of the diffusion operator in the RKHS . . . 166
2.4 Analysis of the estimator . . . 171
2.5 Conclusion and further thoughts . . . 174

1. Statistical estimation of the Poincaré constant and application to sampling multimodal distributions

1.1. Introduction

Sampling is a cornerstone of probabilistic modelling, in particular in the Bayesian framework where statistical inference is rephrased as the estimation of the posterior distribution given the data [Rob07, Mur12]: the representation of this distribution through samples is both flexible, as most interesting quantities can be computed from them (e.g., various moments or quantiles), and practical, as there are many sampling algorithms available depending on the various structural assumptions made on the model. Beyond one-dimensional distributions, a large class of these algorithms are iterative and update samples with a Markov chain which eventually converges to the desired distribution, such as Gibbs sampling or Metropolis-Hastings (or more general Markov chain Monte-Carlo algorithms [GL06, GRS95, DM17]) which are adapted to most situations, or Langevin's algorithm [DM17, RRT17, WT11, MHB17, LS16, BGL14], which is adapted to sampling from densities in $\mathbb{R}^d$.

While these sampling algorithms are provably converging in general settings when the number of iterations tends to infinity, obtaining good explicit convergence rates has been a central focus of study, and is often related to the mixing time of the underlying Markov chain [MT12]. In particular, for sampling from positive densities in $\mathbb{R}^d$, the Markov chain used in Langevin's algorithm can classically be related to a diffusion process, thus allowing links with other communities such as molecular dynamics [LS16]. The main objective of molecular dynamics is to infer macroscopic properties of matter from atomistic models via averages with respect to probability measures dictated by the principles of statistical physics. Hence, it relies on high dimensional and highly multimodal probabilistic models.

When the density is log-concave, sampling can be done in polynomial time with respect to the dimension [MCJ+18, DRVZ17, DM17]. However, in general, sampling with generic algorithms does not scale well with respect to the dimension. Furthermore, the multimodality of the objective measure can trap the iterates of the algorithm in some regions for long durations: this phenomenon is known as metastability. To accelerate the sampling procedure, a common technique in molecular dynamics is to resort to importance sampling strategies where the target probability measure is biased using the image law of the process for some low-dimensional function, known as "reaction coordinate" or "collective variable". Biasing by this low-dimensional probability measure can improve the convergence rate of the algorithms by several orders of magnitude [LRS08, Lel13]. Usually, in molecular dynamics, the choice of a good reaction coordinate is based on physical intuition on the model but this approach has limitations, particularly in the Bayesian context [CLS12]. There have been efforts to numerically find these reaction coordinates [Gke19]. Computations of spectral gaps by approximating directly the diffusion operator work well in low-dimensional settings but scale poorly with the dimension. One popular method is based on diffusion maps [CL06, CBLK06, RZMC11], for which reaction coordinates are built by approximating the entire infinite-dimensional diffusion operator and selecting its first eigenvectors.

In order to assess or find a reaction coordinate, it is necessary to understand the convergence rate of diffusion processes. We first introduce in Section 1.2 Poincaré inequalities and Poincaré constants that control the convergence rate of diffusions to their equilibrium. We then derive in Section 1.3 a kernel method to estimate it and optimize over it to find good low dimensional representation of the data for sampling in Section 1.4. Finally we present in Section 1.5 synthetic examples for which our procedure is able to find good reaction coordinates.

Contributions. In this paper, we make the following contributions:

- We show both theoretically and experimentally that, given sufficiently many samples of a measure, we can estimate its Poincaré constant and thus quantify the rate of convergence of Langevin dynamics.
- By finding projections whose marginal laws have the largest Poincaré constant, we derive an algorithm that captures a low dimensional representation of the data. This knowledge of "difficult to sample directions" can be then used to accelerate dynamics to their equilibrium measure.

1.2. Poincaré Inequalities

1.2.1. Definition

We introduce in this part the main object of this paper which is the Poincaré inequality [BGL14]. Let us consider a probability measure $d\mu$ on $\mathbb{R}^d$ which has a density with respect to the Lebesgue measure. Consider $H^1(\mu)$ the space of functions in $L^2(\mu)$ (i.e., which are square integrable) that also have all their first order derivatives in L^2, that is, $H^1(\mu) = \{f \in L^2(\mu),\ \int_{\mathbb{R}^d} f^2 d\mu + \int_{\mathbb{R}^d} \|\nabla f\|^2 d\mu < \infty\}$.

Definition 3 (*Poincaré inequality and Poincaré constant*)

*The Poincaré constant of the probability measure $d\mu$ is the smallest constant $\mathcal{P}_\mu$ such that for all $f \in H^1(\mu)$ the following Poincaré inequality **(PI)** holds:*

$$\int_{\mathbb{R}^d} f(x)^2 d\mu(x) - \left(\int_{\mathbb{R}^d} f(x) d\mu(x)\right)^2 \leqslant \mathcal{P}_\mu \int_{\mathbb{R}^d} \|\nabla f(x)\|^2 d\mu(x). \tag{94}$$

In Definition 3 we took the largest possible and the most natural functional space $H^1(\mu)$ for which all terms make sense, but Poincaré inequalities can be equivalently defined for subspaces of test functions $\mathcal{H}$ which are dense in $H^1(\mu)$. This will be the case when we derive the estimator of the Poincaré constant in Section 1.3.

Remark 8 (*A probabilistic formulation of the Poincaré inequality.*)

*Let X be a random variable distributed according to the probability measure $d\mu$. **(PI)** can be reformulated as: for all $f \in H^1(\mu)$,*

$$\mathrm{Var}_\mu\left(f(X)\right) \leqslant \mathcal{P}_\mu\, \mathbb{E}_\mu\left[\|\nabla f(X)\|^2\right]. \tag{95}$$

Poincaré inequalities are hence a way to bound the variance from above by the so-called Dirichlet energy $\mathbb{E}\left[\|\nabla f(X)\|^2\right]$ (see [BGL14]).

1.2.2. Consequences of (PI): convergence rate of diffusions

Poincaré inequalities are ubiquitous in various domains such as probability, statistics or partial differential equations (PDEs). For example, in PDEs they play a crucial role for showing the existence of solutions of Poisson equations or Sobolev embeddings [GT01], and they lead in statistics to concentration of measure results [Goz10]. In this paper, the property that we are the most interested in is the convergence rate of diffusions to their stationary measure $d\mu$. In this section, we consider a very general class of measures: $d\mu(x) = \mathrm{e}^{-V(x)}dx$ (called Gibbs measures with potential V), which allows for a clearer explanation. Note that all measures admitting a positive density can be written like this and are typical in Bayesian machine

learning [Rob07] or molecular dynamics [LS16]. Yet, the formalism of this section can be extended to more general cases [BGL14].

Let us consider the overdamped Langevin diffusion in $\mathbb{R}^d$, that is the solution of the following stochastic differential equation (SDE):

$$\mathrm{d}X_t = -\nabla V(X_t)\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t, \tag{96}$$

where $(B_t)_{t\geqslant 0}$ is a d-dimensional Brownian motion. It is well-known [BGL14] that the law of $(X_t)_{t\geqslant 0}$ converges to the Gibbs measure $d\mu$ and that the Poincaré constant controls the rate of convergence to equilibrium in $L^2(\mu)$. Let us denote by $P_t(f)$ the Markovian semi-group associated with the Langevin diffusion $(X_t)_{t\geqslant 0}$. It is defined in the following way: $P_t(f)(x) = \mathbb{E}[f(X_t)|X_0 = x]$. This semi-group satisfies the dynamics

$$\frac{d}{dt}P_t(f) = \mathcal{L}P_t(f),$$

where $\mathcal{L}\phi = \Delta^L\phi - \nabla V\cdot\nabla\phi$ is a differential operator called the infinitesimal generator of the Langevin diffusion (96) (Δ^L denotes the standard Laplacian on $\mathbb{R}^d$). Note that by integration by parts, the semi-group $(P_t)_{t\geqslant 0}$ is reversible with respect to $d\mu$, that is: $-\int f(\mathcal{L}g)\,d\mu = \int \nabla f\cdot\nabla g\,d\mu = -\int(\mathcal{L}f)g\,d\mu$. Let us now state a standard convergence theorem (see e.g. [BGL14, Theorem 2.4.5]), which proves that $\mathcal{P}_\mu$ is the characteristic time of the exponential convergence of the diffusion to equilibrium in $L^2(\mu)$.

Theorem 11 (*Poincaré and convergence to equilibrium*)

With the notation above, the following statements are equivalent:

(i) *μ satisfies a Poincaré inequality with constant $\mathcal{P}_\mu$;*

(ii) *For all f smooth and compactly supported, $\mathrm{Var}_\mu(P_t(f)) \leqslant \mathrm{e}^{-2t/\mathcal{P}_\mu}\mathrm{Var}_\mu(f)$ for all $t\geqslant 0$.*

Proof : The proof is standard. Note that upon replacing f by $f - \int f d\mu$, one can assume that $\int f d\mu = 0$. Then, for all $t\geqslant 0$,

$$\frac{d}{dt}\mathrm{Var}_\mu(P_t(f)) = \frac{d}{dt}\int (P_t(f))^2 d\mu = 2\int P_t(f)(\mathcal{L}P_t(f))d\mu = -2\int \|\nabla P_t(f)\|^2 d\mu \qquad (*)$$

Let us assume (i). With equation $(*)$, we have

$$\frac{d}{dt}\mathrm{Var}_\mu(P_t(f)) = -2\int \|\nabla P_t(f)\|^2 d\mu \leqslant -2\,\mathcal{P}_\mu^{-1}\int (P_t(f))^2 d\mu = -2\,\mathcal{P}_\mu^{-1}\mathrm{Var}_\mu(P_t(f)).$$

The proof is then completed by using Grönwall's inequality.

Let us assume (ii). We write, for $t > 0$,

$$-t^{-1}(\mathrm{Var}_\mu(P_t(f)) - \mathrm{Var}_\mu(f)) \geqslant -t^{-1}(\mathrm{e}^{-2t/\mathcal{P}_\mu} - 1)\mathrm{Var}_\mu(f).$$

By letting t go to 0 and using equation $(*)$,

$$2\mathcal{P}_\mu^{-1}\mathrm{Var}_\mu(f) \leqslant \frac{d}{dt}\mathrm{Var}_\mu(P_t(f))_{t=0} = 2\int \|\nabla f\|^2 d\mu,$$

which shows the converse implication. ∎

Remark 9

Let f be a centered eigenvector of $-\mathcal{L}$ with eigenvalue $\lambda \neq 0$. By the Poincaré inequality,

$$\int f^2 d\mu \leqslant \mathcal{P}_\mu\int \|\nabla f\|^2 d\mu = \mathcal{P}_\mu\int f(-\mathcal{L}f)d\mu = \mathcal{P}_\mu\lambda\int f^2 d\mu,$$

from which we deduce that every non-zero eigenvalue of $-\mathcal{L}$ is larger that $1/\mathcal{P}_\mu$. The best Poincaré

> *constant is thus the inverse of the smallest non zero eigenvalue of $-\mathcal{L}$. The finiteness of the Poincaré constant is therefore equivalent to a* spectral gap *property of $-\mathcal{L}$. Similarly, a discrete space Markov chain with transition matrix P converges at a rate determined by the spectral gap of $I - P$.*

There have been efforts in the past to estimate spectral gaps of Markov chains [HKS15, LP16, QHK+19, WK19, CT19] but these have been done with samples from trajectories of the dynamics. The main difference here is that the estimation will only rely on samples from the stationary measure.

Poincaré constant and sampling. In high dimensional settings (in Bayesian machine learning [Rob07]) or molecular dynamics [LS16] where d can be large – from 100 to 10^7), one of the standard techniques to sample $d\mu(x) = \mathrm{e}^{-V(x)}dx$ is to build a Markov chain by discretizing in time the overdamped Langevin diffusion (96) whose law converges to $d\mu$. According to Theorem 11, the typical time to wait to reach equilibrium is $\mathcal{P}_\mu$. Hence, the larger the Poincaré constant of a probability measure $d\mu$ is, the more difficult the sampling of $d\mu$ is. Note also that V need not be convex for the Markov chain to converge.

1.2.3. Examples

Gaussian distribution. For $d\mu(x) = \frac{1}{(2\pi)^{d/2}}\mathrm{e}^{-\|x\|^2/2}dx$, the Gaussian measure on $\mathbb{R}^d$ of mean 0 and variance 1, it holds for all f smooth and compactly supported,

$$\mathrm{Var}_\mu(f) \leqslant \int_{\mathbb{R}^d} \|\nabla f\|^2 d\mu,$$

and one can show that $\mathcal{P}_\mu = 1$ is the optimal Poincaré constant (see [Che81]). More generally, for a Gaussian measure with covariance matrix Σ, the Poincaré constant is the spectral radius of Σ.

Other examples of analytically known Poincaré constant are $1/d$ for the uniform measure on the unit sphere in dimension d [Led14] and 4 for the exponential measure on the real line [BGL14]. There also exist various criteria to ensure the existence of **(PI)**. We will not give an exhaustive list as our aim is rather to emphasize the link between sampling and optimization. Let us however finish this part with particularly important results.

A measure of non-convexity. Let $d\mu(x) = \mathrm{e}^{-V(x)}dx$. It has been shown in the past decades that the "more convex" V is, the smaller the Poincaré constant is. Indeed, if V is ρ-strongly convex, then the Bakry-Emery criterion [BGL14] tells us that $\mathcal{P}_\mu \leqslant 1/\rho$. If V is only convex, it has been shown that $d\mu$ satisfies also a **(PI)** (with a possibly very large Poincaré constant) [RLM95, Bob99]. Finally, the case where V is non-convex is explored in detail in a one-dimensional setting and it is shown that for potentials V with an energy barrier of height h between two wells, the Poincaré constant explodes exponentially with respect the height h [MS14]. In that spirit, the Poincaré constant of $d\mu(x) = \mathrm{e}^{-V(x)}dx$ can be a quantitative way to quantify how multimodal the distribution $d\mu$ is and hence how non-convex the potential V is [JK17, RRT17].

1.3. Statistical Estimation of the Poincaré Constant

The aim of this section is to provide an estimator of the Poincaré constant of a measure μ when we only have access to n samples of it, and to study its convergence properties. More precisely, given n independent and identically distributed (i.i.d.) samples $(x_1, \dots, x_n)$ of the probability measure $d\mu$, our goal is to estimate $\mathcal{P}_\mu$. We will denote this estimator (function of $(x_1, \dots, x_n)$) by the standard notation $\widehat{\mathcal{P}}_\mu$.

1.3.1. Reformulation of the problem in a reproducing kernel Hilbert Space

Definition and first properties. Let us suppose here that the space of test functions of the **(PI)**, $\mathcal{H}$, is a reproducing kernel Hilbert space (RKHS) associated with a kernel K on $\mathbb{R}^d$ [SS02, STC04]. This has two important consequences:

1. $\mathcal{H}$ is the linear function space $\mathcal{H} = \text{span}\{K(\cdot, x),\ x \in \mathbb{R}^d\}$, and in particular, for all $x \in \mathbb{R}^d$, the function $y \mapsto K(y, x)$ is an element of $\mathcal{H}$ that we will denote by K_x.
2. The reproducing property: $\forall f \in \mathcal{H}$ and $\forall x \in \mathbb{R}^d$, $f(x) = \langle f, K(\cdot, x)\rangle_{\mathcal{H}}$. In other words, function evaluations are equal to dot products with canonical elements of the RKHS.

We make the following mild assumptions on the RKHS:

Assumption 1

The RKHS $\mathcal{H}$ is dense in $H^1(\mu)$.

Note that this is the case for most of the usual kernels: Gaussian, exponential [MXZ06]. As **(PI)** involves derivatives of test functions, we will also need some regularity properties of the RKHS. Indeed, to represent ∇f in our RKHS we need a partial derivative reproducing property of the kernel space.

Assumption 2

K is a Mercer kernel such that $K \in C^2(\mathbb{R}^d \times \mathbb{R}^d)$.

Let us denote by $\partial_i = \partial_{x^i}$ the partial derivative operator with respect to the i-th component of x. It has been shown [Zho08] that under assumption 2, $\forall i \in [\![1, d]\!]$, $\partial_i K_x \in \mathcal{H}$ and that a partial derivative reproducing property holds true: $\forall f \in \mathcal{H}$ and $\forall x \in \mathbb{R}^d$, $\partial_i f(x) = \langle \partial_i K_x, f\rangle_{\mathcal{H}}$. Hence, thanks to assumption 2, ∇f is easily represented in the RKHS. We also need some boundedness properties of the kernel.

Assumption 3

K is a kernel such that $\forall x \in \mathbb{R}^d$, $K(x, x) \leqslant \mathcal{K}$ and[10] $\|\nabla K_x\|^2 \leqslant \mathcal{K}_d$, where $\|\nabla K_x\|^2 := \sum_{i=1}^d \langle \partial_i K_x, \partial_i K_x\rangle = \sum_{i=1}^d \frac{\partial^2 K}{\partial x^i \partial y^i}(x, x)$ (see calculations below), x and y standing respectively for the first and the second variables of $(x, y) \mapsto K(x, y)$.

The equality mentioned in the expression of $\|\nabla K_x\|^2$ arises from the following computation: $\partial_i K_y(x) = \langle \partial_i K_y, K_x\rangle = \partial_{y^i} K(x, y)$ and we can write that for all $x, y \in \mathbb{R}^d$, $\langle \partial_i K_x, \partial_i K_y\rangle = \partial_{x^i}\left(\partial_i K_y(x)\right) = \partial_{x^i}\partial_{y^i} K(x, y)$. Note that, for example, the Gaussian kernel satisfies 1, 2, 3.

A spectral point of view. Let us define the following operators from $\mathcal{H}$ to $\mathcal{H}$:

$$\Sigma = \mathbb{E}\left[K_x \otimes K_x\right], \qquad \mathsf{L} = \mathbb{E}\left[\nabla K_x \otimes_d \nabla K_x\right],$$

and their empirical counterparts,

$$\widehat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^n K_{x_i} \otimes K_{x_i}, \qquad \widehat{\mathsf{L}} = \frac{1}{n}\sum_{i=1}^n \nabla K_{x_i} \otimes_d \nabla K_{x_i},$$

where $\otimes$ is the standard tensor product: $\forall f, g, h \in \mathcal{H}$, $(f \otimes g)(h) = \langle g, h\rangle_{\mathcal{H}} f$ and $\otimes_d$ is defined as follows: $\forall f, g \in \mathcal{H}^d$ and $h \in \mathcal{H}$, $(f \otimes_d g)(h) = \sum_{i=1}^d \langle g_i, h\rangle_{\mathcal{H}} f_i$.

Proposition 11 (*Spectral characterization of the Poincaré constant*)

Suppose that assumptions 1, 2, 3 hold true. Then the Poincaré constant $\mathcal{P}_\mu$ is the maximum of the following

Rayleigh ratio:

$$\mathcal{P}_\mu = \sup_{f \in \mathcal{H} \setminus \mathrm{Ker}(\Delta)} \frac{\langle f, Cf \rangle_{\mathcal{H}}}{\langle f, \mathsf{L} f \rangle_{\mathcal{H}}} = \left\| \mathsf{L}^{-1/2} C \mathsf{L}^{-1/2} \right\|, \tag{97}$$

with $\|\cdot\|$ the operator norm on $\mathcal{H}$ and $C = \Sigma - m \otimes m$ where $m = \int_{\mathbb{R}^d} K_x d\mu(x) \in \mathcal{H}$ is the covariance operator, considering Δ^{-1} as the inverse of Δ restricted to $(\mathrm{Ker}(\Delta))^\perp$.

Note that C and L are symmetric positive semi-definite trace-class operators (see Appendix C.3.2). Note also that $\mathrm{Ker}(\Delta)$ is the set of constant functions, which suggests introducing $\mathcal{H}_0 := (\mathrm{Ker}(\Delta))^\perp = \mathcal{H} \cap L^2_0(\mu)$, where $L^2_0(\mu)$ is the space of $L^2(\mu)$ functions with mean zero with respect to μ. Finally note that $\mathrm{Ker}(\Delta) \subset \mathrm{Ker}(C)$ (see Section C.1 of the Appendix). With the characterization provided by Proposition 11, we can easily define an estimator of the Poincaré constant $\widehat{\mathcal{P}}_\mu$, following standard regularization techniques from kernel methods [SS02, STC04, FBG07].

Definition 4

The estimator $\widehat{\mathcal{P}}^{n,\lambda}_\mu$ of the Poincaré constant is the following:

$$\widehat{\mathcal{P}}^{n,\lambda}_\mu := \sup_{f \in \mathcal{H} \setminus \mathrm{Ker}(\Delta)} \frac{\langle f, \widehat{C} f \rangle_{\mathcal{H}}}{\langle f, (\widehat{\mathsf{L}} + \lambda I) f \rangle_{\mathcal{H}}} = \left\| \widehat{\mathsf{L}}_\lambda^{-1/2} \widehat{C} \widehat{\mathsf{L}}_\lambda^{-1/2} \right\|, \tag{98}$$

with $\widehat{C} = \widehat{\Sigma}_n - \widehat{m} \otimes \widehat{m}$ and where $\widehat{m} = \frac{1}{n} \sum_{i=1}^n K_{x_i}$. $\widehat{C}$ is the empirical covariance operator and $\widehat{\mathsf{L}}_\lambda = \widehat{\mathsf{L}} + \lambda I$ is a regularized empirical version of the operator L restricted to $(\mathrm{Ker}(\Delta))^\perp$ as in Proposition 11.

Note that regularization is necessary as the nullspace of $\widehat{\Delta}$ is no longer included in the nullspace of $\widehat{C}$ so that the Poincaré constant estimates blows up when $\lambda \to 0$. The problem in Equation (98) has a natural interpretation in terms of Poincaré inequality as it corresponds to a regularized **(PI)** for the empirical measure $\widehat{\mu}_n = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$ associated with the i.i.d. samples $x_1, \ldots, x_n$ from $d\mu$. To alleviate the notation, we will simply denote the estimator by $\widehat{\mathcal{P}}_\mu$ until the end of the paper.

1.3.2. Statistical consistency of the estimator

We show that, under some assumptions and by choosing carefully λ as a function of n, the estimator $\widehat{\mathcal{P}}_\mu$ is statistically consistent, i.e., almost surely:

$$\widehat{\mathcal{P}}_\mu \xrightarrow{n \to \infty} \mathcal{P}_\mu.$$

As we regularized our problem, we prove the convergence in two steps: first, the convergence of $\widehat{\mathcal{P}}_\mu$ to the regularized problem $\mathcal{P}^\lambda_\mu = \sup_{f \in \mathcal{H} \setminus \{0\}} \frac{\langle f, Cf \rangle}{\langle f, (\mathsf{L} + \lambda I) f \rangle} = \|\mathsf{L}_\lambda^{-1/2} C \mathsf{L}_\lambda^{-1/2}\|$, which corresponds to controlling the statistical error associated with the estimator $\widehat{\mathcal{P}}_\mu$ (variance); second, the convergence of $\mathcal{P}^\lambda_\mu$ to $\mathcal{P}_\mu$ as λ goes to zero which corresponds to the bias associated with the estimator $\widehat{\mathcal{P}}_\mu$. The next result states the statistical consistency of the estimator when λ is a sequence going to zero as n goes to infinity (typically as an inverse power of n).

Theorem 12 (*Statistical consistency*)

Assume that 1, 2, 3 hold true and that the operator $\Delta^{-1/2} C \Delta^{-1/2}$ is compact on $\mathcal{H}$. Let $(\lambda_n)_{n \in \mathbb{N}}$ be a sequence of positive numbers such that $\lambda_n \to 0$ and $\lambda_n \sqrt{n} \to +\infty$. Then, almost surely,

$$\widehat{\mathcal{P}}_\mu \xrightarrow{n \to \infty} \mathcal{P}_\mu.$$

As already mentioned, the proof is divided into two steps: the analysis of the statistical error for which we have an explicit rate of convergence in probability (see Proposition 12 below) and which requires $n^{-1/2}/\lambda_n \to 0$, and the analysis of the bias for which we need $\lambda_n \to 0$ and the compactness condition (see Proposition 13). Notice that the compactness assumption in Proposition 13 and Theorem 12 is stronger than **(PI)**. Indeed, it can be shown that satisfying **(PI)** is equivalent to having the operator $\Delta^{-1/2}C\Delta^{-1/2}$ bounded whereas to have convergence of the bias we need compactness. Note also that $\lambda_n = n^{-1/4}$ matches the two conditions stated in Theorem 12 and is the optimal balance between the rate of convergence of the statistical error (of order $\frac{1}{\lambda\sqrt{n}}$, see Proposition 12) and of the bias we obtain in some cases (of order λ, see Section C.2 of the Appendix). Note that the rates of convergence do not depend on the dimension d of the problem which is a usual strength of kernel methods and differ from local methods like diffusion maps [CL06, HAL07].

For the statistical error term, it is possible to quantify the rate of convergence of the estimator to the regularized Poincaré constant as shown below.

Proposition 12 (*Analysis of the statistical error*)

Suppose that 1, 2, 3 hold true. For any $\delta \in (0, 1/3)$, and $\lambda > 0$ such that $\lambda \leqslant \|\Delta\|$ and any integer $n \geqslant 15\frac{\mathcal{K}_d}{\lambda}\log\frac{4\,\mathrm{Tr}\mathsf{L}}{\lambda\delta}$, with probability at least $1-3\delta$,

$$\left|\widehat{\mathcal{P}}_\mu - \mathcal{P}_\mu^\lambda\right| \leqslant \frac{8\mathcal{K}}{\lambda\sqrt{n}}\log(2/\delta) + \mathrm{o}\left(\frac{1}{\lambda\sqrt{n}}\right). \tag{99}$$

Note that in Proposition 12 we are only interested in the regime where $\lambda\sqrt{n}$ is large. Lemmas 31 and 32 of the Appendix give explicit and sharper bounds under refined hypotheses on the spectra of C and Δ. Recall also that under assumption 3, C and Δ are trace-class operators (as proved in the Appendix, Section C.3.2) so that $\|\Delta\|$ and $\mathrm{tr}(\Delta)$ are indeed finite. Finally, remark that (99) implies the almost sure convergence of the statistical error by applying the Borel-Cantelli lemma.

Proposition 13 (*Analysis of the bias*)

Assume that 1, 2, 3 hold true, and that the bounded operator $\Delta^{-1/2}C\Delta^{-1/2}$ is compact on $\mathcal{H}$. Then,

$$\lim_{\lambda\to 0}\mathcal{P}_\mu^\lambda = \mathcal{P}.$$

As said above the compactness condition (similar to the one used for convergence proofs of kernel Canonical Correlation Analysis [FBG07]) is stronger than satisfying **(PI)**. The compactness condition adds conditions on the spectrum of $\Delta^{-1/2}C\Delta^{-1/2}$: it is discrete and accumulates at 0. We give more details on this condition in Section C.2 of the Appendix and derive explicit rates of convergence under general conditions. We derive also a rate of convergence for more specific structures (Gaussian case or under an assumption on the support of μ) in Sections C.2 and C.4 of the Appendix.

1.4. Learning a Reaction Coordinate

If the measure μ is multimodal, the Langevin dynamics (96) is trapped for long times in certain regions (modes) preventing it from efficient space exploration. This phenomenon is called *metastability* and is responsible for the slow convergence of the diffusion to its equilibrium [Lel13, LRS08]. Some efforts in the past decade [Lel15] have focused on understanding this multimodality by capturing the behavior of the dynamics at a coarse-grained level, which often have a low-dimensional nature. The aim of this section is to take advantage of the estimation of the Poincaré constant to give a procedure to unravel these dynamically meaningful slow variables called reaction coordinate.

1.4.1. Good Reaction Coordinate

From a numerical viewpoint, a good reaction coordinate can be defined as a low dimensional function $\xi : \mathbb{R}^d \to \mathbb{R}^p$ ($p \ll d$) such that the family of conditional measures $(\mu(\cdot|\xi(x) = r))_{z\in\mathbb{R}^p}$ are "less multimodal" than the measure $d\mu$. This can be fully formalized in particular in the context of free energy techniques such as the adaptive biasing force method, see for example [LRS08]. For more details on mathematical formalizations of metastability, we also refer to [Lel13]. The point of view we will follow in this work is to choose ξ in order to maximize the Poincaré constant of the pushforward distribution $\xi * \mu$. The idea is to capture in $\xi * \mu$ the essential multimodality of the original measure, in the spirit of the two scale decomposition of Poincaré or logarithmic Sobolev constant inequalities [Lel09, MS14, OR07].

1.4.2. Learning a Reaction Coordinate

Optimization problem. Let us assume in this subsection that the reaction coordinate is an orthogonal projection onto a linear subspace of dimension p. Hence ξ can be represented by $\forall x \in \mathbb{R}^d,\ \xi(x) = Ax$ with $A \in \mathcal{S}^{p,d}$ where $\mathcal{S}^{p,d} = \{A \in \mathbb{R}^{p\times d} \text{ s. t. } AA^\top = I_p\}$ is the Stiefel manifold [EAS98]. As discussed in Section 1.4.1, to find a good reaction coordinate we look for ξ for which the Poincaré constant of the pushforward measure $\xi * \mu$ is the largest. Given n samples, let us define the matrix $X = (x_1, \dots, x_n)^\top \in \mathbb{R}^{n\times d}$. We denote by $\widehat{\mathcal{P}}_X$ the estimator of the Poincaré constant using the samples $(x_1, \dots, x_n)$. Hence $\widehat{\mathcal{P}}_{AX^\top}$ defines an estimator of the Poincaré constant of the pushforward measure $\xi * \mu$. Our aim is to find $\operatorname{argmax}_{A\in\mathcal{S}^{p,d}} \widehat{\mathcal{P}}_{AX^\top}$.

Random features. One computational issue with the estimation of the Poincaré constant is that building $\widehat{C}$ and $\widehat{\Delta}$ requires respectively constructing $n \times n$ and $nd \times nd$ matrices. Random features [RR08] avoid this problem by building explicitly features that approximate a translation invariant kernel $K(x, x') = K(x - x')$. More precisely, let M be the number of random features, $(w_m)_{1\leqslant m\leqslant M}$ be random variables independently and identically distributed according to $\mathbb{P}(dw) = \int_{\mathbb{R}^d} \mathrm{e}^{-\mathrm{i}w^\top\delta} K(\delta)d\delta\, dw$ and $(b_m)_{1\leqslant m\leqslant M}$ be independently and identically distributed according to the uniform law on $[0, 2\pi]$, then the feature vector $\phi^M(x) = \sqrt{\frac{2}{M}}\left(\cos(w_1^\top x + b_1), \dots, \cos(w_M^\top x + b_M)\right)^\top \in \mathbb{R}^M$ satisfies $K(x, x') \approx \phi^M(x)^\top\phi^M(x')$. Therefore, random features allow to approximate $\widehat{C}$ and $\widehat{\Delta}$ by $M \times M$ matrices $\widehat{C}^M$ and $\widehat{\mathsf{L}}^M$ respectively. Finally, when these matrices are constructed using the projected samples, i.e. $\left(\cos(w_m^\top Ax_i + b_m)\right)_{\substack{1\leq m\leq M\\ 1\leq i\leq n}}$, we denote them by $\widehat{C}^M_A$ and $\widehat{\mathsf{L}}^M_A$ respectively. Hence, the problem reads

$$\text{Find } \operatorname*{argmax}_{A\in\mathcal{S}^{p,d}} \widehat{\mathcal{P}}_{AX^\top} = \operatorname*{argmax}_{A\in\mathcal{S}^{p,d}} \max_{v\in\mathbb{R}^M\setminus\{0\}} F(A, v)\,, \quad \text{where } F(A, v) := \frac{v^\top \widehat{C}^M_A v}{v^\top(\widehat{\mathsf{L}}^M_A + \lambda I)v}. \tag{100}$$

Algorithm. To solve the non-concave optimization problem (100), our procedure is to do one step of non-Euclidean gradient descent to update A (gradient descent in the Stiefel manifold) and one step by

solving the generalized eigenvalue problem to update v. More precisely, the algorithm reads:

Result: Best linear Reaction Coordinate: $A_* \in \mathcal{S}^{d,p}$
A_0 random matrix in $\mathcal{S}^{d,p}$, $\eta_t > 0$ step-size;
for $t = 0, \dots, T-1$ **do**

- Solve generalized largest eigenvalue problem with matrices $\widehat{C}_{A_t}^M$ and $\widehat{\mathsf{L}}_{A_t}^M$ to get $v^*(A_t)$:

$$v^*(A_t) = \underset{v \in \mathbb{R}^M \setminus \{0\}}{\operatorname{argmax}} \frac{v^\top \widehat{C}_A^M v}{v^\top (\widehat{\mathsf{L}}_A^M + \lambda I) v}.$$

- Do one gradient ascent step: $A_{t+1} = A_t + \eta_t \operatorname{grad}_A F(A, v^*(A_t))$.

end

Algorithm 1: Algorithm to find best linear Reaction Coordinate.

1.5. Numerical experiments

We divide our experiments into two parts: the first one illustrates the convergence of the estimated Poincaré constant as given by Theorem 12 (see Section 1.5.1), and the second one demonstrates the interest of the reaction coordinates learning procedure described in Section 1.4.2 (see Section 1.5.2).

1.5.1. Estimation of the Poincaré constant

In our experiments we choose the Gaussian Kernel $K(x, x') = \exp(-\|x - x'\|^2)$? This induces a RKHS satisfying 1, 2, 3. Estimating $\widehat{\mathcal{P}}_\mu$ from n samples $(x_i)_{i \leqslant n}$ is equivalent to finding the largest eigenvalue for an operator from $\mathcal{H}$ to $\mathcal{H}$. Indeed, we have

$$\widehat{\mathcal{P}}_\mu = \left\| (\widehat{Z}_n^* \widehat{Z}_n + \lambda I)^{-\frac{1}{2}} \widehat{S}_n^* \left(I - \frac{1}{n} \mathbb{1}\mathbb{1}^\top \right) \widehat{S}_n (\widehat{Z}_n^* \widehat{Z}_n + \lambda I)^{-\frac{1}{2}} \right\|_{\mathcal{H}},$$

where $\widehat{Z}_n = \sum_{i=1}^d \widehat{Z}_n^i$ and $\widehat{Z}_n^i$ is the operator from $\mathcal{H}$ to $\mathbb{R}^n$: $\forall g \in \mathcal{H}$, $\widehat{Z}_n^i(g) = \frac{1}{\sqrt{n}} \left(\langle g, \partial_i K_{x_j} \rangle \right)_{1 \leqslant j \leqslant n}$ and $\widehat{S}_n$ is the operator from $\mathcal{H}$ to $\mathbb{R}^n$: $\forall g \in \mathcal{H}$, $\widehat{S}_n(g) = \frac{1}{\sqrt{n}} \left(\langle g, K_{x_j} \rangle \right)_{1 \leqslant j \leqslant n}$. By the Woodbury operator identity, $(\lambda I + \widehat{Z}_n^* \widehat{Z}_n)^{-1} = \frac{1}{\lambda} \left(I - \widehat{Z}_n^* (\lambda I + \widehat{Z}_n \widehat{Z}_n^*)^{-1} \widehat{Z}_n \right)$, and the fact that for any operator $\|T^* T\| = \|T T^*\|$,

$$\begin{aligned}
\widehat{\mathcal{P}}_\mu &= \left\| (\widehat{Z}_n^* \widehat{Z}_n + \lambda I)^{-\frac{1}{2}} \widehat{S}_n^* \left(I - \frac{1}{n} \mathbb{1}\mathbb{1}^\top \right) \widehat{S}_n (\widehat{Z}_n^* \widehat{Z}_n + \lambda I)^{-\frac{1}{2}} \right\|_{\mathcal{H}} \\
&= \left\| (\widehat{Z}_n^* \widehat{Z}_n + \lambda I)^{-\frac{1}{2}} \widehat{S}_n^* \left(I - \frac{1}{n} \mathbb{1}\mathbb{1}^\top \right) \left(I - \frac{1}{n} \mathbb{1}\mathbb{1}^\top \right) \widehat{S}_n (\widehat{Z}_n^* \widehat{Z}_n + \lambda I)^{-\frac{1}{2}} \right\|_{\mathcal{H}} \\
&= \left\| \left(I - \frac{1}{n} \mathbb{1}\mathbb{1}^\top \right) \widehat{S}_n (\widehat{Z}_n^* \widehat{Z}_n + \lambda I)^{-1} \widehat{S}_n^* \left(I - \frac{1}{n} \mathbb{1}\mathbb{1}^\top \right) \right\|_2 \\
&= \frac{1}{\lambda} \left\| \left(I - \frac{1}{n} \mathbb{1}\mathbb{1}^\top \right) \left(\widehat{S}_n \widehat{S}_n^* - \widehat{S}_n \widehat{Z}_n^* (\widehat{Z}_n \widehat{Z}_n^* + \lambda I)^{-1} \widehat{Z}_n \widehat{S}_n^* \right) \left(I - \frac{1}{n} \mathbb{1}\mathbb{1}^\top \right) \right\|_2,
\end{aligned}$$

which is now the largest eigenvalue of a $n \times n$ matrix built as the product of matrices involving the kernel K and its derivatives. Note for the above calculation that we used that $\left(I - \frac{1}{n} \mathbb{1}\mathbb{1}^\top \right)^2 = \left(I - \frac{1}{n} \mathbb{1}\mathbb{1}^\top \right)$.

We illustrate in Figure 18 the rate of convergence of the estimated Poincaré constant to 1 for the Gaussian $\mathcal{N}(0, 1)$ as the number of samples n grows. Recall that in this case the Poincaré constant is equal

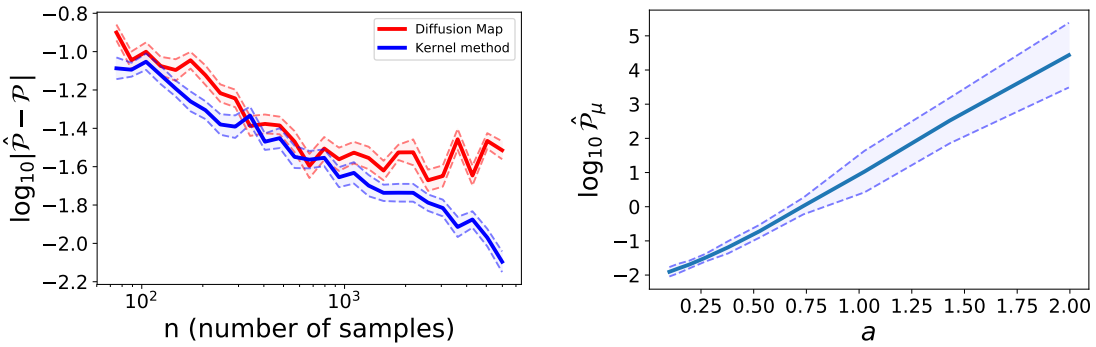

Figure 18: **(Left)** Comparison of the convergences of the kernel-based method described in this paper and diffusion maps in the case of a Gaussian of variance 1 (for each n we took the mean over 50 runs). The dotted lines correspond to standard deviations of the estimator. **(Right)** Exponential growth of the Poincaré constant for a mixture of two Gaussians $\mathcal{N}(\pm\frac{a}{2}, \sigma^2)$ as a function of the distance a between the two Gaussians ($\sigma = 0.1$ and $n = 500$).

to 1 (see Subsection 1.2.3). We compare our prediction to the one given by diffusion maps techniques [CL06]. For our method, in all the experiments we set $\lambda_n = \frac{C_\lambda}{n}$, which is smaller than what is given by Theorem 12, and optimize the constant C_λ with a grid search. Following [HAL07], to find the correct bandwidth ε_n of the kernel involved in diffusion maps, we performed a similar grid search on the constant C_ε for the Diffusion maps with the scaling $\varepsilon_n = \frac{C_\varepsilon}{n^{1/4}}$. Additionally to a faster convergence when n become large, the kernel-based method is more robust with respect to the choice of itss hyperparameter, which is of crucial importance for the quality of diffusion maps. Note also that we derive an explicit convergence rate for the bias in the Gaussian case in Section C.4 of the Appendix. In Figure 18, we also show the growth of the Poincaré constant for a mixture of Gaussians of variances 1 as a function of the distance between the two means of the Gaussians. This is a situation for which the estimation provides an estimate when, up to our knowledge, no precise Poincaré constant is known (even if lower and upper bounds are known [CM10]).

1.5.2. Learning a reaction coordinate

We next illustrate the algorithm described in Section 1.4 to learn a reaction coordinate which, we recall, encodes directions which are difficult to sample. To perform the gradient step over the Stiefel manifold we used Pymanopt [TKW16], a Python library for manifold optimization derived from Manopt [BMAS14] (Matlab). We show here a synthetic two-dimensional example example. We first preprocessed the samples with "whitening", i.e., making it of variance 1 in all directions to avoid scaling artifacts. In both examples, we took $M = 200$ for the number of random features and $n = 200$ for the number of samples.

We show (Figure 19) one synthetic example for which our algorithm found a good reaction coordinate. The samples are taken from a mixture of three Gaussians of means $(0, 0), (1, 1)$ and $(2, 2)$ and covariance $\Sigma = \sigma^2 I$ where $\sigma = 0.1$. The three means are aligned along a line which makes an angle $\theta = \pi/4$ with respect to the x-axis: one expects the algorithm to identify this direction as the most difficult one to sample (see left and center plots of Figure 19). With a few restarts, our algorithm indeed finds the largest Poincaré constant for a projection onto the line parametrized by $\theta = \pi/4$.

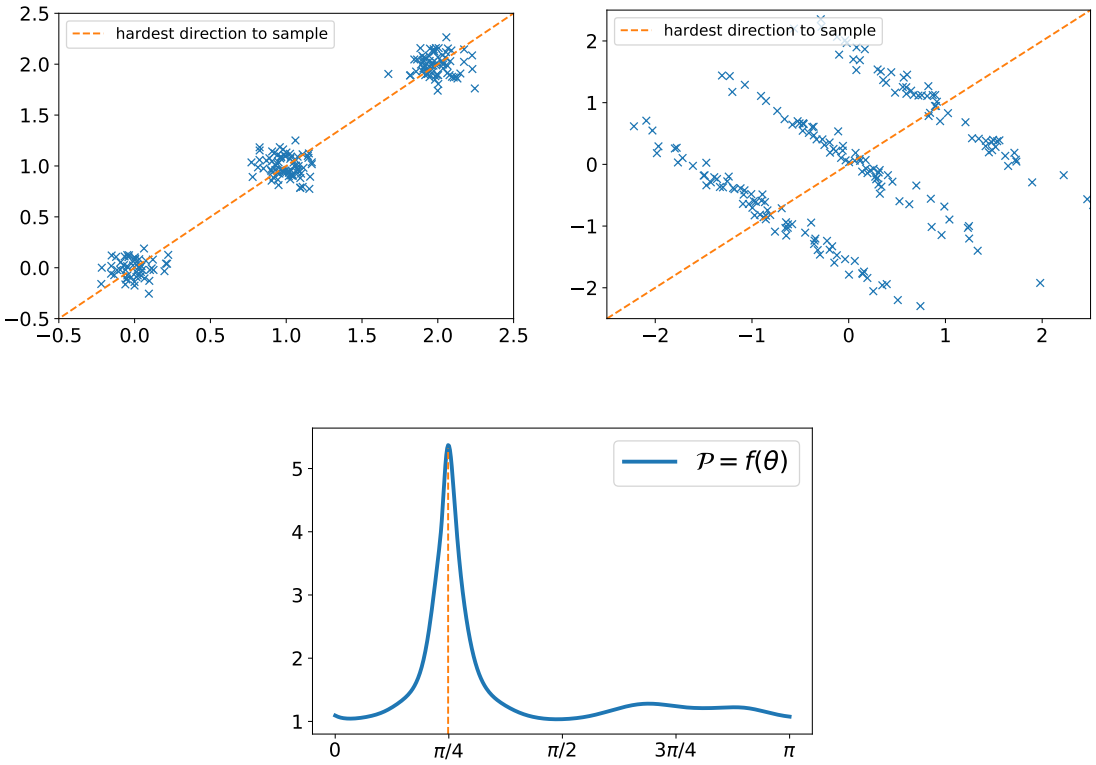

Figure 19: **(Top Left)** Samples of mixture of three Gaussians. **(Top right)** Whiten samples of Gaussian mixture on the left. **(Bottom)** Plot of the Poincaré constant of the projected samples on a line of angle θ.

1.6. Conclusion and Perspectives

In this paper, we have presented an efficient method to estimate the Poincaré constant of a distribution from independent samples, paving the way to learn low-dimensional marginals that are hard to sample (corresponding to the image measure of so-called reaction coordinates). While we have focused on linear projections, learning non-linear projections is important in molecular dynamics and it can readily be done with a well-defined parametrization of the non-linear function and then applied to real data sets, where this would lead to accelerated sampling [Lel15]. Finally, it would be interesting to apply our framework to Bayesian inference [CLS12] and leverage the knowledge of reaction coordinates to accelerate sampling methods.

⋆

⋆ ⋆

C. Appendix of Statistical estimation of the Poincaré constant and application to sampling multimodal distributions

The Appendix is organized as follows. In Section C.1 we prove Propositions 11 and 12. Section C.2 is devoted to the analysis of the bias. We study spectral properties of the diffusion operator L to give sufficient and general conditions for the compactness assumption from Theorem 12 and Proposition 13 to hold. Section C.3 provides concentration inequalities for the operators involved in Proposition 12. We conclude by Section C.4 that gives explicit rates of convergence for the bias when μ is a 1-D Gaussian (this result could be easily extended to higher dimensional Gaussians).

C.1. Proofs of Proposition 11 and 12

Recall that $L^2_0(\mu)$ is the subspace of $L^2(\mu)$ of zero mean functions: $L^2_0(\mu) := \{f \in L^2(\mu),\ \int f(x)d\mu(x) = 0\}$ and that we similarly defined $\mathcal{H}_0 := \mathcal{H} \cap L^2_0(\mu)$. Let us also denote by $\mathbb{R}\mathbb{1}$, the set of constant functions.

Proof Proof of Proposition 11 : The proof is simply the following reformulation of Equation (94). Under assumption 1:

$$\begin{aligned}\mathcal{P}_\mu &= \sup_{f \in H^1(\mu)\setminus\mathbb{R}\mathbb{1}} \frac{\int_{\mathbb{R}^d} f(x)^2 d\mu(x) - \left(\int_{\mathbb{R}^d} f(x)d\mu(x)\right)^2}{\int_{\mathbb{R}^d} \|\nabla f(x)\|^2 d\mu(x)} \\ &= \sup_{f \in \mathcal{H}\setminus\mathbb{R}\mathbb{1}} \frac{\int_{\mathbb{R}^d} f(x)^2 d\mu(x) - \left(\int_{\mathbb{R}^d} f(x)d\mu(x)\right)^2}{\int_{\mathbb{R}^d} \|\nabla f(x)\|^2 d\mu(x)} \\ &= \sup_{f \in \mathcal{H}_0\setminus\{0\}} \frac{\int_{\mathbb{R}^d} f(x)^2 d\mu(x) - \left(\int_{\mathbb{R}^d} f(x)d\mu(x)\right)^2}{\int_{\mathbb{R}^d} \|\nabla f(x)\|^2 d\mu(x)}.\end{aligned}$$

We then simply note that

$$\left(\int_{\mathbb{R}^d} f(x)d\mu(x)\right)^2 = \left(\left\langle f, \int_{\mathbb{R}^d} K_x d\mu(x)\right\rangle_{\mathcal{H}}\right)^2 = \langle f, m\rangle^2_{\mathcal{H}} = \langle f, (m\otimes m)f\rangle_{\mathcal{H}}.$$

Similarly,

$$\int_{\mathbb{R}^d} f(x)^2 d\mu(x) = \langle f, \Sigma f\rangle_{\mathcal{H}} \quad \text{and} \quad \int_{\mathbb{R}^d} \|\nabla f(x)\|^2 d\mu(x) = \langle f, \mathsf{L} f\rangle_{\mathcal{H}}.$$

Note here that $\mathrm{Ker}(\Delta) \subset \mathrm{Ker}(C)$. Indeed, if $f \in \mathrm{Ker}(\Delta)$, then $\langle f, \Delta f\rangle_{\mathcal{H}} = 0$. Hence, μ-almost everywhere, $\nabla f = 0$ so that f is constant and $Cf = 0$. Note also the previous reasoning shows that $\mathrm{Ker}(\Delta)$ is the subset of $\mathcal{H}$ made of constant functions, and $(\mathrm{Ker}(\Delta))^\perp = \mathcal{H} \cap L^2_0(\mu) = \mathcal{H}_0$.

Thus we can write,

$$\mathcal{P}_\mu = \sup_{f \in \mathcal{H}\setminus\mathrm{Ker}(\Delta)} \frac{\langle f, (\Sigma - m\otimes m)f\rangle_{\mathcal{H}}}{\langle f, \mathsf{L} f\rangle_{\mathcal{H}}} = \left\|\mathsf{L}^{-1/2} C \mathsf{L}^{-1/2}\right\|,$$

where we consider Δ^{-1} as the inverse of Δ restricted to $(\mathrm{Ker}(\Delta))^\perp$ and thus get Proposition 11. ■

Proof Proof of Proposition 12 : We refer to Lemmas 31 and 32 in Section C.3 for the explicit bounds. We have the following inequalities:

$$\begin{aligned}\left|\widehat{\mathcal{P}}_\mu - \mathcal{P}_\mu^\lambda\right| &= \left|\left\|\widehat{\mathsf{L}}_\lambda^{-1/2}\widehat{C}\widehat{\mathsf{L}}_\lambda^{-1/2}\right\| - \left\|\mathsf{L}_\lambda^{-1/2}C\mathsf{L}_\lambda^{-1/2}\right\|\right| \\ &\leqslant \left|\left\|\widehat{\mathsf{L}}_\lambda^{-1/2}\widehat{C}\widehat{\mathsf{L}}_\lambda^{-1/2}\right\| - \left\|\widehat{\mathsf{L}}_\lambda^{-1/2}C\widehat{\mathsf{L}}_\lambda^{-1/2}\right\|\right| + \left|\left\|\widehat{\mathsf{L}}_\lambda^{-1/2}C\widehat{\mathsf{L}}_\lambda^{-1/2}\right\| - \left\|\mathsf{L}_\lambda^{-1/2}C\mathsf{L}_\lambda^{-1/2}\right\|\right| \\ &\leqslant \left\|\widehat{\mathsf{L}}_\lambda^{-1/2}(\widehat{C}-C)\widehat{\mathsf{L}}_\lambda^{-1/2}\right\| + \left|\left\|C^{1/2}\widehat{\mathsf{L}}_\lambda^{-1}C^{1/2}\right\| - \left\|C^{1/2}\mathsf{L}_\lambda^{-1}C^{1/2}\right\|\right| \\ &\leqslant \left\|\widehat{\mathsf{L}}_\lambda^{-1/2}(\widehat{C}-C)\widehat{\mathsf{L}}_\lambda^{-1/2}\right\| + \left\|C^{1/2}(\widehat{\mathsf{L}}_\lambda^{-1}-\mathsf{L}_\lambda^{-1})C^{1/2}\right\|.\end{aligned}$$

Consider an event where the estimates of Lemmas 31, 32 and 33 hold for a given value of $\delta > 0$. A simple computation shows that this event has a probability $1-3\delta$ at least. We study the two terms above separately. First, provided that $n \geqslant 15\mathcal{F}_\infty(\lambda)\log\frac{4\operatorname{Tr}\mathsf{L}}{\lambda\delta}$ and $\lambda \in (0, \|\Delta\|]$ in order to use Lemmas 32 and 33,

$$\begin{aligned}\left\|\widehat{\mathsf{L}}_\lambda^{-1/2}(\widehat{C}-C)\widehat{\mathsf{L}}_\lambda^{-1/2}\right\| &= \left\|\widehat{\mathsf{L}}_\lambda^{-1/2}\mathsf{L}_\lambda^{1/2}\mathsf{L}_\lambda^{-1/2}(\widehat{C}-C)\mathsf{L}_\lambda^{-1/2}\mathsf{L}_\lambda^{1/2}\widehat{\mathsf{L}}_\lambda^{-1/2}\right\| \\ &\leqslant \underbrace{\left\|\widehat{\mathsf{L}}_\lambda^{-1/2}\mathsf{L}_\lambda^{1/2}\right\|^2}_{\text{Lemma 33}}\underbrace{\left\|\mathsf{L}_\lambda^{-1/2}(\widehat{C}-C)\mathsf{L}_\lambda^{-1/2}\right\|}_{\text{Lemma 31}} \\ &\leqslant 2\,(\text{Lemma 31}).\end{aligned}$$

For the second term,

$$\begin{aligned}\left\|C^{1/2}(\widehat{\mathsf{L}}_\lambda^{-1}-\mathsf{L}_\lambda^{-1})C^{1/2}\right\| &= \left\|C^{1/2}\widehat{\mathsf{L}}_\lambda^{-1}(\mathsf{L}-\widehat{\mathsf{L}})\mathsf{L}_\lambda^{-1}C^{1/2}\right\| \\ &= \left\|C^{1/2}\mathsf{L}_\lambda^{-1/2}\mathsf{L}_\lambda^{1/2}\widehat{\mathsf{L}}_\lambda^{-1}\mathsf{L}_\lambda^{1/2}\mathsf{L}_\lambda^{-1/2}(\mathsf{L}-\widehat{\mathsf{L}})\mathsf{L}_\lambda^{-1/2}\mathsf{L}_\lambda^{-1/2}C^{1/2}\right\| \\ &\leqslant \underbrace{\left\|\widehat{\mathsf{L}}_\lambda^{-1/2}\mathsf{L}_\lambda^{1/2}\right\|^2}_{\text{Lemma 33}}\underbrace{\left\|C^{1/2}\mathsf{L}_\lambda^{-1/2}\right\|^2}_{\mathcal{P}_\mu^\lambda}\underbrace{\left\|\mathsf{L}_\lambda^{-1/2}(\mathsf{L}-\widehat{\mathsf{L}})\mathsf{L}_\lambda^{-1/2}\right\|}_{\text{Lemma 32}} \\ &\leqslant 2\cdot\mathcal{P}_\mu^\lambda\cdot(\text{Lemma 32}).\end{aligned}$$

The leading order term in the estimate of Lemma 32 is of order $\left(\frac{2\mathcal{K}_d\log(4\operatorname{tr}\mathsf{L}/\lambda\delta)}{\lambda n}\right)^{1/2}$ whereas the leading one in Lemma 31 is of order $\frac{8\mathcal{K}\log(2/\delta)}{\lambda\sqrt{n}}$. Hence, the latter is the dominant term in the final estimation. ■

C.2. Analysis of the bias: convergence of the regularized Poincaré constant to the true one

We begin this section by proving Proposition 13. We then investigate the compactness condition required in the assumptions of Proposition 13 by studying the spectral properties of the diffusion operator L. In Proposition 16, we derive, under some general assumption on the RKHS and usual growth conditions on V, some convergence rate for the bias term.

C.2.1. General condition for consistency: proof of Proposition 13

To prove Proposition 13, we first need a general result on operator norm convergence.

Lemma 27

Let $\mathcal{H}$ be a Hilbert space and suppose that $(A_n)_{n\geqslant 0}$ is a family of bounded operators such that $\forall n \in \mathbb{N}$, $\|A_n\| \leqslant 1$ and $\forall f \in \mathcal{H}$, $A_n f \xrightarrow{n\to\infty} Af$. Suppose also that B is a compact operator. Then, in operator

norm,

$$A_n B A_n^* \xrightarrow{n\to\infty} ABA^*.$$

Proof : Let $\varepsilon > 0$. As B is compact, it can be approximated by a finite rank operator $B_{n_\varepsilon} = \sum_{i=1}^{n_\varepsilon} b_i \langle f_i, \cdot \rangle g_i$, where $(f_i)_i$ and $(g_i)_i$ are orthonormal bases, and $(b_i)_i$ is a sequence of nonnegative numbers with limit zero (singular values of the operator). More precisely, n_ε is chosen so that

$$\|B - B_{n_\varepsilon}\| \leqslant \frac{\varepsilon}{2}.$$

Moreover, ε being fixed, $A_n B_{n_\varepsilon} A_n^* = \sum_{i=1}^{n_\varepsilon} b_i \langle A_n f_i, \cdot \rangle A_n g_i \underset{n\infty}{\longrightarrow} \sum_{i=1}^{n_\varepsilon} b_i \langle A f_i, \cdot \rangle A g_i = A B_{n_\varepsilon} A^*$ in operator norm, so that, for $n \geqslant N_\varepsilon$, with $N_\varepsilon \geqslant n_\varepsilon$ sufficiently large, $\|A_n B_{n_\varepsilon} A_n^* - A B_{n_\varepsilon} A^*\| \leqslant \frac{\varepsilon}{2}$. Finally, as $\|A\| \leqslant 1$, it holds, for $n \geqslant N_\varepsilon$

$$\begin{aligned}\|A_n B_{n_\varepsilon} A_n^* - ABA^*\| &\leqslant \|A_n B_{n_\varepsilon} A_n^* - A B_{n_\varepsilon} A^*\| + \|A(B_{n_\varepsilon} - B)A^*\| \\ &\leqslant \|A_n B_{n_\varepsilon} A_n^* - A B_{n_\varepsilon} A^*\| + \|B_{n_\varepsilon} - B\| \leqslant \varepsilon.\end{aligned}$$

This proves the convergence in operator norm of $A_n B A_n^*$ to ABA^* when n goes to infinity. ∎

We can now prove Proposition 13.

Proof Proof of Proposition 13 : Let $\lambda > 0$, we want to show that

$$\mathcal{P}_\mu^\lambda = \|\Delta_\lambda^{-1/2} C \Delta_\lambda^{-1/2}\| \underset{\lambda\to 0}{\longrightarrow} \|\Delta^{-1/2} C \Delta^{-1/2}\| = \mathcal{P}_\mu.$$

Actually, with Lemma 27, we will show a stronger result which is the norm convergence of the operator $\Delta_\lambda^{-1/2} C \Delta_\lambda^{-1/2}$ to $\Delta^{-1/2} C \Delta^{-1/2}$. Indeed, denoting by $B = \Delta^{-1/2} C \Delta^{-1/2}$ and by $A_\lambda = \Delta_\lambda^{-1/2} \Delta^{1/2}$ both defined on $\mathcal{H}_0$, we have $\Delta_\lambda^{-1/2} C \Delta_\lambda^{-1/2} = A_\lambda B A_\lambda^*$ with B compact and $\|A_\lambda\| \leqslant 1$. Furthermore, let $(\phi_i)_{i\in\mathbb{N}}$ be an orthonormal family of eigenvectors of the compact operator L associated to eigenvalues $(\nu_i)_{i\in\mathbb{N}}$. Then we can write, for any $f \in \mathcal{H}_0$,

$$A_\lambda f = \Delta_\lambda^{-1/2} \Delta^{1/2} f = \sum_{i=0}^{\infty} \sqrt{\frac{\nu_i}{\lambda + \nu_i}} \langle f, \phi_i \rangle_{\mathcal{H}} \, \phi_i \underset{\lambda\to 0}{\longrightarrow} f.$$

Hence by applying Lemma 27, we have the convergence in operator norm of $\Delta_\lambda^{-1/2} C \Delta_\lambda^{-1/2}$ to $\Delta^{-1/2} C \Delta^{-1/2}$, hence in particular the convergence of the norms of the operators. ∎

C.2.2. Introduction of the operator L

In all this section we focus on a distribution $d\mu$ of the form $d\mu(x) = \mathrm{e}^{-V(x)} dx$.

Let us give first a characterization of the function that allows to recover the Poincaré constant, i.e., the function in $H^1(\mu)$ that minimizes $\frac{\int_{\mathbb{R}^d} \|\nabla f(x)\|^2 d\mu(x)}{\int_{\mathbb{R}^d} f(x)^2 d\mu(x) - \left(\int_{\mathbb{R}^d} f(x) d\mu(x)\right)^2}$. We call f_* this function. We recall that we denote by Δ^L the standard Laplacian in $\mathbb{R}^d$: $\forall f \in H^1(\mu)$, $\Delta^L f = \sum_{i=1}^d \frac{\partial^2 f_i}{\partial^2 x i}$. Let us define the operator $\forall f \in H^1(\mu)$, $Lf = -\Delta^L f + \langle \nabla V, \nabla f \rangle$, which is the opposite of the infinitesimal generator of the dynamics (96). We can verify that it is symmetric in $L^2(\mu)$. Indeed by integrations by parts for any

$\forall f, g \in C_c^\infty$,

$$\begin{aligned}
\langle Lf, g\rangle_{L^2(\mu)} &= \int (Lf)(x)g(x)d\mu(x) \\
&= -\int \Delta^L f(x)g(x)\mathrm{e}^{-V(x)}dx + \int \langle \nabla V(x), \nabla f(x)\rangle g(x)\mathrm{e}^{-V(x)}dx \\
&= \int \left\langle \nabla f(x), \nabla\left(g(x)\mathrm{e}^{-V(x)}\right)\right\rangle dx + \int \langle \nabla V(x), \nabla f(x)\rangle g(x)\mathrm{e}^{-V(x)}dx \\
&= \int \langle \nabla f(x), \nabla g(x)\rangle \mathrm{e}^{-V(x)}dx - \int \langle \nabla f(x), \nabla V(x)\rangle g(x)\mathrm{e}^{-V(x)}dx \\
&\qquad + \int \langle \nabla V(x), \nabla f(x)\rangle g(x)\mathrm{e}^{-V(x)}dx \\
&= \int \langle \nabla f(x), \nabla g(x)\rangle d\mu(x).
\end{aligned}$$

The last equality being totally symmetric in f and g, we have the symmetry of the operator L: $\langle Lf, g\rangle_{L^2(\mu)} = \int \langle \nabla f, \nabla g\rangle d\mu = \langle f, Lg\rangle_{L^2(\mu)}$ (for the self-adjointness we refer to [BGL14]). Remark that the same calculation shows that $\nabla^* = -\mathrm{div} + \nabla V \cdot$, hence $L = \nabla^* \cdot \nabla = -\Delta^L + \langle \nabla V, \nabla \cdot\rangle$, where ∇^* is the adjoint of ∇ in $L^2(\mu)$.

Let us call π the orthogonal projector of $L^2(\mu)$ on constant functions: $\pi f : x \in \mathbb{R}^d \mapsto \int f d\mu$. The problem (97) then rewrites:

$$\mathcal{P}^{-1} = \inf_{f \in (H^1(\mu) \cap L_0^2(\mu)) \setminus \{0\}} \frac{\langle Lf, f\rangle_{L^2(\mu)}}{\|(I_{L^2(\mu)} - \pi)f\|^2}, \tag{101}$$

Until the end of this part, to alleviate the notation we omit to mention that the scalar product is the canonical one on $L^2(\mu)$. In the same way, we also denote $\mathbb{1} = I_{L^2(\mu)}$.

Case where $d\mu$ has infinite support. The minimizer of Eq. (101) is not unique but all the minimizer satisfy a eigenvalue property if the potential V goes fast enough at infinity.

Proposition 14 (*Properties of the minimizer*)

If $\lim_{|x|\to\infty} \frac{1}{4}|\nabla V|^2 - \frac{1}{2}\Delta^L V = +\infty$, the problem (101) admits a minimizer in $H^1(\mu)$ and every minimizer f is an eigenvector of L associated with the eigenvalue $\mathcal{P}^{-1}$:

$$Lf = \mathcal{P}^{-1} f. \tag{102}$$

To prove the existence of a minimizer in $H^1(\mu)$, we need the following lemmas.

Lemma 28 (*Criterion for compact embedding of $H^1(\mu)$ in $L^2(\mu)$*)

The injection $H^1(\mu) \hookrightarrow L^2(\mu)$ is compact if and only if the Schrödinger operator $-\Delta^L + \frac{1}{4}|\nabla V|^2 - \frac{1}{2}\Delta^L V$ has compact resolvent.

Proof : See [Gan10, Proposition 1.3] or [RS12, Lemma XIII.65]. ■

Lemma 29 (*A sufficient condition*)

If $\Phi \in C^\infty$ and $\Phi(x) \longrightarrow +\infty$ when $|x| \to \infty$, the Schrödinger operator $-\Delta^L + \Phi$ on $\mathbb{R}^d$ has compact resolvent.

Proof : See [HN05, Section 3] or [RS12, Lemma XIII.67]. ■

Now we can prove Proposition 14.

Proof Proof of Proposition 14 : We first prove that (101) admits a minimizer in $H^1(\mu)$. Indeed, we have,

$$\mathcal{P}^{-1} = \inf_{f\in(H^1\cap L^2_0)\setminus\{0\}} \frac{\langle Lf, f\rangle_{L^2(\mu)}}{\|(\mathbb{1}-\pi)f\|^2} = \inf_{f\in(H^1\cap L^2_0)\setminus\{0\}} J(f), \quad \text{where } J(f) := \frac{\|\nabla f\|^2}{\|f\|^2}.$$

Let $(f_n)_{n\geqslant 0}$ be a sequence of functions in $H^1_0(\mu)$ equipped with the natural H^1-norm such that $(J(f_n))_{n\geqslant 0}$ converges to $\mathcal{P}^{-1}$. As the problem in invariant by rescaling of f, we can assume that $\forall n \geqslant 0,\ \|f_n\|^2_{L^2(\mu)} = 1$. Hence $J(f_n) = \|\nabla f_n\|^2_{L^2(\mu)}$ converges (to $\mathcal{P}^{-1}$). In particular $\|\nabla f_n\|^2_{L^2(\mu)}$ is bounded in $L^2(\mu)$, hence $(f_n)_{n\geqslant 0}$ is bounded in $H^1(\mu)$. Since by Lemma 28 and 29 we have a compact injection of $H^1(\mu)$ in $L^2(\mu)$, it holds, upon extracting a subsequence, that there exists $f \in H^1(\mu)$ such that

$$\begin{cases} f_n \to f & \text{strongly in } L^2(\mu) \\ f_n \rightharpoonup f & \text{weakly in } H^1(\mu). \end{cases}$$

Thanks to the strong $L^2(\mu)$ convergence, $\|f\|^2 = \lim_{n\infty} \|f_n\|^2 = 1$. By the Cauchy-Schwarz inequality and then taking the limit $n \to +\infty$,

$$\|\nabla f\|^2 = \lim_{n\infty} \langle \nabla f_n, \nabla f\rangle \leqslant \lim_{n\infty} \|\nabla f\|\|\nabla f_n\| = \|\nabla f\|\mathcal{P}^{-1}.$$

Therefore, $\|\nabla f\| \leqslant \mathcal{P}^{-1/2}$ which implies that $J(f) \leqslant \mathcal{P}^{-1}$, and so $J(f) = \mathcal{P}^{-1}$. This shows that f is a minimizer of J.

Let us next prove the PDE characterization of minimizers.

A necessary condition on a minimizer f_* of the problem $\inf_{f\in H^1(\mu)}\{\|\nabla f\|_{L^2(\mu)},\ \|f\|^2 = 1\}$ is to satisfy the following Euler-Lagrange equation: there exists $\beta \in \mathbb{R}$ such that:

$$Lf_* + \beta f_* = 0.$$

Plugging this into (101), we have: $\mathcal{P}^{-1} = \langle Lf_*, f_*\rangle = -\beta\langle f_*, f_*\rangle = -\beta\|f_*\|_2^2 = -\beta$. Finally, the equation satisfied by f_* is:

$$Lf = -\Delta^L f_* + \langle \nabla V, \nabla f_*\rangle = \mathcal{P}^{-1} f_*,$$

which concludes the proof. ■

Case where $d\mu$ has compact support We suppose in this section that $d\mu$ has a compact support included in Ω. Without loss of generality we can take a set Ω with a C^∞ smooth boundary $\partial\Omega$. In this case, without changing the result of the variational problem, we can restrict ourselves to functions that vanish at the boundary, namely the Sobolev space $H^1_D(\mathbb{R}^d, d\mu) = \left\{f \in H^1(\mu) \text{ s.t. } f_{|\partial\Omega} = 0\right\}$. Note that, as V is smooth, $H^1(\mu) \supset H^1(\mathbb{R}^d, d\lambda)$ the usual "flat" space equipped with $d\lambda$, the Lebesgue measure. Note also that only in this section the domain of the operator L is $H^2 \cap H^1_D$.

Proposition 15 (*Properties of the minimizer in the compact support case*)

The problem (101) admits a minimizer in H^1_D and every minimizer f satisfies the partial differential equation:

$$Lf = \mathcal{P}^{-1} f. \tag{103}$$

Proof : The proof is exactly the same than the one of Proposition 14 since H^1_D can be compactly injected in L^2 without any additional assumption on V. ■

Let us take in this section $\mathcal{H} = H^d(\mathbb{R}^d,\ d\lambda)$, which is the RKHS associated to the kernel $k(x, x') = \mathrm{e}^{-\|x-x'\|}$. As f_* satisfies (103), from regularity properties of elliptic PDEs, we infer that f_* is $C^\infty(\overline{\Omega})$. By the Whitney extension theorem [Whi34], we can extend f_* defined on $\overline{\Omega}$ to a smooth and compactly supported function in $\Omega' \supset \Omega$ of $\mathbb{R}^d$. Hence $f_* \in C^\infty_c(\mathbb{R}^d) \subset \mathcal{H}$.

Proposition 16

Consider a minimizer f_* of (101). Then

$$\mathcal{P}^{-1} \leqslant \mathcal{P}_\lambda^{-1} \leqslant \mathcal{P}^{-1} + \lambda \frac{\|f_*\|_{\mathcal{H}}^2}{\|f_*\|_{L^2(\mu)}^2}. \tag{104}$$

Proof: First note that f_* has mean zero with respect to $d\mu$. Indeed, $\int f d\mu = \mathcal{P}^{-1} \int Lf d\mu = 0$, by the fact that $d\mu$ is the stationary distribution of the dynamics.

For $\lambda > 0$,

$$\begin{aligned}\mathcal{P}^{-1} \leqslant \mathcal{P}_\lambda^{-1} &= \inf_{f \in \mathcal{H} \setminus \mathbb{R}\mathbb{1}} \frac{\int_{\mathbb{R}^d} \|\nabla f(x)\|^2 d\mu(x) + \lambda \|f\|_{\mathcal{H}}^2}{\int_{\mathbb{R}^d} f(x)^2 d\mu(x) - \left(\int_{\mathbb{R}^d} f(x) d\mu(x)\right)^2} \\ &\leqslant \frac{\int_{\mathbb{R}^d} \|\nabla f_*(x)\|^2 d\mu(x) + \lambda \|f_*\|_{\mathcal{H}}^2}{\int_{\mathbb{R}^d} f_*(x)^2 d\mu(x)} = \mathcal{P}^{-1} + \lambda \frac{\|f_*\|_{\mathcal{H}}^2}{\|f_*\|_{L^2(\mu)}^2},\end{aligned}$$

which provides the result. ■

C.3. Technical inequalities

C.3.1. Concentration inequalities

We first begin by recalling some concentration inequalities for sums of random vectors and operators.

Proposition 17 (*Bernstein's inequality for sums of random vectors*)

Let $z_1, \dots, z_n$ be a sequence of independent identically and distributed random elements of a separable Hilbert space $\mathcal{H}$. Assume that $\mathbb{E}\|z_1\| < +\infty$ and note $\mu = \mathbb{E} z_1$. Let $\sigma, L \geqslant 0$ such that,

$$\forall p \geqslant 2, \qquad \mathbb{E}\,\|z_1 - \mu\|_{\mathcal{H}}^p \leqslant \frac{1}{2} p! \sigma^2 L^{p-2}.$$

Then, for any $\delta \in (0, 1]$,

$$\left\| \frac{1}{n} \sum_{i=1}^n z_i - \mu \right\|_{\mathcal{H}} \leqslant \frac{2L \log(2/\delta)}{n} + \sqrt{\frac{2\sigma^2 \log(2/\delta)}{n}}, \tag{105}$$

with probability at least $1 - \delta$.

Proof: This is a restatement of Theorem 3.3.4 of [Yur95]. ■

Proposition 18 (*Bernstein's inequality for sums of random operators*)

Let $\mathcal{H}$ be a separable Hilbert space and let $X_1, \dots, X_n$ be a sequence of independent and identically distributed self-adjoint random operators on $\mathcal{H}$. Assume that $\mathbb{E}(X_i) = 0$ and that there exist $T > 0$ and S a positive trace-class operator such that $\|X_i\| \leqslant T$ almost surely and $\mathbb{E} X_i^2 \preccurlyeq S$ for any $i \in \{1, \dots, n\}$. Then, for any $\delta \in (0, 1]$, the following inequality holds:

$$\left\| \frac{1}{n} \sum_{i=1}^n X_i \right\| \leqslant \frac{2T\beta}{3n} + \sqrt{\frac{2\|S\|\beta}{n}}, \tag{106}$$

with probability at least $1 - \delta$ and where $\beta = \log \frac{2 \mathrm{Tr} S}{\|S\| \delta}$.

Proof : The theorem is a restatement of Theorem 7.3.1 of [Tro12b] generalized to the separable Hilbert space case by means of the technique in Section 4 of [Sta17]. ∎

C.3.2. Operator bounds

Lemma 30

Under assumptions 2 and 3, Σ, C and Δ are trace-class operators.

Proof : We only prove the result for Δ, the proof for Σ and C being similar. Consider an orthonormal basis $(\phi_i)_{i\in\mathbb{N}}$ of $\mathcal{H}$. Then, as Δ is a positive self adjoint operator,

$$\begin{aligned}\operatorname{tr}\Delta &= \sum_{i=1}^{\infty}\langle\Delta\phi_i,\phi_i\rangle = \sum_{i=1}^{\infty}\mathbb{E}_\mu\left[\sum_{j=1}^{d}\langle\partial_j K_x,\phi_i\rangle^2\right] = \mathbb{E}_\mu\left[\sum_{i=1}^{\infty}\sum_{j=1}^{d}\langle\partial_j K_x,\phi_i\rangle^2\right]\\ &= \mathbb{E}_\mu\left[\sum_{j=1}^{d}\|\partial_j K_x\|^2\right] \leqslant \mathcal{K}_d.\end{aligned}$$

Hence, Δ is a trace-class operator. ∎

The following quantities are useful for the estimates in this section:

$$\mathcal{N}_\infty(\lambda) = \sup_{x\in\operatorname{supp}(\mu)}\left\|\mathsf{L}_\lambda^{-1/2}K_x\right\|_{\mathcal{H}}^2, \text{ and } \quad \mathcal{F}_\infty(\lambda) = \sup_{x\in\operatorname{supp}(\mu)}\left\|\mathsf{L}_\lambda^{-1/2}\nabla K_x\right\|_{\mathcal{H}}^2.$$

Note that under assumption 3, $\mathcal{N}_\infty(\lambda)\leqslant\frac{\mathcal{K}}{\lambda}$ and $\mathcal{F}_\infty(\lambda)\leqslant\frac{\mathcal{K}_d}{\lambda}$. Note also that under refined assumptions on the spectrum of Δ, we could have a better dependence of the latter bounds with respect to λ. Let us now state three useful lemmas to bound the norms of the operators that appear during the proof of Proposition 12.

Lemma 31

For any $\lambda>0$ and any $\delta\in(0,1]$,

$$\begin{aligned}\left\|\mathsf{L}_\lambda^{-1/2}(\widehat{C}-C)\mathsf{L}_\lambda^{-1/2}\right\| \leqslant{}& \frac{4\mathcal{N}_\infty(\lambda)\log\frac{2\operatorname{Tr}\Sigma}{\mathcal{P}_\mu^\lambda\lambda\delta}}{3n} + \left[\frac{2\,\mathcal{P}_\mu^\lambda\,\mathcal{N}_\infty(\lambda)\log\frac{2\operatorname{Tr}\Sigma}{\mathcal{P}_\mu^\lambda\lambda\delta}}{n}\right]^{1/2}\\ &+ 8\mathcal{N}_\infty(\lambda)\left(\frac{\log(\frac{2}{\delta})}{n}+\sqrt{\frac{\log(\frac{2}{\delta})}{n}}\right)\\ &+ 16\mathcal{N}_\infty(\lambda)\left(\frac{\log(\frac{2}{\delta})}{n}+\sqrt{\frac{\log(\frac{2}{\delta})}{n}}\right)^2,\end{aligned}$$

with probability at least $1-\delta$.

Proof Proof of Lemma 31 : We apply some concentration inequality to the operator $\mathsf{L}_\lambda^{-1/2}\widehat{C}\mathsf{L}_\lambda^{-1/2}$ whose mean

is exactly $\mathsf{L}_\lambda^{-1/2} C \mathsf{L}_\lambda^{-1/2}$. The calculation is the following:

$$\begin{aligned}
\left\| \mathsf{L}_\lambda^{-1/2} (\widehat{C} - C) \mathsf{L}_\lambda^{-1/2} \right\| &= \left\| \mathsf{L}_\lambda^{-1/2} \widehat{C} \mathsf{L}_\lambda^{-1/2} - \mathsf{L}_\lambda^{-1/2} C \mathsf{L}_\lambda^{-1/2} \right\| \\
&\leqslant \left\| \mathsf{L}_\lambda^{-1/2} \widehat{\Sigma} \mathsf{L}_\lambda^{-1/2} - \mathsf{L}_\lambda^{-1/2} \Sigma \mathsf{L}_\lambda^{-1/2} \right\| \\
&\qquad + \left\| \mathsf{L}_\lambda^{-1/2} (\widehat{m} \otimes \widehat{m}) \mathsf{L}_\lambda^{-1/2} - \mathsf{L}_\lambda^{-1/2} (m \otimes m) \mathsf{L}_\lambda^{-1/2} \right\| \\
&= \left\| \frac{1}{n} \sum_{i=1}^n \left[(\mathsf{L}_\lambda^{-1/2} K_{x_i}) \otimes (\mathsf{L}_\lambda^{-1/2} K_{x_i}) - \mathsf{L}_\lambda^{-1/2} \Sigma \mathsf{L}_\lambda^{-1/2} \right] \right\| \\
&\qquad + \left\| (\mathsf{L}_\lambda^{-1/2} \widehat{m}) \otimes (\mathsf{L}_\lambda^{-1/2} \widehat{m}) - (\mathsf{L}_\lambda^{-1/2} m) \otimes (\mathsf{L}_\lambda^{-1/2} m) \right\|.
\end{aligned}$$

We estimate the two terms separately.

Bound on the first term: we use Proposition 18. To do this, we bound for $i \in [\![1, n]\!]$:

$$\begin{aligned}
\left\| (\mathsf{L}_\lambda^{-1/2} K_{x_i}) \otimes (\mathsf{L}_\lambda^{-1/2} K_{x_i}) - \mathsf{L}_\lambda^{-1/2} \Sigma \mathsf{L}_\lambda^{-1/2} \right\| &\leqslant \left\| \mathsf{L}_\lambda^{-1/2} K_{x_i} \right\|_{\mathcal{H}}^2 + \left\| \mathsf{L}_\lambda^{-1/2} \Sigma \mathsf{L}_\lambda^{-1/2} \right\| \\
&\leqslant 2 \mathcal{N}_\infty(\lambda),
\end{aligned}$$

and, for the second order moment,

$$\begin{aligned}
&\mathbb{E} \left((\mathsf{L}_\lambda^{-1/2} K_{x_i}) \otimes (\mathsf{L}_\lambda^{-1/2} K_{x_i}) - \mathsf{L}_\lambda^{-1/2} \Sigma \mathsf{L}_\lambda^{-1/2} \right)^2 \\
&\quad = \mathbb{E} \left[\left\| \mathsf{L}_\lambda^{-1/2} K_{x_i} \right\|_{\mathcal{H}}^2 (\mathsf{L}_\lambda^{-1/2} K_{x_i}) \otimes (\mathsf{L}_\lambda^{-1/2} K_{x_i}) \right] - \mathsf{L}_\lambda^{-1/2} \Sigma \mathsf{L}_\lambda^{-1} \Sigma \mathsf{L}_\lambda^{-1/2} \\
&\quad \preccurlyeq \mathcal{N}_\infty(\lambda) \mathsf{L}_\lambda^{-1/2} \Sigma \mathsf{L}_\lambda^{-1/2}.
\end{aligned}$$

We conclude this first part of the proof by some estimation of the constant $\beta = \log \frac{2 \operatorname{Tr}(\Sigma \mathsf{L}_\lambda^{-1})}{\left\| \mathsf{L}_\lambda^{-1/2} \Sigma \mathsf{L}_\lambda^{-1/2} \right\| \delta}$. Using $\operatorname{Tr} \Sigma \mathsf{L}_\lambda^{-1} \leqslant \lambda^{-1} \operatorname{Tr} \Sigma$, it holds $\beta \leqslant \log \frac{2 \operatorname{Tr} \Sigma}{\mathcal{P}_\mu^\lambda \lambda \delta}$. Therefore,

$$\begin{aligned}
&\left\| \frac{1}{n} \sum_{i=1}^n \left[(\mathsf{L}_\lambda^{-1/2} K_{x_i}) \otimes (\mathsf{L}_\lambda^{-1/2} K_{x_i}) - \mathsf{L}_\lambda^{-1/2} \Sigma \mathsf{L}_\lambda^{-1/2} \right] \right\| \\
&\qquad \leqslant \frac{4 \mathcal{N}_\infty(\lambda) \log \frac{2 \operatorname{Tr} \Sigma}{\mathcal{P}_\mu^\lambda \lambda \delta}}{3n} + \left[\frac{2 \, \mathcal{P}_\mu^\lambda \, \mathcal{N}_\infty(\lambda) \log \frac{2 \operatorname{Tr} \Sigma}{\mathcal{P}_\mu^\lambda \lambda \delta}}{n} \right]^{1/2}.
\end{aligned}$$

Bound on the second term. Denote by $v = \mathsf{L}_\lambda^{-1/2} m$ and $\widehat{v} = \mathsf{L}_\lambda^{-1/2} \widehat{m}$. A simple calculation leads to

$$\begin{aligned}
\| \widehat{v} \otimes \widehat{v} - v \otimes v \| &\leqslant \| v \otimes (\widehat{v} - v) \| + \| (\widehat{v} - v) \otimes v \| + \| (\widehat{v} - v) \otimes (\widehat{v} - v) \| \\
&\leqslant 2 \|v\| \|\widehat{v} - v\| + \|\widehat{v} - v\|^2.
\end{aligned}$$

We bound $\|\widehat{v} - v\|$ with Proposition 17. It holds: $\widehat{v} - v = \Delta_\lambda^{-1/2} (\widehat{m} - m) = \frac{1}{n} \sum_{i=1}^n \Delta_\lambda^{-1/2} (K_{x_i} - m) = \frac{1}{n} \sum_{i=1}^n Z_i$, with $Z_i = \Delta_\lambda^{-1/2} (K_{x_i} - m)$. Obviously for any $i \in [\![1, n]\!]$, $\mathbb{E}(Z_i) = 0$, and $\|Z_i\| \leqslant \|\Delta_\lambda^{-1/2} K_{x_i}\| + \|\Delta_\lambda^{-1/2} m\| \leqslant 2 \sqrt{\mathcal{N}_\infty(\lambda)}$. Furthermore,

$$\begin{aligned}
\mathbb{E} \|Z_i\|^2 = \mathbb{E} \left\langle \Delta_\lambda^{-1/2} (K_{x_i} - m), \Delta_\lambda^{-1/2} (K_{x_i} - m) \right\rangle &= \mathbb{E} \left\| \Delta_\lambda^{-1/2} K_{x_i} \right\|^2 - \left\| \Delta_\lambda^{-1/2} m \right\|^2 \\
&\leqslant \mathcal{N}_\infty(\lambda).
\end{aligned}$$

Thus, for $p \geqslant 2$,

$$\mathbb{E} \|Z_i\|^p \leqslant \mathbb{E} \left(\|Z_i\|^{p-2} \|Z_i\|^2 \right) \leqslant \frac{1}{2} p! \left(\sqrt{\mathcal{N}_\infty(\lambda)} \right)^2 \left(2 \sqrt{\mathcal{N}_\infty(\lambda)} \right)^{p-2},$$

hence, by applying Proposition 17 with $L = 2\sqrt{\mathcal{N}_\infty(\lambda)}$ and $\sigma = \sqrt{\mathcal{N}_\infty(\lambda)}$,

$$\begin{aligned}\|\widehat{v} - v\| &\leqslant \frac{4\sqrt{\mathcal{N}_\infty(\lambda)}\log(2/\delta)}{n} + \sqrt{\frac{2\mathcal{N}_\infty(\lambda)\log(2/\delta)}{n}} \\ &\leqslant 4\sqrt{\mathcal{N}_\infty(\lambda)}\left(\frac{\log(2/\delta)}{n} + \sqrt{\frac{\log(2/\delta)}{n}}\right).\end{aligned}$$

Finally, as $\|v\| \leqslant \sqrt{\mathcal{N}_\infty(\lambda)}$,

$$\begin{aligned}\|\widehat{v}\otimes\widehat{v} - v\otimes v\| \leqslant\ & 8\mathcal{N}_\infty(\lambda)\left(\frac{\log(2/\delta)}{n} + \sqrt{\frac{\log(2/\delta)}{n}}\right) \\ &+ 16\mathcal{N}_\infty(\lambda)\left(\frac{\log(2/\delta)}{n} + \sqrt{\frac{\log(2/\delta)}{n}}\right)^2.\end{aligned}$$

This concludes the proof of Lemma 31. ∎

Lemma 32

For any $\lambda \in (0, \|\Delta\|]$ and any $\delta \in (0,1]$,

$$\left\|\mathsf{L}_\lambda^{-1/2}(\widehat{\mathsf{L}} - \mathsf{L})\mathsf{L}_\lambda^{-1/2}\right\| \leqslant \frac{4\mathcal{F}_\infty(\lambda)\log\frac{4\operatorname{Tr}\mathsf{L}}{\lambda\delta}}{3n} + \sqrt{\frac{2\,\mathcal{F}_\infty(\lambda)\log\frac{4\operatorname{Tr}\mathsf{L}}{\lambda\delta}}{n}},$$

with probability at least $1-\delta$.

Proof Proof of Lemma 32 : As in the proof of Lemma 31, we want to apply some concentration inequality to the operator $\mathsf{L}_\lambda^{-1/2}\widehat{\Delta}\mathsf{L}_\lambda^{-1/2}$, whose mean is exactly $\mathsf{L}_\lambda^{-1/2}\Delta\mathsf{L}_\lambda^{-1/2}$. The proof is almost the same as Lemma 31. We start by writing

$$\begin{aligned}\left\|\mathsf{L}_\lambda^{-1/2}(\widehat{\mathsf{L}} - \mathsf{L})\mathsf{L}_\lambda^{-1/2}\right\| &= \left\|\mathsf{L}_\lambda^{-1/2}\widehat{\mathsf{L}}\mathsf{L}_\lambda^{-1/2} - \mathsf{L}_\lambda^{-1/2}\mathsf{L}\mathsf{L}_\lambda^{-1/2}\right\| \\ &= \left\|\frac{1}{n}\sum_{i=1}^n\left[(\mathsf{L}_\lambda^{-1/2}\nabla K_{x_i})\otimes(\mathsf{L}_\lambda^{-1/2}\nabla K_{x_i}) - \mathsf{L}_\lambda^{-1/2}\Delta\mathsf{L}_\lambda^{-1/2}\right]\right\|.\end{aligned}$$

In order to use Proposition 18, we bound for $i \in [\![1,n]\!]$,

$$\begin{aligned}\left\|(\mathsf{L}_\lambda^{-1/2}\nabla K_{x_i})\otimes(\mathsf{L}_\lambda^{-1/2}\nabla K_{x_i}) - \mathsf{L}_\lambda^{-1/2}\mathsf{L}\mathsf{L}_\lambda^{-1/2}\right\| &\leqslant \left\|\mathsf{L}_\lambda^{-1/2}\nabla K_{x_i}\right\|_{\mathcal{H}}^2 + \left\|\mathsf{L}_\lambda^{-1/2}\mathsf{L}\mathsf{L}_\lambda^{-1/2}\right\| \\ &\leqslant 2\mathcal{F}_\infty(\lambda),\end{aligned}$$

and, for the second order moment,

$$\begin{aligned}&\mathbb{E}\left[\left((\mathsf{L}_\lambda^{-1/2}\nabla K_{x_i})\otimes(\mathsf{L}_\lambda^{-1/2}\nabla K_{x_i}) - \mathsf{L}_\lambda^{-1/2}\mathsf{L}\mathsf{L}_\lambda^{-1/2}\right)^2\right] \\ &= \mathbb{E}\left[\left\|\mathsf{L}_\lambda^{-1/2}\nabla K_{x_i}\right\|_{\mathcal{H}}^2(\mathsf{L}_\lambda^{-1/2}\nabla K_{x_i})\otimes(\mathsf{L}_\lambda^{-1/2}\nabla K_{x_i})\right] - \mathsf{L}_\lambda^{-1/2}\mathsf{L}\mathsf{L}_\lambda^{-1}\mathsf{L}\mathsf{L}_\lambda^{-1/2} \\ &\preccurlyeq \mathcal{F}_\infty(\lambda)\mathsf{L}_\lambda^{-1/2}\mathsf{L}\mathsf{L}_\lambda^{-1/2}.\end{aligned}$$

We conclude by some estimation of $\beta = \log\frac{2\operatorname{Tr}(\mathsf{L}\mathsf{L}_\lambda^{-1})}{\|\mathsf{L}_\lambda^{-1}\mathsf{L}\|\delta}$. Since $\operatorname{Tr}(\mathsf{L}\mathsf{L}_\lambda^{-1}) \leqslant \lambda^{-1}\operatorname{Tr}\mathsf{L}$ and for $\lambda \leqslant \|\mathsf{L}\|$, $\|\mathsf{L}_\lambda^{-1}\mathsf{L}\| \geqslant 1/2$, it follow that $\beta \leqslant \log\frac{4\operatorname{Tr}\mathsf{L}}{\lambda\delta}$. The conclusion then follows from (106). ∎

Lemma 33 (*Bounding operators*)

For any $\lambda > 0$, $\delta \in (0,1)$, and $n \geqslant 15\mathcal{F}_\infty(\lambda)\log\frac{4\operatorname{Tr}\mathsf{L}}{\lambda\delta}$, it holds with probability at least $1-\delta$:

$$\left\|\widehat{\mathsf{L}}_\lambda^{-1/2}\mathsf{L}_\lambda^{1/2}\right\|^2 \leqslant 2,$$

The proof of this result relies on the following lemma (see proof in [RR17, Proposition 8]).

Lemma 34

Let $\mathcal{H}$ be a separable Hilbert space, A and B two bounded self-adjoint positive linear operators on $\mathcal{H}$ and $\lambda > 0$. Then

$$\left\|(A+\lambda I)^{-1/2}(B+\lambda I)^{1/2}\right\| \leqslant (1-\beta)^{-1/2},$$

with $\beta = \lambda_{\max}\left((B+\lambda I)^{-1/2}(B-A)(B+\lambda I)^{-1/2}\right) < 1$, where $\lambda_{\max}(O)$ is the largest eigenvalue of the self-adjoint operator O.

We can now write the proof of Lemma 33.

Proof Proof of Lemma 33 : Thanks to Lemma 34, we see that

$$\left\|\widehat{\mathsf{L}}_\lambda^{-1/2}\mathsf{L}_\lambda^{1/2}\right\|^2 \leqslant \left(1-\lambda_{\max}\left(\mathsf{L}_\lambda^{-1/2}(\widehat{\Delta}-\Delta)\mathsf{L}_\lambda^{-1/2}\right)\right)^{-1},$$

and as $\left\|\mathsf{L}_\lambda^{-1/2}(\widehat{\Delta}-\Delta)\mathsf{L}_\lambda^{-1/2}\right\| < 1$, we have:

$$\left\|\widehat{\mathsf{L}}_\lambda^{-1/2}\mathsf{L}_\lambda^{1/2}\right\|^2 \leqslant \left(1-\left\|\mathsf{L}_\lambda^{-1/2}(\widehat{\Delta}-\Delta)\mathsf{L}_\lambda^{-1/2}\right\|\right)^{-1}.$$

We can then apply the bound of Lemma 32 to obtain that, if λ is such that $\frac{4\mathcal{F}_\infty(\lambda)\log\frac{4\operatorname{Tr}\mathsf{L}}{\lambda\delta}}{3n}+\sqrt{\frac{2\,\mathcal{F}_\infty(\lambda)\log\frac{4\operatorname{Tr}\mathsf{L}}{\lambda\delta}}{n}} \leqslant \frac{1}{2}$, then $\left\|\widehat{\mathsf{L}}_\lambda^{-1/2}\mathsf{L}_\lambda^{1/2}\right\|^2 \leqslant 2$ with probability $1-\delta$. The condition on λ is satisfied when $n \geqslant 15\mathcal{F}_\infty(\lambda)\log\frac{4\operatorname{Tr}\mathsf{L}}{\lambda\delta}$. ■

C.4. Calculation of the bias in the Gaussian case

We can derive a rate of convergence when μ is a one-dimensional Gaussian. Hence, we consider the one-dimensional distribution $d\mu$ as the normal distribution with mean zero and variance $1/(4a)$. Let $b > 0$, we consider also the following approximation $\mathcal{P}_\kappa^{-1} = \inf_{f\in\mathcal{H}} \frac{\mathbb{E}_\mu(f'^2)+\kappa\|f\|_{\mathcal{H}}^2}{\operatorname{var}_\mu(f)}$ where $\mathcal{H}$ is the RKHS associated with the Gaussian kernel $\exp(-b(x-y)^2)$. Our goal is to study how $\mathcal{P}_\kappa$ tends to $\mathcal{P}$ when κ tends to zero.

Proposition 19 (*Rate of convergence for the bias in the one-dimensional Gaussian case*)

If $d\mu$ is a one-dimensional Gaussian of mean zero and variance $1/(4a)$ there exists $A > 0$ such that, if $\lambda \leqslant A$, it holds

$$\mathcal{P}^{-1} \leqslant \mathcal{P}_\lambda^{-1} \leqslant \mathcal{P}^{-1}(1+B\lambda\ln^2(1/\lambda)), \tag{107}$$

where A and B depend only on the constant a.

We will show it by considering a specific orthonormal basis of $L^2(\mu)$, where all operators may be expressed simply in closed form.

C.4.1. An orthonormal basis of $L^2(\mu)$ and $\mathcal{H}$

We begin by giving an explicit a basis of $L^2(\mu)$ which is also a basis of $\mathcal{H}$.

Proposition 20 (*Explicit basis*)

We consider

$$f_i(x) = \left(\frac{c}{a}\right)^{1/4}\left(2^i i!\right)^{-1/2}\mathrm{e}^{-(c-a)x^2}H_i\left(\sqrt{2c}x\right),$$

where H_i is the i-th Hermite polynomial, and $c = \sqrt{a^2 + 2ab}$. Then,

- *$(f_i)_{i \geqslant 0}$ is an orthonormal basis of $L^2(\mu)$;*
- *$\tilde{f}_i = \lambda_i^{1/2} f_i$ forms an orthonormal basis of $\mathcal{H}$, with $\lambda_i = \sqrt{\frac{2a}{a+b+c}} \left(\frac{b}{a+b+c}\right)^i$.*

Proof : We can check that this is indeed an orthonormal basis of $L^2(\mu)$:

$$\begin{aligned}
\langle f_k, f_m \rangle_{L^2(\mu)} &= \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi/4a}} \mathrm{e}^{-2ax^2} \left(\frac{c}{a}\right)^{1/2} \mathrm{e}^{-2(c-a)x^2} \left(2^k k!\right)^{-1/2} \left(2^m m!\right)^{-1/2} H_k(\sqrt{2c}x) H_m(\sqrt{2c}x) dx \\
&= \sqrt{2c/\pi} \left(2^k k!\right)^{-1/2} \left(2^m m!\right)^{-1/2} \int_{\mathbb{R}} \mathrm{e}^{-2cx^2} H_k(\sqrt{2c}x) H_m(\sqrt{2c}x) dx \\
&= \delta_{mk},
\end{aligned}$$

using properties of Hermite polynomials. Considering the integral operator $T : L^2(\mu) \to L^2(\mu)$, defined as $Tf(y) = \int_{\mathbb{R}} \mathrm{e}^{-b(x-y)^2} f(x) d\mu(x)$, we have:

$$\begin{aligned}
Tf_k(y) &= \left(\frac{c}{a}\right)^{1/4} \left(2^k k!\right)^{-1/2} \int_{\mathbb{R}} \mathrm{e}^{-(c-a)x^2} H_k(\sqrt{2c}x) \frac{1}{\sqrt{2\pi/4a}} \mathrm{e}^{-2ax^2} \mathrm{e}^{-b(x-y)^2} dx \\
&= \left(\frac{c}{a}\right)^{1/4} \left(2^k k!\right)^{-1/2} \mathrm{e}^{-by^2} \frac{1}{\sqrt{2\pi/4a}} \frac{1}{\sqrt{2c}} \int_{\mathbb{R}} \mathrm{e}^{-(a+b+c)x^2} H_k(\sqrt{2c}x) \mathrm{e}^{2bxy} \sqrt{2c} dx \\
&= \left(\frac{c}{a}\right)^{1/4} \left(2^k k!\right)^{-1/2} \mathrm{e}^{-by^2} \frac{1}{\sqrt{2\pi/4a}} \frac{1}{\sqrt{2c}} \int_{\mathbb{R}} \mathrm{e}^{-\frac{a+b+c}{2c}x^2} H_k(x) \mathrm{e}^{\frac{2b}{\sqrt{2c}}xy} dx.
\end{aligned}$$

We consider u such that $\frac{1}{1-u^2} = \frac{a+b+c}{2c}$, that is, $1 - \frac{2c}{a+b+c} = \frac{a+b-c}{a+b+c} = \frac{b^2}{(a+b+c)^2} = u^2$, which implies that $u = \frac{b}{a+b+c}$; and then $\frac{2u}{1-u^2} = \frac{b}{c}$.

Thus, using properties of Hermite polynomials (see Section C.4.4), we get:

$$\begin{aligned}
Tf_k(y) &= \left(\frac{c}{a}\right)^{1/4} \left(2^k k!\right)^{-1/2} \mathrm{e}^{-by^2} \frac{1}{\sqrt{2\pi/4a}} \frac{1}{\sqrt{2c}} \sqrt{\pi} \sqrt{1-u^2} H_k(\sqrt{2c}y) \exp\left(\frac{u^2}{1-u^2} 2cy^2\right) u^k \\
&= \left(\frac{c}{a}\right)^{1/4} \left(2^k k!\right)^{-1/2} \frac{1}{\sqrt{2\pi/4a}} \frac{1}{\sqrt{2c}} \sqrt{\pi} \frac{\sqrt{2c}}{\sqrt{a+b+c}} H_k(\sqrt{2c}y) \exp(buy^2 - by^2) u^k \\
&= \left(\frac{c}{a}\right)^{1/4} \left(2^k k!\right)^{-1/2} \frac{\sqrt{2a}}{\sqrt{a+b+c}} H_k(\sqrt{2c}y) \exp\left(-by^2 + 2cy^2 \left(-1 + \frac{1}{1-u^2}\right)\right) u^k \\
&= \frac{\sqrt{2a}}{\sqrt{a+b+c}} \left(\frac{b}{a+b+c}\right)^k f_k(y) \\
&= \lambda_k f_k(y).
\end{aligned}$$

This implies that $(\tilde{f}_i)$ is an orthonormal basis of $\mathcal{H}$. ■

We can now rewrite our problem in this basis, which is the purpose of the following lemma:

Lemma 35 (*Reformulation of the problem in the basis*)

Let $(\alpha_i)_i \in \ell^2(\mathbb{N})$. For $f = \sum_{i=0}^{\infty} \alpha_i f_i$, we have:

- $\|f\|_{\mathcal{H}}^2 = \sum_{i=0}^{\infty} \alpha_i^2 \lambda_i^{-1} = \alpha^\top \operatorname{Diag}(\lambda)^{-1} \alpha$;
- $\operatorname{var}_\mu(f(x)) = \sum_{i=0}^{\infty} \alpha_i^2 - \left(\sum_{i=0}^{\infty} \eta_i \alpha_i\right)^2 = \alpha^\top (I - \eta\eta^\top) \alpha$;

- $\mathbb{E}_\mu f'(x)^2 = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \alpha_i\alpha_j (M^\top M)_{ij} = \alpha^\top M^\top M\alpha,$

where η is the vector of coefficients of $\mathbf{1}_{L^2(\mu)}$ and M the matrix of coordinates of the derivative operator in the (f_i) basis. The problem can be rewritten under the following form:

$$\mathcal{P}_\kappa^{-1} = \inf_\alpha \frac{\alpha^\top (M^\top M + \kappa \operatorname{Diag}(\lambda)^{-1})\alpha}{\alpha^\top (I - \eta\eta^\top)\alpha}, \tag{108}$$

where

- $\forall k \geqslant 0, \eta_{2k} = \left(\frac{c}{a}\right)^{1/4} \sqrt{\frac{2a}{a+c}} \left(\frac{b}{a+b+c}\right)^k \frac{\sqrt{(2k)!}}{2^k k!}$ *and* $\eta_{2k+1} = 0$
- $\forall i \in \mathbb{N}, \left(M^\top M\right)_{ii} = \frac{1}{c}\left(2i(a^2+c^2) + (a-c)^2\right)$ *and* $\left(M^\top M\right)_{i,i+2} = \frac{1}{c}\left((a^2-c^2)\sqrt{(i+1)(i+2)}\right).$

Proof : Covariance operator. Since (f_i) is orthonormal for $L^2(\mu)$, we only need to compute for each i, $\eta_i = \mathbb{E}_\mu f_i(x)$, as follows (and using properties of Hermite polynomials):

$$\begin{aligned}
\eta_i = \langle 1, f_i\rangle_{L^2(\mu)} &= \left(\frac{c}{a}\right)^{1/4} \left(2^i i!\right)^{-1/2} \int_{\mathbb{R}} \mathrm{e}^{-(c-a)x^2} H_i(\sqrt{2c}x)\mathrm{e}^{-2ax^2}\sqrt{2a/\pi}dx \\
&= \left(\frac{c}{a}\right)^{1/4} \left(2^i i!\right)^{-1/2} \sqrt{a/(\pi c)} \int_{\mathbb{R}} \mathrm{e}^{-\frac{a+c}{2c}x^2} H_i(x)dx \\
&= \left(\frac{c}{a}\right)^{1/4} \left(2^i i!\right)^{-1/2} \sqrt{\frac{2a}{a+c}} \left(\frac{c-a}{c+a}\right)^{i/2} H_i(0)\mathrm{i}^i.
\end{aligned}$$

This is only non-zero for i even, and

$$\begin{aligned}
\eta_{2k} &= \left(\frac{c}{a}\right)^{1/4} \left(2^{2k}(2k)!\right)^{-1/2} \sqrt{\frac{2a}{a+c}} \left(\frac{c-a}{c+a}\right)^k H_{2k}(0)(-1)^k \\
&= \left(\frac{c}{a}\right)^{1/4} \left(2^{2k}(2k)!\right)^{-1/2} \sqrt{\frac{2a}{a+c}} \left(\frac{c-a}{c+a}\right)^k \frac{(2k)!}{k!} \\
&= \left(\frac{c}{a}\right)^{1/4} \sqrt{\frac{2a}{a+c}} \left(\frac{c-a}{c+a}\right)^k \frac{\sqrt{(2k)!}}{2^k k!} \\
&= \left(\frac{c}{a}\right)^{1/4} \sqrt{\frac{2a}{a+c}} \left(\frac{b}{a+b+c}\right)^k \frac{\sqrt{(2k)!}}{2^k k!}.
\end{aligned}$$

Note that we must have $\sum_{i=0}^\infty \eta_i^2 = \|1\|^2_{L^2(\mu)} = 1$, which can indeed be checked —the shrewd reader will recognize the entire series development of $(1-z^2)^{-1/2}$.

Derivatives. We have, using the recurrence properties of Hermite polynomials:

$$f_i' = \frac{a-c}{\sqrt{c}}\sqrt{i+1} f_{i+1} + \frac{a+c}{\sqrt{c}}\sqrt{i} f_{i-1},$$

for $i > 0$, while for $i = 0$, $f_0' = \frac{a-c}{\sqrt{c}} f_1$. Thus, if M is the matrix of coordinates of the derivative operator in the basis (f_i), we have $M_{i+1,i} = \frac{a-c}{\sqrt{c}}\sqrt{i+1}$ and $M_{i-1,i} = \frac{a+c}{\sqrt{c}}\sqrt{i}$. This leads to

$$\langle f_i', f_j'\rangle_{L^2(\mu)} = (M^\top M)_{ij}.$$

We have

$$\begin{aligned}(M^\top M)_{ii} &= \langle f_i', f_i' \rangle_{L^2(\mu)} \\ &= \frac{1}{c}\Big((i+1)(a-c)^2 + i(a+c)^2\Big) \\ &= \frac{1}{c}\Big(2i(a^2+c^2) + (a-c)^2\Big) \text{ for } i \geqslant 0, \\ (M^\top M)_{i,i+2} &= \langle f_i', f_{i+2}' \rangle_{L^2(\mu)} \\ &= \frac{1}{c}\Big((a^2-c^2)\sqrt{(i+1)(i+2)}\Big) \text{ for } i \geqslant 0.\end{aligned}$$

Note that we have $M\eta = 0$ as these are the coordinates of the derivative of the constant function (this can be checked directly by computing $(M\eta)_{2k+1} = M_{2k+1,2k}\eta_{2k} + M_{2k+1,2k+2}\eta_{2k+2}$). ■

C.4.2. Unregularized solution

Recall that we want to solve $\mathcal{P}^{-1} = \inf_f \frac{\mathbb{E}_\mu f'(x)^2}{\mathrm{var}_\mu(f(x))}$,. The following lemma characterizes the optimal solution completely.

Lemma 36 (*Optimal solution for one dimensional Gaussian*)

We know that the solution of the Poincaré problem is $\mathcal{P}^{-1} = 4a$ which is attained for $f_*(x) = x$. The decomposition of f_* is the basis $(f_i)_i$ is given by $f_* = \sum_{i\geqslant 0} \nu_i f_i$, where $\forall k \geqslant 0$, $\nu_{2k} = 0$ and $\nu_{2k+1} = \left(\frac{c}{a}\right)^{1/4} \frac{\sqrt{a}}{2c}\left(\frac{2c}{a+c}\right)^{3/2}\left(\frac{b}{a+b+c}\right)^k \frac{\sqrt{(2k+1)!}}{2^k k!}$.

Proof : We thus need to compute:

$$\begin{aligned}\nu_i &= \langle f_*, f_i \rangle_{L^2(\mu)} \\ &= \left(\frac{c}{a}\right)^{1/4}\left(2^i i!\right)^{-1/2}\int_{\mathbb{R}} \mathrm{e}^{-(c-a)x^2} H_i(\sqrt{2c}x)\mathrm{e}^{-2ax^2}\sqrt{2a/\pi}x dx \\ &= \left(\frac{c}{a}\right)^{1/4}\left(2^i i!\right)^{-1/2}\sqrt{2a/\pi}\int_{\mathbb{R}} \mathrm{e}^{-(c+a)x^2} H_i(\sqrt{2c}x)x dx \\ &= \left(\frac{c}{a}\right)^{1/4}\left(2^i i!\right)^{-1/2}\sqrt{2a/\pi}\frac{1}{2c}\int_{\mathbb{R}} \mathrm{e}^{-\frac{c+a}{2c}x^2} H_i(x)x dx \\ &= \left(\frac{c}{a}\right)^{1/4}\left(2^i i!\right)^{-1/2}\sqrt{2a/\pi}\frac{1}{4c}\int_{\mathbb{R}} \mathrm{e}^{-\frac{c+a}{2c}x^2} [H_{i+1}(x) + 2iH_{i-1}(x)] dx \\ &= \left(\frac{c}{a}\right)^{1/4}\left(2^i i!\right)^{-1/2}\sqrt{2a/\pi}\frac{\sqrt{\pi}}{4c}\sqrt{\frac{2c}{a+c}}\Big(\big(\frac{c-a}{c+a}\big)^{(i+1)/2} H_{i+1}(0)\mathrm{i}^{i+1} \\ &\qquad + 2i\big(\frac{c-a}{c+a}\big)^{(i-1)/2} H_{i-1}(0)\mathrm{i}^{i-1}\Big),\end{aligned}$$

which is only non-zero for i odd. We have:

$$\begin{aligned}
\nu_{2k+1} &= \Big(\frac{c}{a}\Big)^{1/4}\big(2^{2k+1}(2k+1)!\big)^{-1/2}\sqrt{2a/\pi}\frac{\sqrt{\pi}}{4c}\sqrt{\frac{2c}{a+c}}\Big(\big(\frac{c-a}{c+a}\big)^{k+1}H_{2k+2}(0)(-1)^{k+1} \\
&\qquad\qquad + 2(2k+1)\big(\frac{c-a}{c+a}\big)^{k}H_{2k}(0)(-1)^{k}\Big) \\
&= \Big(\frac{c}{a}\Big)^{1/4}\big(2^{2k+1}(2k+1)!\big)^{-1/2}\sqrt{2a/\pi}\frac{\sqrt{\pi}}{4c}\sqrt{\frac{2c}{a+c}}\Big(\big(\frac{c-a}{c+a}\big)^{k+1}H_{2k+2}(0)(-1)^{k+1} \\
&\qquad\qquad + 2(2k+1)\big(\frac{c-a}{c+a}\big)^{k}H_{2k}(0)(-1)^{k}\Big) \\
&= \Big(\frac{c}{a}\Big)^{1/4}\big(2^{2k+1}(2k+1)!\big)^{-1/2}\sqrt{2a/\pi}\frac{\sqrt{\pi}}{4c}\sqrt{\frac{2c}{a+c}}\big(\frac{c-a}{c+a}\big)^{k}(-1)^{k} \\
&\qquad\qquad \Big(\big(\frac{c-a}{c+a}\big)2(2k+1)H_{2k}(0) + 2(2k+1)H_{2k}(0)\Big) \\
&= \Big(\frac{c}{a}\Big)^{1/4}\big(2^{2k+1}(2k+1)!\big)^{-1/2}\sqrt{2a/\pi}\frac{\sqrt{\pi}}{4c}\sqrt{\frac{2c}{a+c}}\big(\frac{c-a}{c+a}\big)^{k}(-1)^{k}2(2k+1)H_{2k}(0)\frac{2c}{c+a} \\
&= \Big(\frac{c}{a}\Big)^{1/4}\big(2^{2k+1}(2k+1)!\big)^{-1/2}\sqrt{a}\frac{1}{c\sqrt{2}}\big(\frac{2c}{a+c}\big)^{3/2}\big(\frac{c-a}{c+a}\big)^{k}(-1)^{k}(2k+1)H_{2k}(0) \\
&= \Big(\frac{c}{a}\Big)^{1/4}\big(2^{2k+1}(2k+1)!\big)^{-1/2}\sqrt{a}\frac{1}{c\sqrt{2}}\big(\frac{2c}{a+c}\big)^{3/2}\big(\frac{c-a}{c+a}\big)^{k}(2k+1)\frac{(2k)!}{k!} \\
&= \Big(\frac{c}{a}\Big)^{1/4}\frac{\sqrt{a}}{2c}\big(\frac{2c}{a+c}\big)^{3/2}\big(\frac{c-a}{c+a}\big)^{k}\frac{\sqrt{(2k+1)!}}{2^k k!} \\
&= \Big(\frac{c}{a}\Big)^{1/4}\frac{\sqrt{a}}{2c}\big(\frac{2c}{a+c}\big)^{3/2}\big(\frac{b}{a+b+c}\big)^{k}\frac{\sqrt{(2k+1)!}}{2^k k!}.
\end{aligned}$$

■

Note that we have:

$$\begin{aligned}
\mu^\top\nu &= \langle 1, f_*\rangle_{L^2(\mu)} = 0 \\
\|\nu\|^2 &= \|f_*\|^2_{L^2(\mu)} = \frac{1}{4a} \\
M^\top M\nu &= 4a\nu.
\end{aligned}$$

The first equality if obvious from the odd/even sparsity patterns. The third one can be checked directly. The second one can probably be checked by another shrewd entire series development.

If we had $\nu^\top \operatorname{Diag}(\lambda)^{-1}\nu$ finite, then we would have

$$\mathcal{P}^{-1} \leqslant \mathcal{P}_\kappa^{-1} \leqslant \mathcal{P}^{-1}\left(1 + \kappa\cdot\nu^\top\operatorname{Diag}(\lambda)^{-1}\nu\right),$$

which would very nice and simple. Unfortunately, this is not true (see below).

Some further properties for ν We have: $\frac{c-a}{c+a} = \frac{b}{a+b+c}$, and the following equivalent $\frac{\sqrt{\sqrt{k}(2k/e)^{2k+1}}}{2^k\sqrt{k}(k/e)^k} \sim \frac{k^{1/4+k+1/2}}{k^{k+1/2}} \sim k^{1/4}$ (up to constants). Thus

$$|\nu_{2k+1}^2\lambda_{2k+1}^{-1}| \leqslant \Big(\frac{c}{a}\Big)^{1/2}\frac{a}{c^2}\big(\frac{2c}{a+c}\big)^3\big(\frac{b}{a+b+c}\big)^{2k-2k-1}\sqrt{\frac{a+b+c}{2a}}\sqrt{k} = \Theta(\sqrt{k})$$

hence,

$$\sum_{k=0}^{2m+1}\nu_k^2\lambda_k^{-1} \sim \Theta(m^{3/2}).$$

Consequently, $\nu^\top \operatorname{Diag}(\lambda)^{-1}\nu = +\infty$.

Note that we have the extra recursion

$$\nu_k = \frac{1}{\sqrt{4c}}\big[\sqrt{k+1}\eta_{k+1} + \sqrt{k}\eta_{k-1}\big].$$

C.4.3. Truncation

We are going to consider a truncated version α, of ν, with only the first $2m+1$ elements. That is $\alpha_k = \nu_k$ for $k \leqslant 2m+1$ and 0 otherwise.

Lemma 37 (*Convergence of the truncation*)

Consider $g^m = \sum_{k=0}^{\infty} \alpha_k f_k = \sum_{k=0}^{2m+1} \nu_k f_k$, recall that $u = \frac{b}{a+b+c}$. For $m \geqslant max\{-\frac{3}{4\ln u}, \frac{1}{6c}\}$, we have the following:

(i) $\big|\|\alpha\|^2 - \frac{1}{4a}\big| \leqslant Lmu^{2m}$

(ii) $\alpha^\top \eta = 0$

(iii) $\big|\alpha^\top M^\top M\alpha - 1\big| \leqslant Lm^2u^{2m}$

(iv) $\alpha^\top \operatorname{Diag}(\lambda)^{-1}\alpha \leqslant Lm^{3/2}$,

where L depends only on a, b, c.

Proof : We show successively the four estimations.

(i) Let us calculate $\|\alpha\|^2$. We have: $\|\alpha\|^2 - \frac{1}{4a} = \|\alpha\|^2 - \|\nu\|^2 = \sum_{k=m+1}^{\infty} \nu_{2k+1}^2$. Recall that $u = \frac{b}{a+b+c} \leqslant 1$, by noting $A = \left(\frac{c}{a}\right)^{1/4} \frac{\sqrt{a}}{2c}\left(\frac{2c}{a+c}\right)^{3/2}$, we have

$$\|\alpha\|^2 - \frac{1}{4a} = A^2 \sum_{k=m+1}^{\infty} \frac{(2k+1)!}{(2^k k!)^2} u^{2k}.$$

Now by Stirling inequality:

$$\begin{aligned}
\frac{(2k+1)!}{(2^k k!)^2} u^{2k} &\leqslant \frac{e\,(2k+1)^{2k+1+1/2}\,\mathrm{e}^{-(2k+1)}}{(\sqrt{2\pi}2^k k^{k+1/2}\mathrm{e}^{-k})^2} u^{2k} \\
&= \frac{\sqrt{2}}{\pi}\left(1+\frac{1}{2k}\right)^{2k+1}\left(k+\frac{1}{2}\right)^{1/2} u^{2k}. \\
&\leqslant \frac{4e}{\pi}\sqrt{k}u^{2k}.
\end{aligned}$$

And for $m \geqslant -\frac{1}{4\ln u}$,

$$\begin{aligned}
\sum_{m+1}^{\infty} \sqrt{k}u^{2k} &\leqslant \int_m^{\infty} \sqrt{x}u^{2x}dx \\
&\leqslant \int_m^{\infty} xu^{2x}dx \\
&= u^{2m}\frac{(1-2m\ln u)}{(2\ln u)^2} \\
&\leqslant \frac{mu^{2m}}{\ln(1/u)}.
\end{aligned}$$

Hence finally:

$$\left|\|\alpha\|^2 - \frac{1}{4a}\right| \leqslant \frac{4A^2e}{\pi\ln(1/u)} mu^{2m}.$$

(ii) is straightforward because of the odd/even sparsity of ν and η.

(iii) Let us calculate $\|M\alpha\|^2$. We have:

$$\begin{aligned}
\|M\alpha\|^2 - 1 &= \|M\alpha\|^2 - \|M\nu\|^2 \\
&= \sum_{k,j\geqslant m+1} \nu_{2k+1}\nu_{2j+1}\left(M^\top M\right)_{2k+1,2j+1} \\
&= \sum_{k=m+1}^{\infty} \nu_{2k+1}^2\left(M^\top M\right)_{2k+1,2k+1} + 2\sum_{k=m+1}^{\infty} \nu_{2k+1}\nu_{2k+3}\left(M^\top M\right)_{2k+1,2k+3} \\
&= \frac{A^2}{c}\sum_{k=m+1}^{\infty} \frac{(2k+1)!}{(2^k k!)^2}\left(2(2k+1)(a^2+c^2)+(a-c)^2\right)u^{2k} \\
&\qquad - \frac{2A^2ab}{c}\sum_{k=m+1}^{\infty} \frac{\sqrt{(2k+1)!}}{(2^k k!)}\frac{\sqrt{(2k+3)!}}{(2^{k+1}(k+1)!)}\sqrt{(2k+2)(2k+3)}u^{2k+1}.
\end{aligned}$$

Let us call the two terms u_m and v_m respectively. For the first term, when $m \geqslant \max\{-\frac{3}{4\ln u}, \frac{1}{6c}\}$ a calculation as in **(i)** leads to:

$$\begin{aligned}
|u_m| &\leqslant \frac{24A^2e(u^2+c^2)}{\pi c}\int_m^\infty x\sqrt{x}u^{2x}dx + \frac{(a-c)^2}{c}\left(\|\alpha\|^2 - \|\nu\|^2\right) \\
&\leqslant \frac{24A^2e(u^2+c^2)}{\pi c}\int_m^\infty x^2u^{2x}dx - \frac{4A^2e}{\pi\ln u}mu^{2m} \\
&= -\frac{24A^2e(u^2+c^2)}{\pi c}\frac{u^{2m}(2m\ln u(2m\ln(u)-2)+2)}{8\ln^3(u)} - \frac{4A^2e}{\pi\ln u}mu^{2m} \\
&\leqslant -\frac{12A^2e(a^2+c^2)}{\pi c\ln(u)}m^2u^{2m} - \frac{4A^2e}{\pi\ln u}mu^{2m} \\
&\leqslant -\frac{4A^2e}{\pi\ln u}\left(\frac{3(a^2+c^2)}{c}m+1\right)mu^{2m} \\
&\leqslant \frac{24A^2ce}{\pi\ln(1/u)}m^2u^{2m}.
\end{aligned}$$

and for the second term, applying another time Stirling inequality, we get:

$$\begin{aligned}
\frac{\sqrt{(2k+1)!}}{2^k k!}\frac{\sqrt{(2k+3)!}}{2^{k+1}(k+1)!}u^{2k+1} &\leqslant \frac{\mathrm{e}^{1/2}(2k+1)^{k+3/4}\mathrm{e}^{-(k+1/2)}}{\sqrt{2\pi}2^k k^{k+1/2}\mathrm{e}^{-k}}\frac{\mathrm{e}^{1/2}(2k+3)^{k+7/4}\mathrm{e}^{-(k+3/2)}}{\sqrt{2\pi}2^{k+1}(k+1)^{k+3/2}\mathrm{e}^{-(k+1)}}u^{2k+1} \\
&\leqslant \frac{(2k+1)^{k+3/4}}{\sqrt{2\pi}2^k k^{k+1/2}}\frac{(2k+3)^{k+7/4}}{\sqrt{2\pi}2^{k+1}(k+1)^{k+3/2}}u^{2k+1} \\
&= \frac{\sqrt{2}}{\pi}\frac{\left(1+\frac{1}{2k}\right)^{k+3/4}\left(1+\frac{3}{2k}\right)^{k+7/4}}{\left(1+\frac{1}{k}\right)^{k+3/2}}\sqrt{k}u^{2k+1} \\
&\leqslant \frac{\sqrt{2}}{\pi}\frac{\left(1+\frac{3}{2k}\right)^{2k}\left(1+\frac{3}{2k}\right)^{5/2}}{\left(1+\frac{1}{k}\right)^{k}\left(1+\frac{1}{k}\right)^{3/2}}\sqrt{k}u^{2k+1} \\
&\leqslant \frac{\sqrt{2}}{\pi}\left(1+\frac{3}{2k}\right)^{2k}\left(1+\frac{3}{2k}\right)^{5/2}\sqrt{k}u^{2k+1} \\
&\leqslant \frac{15e^3}{\pi}\sqrt{k}u^{2k+1}.
\end{aligned}$$

Hence, as $\sum_{k\geqslant m+1}\sqrt{k}u^{2k+1} \leqslant -\frac{mu^{2m+1}}{\ln u}$, we have $|v_m| \leqslant \frac{30A^2abe^3}{\pi c\ln(1/u)}mu^{2m}$.

(iv) Let us calculate $\alpha^\top \operatorname{Diag}(\lambda)^{-1}\alpha$. We have:

$$\begin{aligned}\alpha^\top \operatorname{Diag}(\lambda)^{-1}\alpha &= \sum_{k=0}^{m} \nu_{2k+1}^2 \lambda_{2k+1}^{-1} \\ &= A^2\sqrt{\frac{bu}{2a}}\sum_{k=0}^{m}\frac{(2k+1)!}{(2^k k!)^2}u^{2k}u^{-(2k+1)} \\ &= A^2\sqrt{\frac{b}{2au}}\sum_{k=0}^{m}\frac{(2k+1)!}{(2^k k!)^2} \\ &\leqslant \frac{4A^2e\sqrt{b}}{\pi\sqrt{2au}}\sum_{k=0}^{m}\sqrt{k} \\ &\leqslant \frac{8A^2e\sqrt{b}}{\pi\sqrt{2au}}m^{3/2}.\end{aligned}$$

(Final constant.) By taking $L = \max\left\{\frac{4A^2e}{\pi\ln(1/u)}, \frac{48A^2ce}{\pi\ln(1/u)}, \frac{60A^2abe^3}{\pi c\ln(1/u)}, \frac{8A^2e\sqrt{b}}{\pi\sqrt{2au}}\right\}$, we have proven the lemma. ■

We can now state the principal result of this section:

Proposition 21 (*Rate of convergence for the bias*)

If $\kappa \leqslant \min\{a^2, 1/5, u^{1/(3c)}\}$ *and such that* $\ln(1/\kappa)\kappa \leqslant \frac{\ln(1/u)}{2aL}$, *then*

$$\mathcal{P}^{-1} \leqslant \mathcal{P}_\kappa^{-1} \leqslant \mathcal{P}^{-1}\left(1 + \frac{L}{2\ln^2(1/u)}\kappa\ln^2(1/\kappa)\right). \tag{109}$$

Proof : The first inequality $\mathcal{P}^{-1} \leqslant \mathcal{P}_\kappa^{-1}$ is obvious. On the other side,

$$\mathcal{P}_\kappa^{-1} = \inf_\beta \frac{\beta^\top(M^\top M + \kappa\operatorname{Diag}(\lambda)^{-1})\beta}{\beta^\top(I-\eta\eta^\top)\beta} \leqslant \frac{\alpha^\top(M^\top M + \kappa\operatorname{Diag}(\lambda)^{-1})\alpha}{\alpha^\top(I-\eta\eta^\top)\alpha},$$

With the estimates of Lemma 37, we have for $mu^{2m} < \frac{1}{4aL}$:

$$\begin{aligned}\mathcal{P}_\kappa^{-1} &\leqslant \frac{1 + Lm^2u^{2m} + \kappa Lm^{3/2}}{\frac{1}{4a} - Lmu^{2m}} \\ &\leqslant \mathcal{P}^{-1}(1 + Lm^2u^{2m} + \kappa Lm^{3/2}).\end{aligned}$$

Let us take $m = \frac{\ln(1/\kappa)}{2\ln(1/u)}$.Then

$$\begin{aligned}\mathcal{P}_\kappa^{-1} &\leqslant \mathcal{P}^{-1}(1 + \kappa L\frac{\ln^2(1/\kappa)}{4\ln^2(1/u)} + \kappa L\frac{\ln^{3/2}(1/\kappa)}{2^{3/2}\ln^{3/2}(1/u)}) \\ &\leqslant \mathcal{P}^{-1}\left(1 + \kappa L\frac{\ln^2(1/\kappa)}{2\ln^2(1/u)}\right),\end{aligned}$$

as soon as $\kappa \leqslant a^2$. Note also that the condition $mu^{2m} < \frac{1}{4aL}$ can be rewritten in terms of m as $\kappa\ln(1/\kappa) < \frac{\ln(1/u)}{2aL}$. The other conditions of Lemma 37 are $\kappa \leqslant \mathrm{e}^{-3/2} \sim 0.22$ and $\kappa \leqslant u^{1/(3c)}$ ■

C.4.4. Facts about Hermite polynomials

Orthogonality. We have:

$$\int_{\mathbb{R}} \mathrm{e}^{-x^2} H_k(x)H_m(x) = 2^k k!\sqrt{\pi}\delta_{km}.$$

Recurrence relations. We have:

$$H_i'(x) = 2iH_{i-1}(x),$$

and

$$H_{i+1}(x) = 2xH_i(x) - 2iH_{i-1}(x).$$

Mehler's formula. We have:

$$\sum_{k=0}^{\infty} \frac{H_k(x)\mathrm{e}^{-x^2/2} H_k(y)\mathrm{e}^{-y^2/2}}{2^k k!\sqrt{\pi}} u^k = \frac{1}{\sqrt{\pi}} \frac{1}{\sqrt{1-u^2}} \exp\Big(\frac{2u}{1+u}xy - \frac{u^2}{1-u^2}(x-y)^2 - \frac{x^2}{2} - \frac{y^2}{2}\Big).$$

This implies that the functions $x \mapsto \frac{1}{\sqrt{\pi}} \frac{1}{\sqrt{1-u^2}} \exp\Big(\frac{2u}{1+u}xy - \frac{u^2}{1-u^2}(x-y)^2 - \frac{x^2}{2} - \frac{y^2}{2}\Big)$ has coefficients $\frac{H_k(y)\mathrm{e}^{-y^2/2}}{\sqrt{2^k k!\sqrt{\pi}}} u^k$ in the orthonormal basis $(x \mapsto \frac{H_k(x)\mathrm{e}^{-x^2/2}}{\sqrt{2^k k!\sqrt{\pi}}})$ of $L_2(dx)$.

Thus

$$\int_{\mathbb{R}} \frac{1}{\sqrt{\pi}} \frac{1}{\sqrt{1-u^2}} \exp\Big(\frac{2u}{1+u}xy - \frac{u^2}{1-u^2}(x-y)^2 - \frac{x^2}{2} - \frac{y^2}{2}\Big) \frac{H_k(x)\mathrm{e}^{-x^2/2}}{\sqrt{2^k k!\sqrt{\pi}}} dx = \frac{H_k(y)\mathrm{e}^{-y^2/2}}{\sqrt{2^k k!\sqrt{\pi}}} u^k,$$

that is

$$\int_{\mathbb{R}} \exp\Big(\frac{2u}{1+u}xy - \frac{u^2}{1-u^2}(x-y)^2 - x^2\Big) H_k(x)dx = \sqrt{\pi}\sqrt{1-u^2} H_k(y) u^k.$$

This implies:

$$\int_{\mathbb{R}} \exp\Big(\frac{2u}{1-u^2}xy - \frac{x^2}{1-u^2}\Big) H_k(x)dx = \sqrt{\pi}\sqrt{1-u^2} H_k(y) \exp(\frac{u^2}{1-u^2}y^2) u^k$$

For $y = 0$, we get

$$\int_{\mathbb{R}} \exp\Big(-\frac{x^2}{1-u^2}\Big) H_k(x)dx = \sqrt{\pi}\sqrt{1-u^2} H_k(0) u^k.$$

Another consequence is that

$$\begin{aligned}
\sum_{k=0}^{\infty} \frac{H_k(x)H_k(y)}{2^k k!\sqrt{\pi}} u^k &= \frac{1}{\sqrt{\pi}} \frac{1}{\sqrt{1-u^2}} \exp\Big(\frac{2u(1-u)+2u^2}{1-u^2}xy - \frac{u^2}{1-u^2}(x^2+y^2)\Big) \\
&= \frac{1}{\sqrt{\pi}} \frac{1}{\sqrt{1-u^2}} \exp\Big(\frac{2u}{1-u^2}xy - \frac{u}{1-u^2}(x^2+y^2) + \frac{u}{1+u}(x^2+y^2)\Big) \\
&= \frac{1}{\sqrt{\pi}} \frac{1}{\sqrt{1-u^2}} \exp\Big(-\frac{u}{1-u^2}(x-y)^2\Big) \exp\Big(\frac{u}{1+u}(x^2+y^2)\Big) \\
&= \frac{1}{\sqrt{\pi}} \frac{\sqrt{u}}{\sqrt{1-u^2}} \exp\Big(-\frac{u}{1-u^2}(x-y)^2\Big) \frac{1}{\sqrt{u}} \exp\Big(\frac{u}{1+u}(x^2+y^2)\Big).
\end{aligned}$$

Thus, when u tends to 1, as a function of x, this tends to a Dirac at y times e^{y^2}.

⋆

⋆ ⋆

2. Statistical estimation of Laplacian and application to dimensionality reduction

2.1. Introduction

One of the reasons of the success of learning with Reproducing Kernel Hilbert Spaces (RKHS) is that they naturally select problem-adapted basis of test functions. Even more interestingly, leveraging the underlying regularity of the target function, RKHS have the ability to circumvent the curse of dimensionality. This is exactly where all techniques resting on local averages fail: approximating a problem will always be cursed by the high-dimension d, because one will need $\sim n^{-1/d}$ points to perform well. This main difference echoes in the nature of the kernels: *pointwise positive kernels* in the non-parametric estimation literature [Nad64] and *positive semi-definite (PSD) kernels* in modern kernel learning [SC08, SS00].

Solving a problem with PSD kernels that used to be tackled with local techniques is at the heart of this work. Indeed, we estimate the diffusion operator related to a measure μ through its principal eigenelements. When cast into an *unsupervised* learning problem, this can be seen as a dimensionality reduction technique resting on the diffusive nature of the data. This is exactly what the celebrated procedure of Diffusion maps [CL06] is used for: it finds the slowest diffusion directions, giving a precious information to understand the structure of the samples [CBLK06]. However, as introduced before, Diffusion maps are based on a local construction that scales poorly with the dimension [AT07] and do not benefit of all the recent work on PSD kernels that tackles potential high-dimensional settings.

Let us explain the fundamental difference between the approach of this work and of Diffusion maps. When we want to estimate the diffusion operator

$$\mathcal{L} := -\Delta + \langle \nabla V, \nabla \cdot \rangle, \tag{110}$$

one of the difficult aspect is to approximate differential operators. While, currently, people use local kernel smoothing techniques, our approach is very different. It leverages the reproducing property of derivative in RKHS to circumvent this difficulty: this is a well-known strategy in numerical analysis for Partial Differential Equations called *meshless methods* [SW06].

In another direction, it is very interesting to note that in a very nice article [Sal98], Salinelli showed that considering the first eigenvectors of $\mathcal{L}$ was *the good way* of generalizing the Principal Components Analysis [Pea01, Hot33] procedure to non-linear principal components. At this time, (i) neither the theory behind diffusions and weighted Sobolev spaces (ii) nor the theory of RKHS were mature. Hence, he clearly explained (i) that the theoretical framework of his analysis was quite poor but could be extended (ii) that at this point solving numerically the problem was impossible in high-dimension as it necessitates to discretize the Laplacian. Quite surprisingly, the literature of Diffusion maps seems to have forgotten Salinelli's seminal contribution. Our work can be considered as a natural continuation of his: pushing further the theoretical comprehension of this *Non-linear Principal Components* with modern tools and giving a way to solve it efficiently thanks to PSD kernels.

Note also that this work is also strongly related to the previous one [PVBL+20] whose aim was to estimate the first non-zero eigenvalue of $\mathcal{L}$ (this is saying, its spectral gap). Besides being more mature, the focus of this work is quite different: while previously we leveraged the estimation of the eigenvalue to find reduced order models in physical systems, we will focus here on the estimation of the whole spectrum of $\mathcal{L}$ and try to be more precise regarding its convergence properties.

Finally, note that even if the story behind it is almost complete, this work is still unfinished. As the aim is to compare our procedure to Diffusion maps, we would like to be end-to-end, showing explicitly

how and why both theoretically and empirically, our work could be more robust when the dimension or number of samples grows. An explicit discussion at the end, in subsection 2.4.3, details where exactly this work stands.

2.2. Diffusion operator

Consider a probability measure $d\mu$ on $\mathbb{R}^d$ which has a density with respect to the Lebesgue measure and can be written under the following form: $d\mu(x) = \mathrm{e}^{-V(x)}dx$. Consider $H^1(\mu)$ the space of functions in $L^2(\mu)$ (i.e., which are square integrable) that also have all their first order derivatives in L^2, that is, $H^1(\mu) = \{f \in L^2(\mu),\ \int_{\mathbb{R}^d} f^2 d\mu + \int_{\mathbb{R}^d} \|\nabla f\|^2 d\mu < \infty\}$.

The aim of this work is to estimate the diffusion operator associated with the measure μ, given access $x_1, \dots, x_n$, n i.i.d. samples distributed according to μ. With test functions ϕ smooth enough,

$$\mathcal{L}\phi := -\Delta\phi + \nabla V \cdot \nabla\phi. \tag{111}$$

Note that this is the natural extension of a previous published article (see Part III, Section 1 of this thesis). In this work, we will extend the results of [PVBL+20] by showing that the the same procedure leads in fact to the estimation of the full spectrum of the diffusion operator. We will also relax some assumptions needed to show consistency of the estimator. A major difference with [PVBL+20] is that we will focus on the construction of the estimator and dig deeper in the convergence theorems forgetting the reaction coordinate estimation discussion.

2.2.1. Langevin diffusion

Let us consider the overdamped Langevin diffusion in $\mathbb{R}^d$, that is the solution of the following Stochastic Differential Equation:

$$\mathrm{d}X_t = -\nabla V(X_t)\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t, \tag{112}$$

where $(B_t)_{t\geqslant 0}$ is a d-dimensional Brownian motion. It is well-known [BGL14] that the law of $(X_t)_{t\geqslant 0}$ converges to the Gibbs measure $d\mu$ and that the Poincaré constant (see Proposition 10 below) controls the rate of convergence to equilibrium in $L^2(\mu)$. Let us denote by $P_t(f)$ the Markovian semi-group associated with the Langevin diffusion $(X_t)_{t\geqslant 0}$. It is defined in the following way: $P_t(f)(x) = \mathbb{E}[f(X_t)|X_0 = x]$. This semi-group satisfies the dynamics

$$\frac{d}{dt}P_t(f) = -\mathcal{L}P_t(f),$$

where $\mathcal{L}\phi = -\Delta\phi + \nabla V \cdot \nabla\phi$ is a differential operator called the infinitesimal generator of the Langevin diffusion (112) (Δ denotes the standard Laplacian on $\mathbb{R}^d$). Note that by integration by parts, the semi-group $(P_t)_{t\geqslant 0}$ is reversible with respect to $d\mu$, that is: $\int f(\mathcal{L}g)\,d\mu = \int \nabla f \cdot \nabla g\,d\mu = \int (\mathcal{L}f)g\,d\mu$. This also shows that $\mathcal{L}$ is a symmetric positive definite operator on $H^1(\mu)$.

Remark 10 (*Link with Poincaré constant*)

Let us call π the orthogonal projector of $L^2(\mu)$ on constant functions: $\pi f : x \in \mathbb{R}^d \mapsto \int f d\mu$. The first non-zero eigenvalue of the operator $\mathcal{L}$ can be define as:

$$\mathcal{P}^{-1} = \inf_{f\in(H^1(\mu)\cap L^2_0(\mu))\setminus\{0\}} \frac{\langle f, \mathcal{L}f\rangle_{L^2(\mu)}}{\|(I_{L^2(\mu)} - \pi)f\|^2_{L^2(\mu)}}, \tag{113}$$

where $\mathcal{P}$ is also known as the Poincaré constant of the distribution $d\mu$, see [PVBL+20] fur more details.

2.2.2. Some useful properties of the diffusion operator

Positive semi-definiteness. The first property that we saw is symmetry and positiveness of $\mathcal{L}$ in $H^1(\mu)$. It comes from the following integration by part identity:

$$\int f(\mathcal{L}g)\,d\mu = \int \nabla f \cdot \nabla g\,d\mu = \int (\mathcal{L}f)g\,d\mu,$$

showing that the quadratic form induced by $\mathcal{L}$ is also the Dirichlet energy

$$\langle \mathcal{L}f, f\rangle_{L^2(\mu)} = \int \|\nabla f\|^2 d\mu =: \mathcal{E}(f).$$

Link with Schrödinger operator. In PDE, we say that an operator is of Schrödinger type if it is the sum of the Laplacian and a multiplicative operator, this comes from the fact that this is the type of operator that governs the dynamics of quantum systems. Here, let us define the Schrödinger operator $\widetilde{\mathcal{L}} := -\Delta + \mathcal{V}$, where $\mathcal{V} := \frac{1}{2}\Delta V - \frac{1}{4}\|\nabla V\|^2$. We can show that $\widetilde{\mathcal{L}}$ and $\mathcal{L}$ and conjugate to each other: indeed, a rapid calculation shows that

$$\widetilde{\mathcal{L}} = e^{-V/2}\mathcal{L}\left[e^{V/2}\,\cdot\right].$$

As Schrödinger operators are well-studied, we can infer from this fact interesting properties on the spectrum of $\mathcal{L}$. Indeed,

$$(\lambda, u) \text{ eigenelements of } \widetilde{\mathcal{L}} \Leftrightarrow (\lambda, e^{V/2}u) \text{ eigenelements of } \mathcal{L},$$

And we have also the following equality for f smooth enough:

$$\frac{\langle \mathcal{L}f, f\rangle_{L^2(\mu)}}{\|f\|^2_{L^2(\mu)}} = \frac{\langle \widetilde{\mathcal{L}}f, f\rangle_{L^2(\mathbb{R}^d)}}{\|f\|^2_{L^2(\mathbb{R}^d)}}.$$

Spectrum of $\mathcal{L}$. The most important property that we can infer from this is the nature of the spectrum of $\mathcal{L}$. Indeed, it is well known [RS12] that if $\mathcal{V}$ is locally integrable, bounded from below and coercive ($\mathcal{V}(x) \longrightarrow +\infty$, when $\|x\| \to +\infty$), then the Schrödinger operator has a compact resolvent. In particular,

- **Assumption 0. Spectrum of $\mathcal{L}$:** Assume that $\frac{1}{2}\Delta V(x) - \frac{1}{4}\|\nabla V\|^2 \longrightarrow +\infty$, when $\|x\| \to +\infty$.

Assumption 0 implies that $\mathcal{L}$ has a compact resolvent. This also implies that $\mathcal{L}$ has a purely discrete spectrum and a complete set of eigenfunctions. Note that this assumption implies also a spectral gap for the diffusion operator $\mathcal{L}$ and hence that a Poincaré inequality holds. Throughout this work and even if not clearly stated, we will assume Assumption 0. For further discussions on the spectrum of $\mathcal{L}$, we refer to [BGL14, HN05].

2.3. Approximation of the diffusion operator in the RKHS

Let $\mathcal{H}$ be a RKHS with kernel K. Let us suppose, as in the precedent Section 1 mild assumptions on the kernel for the problem:

- **Assumption 1. Universality**: $\mathcal{H}$ is dense in $H^1(\mu)$.
- **Assumption 2. Regularity**: K is a psd kernel at least twice continuously derivable.
- **Assumption 3. Smoothness**: K and its derivative are bounded functions in $\mathcal{H}$.

2.3.1. Embedding the diffusion operator in the RKHS

Let us define the following operators from $\mathcal{H}$ to $\mathcal{H}$:

$$\Sigma = \mathbb{E}_\mu \left[K_X \otimes K_X \right], \qquad \mathsf{L} = \mathbb{E}_\mu \left[\nabla K_X \otimes_d \nabla K_X \right],$$

where $\otimes$ is the standard tensor product: $\forall f, g, h \in \mathcal{H}$, $(f \otimes g)(h) = \langle g, h \rangle_{\mathcal{H}} f$ and $\otimes_d$ is defined as follows: $\forall f, g \in \mathcal{H}^d$ and $h \in \mathcal{H}$, $(f \otimes_d g)(h) = \sum_{i=1}^d \langle g_i, h \rangle_{\mathcal{H}} f_i$. Define also S, the injection from $\mathcal{H}$ to $L^2(\mu)$, and its adjoint S^* from $L^2(\mu)$ to $\mathcal{H}$:

$$\forall f \in \mathcal{H}, \ \ \mathsf{S}f(x) = \langle f, K_x \rangle_{\mathcal{H}} = f(x), \qquad \forall f \in L^2(\mu), \ \ \mathsf{S}^* f(x) = \mathbb{E}_\mu \left[K(x, X) f(X) \right].$$

Note that $\mathsf{S}^*\mathsf{S} = \Sigma$. And denote also the mean function $m := S^* \mathbb{1} = \mathbb{E}_\mu[K_X]$ and the centered covariance operator $\mathsf{C} = \Sigma - m \otimes m$. Note that in the sequel, transforming f to its centered version $f - \int f d\mu$ does not change anything, thus we can assume that f has mean 0. This allows to consider that Σ and C are the same operators.

With these definitions, we can represent the diffusion operator $\mathcal{L}$ in the RKHS, this is the statement of the following proposition.

Proposition 22 (*Embedding of the diffusion operator in the RKHS*)

With natural assumptions 1, 2, 3*, on the kernel, we have the following equality on* $\mathcal{H}$*:*

$$\mathsf{L} = S^* \mathcal{L} S \tag{114}$$

Proof : For $z \in \mathbb{R}^d$, $f \in \mathcal{H}$,

$$\begin{aligned}
\langle \mathsf{L}f, K_z \rangle_{\mathcal{H}} &= \int \nabla f(x) \cdot \nabla_x K(x, z) d\mu(x) \\
&= -\int \Delta f(x) K(x, z) d\mu(x) + \int \nabla f(x) \cdot \nabla V(x) K(x, z) d\mu(x) \\
&= \langle \mathcal{L}\mathsf{S}f, \mathsf{S}K_z \rangle_{L^2(\mu)} \\
&= \langle \mathsf{S}^* \mathcal{L} \mathsf{S} f, K_z \rangle_{\mathcal{H}},
\end{aligned}$$

hence the equality between operators. ■

We want to construct an approximation of the eigenelements of $\mathcal{L}$ with domain $H^1(\mu) \cap L^2_0(\mu)$. Note that this operator is invertible by the spectral gap Assumption 0. In the following we will approximate the eigenelements of $\mathcal{L}^{-1}$. First we give a representation of $\mathcal{L}^{-1}$ in the RKHS $\mathcal{H}$, then we construct an operator on $\mathcal{H}$ that have the same eigenelements of $\mathcal{L}^{-1}$. Indeed, if we denote L^{-1} the inverse of L restricted on $(\mathrm{Ker}\mathsf{L})^\perp$, we have:

Proposition 23 (*Representation of the inverse of the diffusion operator in the RKHS*)

On $H^1(\mu) \cap L^2_0(\mu)$*, we have the equality between operators:*

$$\mathcal{L}^{-1} = \mathsf{S}\mathsf{L}^{-1}\mathsf{S}^*. \tag{115}$$

Proof : Let $g \in \mathrm{Ran}\, \mathsf{S}$, there exists $f \in \mathcal{H}$ such that $g = \mathsf{S}f$. Let us calculate:

$$\mathsf{S}\mathsf{L}^{-1}\mathsf{S}^* \mathcal{L} g = \mathsf{S}\mathsf{L}^{-1}\mathsf{S}^* \mathcal{L} \mathsf{S} f = \mathsf{S}\mathsf{L}^{-1}\mathsf{L} f = \mathsf{S} f = g.$$

And as $\mathcal{L}$ is invertible on $\mathrm{Ran}\, \mathsf{S} \cap H^1(\mu) \cap L^2_0(\mu)$, the left and right inverse are the same. Hence, $\mathcal{L}^{-1}$ and $\mathsf{S}\mathsf{L}^{-1}\mathsf{S}^*$ are equal on $\mathrm{Ran}\, \mathsf{S}$. Furthermore we can notice that $\mathcal{L}^{-1}$ and $\mathsf{S}\mathsf{L}^{-1}\mathsf{S}^*$ are bounded on $L^2(\mu)$.

Indeed,

$$\begin{aligned}\mathcal{P} = \sup_{f\in(\mathrm{Ker}\mathsf{L})^\perp} \frac{\langle f, \mathsf{S}^*\mathsf{S}f\rangle_{\mathcal{H}}}{\langle f, \mathsf{L}f\rangle_{\mathcal{H}}} \geqslant \sup_{f\in(\mathrm{Ker}\mathsf{L})^\perp} \frac{\langle \mathsf{L}^{-1/2}f, \mathsf{S}^*\mathsf{S}\mathsf{L}^{-1/2}f\rangle_{\mathcal{H}}}{\langle \mathsf{L}^{-1/2}f, \mathsf{L}\mathsf{L}^{-1/2}f\rangle_{\mathcal{H}}} &= \sup_{f\in(\mathrm{Ker}\mathsf{L})^\perp} \frac{\langle f, \mathsf{L}^{-1/2}\mathsf{S}^*\mathsf{S}\mathsf{L}^{-1/2}f\rangle_{\mathcal{H}}}{\|f\|^2_{\mathcal{H}}} \\ &= \|\mathsf{L}^{-1/2}\mathsf{S}^*\mathsf{S}\mathsf{L}^{-1/2}\|_{\mathcal{H}} \\ &= \|\mathsf{S}\mathsf{L}^{-1}\mathsf{S}^*\|_{L^2(\mu)}.\end{aligned}$$

As $\mathcal{L}^{-1}$ and $\mathsf{S}\mathsf{L}^{-1}\mathsf{S}^*$ are equal and continuous on $\mathrm{Ran}\,\mathsf{S}$, they are also equal on its closure. ■

Note that we used that for universal kernels (Assumption 3), $\overline{\mathrm{Ran}\,\mathsf{S}} = L^2(\mu)$, see [MXZ06] for further details. Note also that most of the used kernels have this property: this is, for example, the case for the Gaussian and Laplace kernels.

Now, thank to Proposition 23, we have a representation of $\mathcal{L}^{-1}$ in the RKHS $\mathcal{H}$. But what we really would like is an operator on $\mathcal{H}$ that as the same eigenelements as $\mathcal{L}^{-1}$. Having such a representation would allow for numerical computations. For this we need the following Lemma.

Lemma 38 (*Link between $\mathsf{A}^*\mathsf{A}$ and $\mathsf{A}\mathsf{A}^*$ in the compact case.*)

Let $\mathcal{H}_1$ and $\mathcal{H}_2$ two Hilbert spaces. Let A be an operator from $\mathcal{H}_1$ to $\mathcal{H}_2$ such that $\mathsf{A}^*\mathsf{A}$ is a self-adjoint compact operator on $\mathcal{H}_1$. Then,

(i) A is a bounded operator from $\mathcal{H}_1$ to $\mathcal{H}_2$.

(ii) $\mathsf{A}\mathsf{A}^*$ is a self-adjoint compact operator on $\mathcal{H}_2$ with the same spectrum as $\mathsf{A}\mathsf{A}^*$.

(iii) If $\lambda \neq 0$ is an eigenvalue of $\mathsf{A}^*\mathsf{A}$ with eigenvector $u \in \mathcal{H}_1$, then λ is an eigenvalue of $\mathsf{A}\mathsf{A}^*$ with eigenvector $\mathsf{A}u \in \mathcal{H}_2$.

Proof : First let us notice that A is necessarily bounded. Indeed, let $u \in \mathcal{H}_1$,

$$\|\mathsf{A}u\|^2_{\mathcal{H}_2} = \langle \mathsf{A}u, \mathsf{A}u\rangle_{\mathcal{H}_2} = \langle \mathsf{A}^*\mathsf{A}u, u\rangle_{\mathcal{H}_1} \leqslant \|\mathsf{A}^*\mathsf{A}u\|_{\mathcal{H}_1}\|u\|_{\mathcal{H}_1} \leqslant \|\mathsf{A}^*\mathsf{A}\|\,\|u\|^2_{\mathcal{H}_1}.$$

Hence, $\|\mathsf{A}\| \leqslant \sqrt{\|\mathsf{A}^*\mathsf{A}\|}$.

Second, as $\mathsf{A}^*\mathsf{A}$ is self-adjoint and compact on $\mathcal{H}_1$, there exist $(\psi_i)_{i\in\mathbb{N}}$ an orthonormal basis of $\mathcal{H}_1$ and a sequence of reals $(\lambda_i)_{i\in\mathbb{N}}$ such that:

$$\mathsf{A}^*\mathsf{A} = \sum_{i\geqslant 0} \lambda_i \psi_i \otimes \psi_i,$$

where the infinite sum stands for the strong convergence of operators. Now, by composing on the left side by A^* and on the right side by A^*, we get:

$$(\mathsf{A}\mathsf{A}^*)^2 = \mathsf{A}\mathsf{A}^*\mathsf{A}\mathsf{A}^* = \sum_{i\geqslant 0} \lambda_i (\mathsf{A}\psi_i)\otimes(\mathsf{A}\psi_i) = \sum_{i\geqslant 0} \lambda_i^2 (\mathsf{A}\frac{\psi_i}{\sqrt{\lambda_i}})\otimes(\mathsf{A}\frac{\psi_i}{\sqrt{\lambda_i}}).$$

Hence, $\mathsf{A}\mathsf{A}^* = \sum_{i\geqslant 0} \lambda_i (\mathsf{A}\frac{\psi_i}{\sqrt{\lambda_i}})\otimes(\mathsf{A}\frac{\psi_i}{\sqrt{\lambda_i}})$ and is a compact operator. We can of course check that $\left(\mathsf{A}\frac{\psi_i}{\sqrt{\lambda_i}}\right)_{i\in\mathbb{N}}$ is an orthonormal basis of $\mathcal{H}_2$: $\left\langle \mathsf{A}\frac{\psi_i}{\sqrt{\lambda_i}}, \mathsf{A}\frac{\psi_j}{\sqrt{\lambda_j}}\right\rangle = (\lambda_i\lambda_j)^{-1/2}\langle\psi_i, \mathsf{A}^*\mathsf{A}\psi_j\rangle = \sqrt{\frac{\lambda_j}{\lambda_i}}\langle\psi_i,\psi_j\rangle = \delta_{ij}$. ■

We can now state the following important Proposition. This is a clear consequence of the previous Lemma 38.

Proposition 24 (*Eigenelements of $\mathcal{L}^{-1}$ as function in the RKHS*)

Decompose: $\mathcal{L}^{-1} = \mathsf{S}\mathsf{L}^{-1}\mathsf{S}^* = \mathsf{S}\mathsf{L}^{-1/2}\mathsf{L}^{-1/2}\mathsf{S}^*$, then,

(i) $\mathsf{S}\mathsf{L}^{-1/2}$ is a bounded operator from $\mathcal{H}$ to $L^2(\mu)$.

(ii) $\mathsf{L}^{-1/2}\mathsf{C}\mathsf{L}^{-1/2}$ *is a self-adjoint compact operator on* $\mathcal{H}$ *with the same spectrum as* $\mathcal{L}^{-1}$.

(iii) If $\lambda \neq 0$ *is an eigenvalue of* $\mathsf{L}^{-1/2}\mathsf{C}\mathsf{L}^{-1/2}$ *with eigenvector* $u \in \mathcal{H}$*, then* λ *is an eigenvalue of* $\mathcal{L}^{-1}$ *with eigenvector* $\mathsf{S}\mathsf{L}^{-1/2}u \in L^2(\mu)$.

This proposition will allow us to approximate the eigenelements of $\mathcal{L}^{-1}$ with the ones of the operator $\mathsf{L}^{-1/2}\mathsf{C}\mathsf{L}^{-1/2}$ (that is well-defined only on $\mathcal{H}$) with a finite set of samples.

2.3.2. Definition of the estimator

Empirical operators. As in [PVBL+20], define the empirical counterpart of L and C, where the empirical operator are defined by replacing expectation with respect to μ by expectations with respect to its empirical measure $\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ with $x_1, \dots, x_n$ are n i.i.d samples distributed according to $d\mu$.

$$\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^n K_{x_i} \otimes K_{x_i}, \qquad \widehat{\mathsf{L}} = \frac{1}{n}\sum_{i=1}^n \nabla K_{x_i} \otimes_d \nabla K_{x_i},$$

and $\widehat{\mathsf{C}} = \widehat{\Sigma} - \widehat{m} \otimes \widehat{m}$. Hence, one could be tempted to define our estimator as the empirical diffusion operator as

$$\widehat{\mathsf{L}}^{-1/2}\widehat{\mathsf{C}}\widehat{\mathsf{L}}^{-1/2}.$$

However, this definition carries two main problems:

(i) If $f \in \operatorname{Ker}\widehat{\mathsf{L}}$, i.e. for all $i \leqslant n$, $\nabla f(X_i) = 0$ and at the same time $f \notin \operatorname{Ker}\widehat{\mathsf{C}}$, i.e. $\exists i \neq j$, such that $f(X_i) \neq f(X_j)$, then $\|\widehat{\mathsf{L}}^{-1/2}\widehat{\mathsf{C}}\widehat{\mathsf{L}}^{-1/2}f\| = +\infty$. This is an *overfitting-type issue.*

(ii) Another problem is related to the fact that finding the eigenelements of $\mathsf{L}^{-1/2}\mathsf{C}\mathsf{L}^{-1/2}$ is equivalent to solving the generalized eigenvalue problem: $\mathsf{C}f = \sigma\mathsf{L}f$. Such systems are known to be numerically unstable as mentioned in [Cra76]. This would be especially the case when replacing the operators by their empirical counterpart. This is a *stability issue.*

Regularization. These two concerns recall the pitfall of overfitting for regression tasks. Hence, as for kernel ridge regression, a natural idea is to regularize with some parameter λ. The main drawback is that it induces a bias in our estimation: the acute reader will recognize that the bigger the λ the closer the problem is to kernel-PCA [MSS+99] (this point of view can be further studied but we leave this for future work at this point). This leads to the following definition of our estimator and its empirical counterpart:

Definition 5 (*Definition of the estimator*)

Under Assumptions 1, 2, 3*, we define the two estimators of the inverse diffusion operator* $\mathcal{L}^{-1}$*:*

$$\textit{Biased estimator:} \qquad (\mathsf{L} + \lambda I)^{-1/2}\mathsf{C}(\mathsf{L} + \lambda I)^{-1/2} \tag{116}$$

$$\textit{Empirical estimator:} \qquad (\widehat{\mathsf{L}} + \lambda I)^{-1/2}\widehat{\mathsf{C}}(\widehat{\mathsf{L}} + \lambda I)^{-1/2}. \tag{117}$$

In the following, to shorten notations, let us define $\mathsf{L}_\lambda = \mathsf{L} + \lambda I$ and $\widehat{\mathsf{L}}_\lambda = \widehat{\mathsf{L}} + \lambda I$.

When analyzing the performances of our empirical estimator, we will draw a particular attention to the dependency on the dimension: the goal here is to show that our method scales better that diffusion maps [CL06, HAL07] with the dimension and benefits from the aspects of positive definite kernel methods: takes into account the regularity of the optimum, scales well with the dimension, large number of data can be leveraged by usual kernel techniques (subsampling, kernel features).

2.3.3. What do we want to control?

Requirements of the problem. The natural and general goal of the present work is to give an approximation of the diffusion operator. But there are in fact more precise objects that we may want to have an approximation of:

- **Operator.** Convergence to $\mathcal{L}$ the operator itself. Either its representation in $\mathcal{H}$ either in $H^1(\mu)$.
- **Semigroup.** In fact, as $\mathcal{L}$ is the infinitesimal generator of the dynamics, we can be interested in the convergence to the associated semigroups $e^{t\mathcal{L}}$.
- **Eigenvectors.** As one of the main application of this estimator could be the computation of a low-dimensional embedding of the data through the eigenvectors of $\mathcal{L}$, we are directly interested in the convergence to the eigenvectors. Either eigenvector per eigenvector, either finite dimensional subspaces spanned by few of them. Note that we are mostly interested in the small eigenvalues of $\mathcal{L}$ because they are those governing the behavior of the dynamics.
- **Eigenvalues.** As it has already been done in previous work for the top eigenvalue [PVBL+20], one would like to approximate the sequence of eigenvalues. Another application is the construction of some diffusion distance (see [CL06]) for clustering.

Types of convergence. Let us first list the different convergences we can have for our problem in terms of operators and in term of functions.

- **Operator convergences.** Now that we have the estimator of our operator, one question that is important (as we are dealing with infinite dimensional spaces) is in what norm do we want to control our estimation. Indeed, we have several possibilities in the type of convergence if we want to control the convergence of operator T_n to T. Here is a non-exhaustive list:

  (i) Operator norm : $\|\mathsf{T}_n - \mathsf{T}\|_{\mathcal{H}} \longrightarrow 0$.

  (ii) Strong convergence (pointwise): $\forall f \in \mathcal{H}$, $\|\mathsf{T}_n f - \mathsf{T} f\|_{\mathcal{H}} \longrightarrow 0$.

  (iii) Other types of weak convergence : $\forall f,\ g \in \mathcal{H}$, $\langle g, \mathsf{T}_n f\rangle_{\mathcal{H}} \longrightarrow \langle g, \mathsf{T} f\rangle_{\mathcal{H}}$.

  Of course some of them are stronger than other ones as (i) $\Rightarrow$ (ii) $\Rightarrow$ (iii). Note also that the convergence of operators can be done either for the representation of $\mathcal{L}^{-1}$ in $\mathcal{H}$ either directly in $L^2(\mu)$.

- **Convergences of eigenelements.** In our problem, at one point, we will try to estimate the eigenelements (eigenvectors and related eigenvalue) of the $\mathcal{L}$. Note this will be done by estimating the eigenelements of the representation of $\mathcal{L}^{-1}$ in $\mathcal{H}$: $\mathsf{L}^{-1/2}\mathsf{C}\mathsf{L}^{-1/2}$. Hence, we can either compare the resulting eigenvector in $\mathcal{H}$ either their mapping in $L^2(\mu)$ by the operator from $\mathcal{H}$ to $L^2(\mu)$: $u \to S\mathsf{L}^{-1/2}u$.

Previous results. In all the previous works: Belkin, Audibert [HAL07] and Coifman [CL06], proved the convergence of the estimated operator. However, the convergence theorem are only for given *pointwise*, for *bounded domains* and have a *very bad dependency in the dimension* $\sim n^{-1/d}$. We will try to overpass these three limiting results with our method. Please note that the operator norm convergence to the diffusion operator will *imply all the other convergences:*

- **Semigroup.** Because of the fact that: $\|e^B - e^A\| \leq \|B - A\| e^{\max\{\|A\|, \|B\|\}}$.
- **Eigenvectors and eigenvalues.** Directly by perturbation theory arguments. Refined bounds can also be discussed if one want to approximate k-dimensional subspaces.

2.4. Analysis of the estimator

As said earlier, to shorten the notations, let us define for an operator A, the operator $A_\lambda := A + \lambda I$. We will split the problem in two: bias and variance as follows.

$$\left\|\widehat{\mathsf{L}}_\lambda^{-1/2}\widehat{\mathsf{C}}\widehat{\mathsf{L}}_\lambda^{-1/2} - \mathsf{L}^{-1/2}\mathsf{C}\mathsf{L}^{-1/2}\right\| \leqslant \underbrace{\left\|\widehat{\mathsf{L}}_\lambda^{-1/2}\widehat{\mathsf{C}}\widehat{\mathsf{L}}_\lambda^{-1/2} - \mathsf{L}_\lambda^{-1/2}\mathsf{C}\mathsf{L}_\lambda^{-1/2}\right\|}_{\text{variance}} + \underbrace{\left\|\mathsf{L}_\lambda^{-1/2}\mathsf{C}\mathsf{L}_\lambda^{-1/2} - \mathsf{L}^{-1/2}\mathsf{C}\mathsf{L}^{-1/2}\right\|}_{\text{bias}}$$

The variance term corresponds to the statistical error coming from the fact that we have only access to a finite set of n samples of the distribution μ. The bias comes from the introduction of a regularization of the operator L scaled by λ. We first derive bounds for the variance term.

2.4.1. Variance analysis

Proposition 25 (*Analysis of the statistical error*)

Suppose assumptions 1, 2, 3 *hold true. For any* $\delta \in (0, 1/3)$, *and* $\lambda > 0$ *such that* $\lambda \leqslant \|\mathsf{L}\|$ *and any integer* $n \geqslant 15\frac{\mathcal{K}_d}{\lambda}\log\frac{4\operatorname{Tr}\mathsf{L}}{\lambda\delta}$, *with probability at least* $1 - 3\delta$,

$$\left\|\widehat{\mathsf{L}}_\lambda^{-1/2}\widehat{\mathsf{C}}\widehat{\mathsf{L}}_\lambda^{-1/2} - \mathsf{L}_\lambda^{-1/2}\mathsf{C}\mathsf{L}_\lambda^{-1/2}\right\| \leqslant \frac{8\mathcal{K}}{\lambda\sqrt{n}}\log(2/\delta) + \mathrm{o}\left(\frac{1}{\lambda\sqrt{n}}\right). \tag{118}$$

Note that this is the exact same Proposition than the one in [PVBL+20], hence, the reader can easily refer to it for the detailed proof based on concentration of empirical operators. Note also that in Proposition 25, we are only interested in the regime where $\lambda\sqrt{n}$ is large but a non-asymptotic result can be given: Lemmas 31 and 32 of the Appendix 1.6 give explicit and sharper bounds under refined hypotheses on the spectra of C and L. Note also the following facts on the variance bound:

- It is dimension-free.
- The bound is in operator norm which is a *strong* bound for the operator convergence as it implies many others: eigenvalue and eigenvector convergence by perturbation theory results, bound on the associated semi-group, pointwise convergence or other forms of weak convergence.

2.4.2. Bias analysis: consistency of the estimator

The bias analysis is a little bit trickier although all objects are now deterministic. We know that $\mathsf{L}_\lambda^{-1/2}\mathsf{C}\mathsf{L}_\lambda^{-1/2}$ is a compact operator (C is compact and $\mathsf{L}_\lambda^{-1/2}$ bounded) so that its spectrum is discrete and is formed by isolated points except from 0. On the same manner [RS12, Theorem XIII.67] the inverse of the diffusion operator $\mathcal{L}^{-1}$ is compact so that we can talk of the approximation of the k-th eigenelement of $\mathcal{L}^{-1}$ by the one of $\mathsf{L}_\lambda^{-1/2}\mathsf{C}\mathsf{L}_\lambda^{-1/2}$ as λ goes to 0 (or eigenspaces if the eigenvalues are not isolated).

Consistency of the estimator. First, if we are only interested in consistency of the estimator and not on rates of convergence we have the following consistency result:

Proposition 26 (*Convergence of the bias*)

Under assumptions 1, 2 *and* 3, *we have the following convergence in operator norm:*

$$\left\|\mathsf{L}_\lambda^{-1/2}\mathsf{C}\mathsf{L}_\lambda^{-1/2} - \mathsf{L}^{-1/2}\mathsf{C}\mathsf{L}^{-1/2}\right\| \underset{\lambda\to 0}{\longrightarrow} 0 \tag{119}$$

Note that this result was stated under the assumption that $\mathsf{L}^{-1/2}\mathsf{C}\mathsf{L}^{-1/2}$ was compact in the previous Section 1. With the abstract framework developed in this section, we showed flawlessly that $\mathsf{L}^{-1/2}\mathsf{C}\mathsf{L}^{-1/2}$ was compact (Proposition 24), combining this new result with the proof of Proposition 13 of the previous section, we show Proposition 26. Together with the convergence result of the variance above, this shows that our estimate is statistically consistent when taking a sequence of regularizers such that $\lambda_n \to 0$ and $\lambda_n\sqrt{n} \to +\infty$.

2.4.3. Where this work stands: the bias analysis and the possibility of deriving dimensionless rates of convergence.

For the whole story to be totally complete, and to highlight the differences between our estimator and Diffusion maps [CL06] (especially in high-dimension), we need to guarantee fast and explicit convergence rates. This is the case for the statistical variance term as shown in Proposition 25, but as said before, the bias term is a lot trickier and necessitates some deeper knowledge on the spectrum of the diffusion operator $\mathcal{L}$. In this subsection, we will not show rigorous results as this is exactly where our current reflection lies. Yet, we will try to explain why deriving these explicit rates is a hard task and how it could be done in the future.

Convergence rates. Now, let us try to derive some rates of convergence for the spectrum of the biased operator to the true one. We will focus first only on the top eigenvalue and eigenvector of the operator as we can derive the same analysis with some $\min - \max$ Courant-Fisher principle for the other eigenelements.

Let us denote by $f_* \in H^1(\mu)$ the first eigenvector of the diffusion operator to approximate and $\sigma_* > 0$ its associated eigenvalue : $\mathcal{L}f_* = \sigma_* f_*$. Similarly, let us denote $f_{\mathcal{H}}$ the largest eigenvector of $\mathsf{L}_\lambda^{-1/2}\mathsf{C}\mathsf{L}_\lambda^{-1/2}$ and $\sigma_{\mathcal{H}} > 0$ its associated eigenvalue. As eigenvectors are defined up to a multiplicative constant, we can normalize them such that $\|f_*\|_2^2 = \|f_{\mathcal{H}}\|_2^2 = 1$, where the $\|\cdot\|_2$ norm is the $L^2(\mu)$ usual norm. Let us stress out that the whole story can be seen as the approximation property of the eigenfunctions of $\mathcal{L}$ by functions of the RKHS.

If f_* belongs to the native RKHS space. As in the regression problem (either for kernel ridge regression or stochastic gradient descent as described in Part III Section 2), if the function to be approximated, f_*, lies in the RKHS $\mathcal{H}$, then the problem is easy and the rates of convergence are dimensionless! More precisely we can show that:

$$|\sigma_{\mathcal{H}} - \sigma^*| \leqslant \lambda \|f_*\|_{\mathcal{H}}^2, \tag{120}$$

Hence, if $f_* \in \mathcal{H}$, we have convergence of the biased operator at rate λ! The problem is that when μ has whole support in $\mathbb{R}^d$ then we expect the function f_* to be regular but to not decrease at infinity. For example, when μ is Gaussian, the eigenfunctions of the diffusion operator $\mathcal{L}$ are the Hermite polynomials that do not belong to generic RKHS such as Gaussian or Laplace. However, this case is also quite informative as when μ has a compact support, and the potential is smooth enough (this is what is supposed in [HAL07, CL06]), then we can always take a Sobolev kernel and have $f_* \in \mathcal{H}$. In this case, to go further and really try to not hide the curse of dimensionality, one could try to write explicitly $\|f_*\|_{\mathcal{H}}^2$ as a function of the dimension. This will depend on the tensorisation properties of the measure μ but we leave this for a future work.

If f_* does not belong to the native RKHS space. If $f_* \notin \mathcal{H}$, then it becomes complicated. The first idea is to abstractly assume some approximation property of the function f_* by the RKHS with respect to the problem. It is exactly the purpose of the *source condition* in our earlier work: we formulate a general assumption to quantify how we can really approach a function outside the RKHS (see Eq. (46) in the RKHS part of the Introduction for a precise definition).

However, here, the function f_* can be defined explicitly and has some interesting properties that we could leverage: we know that it lies in $H^1(\mu)$ and that being an eigenfunction of $\mathcal{L}$, it has some strong

regularity. Hence, we can go deeper in the analysis and prefer a more constructive approach. Let us fix $\varepsilon > 0$, the idea is to build a function in $\mathcal{H}$, namely $\widehat{f}_{\mathcal{H}}$ such that:

(i) $\widehat{f}_{\mathcal{H}} \sim f_*$ in $H^1(\mu)$. Typically, $\left\|\widehat{f}_{\mathcal{H}} - f_*\right\|^2_{H^1(\mu)} \leqslant \varepsilon$.

(ii) $\|\widehat{f}_{\mathcal{H}}\|_{\mathcal{H}}$ is not too large. Typically, $\left\|\widehat{f}_{\mathcal{H}}\right\|_{\mathcal{H}} \leqslant \varepsilon^{-\delta}$, with $\delta > 0$.

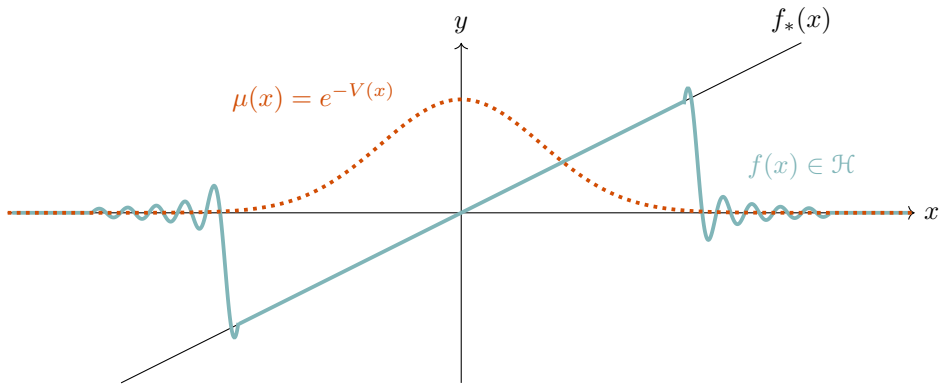

Figure 20: Schematic figure showing the way to approximate the first eigenvector of the diffusion operator by a function in an RKHS. The example shows this in the case of a Gaussian measure in $\mathbb{R}$ for which the first eigenfunction is the monome $f_* \ : \ x \to x$.

This is represented is Figure 20 and is often the way to derive bounds for general functional inequalities when changing spaces of *test functions*, see [BGL14, p378] for such an example. In this case for the eigenvalue, we have a similar bound as previously:

$$|\sigma_{\mathcal{H}} - \sigma^*| \lessapprox \lambda \|\widehat{f}_{\mathcal{H}}\|^2_{\mathcal{H}}, \tag{121}$$

And for the eigenfunction bound, if $\angle^*_{\mathcal{H}}$ is the acute angle between $f_{\mathcal{H}}$ and the eigenspace associated with eigenvalue σ_*, then

$$\sin\left(\angle^*_{\mathcal{H}}\right) \lessapprox \frac{2\lambda \|\widehat{f}_{\mathcal{H}}\|^2_{\mathcal{H}}}{\mathsf{dist}(\sigma_*, \text{spectrum})}, \tag{122}$$

where $\mathsf{dist}(\sigma_*, \text{spectrum})$ is the eigengap between the eigenspace associated to σ_* and the rest of the spectrum.

$\star$

$\star \quad \star$

What is proven and what remains to be: a battle for completeness. We hope to have convinced the reader that the story behind the estimation of the whole diffusion operator is almost done. The fact that remains to be correctly stated is the approximation property of our RKHS with respect to the eigenfunctions of the operator. To conclude this part, we really would like to put emphasis that often in the learning RKHS literature, the approximation properties that could carry the curse of dimensionality is hidden through some technical assumption (e.g. source condition). Here, as we have a strong prior on the function to be approximated, we would like to try to be as precise as possible and not rest on some technical assumption without proper intuitive mathematical content.

2.5. Conclusion and further thoughts

Comparison to Diffusion maps. In this work, we tried to prove that we could estimate the eigenelements of the diffusion operator. This construction relies on positive kernel methods on the contrary to older methods relying on local averaging techniques. This could lead to efficient and robust estimation in higher dimension in comparison to Diffusion maps [CL06]. However, to really compare to this celebrated work, further paths should be explored, and one of the most important is that Diffusion maps can have access to the geometric structure of the data with an appropriate reweighting. Is this possible to do the same with our technique?

The kernel choice. Another discussion that we only sketched is the crucial choice of the kernel in such a situation. As recalled throughout the Introduction, the art of kernel engineering is a central question to apply psd kernel methods. For this problem, the kernel should be respecting two guidelines: (i) it should have a Mercer decomposition with explicit features (random or appropriately chosen features). Indeed, we need to build the covariance matrix with respect to the derivative kernel $\widehat{\mathsf{L}}$ that can be very costly if we do not have explicit features. More precisely if we approximate our kernel with M features we go from size $O(nM)$ to $O(n^2 d)$ matrix: this saves a huge cost. (ii) As explained in the last subsection, the RKHS should be chosen to approximate well the eigenfunctions of the diffusion operator in $H^1(\mu)$. Remark that the weakness of the RKHS comes this time from the decreasing at infinity and not from the regularity. This should enable dimensionless approach in most of the cases.

Tensorisation property. To go as deep as it can be, an important property of eigenelements of the diffusion operator is its tensorisation nature: if μ is a tensor product measure over $\mathbb{R}^d$ then the eigenfunctions will have also a tensorized form and lie in low-dimensional spaces. This should be taken into account when designing the kernel: it should leverage this possible tensorisation property to exploit the low-dimensional structure of the object.

Other works. A recent literature in applied probability aims at estimating the spectral gaps of Markov Chain given the first n iterates of it [HKS15]. Our procedure seems to do exactly the same, and understanding the difference between our algorithm and theirs is something we have to explore. This will require to adapt a bit our algorithm and change our i.i.d. assumption on the samples to a Markovian one.

⋆

⋆ ⋆

Part IV

Conclusion and future work

1. Summary of the thesis

In the course of this manuscript, we have investigated two topics. First, we focused on the convergence properties of the stochastic gradient descent algorithm in Hilbert spaces. Second, we have studied dimensionality-reduction techniques through the estimation of the Laplacian operator associated to the data. In this context, we also have come up with a new algorithm to estimate this latter operator and explain that leveraging this knowledge could lead to accelerate Monte Carlo Markov Chains in high-dimension. Note that for these two directions, even if my background and personal preferences has driven me more to modelling and theoretical questions, I have also been concerned with the numerical efficiency of the algorithms involved and their implementation.

To summarize more precisely the contributions of this thesis, let us dive first into the optimization framework at stake when it comes to solve high-dimensional supervised learning problems. In supervised learning, stochastic gradient methods are ubiquitous, both from the practical side, as a simple algorithm that can learn from a single or a few passes over the data, and from the theoretical side, as it leads to optimal rates for estimation problems in a variety of situations.

In this context, we first showed that, under a margin assumption [AT07], for classification problems, the classification error of the averaged iterates of stochastic gradient descent converge exponentially fast to the best achievable error although the regression error had only a $O(1/n)$ convergence rate, hence establishing theoretically a largely observed behavior in practice.

One of the efforts in theoretical machine learning is to understand or improve the empirical rules of practitioners. However, until our second contribution, theory predicted that practitioners should stop the stochastic gradient descent iterations after having only seen once the whole data-set ($\sim n$ iterations). The second contribution of this thesis establishes some map on the hardness of a learning problem and showed that for hard problems, it was necessary to stop stochastic gradient descent only after several passes over the whole data-set reconciling on this aspect theory and practice. Stating optimality in the context of non-parametric regression requires both a source condition (which quantifies the smoothness of the optimal prediction function) and a capacity condition (related to the eigenvalue decay of the covariance operators).

We show that multiple-pass averaging, combined with larger step sizes than traditional approaches, allows to get this optimal behavior.

In a second direction, we worked on a new kernel-based algorithm for unsupervised learning. This project was initiated by discussions with researchers in molecular dynamics (Tony Lelièvre and Gabriel Stoltz) to find reaction coordinates in molecular systems which are central objects to simulate efficiently and understand such systems. This project gave birth to a new practical algorithm that could have a large area of applications. Indeed, based on a variational formulation of the underlying diffusion operator that generated the samples, we designed a dimensionality-reduction algorithm. One of the main result of this project is that we estimated with a kernel-based method (thus avoiding the curse of dimensionality and hard parameter tuning as in previous approaches [CL06]) the infinitesimal operator of the overdamped Langevin diffusion (or general Laplacian). The eigenelements of such an operator are the cornerstone of clustering or dimensionality reduction techniques.

⋆
⋆ ⋆

2. Perspectives

We have already largely developed in the introduction possible future works, by putting emphasis on interesting unexplored directions. Among them, we have already spoken of my personal interest for the continuous dynamics behind SGD iterations in subsection 2.2.2, the art of engineering kernels that can mimic the expressivity of neural networks in subsection 3.3.2, or the link between statistical physics and Machine Learning in subsection 4.1.2. In this perspective part, even if it is a crucial point and a possible future direction in my academic career, I will not dwell into the kernel topic. I refer to 3.3.2 for detail discussions on it.

To expose clearly the future works that I will be probably conducting during the following months, let us divide them in the two directions taken during this thesis.

Stochastic gradient descent. Let us explain the questions triggered by our contribution on this aspect.

1. *Optimality for classification problems.* In the second work on stochastic gradient descent, the fundamental question was optimality of SGD, whereas we only showed "fast rates" for the classification error in the first work. A natural idea would be to merge these two works wondering if we can reach optimality with SGD for classification problems under low noise conditions.

2. *Continuous-time SGD dynamics.* The Langevin dynamics seems very related to stochastic gradient descent with additive noise as we have seen in Section 2 of this introduction. In this same direction, understanding continuous versions of the algorithms often leads to a deeper comprehension of the behavior of the systems. This could lead to the study of the continuous counterpart of the stochastic gradient descent algorithm [LT19]. Continuing further the discussion introduced in subsection 2.2.2 is one of the first questions I will be studying in the short-time future.

3. *Noiseless setting.* Lately, the community has been curious about the "noiseless setting" [BBG20, VBS19] where the only source of noise in the recursion is multiplicative. In this setting, there is some lack of knowledge, in particular concerning minimax rates or how the implicit bias of SGD can lead to good generalization error. Digging in this direction seems to be promising to understand the impact of the noise is ML models and the good generalization properties of SGD.

Estimation of Laplacian and dimensionality reduction. We distinguish here three possible extensions of our work:

1. *Finishing current work !* The first thing would obviously be to finish the work presented in Section 2 of Part III. Showing its applicability in molecular dynamics and relevance in the statistics community is the first thing I will consider doing.

2. *Approximation of the Laplace-Beltrami operator.* We have developed a kernel method to build from samples the diffusion operator associated with the measure that produced them. For data distributed according to a particular geometry, it seems possible to extend our procedure to construct an approximation of the Laplace-Beltrami operator associated with the sub-manifold to which these data belong. The Laplace-Beltrami operator estimation is the cornerstone of geometric data analysis approaches and clustering methods widely used in practice [VL07]. The construction of these operators from psd kernels would possibly allow to avoid the curse of dimensionality, on the contrary of currently used local averaging methods [HAL07, CL06]. Showing the consistency of such estimates of operators or of their spectral elements (eigenvectors and values) requires detailed knowledge of spectral approximation methods.

3. *Acceleration of sampling procedures.* As discussed in the introduction, one of the main problems of sampling large dimensional non-strongly logarithmic concave measures is that the dynamics used to sample the said measure can get stuck for a long time in localized modes of the distribution. This phenomenon is called metastability (see Section 4.3 for details). To speed up such sampling procedures, one of the common techniques in molecular dynamics is to use importance sampling strategies by biasing the target measure by a low-dimensional function representing the slow directions of diffusion (free energy associated with a reaction coordinate). The estimation of these reaction coordinates therefore allows the use of acceleration methods for the sampling of multimodal measurements. If in this thesis, we implemented numerically the estimation of coordinates of linear and low-dimensional reactions in an idealized framework (mixture of three Gaussians in dimension 2), the objective is to develop this procedure in more realistic cases corresponding to real molecular systems (with applications in pharmacology) or to sample the *a posteriori* laws of the parameters of Gaussian mixtures within the framework of Bayesian inference [CLS12]. Two important obstacles must be removed for the practical application of the procedure: being able to estimate reaction coordinates (i) in high dimension and (ii) that could potentially be non-linear.

Learning/statistical physics interaction. In the medium term, I would like to work on problems at the interface between statistical physics, molecular dynamics and learning. Indeed, in addition to the approaches based on the estimation of a diffusion operator mentioned above, many links exist between statistical learning and statistical physics. Moreover, even if the objectives of the two disciplines are different, the dynamics under study in statistical physics can be seen as the continuous counterpart of the stochastic gradient descent. The large dimension and the metastability of such dynamics, especially in the case of Bayesian inference in large dimension are important aspects common to both disciplines and the acceleration techniques developed in molecular dynamics could be extensively used to solve learning problems. More importantly, one active field of study in statistical physics and Machine Learning is to try to accelerate the stochastic dynamics at hand. These techniques rely on changing it without affecting the invariant measure while accelerating the convergence to equilibrium. We could (i) study Kinetic versions of Langevin dynamics [MCC+19], (ii) add some drift term or (iii) some non-reversibility through birth-and-death processes [LLN19]. How do these techniques transfer for stochastic optimization purposes ?

⋆
⋆ ⋆

References

[ADT20] Alnur Ali, Edgar Dobriban, and Ryan J Tibshirani. The implicit regularization of stochastic gradient flow for least squares. *arXiv preprint arXiv:2003.07802*, 2020.

[AF03] Robert A. Adams and John J.F. Fournier. *Sobolev spaces*, volume 140. Academic Press, 2003.

[AMP00] R. Aguech, E. Moulines, and P. Priouret. On a perturbation approach for the analysis of stochastic tracking algorithms. *SIAM J. Control and Optimization*, 39(3):872–899, 2000.

[Aro50] Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.

[AT07] Jean-Yves Audibert and Alexandre B. Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of statistics*, 35(2):608–633, 2007.

[AZO14] Zeyuan Allen-Zhu and Lorenzo Orecchia. Linear coupling: An ultimate unification of gradient and mirror descent. *arXiv preprint arXiv:1407.1537*, 2014.

[Bac14] Francis Bach. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *Journal of Machine Learning Research*, 15(19):595–627, 2014.

[Bac17] Francis Bach. On the equivalence between kernel quadrature rules and random feature expansions. *Journal of Machine Learning Research*, 18(21):1–38, 2017.

[BB08] Léon Bottou and Olivier Bousquet. The tradeoffs of large scale learning. In *Advances in neural information processing systems*, pages 161–168, 2008.

[BBG20] Raphaël Berthier, Francis Bach, and Pierre Gaillard. Tight nonparametric convergence rates for stochastic gradient descent under the noiseless linear model, 2020.

[BBM05] Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *Ann. Statist.*, 33(4):1497–1537, 08 2005.

[BBV04] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[BCN18] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

[BGL14] Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and Geometry of Markov Diffusion Operators*. Springer, 2014.

[Bis95] Chris M Bishop. Training with noise is equivalent to tikhonov regularization. *Neural computation*, 7(1):108–116, 1995.

[BJM06] Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

[BK16] Gilles Blanchard and Nicole Krämer. Convergence rates of kernel conjugate gradient for random design regression. *Analysis and Applications*, 14(06):763–794, 2016.

[BLC05] L. Bottou and Y. Le Cun. On-line learning for very large data sets. *Applied Stochastic Models in Business and Industry*, 21(2):137–151, 2005.

[BLJ04] Francis R Bach, Gert RG Lanckriet, and Michael I Jordan. Multiple kernel learning, conic duality, and the smo algorithm. In *Proceedings of the twenty-first international conference on Machine learning*, page 6, 2004.

[BLS15] Sébastien Bubeck, Yin Tat Lee, and Mohit Singh. A geometric alternative to nesterov's accelerated gradient descent. *arXiv preprint arXiv:1506.08187*, 2015.

[BM11] F. Bach and E. Moulines. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2011.

[BM13] Francis Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$. In *Advances in Neural Information Processing Systems (NIPS)*, pages 773–781, 2013.

[BM17] Gilles Blanchard and Nicole Mücke. Optimal rates for regularization of statistical inverse learning problems. *Foundations of Computational Mathematics*, pages 1–43, 2017.

[BMAS14] Nicolas Boumal, Bamdev Mishra, P.-A. Absil, and Rodolphe Sepulchre. Manopt, a Matlab toolbox for optimization on manifolds. *Journal of Machine Learning Research*, 15:1455–1459, 2014.

[Bob99] Sergey G. Bobkov. Isoperimetric and analytic inequalities for log-concave probability measures. *Ann. Probab.*, 27:1903–1921, 1999.

[Boy04] Stephen Boyd. Subgradient methods. 2004.

[BRH13] Nawaf Bou-Rabee and Martin Hairer. Nonasymptotic mixing of the mala algorithm. *IMA Journal of Numerical Analysis*, 33(1):80–110, 2013.

[CB20] Lenaic Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. *arXiv preprint arXiv:2002.04486*, 2020.

[CBLK06] Ronald Coifman, Nadler Boaz, Stéphane Lafon, and Ioannis Kevrekidis. Diffusion maps, spectral clustering and reaction coordinates of dynamical systems. *Applied and Computational Harmonic Analysis*, 21(12):113–127, 2006.

[CBR18] Carlo Ciliberto, Francis Bach, and Alessandro Rudi. Localized structured prediction. *arXiv preprint arXiv:1806.02402*, 2018.

[CCS+19] Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-sgd: Biasing gradient descent into wide valleys. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124018, 2019.

[CDV07] Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

[CFG+20] Benjamin Charlier, Jean Feydy, Joan Alexis Glaunès, François-David Collin, and Ghislain Durif. Kernel operations on the gpu, with autodiff, without memory overflows, 2020.

[Che81] Herman Chernoff. A note on an inequality involving the normal distribution. *Ann. Probab.*, 9(3):533–535, 1981.

[CL06] Ronald R. Coifman and Stéphane Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1), 2006.

[CLS12] Nicolas Chopin, Tony Lelièvre, and Gabriel Stoltz. Free energy methods for Bayesian inference: efficient exploration of univariate Gaussian mixture posteriors. *Statistics and Computing*, 22(4):897–916, 2012.

[CM10] Djalil Chafaï and Florent Malrieu. On fine properties of mixtures with respect to concentration of measure and Sobolev type inequalities. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 46:72–96, 2010.

[Cra76] Charles R Crawford. A stable generalized eigenvalue problem. *SIAM Journal on Numerical Analysis*, 13(6):854–860, 1976.

[CRR16] Carlo Ciliberto, Lorenzo Rosasco, and Alessandro Rudi. A consistent regularization approach for structured prediction. In *Advances in Neural Information Processing Systems*, 2016.

[CRR18] Luigi Carratino, Alessandro Rudi, and Lorenzo Rosasco. Learning with sgd and random features. *arXiv preprint arXiv:1807.06343*, 2018.

[CRRP17] Carlo Ciliberto, Alessandro Rudi, Lorenzo Rosasco, and Massimiliano Pontil. Consistent multitask learning with nonlinear output relations. In *Advances in Neural Information Processing Systems*, pages 1986–1996, 2017.

[CT19] Richard Combes and Mikael Touati. Computationally efficient estimation of the spectral gap of a markov chain. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 3(1):7, 2019.

[Dal17] Arnak Dalalyan. Further and stronger analogy between sampling and optimization: Langevin monte carlo and gradient descent. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 678–689. PMLR, 07–10 Jul 2017.

[Dau92] Ingrid Daubechies. *Ten lectures on wavelets*. SIAM, 1992.

[DB15] A. Défossez and F. Bach. Constant step size least-mean-square: Bias-variance trade-offs and optimal sampling distributions. In *Proc. AISTATS*, 2015.

[DB16] Aymeric Dieuleveut and Francis Bach. Nonparametric stochastic approximation with large step-sizes. *The Annals of Statistics*, 44(4):1363–1399, 2016.

[DBLJ14] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems*, 2014.

[DDB17] Aymeric Dieuleveut, Alain Durmus, and Francis Bach. Bridging the gap between constant step size stochastic gradient descent and markov chains. *arXiv preprint arXiv:1707.06386*, 2017.

[DFB17] Aymeric Dieuleveut, Nicolas Flammarion, and Francis Bach. Harder, better, faster, stronger convergence rates for least-squares regression. *Journal of Machine Learning Research*, pages 1–51, 2017.

[DGL13] Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*, volume 31. Springer Science & Business Media, 2013.

[Die17] Aymeric Dieuleuveut. *Stochastic approximation in Hilbert spaces*. PhD thesis, ENS - INRIA, 2017.

[DM17] Alain Durmus and Eric Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *Annals of Applied Probability*, 27(3):1551–1587, 2017.

[DM19] Alain Durmus and Eric Moulines. High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *Bernoulli*, 25(4A):2854–2882, 2019.

[DRVZ17] Alain Durmus, Gareth O. Roberts, Gilles Vilmart, and Konstantinos C. Zygalakis. Fast Langevin based algorithm for mcmc in high dimensions. *Ann. Appl. Probab.*, 27(4):2195–2237, 08 2017.

[DVRT14] Ernesto De Vito, Lorenzo Rosasco, and Alessandro Toigo. Learning sets with separating kernels. *Applied and Computational Harmonic Analysis*, 37(2):185–217, 2014.

[EAS98] Alan Edelman, Tomás A. Arias, and Steven T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

[EK+10] Noureddine El Karoui et al. The spectrum of kernel random matrices. *The Annals of Statistics*, 38(1):1–50, 2010.

[FBG07] Kenji Fukumizu, Francis Bach, and Arthur Gretton. Statistical consistency of kernel canonical correlation analysis. *Journal of Machine Learning Research*, 8:361–383, 2007.

[FBJ04] Kenji Fukumizu, Francis Bach, and Michael I. Jordan. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *Journal of Machine Learning Research*, 5(Jan):73–99, 2004.

[FS01] Daan Frenkel and Berend Smit. *Understanding Molecular Simulation: From Algorithms to Applications*, volume 1. Elsevier, 2001.

[FS17] Simon Fischer and Ingo Steinwart. Sobolev norm learning rates for regularized least-squares algorithm. Technical Report 1702.07254, arXiv, 2017.

[FW98] Mark Iosifovich Freidlin and Alexander D Wentzell. Random perturbations. In *Random perturbations of dynamical systems*, pages 15–43. Springer, 1998.

[G+] Andreas Griewank et al. On automatic differentiation.

[Gan10] Klaus Gansberger. An idea on proving weighted Sobolev embeddings. *arXiv:1007.3525.*, 2010.

[GBR+12] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

[Gke19] Paraskevi Gkeka. Machine learning force field and coarse-grained variables in molecular dynamics: application to materials and biological systems. *Preprint*, 2019.

[GL06] Dani Gamerman and Hedibert Lopes. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Chapman and Hall/CRC, 2006.

[GOP15] Mert Gürbüzbalaban, Asu Ozdaglar, and Pablo Parrilo. Why random reshuffling beats stochastic gradient descent. Technical Report 1510.08560, arXiv, 2015.

[Goz10] Nathael Gozlan. Poincaré inequalities and dimension free concentration of measure. *Ann. Inst. H. Poincaré Probab. Statist.*, 46(3):708–739, 2010.

[GRO+08] L Lo Gerfo, L Rosasco, F Odone, E De Vito, and A Verri. Spectral algorithms for supervised learning. *Neural Computation*, 20(7):1873–1897, 2008.

[GRS95] Walter Gilks, Sylvia Richardson, and David Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Chapman and Hall/CRC, 1995.

[GT01] David Gilbarg and Neil S. Trudinger. *Elliptic Partial Differential Equations of Second Order*. Springer Berlin Heidelberg, 2001.

[Gu13] Chong Gu. *Smoothing spline ANOVA models*, volume 297. Springer Science & Business Media, 2013.

[HAL07] Matthias Hein, Jean-Yves Audibert, and Ulrike von Luxburg. Graph laplacians and their convergence on random neighborhood graphs. *Journal of Machine Learning Research*, pages 1325–1368, 2007.

[HKS15] Daniel Hsu, Aryeh Kontorovich, and Csaba Szepesvári. Mixing time estimation in reversible markov chains from a single sample path. In *Advances in neural information processing systems*, pages 1459–1467, 2015.

[HN92] Robert Hecht-Nielsen. Theory of the backpropagation neural network. In *Neural networks for perception*, pages 65–93. Elsevier, 1992.

[HN05] Bernard Helffer and Francis Nier. Hypoelliptic estimates and spectral theory for Fokker-Planck operators and Witten Laplacians. *Lecture Notes in Mathematics*, 1862, 2005.

[Hot33] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

[HRS16] Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, 2016.

[HTF09] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[IL67] Alekse Grigorevich Ivakhnenko and Valentin Grigorevich Lapa. Cybernetics and forecasting techniques. 1967.

[JGH18] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8571–8580, 2018.

[JK17] Prateek Jain and Purushottam Kar. Non-convex optimization for machine learning. *Foundations and Trends in Machine Learning*, 10(3-4):142–336, 2017.

[JKK+16] Prateek Jain, Sham M. Kakade, Rahul Kidambi, Praneeth Netrapalli, and Aaron Sidford. Parallelizing stochastic approximation through mini-batching and tail-averaging. Technical Report 1610.03774, arXiv, 2016.

[JNN19] Prateek Jain, Dheeraj Nagaraj, and Praneeth Netrapalli. Making the last iterate of sgd information theoretically optimal. In *Conference on Learning Theory*, pages 1752–1755, 2019.

[JZ13] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, 2013.

[KB05] Vladimir Koltchinskii and Olexandra Beznosova. Exponential convergence rates in classification. In *International Conference on Computational Learning Theory*. Springer, 2005.

[KP13] Peter E Kloeden and Eckhard Platen. *Numerical solution of stochastic differential equations*, volume 23. Springer Science & Business Media, 2013.

[KT09] Sham M. Kakade and Ambuj Tewari. On the generalization ability of online strongly convex programming algorithms. In *Advances in Neural Information Processing Systems*, 2009.

[Lan12] Guanghui Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1-2):365–397, 2012.

[LBH15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[Led14] Michel Ledoux. Concentration of measure and logarithmic Sobolev inequalities. *Séminaire de Probas XXXIII*, pages 120–216, 2014.

[Lel09] Tony Lelièvre. A general two-scale criteria for logarithmic Sobolev inequalities. *Journal of Functional Analysis*, 256(7):2211 – 2221, 2009.

[Lel13] Tony Lelièvre. Two mathematical tools to analyze metastable stochastic processes. In *Numerical Mathematics and Advanced Applications 2011*, pages 791–810, Berlin, Heidelberg, 2013. Springer.

[Lel15] Tony Lelièvre. Accelerated dynamics: Mathematical foundations and algorithmic improvements. *The European Physical Journal Special Topics*, 224(12):2429–2444, 2015.

[LJSB12] Simon Lacoste-Julien, Mark Schmidt, and Francis Bach. A simpler approach to obtaining an o (1/t) convergence rate for the projected stochastic subgradient method. *arXiv preprint arXiv:1212.2002*, 2012.

[LLN19] Yulong Lu, Jianfeng Lu, and James Nolen. Accelerating Langevin sampling with birth-death. *arXiv preprint arXiv:1905.09863*, 2019.

[LM+16] Guillaume Lecué, Shahar Mendelson, et al. Performance of empirical risk minimization in linear aggregation. *Bernoulli*, 22(3):1520–1534, 2016.

[LP16] David Levin and Yuval Peres. Estimating the spectral gap of a reversible markov chain from a short trajectory. *arXiv preprint 1612.05330*, 2016.

[LR17] Junhong Lin and Lorenzo Rosasco. Optimal rates for multi-pass stochastic gradient methods. *Journal of Machine Learning Research*, 18(97):1–47, 2017.

[LRRC18] Junhong Lin, Alessandro Rudi, Lorenzo Rosasco, and Volkan Cevher. Optimal rates for spectral algorithms with least-squares regression over hilbert spaces. *Applied and Computational Harmonic Analysis*, 2018.

[LRS08] Tony Lelièvre, Mathias Rousset, and Gabriel Stoltz. Long-time convergence of an adaptive biasing force method. *Nonlinearity*, 21(6):1155, 2008.

[LS16] Tony Lelièvre and Gabriel Stoltz. Partial differential equations and stochastic methods in molecular dynamics. *Acta Numerica*, 25:681–880, 2016.

[LT19] Qianxiao Li and Cheng Tai. Stochastic modified equations and dynamics of stochastic gradient algorithms I: Mathematical foundations. *J. Mach. Learn. Res.*, 20:40–1, 2019.

[MCC+19] Yi-An Ma, Niladri Chatterji, Xiang Cheng, Nicolas Flammarion, Peter Bartlett, and Michael I Jordan. Is there an analog of Nesterov acceleration for MCMC? *arXiv preprint arXiv:1902.00996*, 2019.

[MCJ+18] Yi-An Ma, Yuansi Chen, Chi Jin, Nicolas Flammarion, and Michael I. Jordan. Sampling can be faster than optimization. *arXiv preprint arXiv:1612.05330*, 2018.

[MCRR20] Giacomo Meanti, Luigi Carratino, Lorenzo Rosasco, and Alessandro Rudi. Kernel methods through the roof: handling billions of points efficiently. 2020.

[MFSS17] K Muandet, K Fukumizu, B Sriperumbudur, and B Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning*, 10(1-2):1–144, 2017.

[MHB17] Stephan Mandt, Matthew D. Hoffman, and David M. Blei. Stochastic gradient descent as approximate Bayesian inference. *J. Mach. Learn. Res.*, 18(1):4873–4907, January 2017.

[MKHS14] Julien Mairal, Piotr Koniusz, Zaid Harchaoui, and Cordelia Schmid. Convolutional kernel networks. In *Advances in neural information processing systems*, pages 2627–2635, 2014.

[MM19] Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355*, 2019.

[MS14] Georg Menz and André Schlichting. Poincaré and logarithmic Sobolev inequalities by decomposition of the energy landscape. *Ann. Probab.*, 42(5):1809–1884, 09 2014.

[MSS+99] Sebastian Mika, Bernhard Schölkopf, Alex J Smola, Klaus-Robert Müller, Matthias Scholz, and Gunnar Rätsch. Kernel pca and de-noising in feature spaces. In *Advances in neural information processing systems*, pages 536–542, 1999.

[MT99] Enno Mammen and Alexandre Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27(6):1808–1829, 1999.

[MT12] Sean P. Meyn and Richard L. Tweedie. *Markov Chains and Stochastic Stability*. Springer Science & Business Media, 2012.

[MT20] Per-Gunnar Martinsson and Joel Tropp. Randomized numerical linear algebra: Foundations & algorithms. *arXiv preprint arXiv:2002.01387*, 2020.

[Mur12] Kevin P. Murphy. *Machine Learning: a Probabilistic Perspective*. MIT Press, 2012.

[MXZ06] Charles A Micchelli, Yuesheng Xu, and Haizhang Zhang. Universal kernels. *Journal of Machine Learning Research*, 7(Dec):2651–2667, 2006.

[Nad64] Elizbar A Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142, 1964.

[Nes83] Yurii Nesterov. A method for unconstrained convex minimization problem with the rate of convergence o (1/kˆ 2). In *Doklady an ussr*, volume 269, pages 543–547, 1983.

[Nes12] Yu Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

[NJLS09] 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.

[NV08] Y. Nesterov and J. P. Vial. Confidence level solutions for stochastic programming. *Automatica*, 44(6):1559–1568, 2008.

[NY83] A. S. Nemirovski and D. B. Yudin. *Problem complexity and method efficiency in optimization*. John Wiley, 1983.

[OBLJ17] Anton Osokin, Francis Bach, and Simon Lacoste-Julien. On structured prediction theory with calibrated convex surrogate losses. In *Advances in Neural Information Processing Systems*, 2017.

[Opf06] Roland Opfer. Multiscale kernels. *Advances in computational mathematics*, 25(4):357–380, 2006.

[OR07] Felix Otto and Maria G. Reznikoff. A new criterion for the logarithmic Sobolev inequality and two applications. *Journal of Functional Analysis*, 243(1):121–157, 2007.

[Pea01] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

[Pfl86] Georg Ch Pflug. Stochastic minimization with constant step-size: asymptotic laws. *SIAM Journal on Control and Optimization*, 24(4):655–666, 1986.

[PGC+17] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[Pin94] Iosif Pinelis. Optimum bounds for the distributions of martingales in banach spaces. *The Annals of Probability*, pages 1679–1706, 1994.

[PJ92] B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

[Pol64] Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

[PVBL+20] Loucas Pillaud-Vivien, Francis Bach, Tony Lelièvre, Alessandro Rudi, and Gabriel Stoltz. Statistical estimation of the Poincaré constant and application to sampling multimodal distributions. In *International Conference on Artificial Intelligence and Statistics*, pages 2753–2763, 2020.

[PVRB18] Loucas Pillaud-Vivien, Alessandro Rudi, and Francis Bach. Exponential convergence of testing error for stochastic gradient methods. In *Proceedings of the 31st Conference On Learning Theory*, volume 75, pages 250–296, 2018.

[QHK+19] Qian Qin, James P Hobert, Kshitij Khare, et al. Estimating the spectral gap of a trace-class markov operator. *Electronic Journal of Statistics*, 13(1):1790–1822, 2019.

[RBV10] Lorenzo Rosasco, Mikhail Belkin, and Ernesto De Vito. On learning with integral operators. *Journal of Machine Learning Research*, 11(Feb):905–934, 2010.

[RCDVR14] Alessandro Rudi, Guillermo D Canas, Ernesto De Vito, and Lorenzo Rosasco. Learning sets and subspaces. *Regularization, Optimization, Kernels, and Support Vector Machines*, pages 337–357, 2014.

[RCR13] Alessandro Rudi, Guillermo D Canas, and Lorenzo Rosasco. On the sample complexity of subspace learning. In *Advances in Neural Information Processing Systems*, pages 2067–2075, 2013.

[RCR15] Alessandro Rudi, Raffaello Camoriano, and Lorenzo Rosasco. Less is more: Nyström computational regularization. In *Advances in Neural Information Processing Systems*, pages 1657–1665, 2015.

[RCR17] Alessandro Rudi, Luigi Carratino, and Lorenzo Rosasco. Falkon: An optimal large scale kernel method. In *Advances in Neural Information Processing Systems*, pages 3888–3898, 2017.

[RLM95] Kannan Ravindran, Lovasz Laszlo, and Simonovits Miklos. Isoperimetric problems for convex bodies and a localization lemma. *Discrete Comput. Geom.*, 13(3):541–559, 1995.

[RM51] H. Robbins and S. Monro. A stochastic approximation method. *Ann. Math. Statistics*, 22:400–407, 1951.

[Rob07] Christian Robert. *The Bayesian Choice: from Decision-theoretic Foundations to Computational Implementation*. Springer Science & Business Media, 2007.

[RR08] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pages 1177–1184, 2008.

[RR17] Alessandro Rudi and Lorenzo Rosasco. Generalization properties of learning with random features. In *Advances in Neural Information Processing Systems*, pages 3218–3228, 2017.

[RRT17] Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. In Satyen Kale and Ohad Shamir, editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 1674–1703, Amsterdam, Netherlands, 07–10 Jul 2017. PMLR.

[RS12] Michael Reed and Barry Simon. *Methods of Modern Mathematical Physics: Functional Analysis*, volume IV. Elsevier, 2012.

[RSB12] Nicolas L. Roux, Mark Schmidt, and Francis Bach. A stochastic gradient method with an exponential convergence rate for finite training sets. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

[RSS12] Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 1571–1578, 2012.

[RT14] Peter Richtárik and Martin Takáč. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming*, 144(1-2):1–38, 2014.

[RV15] Lorenzo Rosasco and Silvia Villa. Learning with incremental iterative regularization. In *Advances in Neural Information Processing Systems*, pages 1630–1638, 2015.

[RZMC11] Mary A. Rohrdanz, Wenwei Zheng, Mauro Maggioni, and Cecilia Clementi. Determination of reaction coordinates via locally scaled diffusion map. *The Journal of Chemical Physics*, 134(12):124116, 2011.

[Sal98] Ernesto Salinelli. Nonlinear principal components i. absolutely continuous random variables with positive bounded densities. *Ann. Statist.*, 26(2):596–616, 04 1998.

[SBC16] Weijie Su, Stephen Boyd, and Emmanuel J Candes. A differential equation for modeling nesterov's accelerated gradient method: theory and insights. *The Journal of Machine Learning Research*, 17(1):5312–5354, 2016.

[SC08] Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

[SD16] Aman Sinha and John C Duchi. Learning kernels with random features. In *Advances in Neural Information Processing Systems*, pages 1298–1306, 2016.

[SGSS07] Alex Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A hilbert space embedding for distributions. In *International Conference on Algorithmic Learning Theory*, pages 13–31. Springer, 2007.

[Sha11] Ohad Shamir. Making gradient descent optimal for strongly convex stochastic optimization. *CoRR*, abs/1109.5647, 2011.

[Sha16] Ohad Shamir. Without-replacement sampling for stochastic gradient methods. In *Advances in Neural Information Processing Systems 29*, pages 46–54, 2016.

[SHN+18] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

[SHS06] Ingo Steinwart, Don Hush, and Clint Scovel. An explicit description of the reproducing kernel hilbert spaces of gaussian rbf kernels. *IEEE Transactions on Information Theory*, 52(10):4635–4643, 2006.

[SHS09] Ingo Steinwart, Don R. Hush, and Clint Scovel. Optimal rates for regularized least squares regression. In *Proc. COLT*, 2009.

[SL13] Mark Schmidt and Nicolas Le Roux. Fast convergence of stochastic gradient descent under a strong growth condition. Technical Report 1308.6370, arXiv, 2013.

[Sol98] Mikhail V Solodov. Incremental gradient algorithms with stepsizes bounded away from zero. *Computational Optimization and Applications*, 11(1):23–35, 1998.

[SRL10] Gabriel Stoltz, Mathias Rousset, and Tony Lelièvre. *Free Energy Computations: A Mathematical Perspective*. World Scientific, 2010.

[SS00] Alex J Smola and Bernhard Schölkopf. Sparse greedy matrix approximation for machine learning. 2000.

[SS02] B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, 2002.

[SSBD14] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[SSSS07] 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)*, 2007.

[Sta17] Minsker Stanislav. On some extensions of Bernstein's inequality for self-adjoint operators. *Statistics and Probability Letters*, 127:111–119, 2017.

[STC04] John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[STS16] Ingo Steinwart, Philipp Thomann, and Nico Schmid. Learning with hierarchical gaussian kernels. *arXiv preprint arXiv:1612.00824*, 2016.

[SV09] Thomas Strohmer and Roman Vershynin. A randomized kaczmarz algorithm with exponential convergence. *Journal of Fourier Analysis and Applications*, 15(2):262, 2009.

[SW06] Robert Schaback and Holger Wendland. Kernel techniques: from machine learning to meshless methods. *Acta numerica*, 15:543, 2006.

[Tal94] Michel Talagrand. Sharper bounds for gaussian and empirical processes. *The Annals of Probability*, pages 28–76, 1994.

[TCKG05] B. Taskar, V. Chatalbashev, D. Koller, and C. Guestrin. Learning structured prediction models: A large margin approach. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2005.

[TKW16] James Townsend, Niklas Koep, and Sebastian Weichwald. Pymanopt: A python toolbox for optimization on manifolds using automatic differentiation. *Journal of Machine Learning Research*, 17(137):1–5, 2016.

[Tro12a] Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

[Tro12b] Joel A. Tropp. User-friendly tools for random matrices: an introduction. NIPS Tutorials, 2012.

[Tsy08] Alexandre B Tsybakov. *Introduction to nonparametric estimation*. Springer Science & Business Media, 2008.

[Vap13] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

[VBS19] Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1195–1204, 2019.

[VDVW96] AW Van Der Vaart and JA Wellner. Weak convergence and empirical processes: With applications to statistics. *Springer*, 58:59, 1996.

[Vil09] Cédric Villani. *Hypocoercivity*. American Mathematical Soc., 2009.

[VL07] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

[VLA87] Peter JM Van Laarhoven and Emile HL Aarts. Simulated annealing. In *Simulated annealing: Theory and applications*, pages 7–15. Springer, 1987.

[Wah90] G. Wahba. *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics, 1990.

[Wai19] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

[Wen95] Holger Wendland. Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree. *Advances in computational Mathematics*, 4(1):389–396, 1995.

[Wen04] Holger Wendland. *Scattered data approximation*, volume 17. Cambridge university press, 2004.

[Whi34] Hassler Whitney. Analytic extensions of differentiable functions defined in closed sets. *Transactions of the American Mathematical Society*, 36:63–89, 1934.

[Wib19] Andre Wibisono. Proximal langevin algorithm: Rapid convergence under isoperimetry. *arXiv preprint arXiv:1911.01469*, 2019.

[WK19] Geoffrey Wolfer and Aryeh Kontorovich. Estimating the mixing time of ergodic markov chains. *arXiv preprint arXiv:1902.01224*, 2019.

[Wri15] Stephen J Wright. Coordinate descent algorithms. *Mathematical Programming*, 151(1):3–34, 2015.

[WT11] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688, 2011.

[Xia10] L. Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 9:2543–2596, 2010.

[YP08] Yiming Ying and Massimiliano Pontil. Online gradient descent learning algorithms. *Foundations of Computational Mathematics*, 8(5):561–596, Oct 2008.

[YRC07] Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007.

[Yur95] Vadim Vladimirovich Yurinsky. *Gaussian and Related Approximations for Distributions of Sums*, pages 163–216. Springer Berlin Heidelberg, 1995.

[ZBH+16] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. Technical Report 1611.03530, arXiv, 2016.

[Zha04] Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, pages 56–85, 2004.

[Zho08] Ding-Xuan Zhou. Derivative reproducing properties for kernel methods in learning theory. *Journal of Computational and Applied Mathematics*, 220(1):456–463, 2008.