Exponential Convergence of Testing Error for Stochastic Gradient Methods

Loucas Pillaud-Vivien LOUCAS.PILLAUD-VIVIEN@INRIA.FR
Alessandro Rudi ALESSANDRO.RUDI@INRIA.FR
Francis Bach FRANCIS.BACH@INRIA.FR
INRIA - Département d'Informatique de l'ENS
Ecole Normale Supérieure, CNRS, INRIA, PSL Research University

Editors: Sebastien Bubeck, Vianney Perchet and Philippe Rigollet

Abstract

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. To achieve these rates of convergence we show sharper high-probability bounds with respect to the number of observations for stochastic gradient descent.
Keywords: SGD, positive-definite kernels, margin condition, binary classification.

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 (Bottou and Le Cun, 2005), and from the theoretical side, as it leads to optimal rates for estimation problems in a variety of situations (Nemirovski and Yudin, 1983; Polyak and Juditsky, 1992).

They follow a simple principle (Robbins and Monro, 1951): 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 (see, e.g., Polyak and Juditsky, 1992; Nesterov and Vial, 2008; Nemirovski et al., 2009; Shalev-Shwartz et al., 2007; Xiao, 2010; Bach and Moulines, 2011, 2013; Dieuleveut et al., 2017).

On the one hand, these rates are optimal for the estimation of the minimizer of a function given access to noisy gradients (Nemirovski and Yudin, 1983), 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 (Le Roux et al., 2012), SVRG (Johnson and Zhang, 2013), SAGA (Defazio et al., 2014), 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 (Solodov, 1998; Schmidt and Le Roux, 2013). 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 $O(1/n)$ (Dieuleveut and Bach, 2016; Dieuleveut et al., 2017), 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 (Zhang, 2004; Bartlett et al., 2006), 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 (see, e.g., Mammen and Tsybakov, 1999). 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 (Bartlett et al., 2006).

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 (Audibert and Tsybakov, 2007; Koltchinskii and Beznosova, 2005). Our main contribution is to show that stochastic gradient descent also achieves similar rates (see an empirical illustration in Figure 2 in the Appendix A). 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 (Bach and Moulines, 2013).

The paper is organized as follows: in Section 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 3. In Section 4, we present our version of stochastic gradient descent, with the use of tail averaging (Jain et al., 2016), and provide new deviation inequalities, which we apply in Section 5 to our learning problem, leading to exponential convergence rates for the testing errors. We conclude in Section 6 by providing several avenues for future work. Finally, synthetic experiments illustrating our results can be found in Section A 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 9 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 7 for stochastic gradient descent for least-square 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 Caponnetto and De Vito (2007); Dieuleveut and Bach (2016).

2. Problem Set-up

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

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 (Mammen and Tsybakov, 1999; Audibert and Tsybakov, 2007; Koltchinskii and Beznosova, 2005), 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 J and more precisely in Corollary 27. 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 $g \in \mathcal{H}$, $g(x) = \langle g, K_x \rangle_{\mathcal{H}}$ (Shawe-Taylor and Cristianini, 2004; Schölkopf and Smola, 2002). 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.*

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 (Caponnetto and De Vito, 2007) 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,* $\operatorname{sign}(\mathbb{E}(y|x)) g_\lambda(x) \geqslant \frac{\delta}{2}$.

In the assumption above, we could replace $\delta/2$ by any multiplicative constants in $(0, 1)$ timesδ (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 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.

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* $\mathcal{R}^* = \mathcal{R}(\mathbb{E}\,(y|x))$ which is the minimum achievable classification error (Devroye et al., 2013).

A well known approach to bound the testing errors by testing losses is *via transfer functions*. In particular we recall the following result (Devroye et al., 2013; Bartlett et al., 2006), let $g_*(x)$ be equal to $\mathbb{E}\,(y|x)$ a.e., then

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

with $\phi(u) = \sqrt{u}$ (or $\phi(u) = u^\beta$, with $\beta \in [1/2, 1]$, depending on some properties of ρ (Bartlett et al., 2006). 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 (see Caponnetto and De Vito, 2007).

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 C). This will lead to exponential rates when the distribution satisfies a margin condition **(A1)** as we prove in the next section and in Section 5. Note also that for the sake of completeness we recalled in Appendix D 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)**, let $\widehat{g} \in \mathcal{H}$ be a random function such that $\left\|\widehat{g} - g_\lambda\right\|_{\mathcal{H}} < \frac{\delta}{2R}$, with probability at least $1 - q$. Then*

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

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

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 $g_*(x) = \mathbb{E}\,(y|x)$ a.e. on the support of $\rho_{\mathcal{X}}$, is sufficient to have **(A4)** (proof in Appendix E.1).

Proposition 2 *Under* **(A1)**, *assume that there exists* $g_* \in \mathcal{H}$ *such that* $g_*(x) := \mathbb{E}(y|x)$ *on the support of* $\rho_{\mathcal{X}}$, *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 $\rho_{\mathcal{X}}$ 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 Adams and Fournier, 2003, and Appendix E.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 $\mathbb{E}(y|x)$ on $\mathcal{S}_+$ and the second matching $\mathbb{E}(y|x)$ on $\mathcal{S}_-$ and if the kernel satisfies **(A6)**, then **(A4)** is satisfied.

Proposition 3 *Under* **(A1)**, **(A5)**, **(A6)**, *if there exist two functions* $g_+^*, g_-^* \in W^{s,2}$ *such that* $g_+^*(x) = \mathbb{E}(y|x)$ *on* $\mathcal{S}_+$ *and* $g_-^*(x) = \mathbb{E}(y|x)$ *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 E.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 E Figure 4).*

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 (De Vito et al., 2014; Rudi et al., 2014) instead of **(A6)**.

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

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

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 $\tilde{g}_*$).

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 Fukumizu et al., 2004, and references therein), we have the optimality conditions for g_λ and $\tilde{g}_*$:

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

see Caponnetto and De Vito (2007) or Appendix F.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:

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

We are going to consider Polyak-Ruppert averaging (Polyak and Juditsky, 1992), 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 Jain et al. (2016). 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 I.

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 G. In the next subsection we reformulate the recursion in Eq. (1) as a least-squares recursion converging to g_λ.

4.1. Reformulation as noisy recursion

We can first reformulate the SGD recursion equation in Eq. (1) 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 F.2.):

Lemma 4 *The SGD recursion can be rewritten as follows:*

$$g_n - g_\lambda = \big[I - \gamma_n(K_{x_n} \otimes K_{x_n} + \lambda I)\big](g_{n-1} - g_\lambda) + \gamma_n \varepsilon_n, \qquad (2)$$

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 22 of Appendix G):

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

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

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.

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:

$$\eta_n = (I - \gamma H_n)\eta_{n-1} + \gamma_n \varepsilon_n, \tag{3}$$

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* $\mathbb{E}\|\varepsilon_n\|^2_{\mathcal{H}} \leqslant \operatorname{Tr} C$.

(H4) *For all* $n \geqslant 1$, $\mathbb{E}\Big[H_n C H^{-1} H_n\Big] \preccurlyeq \gamma_0^{-1} C$ *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., Défossez and Bach, 2015). Hence all of our bounds will depend on these two. See more details in Appendix G.

4.3. Non-averaged SGD

In this section, we prove results for the recursion defined by Eq. (3) in the case where for $\alpha \in [0, 1]$, $\gamma_n = \gamma/n^\alpha$. These results extend the ones of Bach and Moulines (2011) by providing deviation inequalities, but are limited to least-squares. For general loss functions and in the strongly-convex case, see also Kakade and Tewari (2009).

Theorem 5 (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. (3). 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* $\mathbb{P}(V_n \geqslant t) \leqslant 2\exp\left(-\dfrac{t^2}{8\gamma \operatorname{Tr} C/\lambda + \gamma c^{1/2} t} \cdot n^\alpha\right)$.

3. See Appendix Section G Lemma 19 for more details.

We can make the following observations:

- The proof technique (see Appendix G.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 5 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 20). 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 Bach and Moulines (2011), 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.

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 Dieuleveut et al. (2017) to deviation inequalities which are sharper than the ones from Bach and Moulines (2013).

Theorem 6 (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. (3): $\bar{\eta}_n = \frac{1}{n+1}\sum_{i=0}^{n} \eta_i$. Assume $\eta_0 = 0$. We have for $t > 0, n \geqslant 1$:*

$$\mathbb{P}\left(\left\|A^{1/2}\bar{\eta}_n\right\|_{\mathcal{H}} \geqslant t\right) \leqslant 2\exp\left[-\frac{(n+1)t^2}{E_t}\right], \tag{4}$$

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

$$E_t = 4\,\mathrm{Tr}(AH^{-2}C) + \frac{2c^{1/2}\|A^{1/2}\|_{op}}{3\lambda}\cdot t. \tag{5}$$

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 I.1 Theorem 24) 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 Jain et al., 2016; Shamir, 2011). 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 7 (Convergence of tail-averaged SGD) *Assume (**H1**), (**H2**), (**H3**), (**H4**), (**H5**) and consider the tail-average of the sequence defined in Eq. (3): $\bar{\eta}_n^{tail} = \frac{1}{\lfloor n/2 \rfloor}\sum_{i=\lfloor n/2 \rfloor}^{n} \eta_i$. We have for $t > 0, n \geqslant 1$:*

$$\left\| A^{1/2}\bar{\eta}_n^{tail} \right\|_{\mathcal{H}} \leqslant (1-\gamma\lambda)^{n/2}\|A^{1/2}\|_{op}\|\eta_0\|_{\mathcal{H}} + L_n \quad , \textit{with} \tag{6}$$

$$\mathbb{P}(L_n \geqslant t) \leqslant 4\exp\left(-(n+1)t^2/(4E_t)\right), \tag{7}$$

where L_n is defined in the proof (see Appendix G.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 G.2 and G.3 for the detailed proofs) relies on concentration inequality of Bernstein type. Indeed, we notice that (in the setting of Theorem 6) $\bar{\eta}_n$ 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 Dieuleveut et al., 2017) 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 Shamir, 2011), 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 Caponnetto and De Vito, 2007; Dieuleveut and Bach, 2016).
- The complete proof for the full average is written in Appendix I.1 and more precisely in Theorem 24. In this case the initial conditions are not forgotten exponentially fast though.

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. (2). 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 H).

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 8 *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. (2), as soon as n satisfies the inequality $\exp\left(-\frac{\gamma\lambda}{1-\alpha}\left((n+1)^{1-\alpha}-1\right)\right) \leqslant \delta/(5R\|g_0 - g_\lambda\|_{\mathcal{H}})$, 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 $C_R = 2^{\alpha+7}\gamma R^2 \operatorname{Tr}\Sigma\left(1 + \|\tilde{g}_* - g_\lambda\|_\infty^2\right)/\lambda + 8\gamma R^2\delta(1+2\|\tilde{g}_* - g_\lambda\|_\infty)/3$, and in particular

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

Note that Theorem 8 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 H.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 5 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 Kakade and Tewari (2009).
- 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: $\mathbb{E}\left[\mathcal{R}(g_n) - \mathcal{R}^*\right] \lesssim 2\exp\left(-\lambda\delta^2 n^\alpha/R^2\right)$.
- When the condition on n is not met, then we still have the usual bound obtained by taking directly the excess loss (Bartlett et al., 2006) but we lose exponential convergence.

5.2. Tail averaged SGD with constant step-size

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

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

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

4. The full averaging result corresponding to Theorem 9 is proved in Appendix I.2, Theorem 25.

Theorem 9 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 H.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 7 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: $\mathbb{E}\left[\mathcal{R}(g_n) - \mathcal{R}^*\right] \lesssim 2\exp\left(-\lambda^2\delta^2 n/R^4\right)$. Note that the λ^2 could be made much smaller with assumptions on the decrease of eigenvalues of Σ (it has been shown Caponnetto and De Vito, 2007, that if the decay happens at speed $1/n^\beta$: $\operatorname{Tr}\Sigma(\Sigma+\lambda I)^{-2} \leqslant \lambda^{-1}\operatorname{Tr}\Sigma(\Sigma+\lambda I)^{-1} \leqslant R^2/\lambda^{1+1/\beta}$).
- 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 I.2 and more precisely in Theorem 25.

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 J). 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 (Ciliberto et al., 2016; Osokin et al., 2017) or without (Taskar et al., 2005); (c) in our experiments, we use regularization, but we have experimented with unregularized recursions, which do exhibit fast convergence, but for which proofs are usually harder (Dieuleveut and Bach, 2016); finally, (d) in order to avoid the $O(n^2)$ complexity, extending the results of Rudi et al. (2017); Rudi and Rosasco (2017) would lead to a subquadratic complexity.

Acknowledgments

We acknowledge support from the European Research Council (grant SEQUOIA 724063). We would like to thank Raphaël Berthier for useful discussions.

References

Robert A. Adams and John J.F. Fournier. *Sobolev spaces*, volume 140. Academic Press, 2003.

Jean-Yves Audibert and Alexandre B. Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of statistics*, 35(2):608–633, 2007.

F. Bach and E. Moulines. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2011.

F. Bach and E. Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$. In *Advances in Neural Information Processing Systems (NIPS)*, 2013.

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.

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.

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

Carlo Ciliberto, Lorenzo Rosasco, and Alessandro Rudi. A consistent regularization approach for structured prediction. In *Advances in Neural Information Processing Systems*, 2016.

Ernesto De Vito, Lorenzo Rosasco, and Alessandro Toigo. Learning sets with separating kernels. *Applied and Computational Harmonic Analysis*, 37(2):185–217, 2014.

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.

A. Défossez and F. Bach. Constant step size least-mean-square: Bias-variance trade-offs and optimal sampling distributions. In *Proc. AISTATS*, 2015.

Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*, volume 31. Springer Science & Business Media, 2013.

Aymeric Dieuleveut and Francis Bach. Nonparametric stochastic approximation with large step-sizes. *The Annals of Statistics*, 44(4):1363–1399, 2016.

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.

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.

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.

R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, 2013.

Sham M. Kakade and Ambuj Tewari. On the generalization ability of online strongly convex programming algorithms. In *Advances in Neural Information Processing Systems*, 2009.

Vladimir Koltchinskii and Olexandra Beznosova. Exponential convergence rates in classification. In *International Conference on Computational Learning Theory*. Springer, 2005.

Nicolas Le 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*, 2012.

Enno Mammen and Alexandre Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27(6):1808–1829, 1999.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

A. S. Nemirovski and D. B. Yudin. *Problem complexity and method efficiency in optimization*. John Wiley, 1983.

Y. Nesterov and J. P. Vial. Confidence level solutions for stochastic programming. *Automatica*, 44 (6):1559–1568, 2008.

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.

Iosif Pinelis. Optimum bounds for the distributions of martingales in banach spaces. *The Annals of Probability*, pages 1679–1706, 1994.

B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

H. Robbins and S. Monro. A stochastic approximation method. *Ann. Math. Statistics*, 22:400–407, 1951.

Lorenzo Rosasco, Mikhail Belkin, and Ernesto De Vito. On learning with integral operators. *Journal of Machine Learning Research*, 11(Feb):905–934, 2010.

Alessandro Rudi and Lorenzo Rosasco. Generalization properties of learning with random features. In *Advances in Neural Information Processing Systems*, 2017.

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.

Alessandro Rudi, Luigi Carratino, and Lorenzo Rosasco. FALKON: An optimal large scale kernel method. In *Advances in Neural Information Processing Systems 30*. 2017.

Mark Schmidt and Nicolas Le Roux. Fast convergence of stochastic gradient descent under a strong growth condition. Technical Report 1308.6370, arXiv, 2013.

B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, 2002.

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.

Ohad Shamir. Making gradient descent optimal for strongly convex stochastic optimization. *CoRR*, abs/1109.5647, 2011.

J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

Mikhail V Solodov. Incremental gradient algorithms with stepsizes bounded away from zero. *Computational Optimization and Applications*, 11(1):23–35, 1998.

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.

L. Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 9:2543–2596, 2010.

Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007.

Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, pages 56–85, 2004.

Organization of the Appendix

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

B. *Probabilistic lemmas*
where concentration inequalities in Hilbert spaces used in section G are recalled.

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

D. *Proofs of Exponential rates for Kernel Ridge Regression*
where exponential rates for Kernel Ridge Regression are proven (Theorem 13).

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

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

G. *Proof of stochastic gradient descent results*
where high probability bounds for the general SGD recursion are shown (Theorems 5 and 6).

H. *Exponentially convergent SGD for classification error*
where exponential convergence of test error are shown (Theorems 8 and 9).

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

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

Appendix A. 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 $X \sim U\left([0, (1-\varepsilon)/2] \cup [(1+\varepsilon)/2, 1]\right)$ 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 (Adams and Fournier, 2003).
- **(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 $[(1+\varepsilon)/2,1]$.

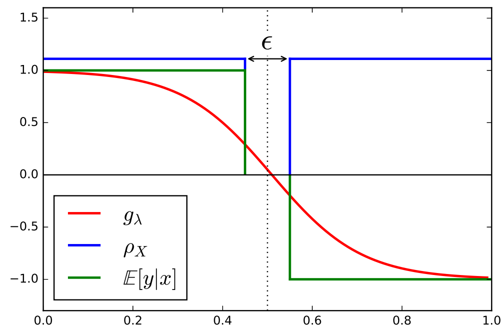

Figure 1: Representing the $\rho_{\mathcal{X}}$ density (uniform with ε-margin), the best estimator, i.e., $\mathbb{E}(x|y)$ and g_λ used for the simulations ($\lambda = 0.01$).

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 :

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

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 2 and 3 right).
- We recover the results of Dieuleveut et al. (2017) that show convergence at speed $1/n$ for the loss (Figure 2 left). We adapted the scale to compare with the error plot.
- For Figure 3 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 $n = 2000$.

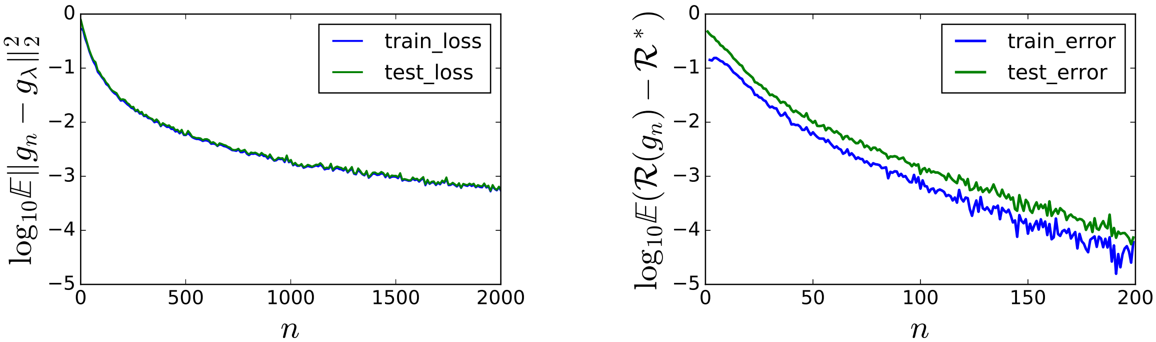

Figure 2: Showing linear convergence for the L^{01} errors in the case of margin of width ε. **Left** figure corresponds to the test and training loss in the averaged case whereas the **right** one corresponds to the error in the same setting. Note that the y-axis is the same while the x-axis is different 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 : $\varepsilon = 0.05$, $\gamma = 0.25$, $\lambda = 0.01$.

We can make the following observations:

First remark that between plots of losses and errors (Figure 2 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 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.

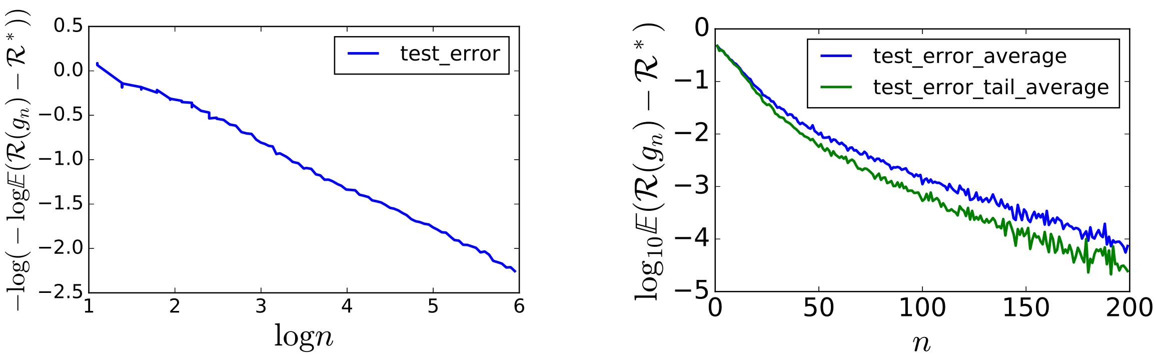

Figure 3: **Left** plot shows the error in the non-averaged case for $\gamma_n = \gamma/\sqrt{n}$ and **right** compares the test error between averaged and tail averaged case. We took the following parameters : $\varepsilon = 0.05$, $\gamma = 0.25$, $\lambda = 0.01$.

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).

Appendix B. Probabilistic lemmas

In this section we recall two fundamental results for concentration inequalities in Hilbert spaces shown in Pinelis (1994).

Proposition 10 *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 $t\geqslant 0$, $n\geqslant 1$,*

$$\mathbb{P}\left(\left\|\sum_{k=1}^n X_k\right\|\geqslant t\right)\leqslant 2\exp\left(\frac{t}{a_n}-\left(\frac{t}{a_n}+\frac{b_n^2}{a_n^2}\right)\ln\left(1+\frac{ta_n}{b_n}\right)\right). \tag{8}$$

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 Pinelis (1994) to the martingale $(f_{j\wedge n})$, we have the result. ■

Corollary 11 *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 $t\geqslant 0$, $n\geqslant 1$,*

$$\mathbb{P}\left(\left\|\sum_{k=1}^n X_k\right\|\geqslant t\right)\leqslant 2\exp\left(-\frac{t^2}{2\left(b_n^2+a_nt/3\right)}\right). \tag{9}$$

Proof We apply 10 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:

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

■

Appendix C. 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 $\|\widehat{g} - g_\lambda\|_\mathcal{H} < \delta/(2R)$. Note that for any $f \in \mathcal{H}$,

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

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

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

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

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

■

Appendix D. Exponential rates for Kernel Ridge Regression

D.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 (Caponnetto and De Vito, 2007) 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 Fukumizu et al., 2004, 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 Caponnetto and De Vito, 2007)

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

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 Caponnetto and De Vito (2007); here we rederive it in a subcase (see below for the proof).

Lemma 12 *Under assumption* **(A2)**, **(A3)** *for any* $\lambda > 0$, *note* $u_n = \| \frac{1}{n} \sum_{i=1}^{n} y_i K_{x_i} - \mathbb{E}\left[y K_x \right] \|_{\mathcal{H}}$ *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 12 and then applying Lemma 1, we obtain the following exponential bound for kernel ridge regression (see complete proof below):

Theorem 13 *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, $\mathbb{E}\left[\mathcal{R}(\widehat{g}_\lambda) - \mathcal{R}^* \right] \leq 4 \exp\left(-C_0 \lambda^4 \delta^2 n / R^8 \right)$, *with* $C_0^{-1} := 72(1 + \lambda R^2)^2$.

The result above is a refinement of Thm. 2.6 from Yao et al. (2007). We improved the dependency in n and removed the requirements that $g^* \in \mathcal{H}$ or $g^* = \Sigma^r w$ for a $w \in L^2(d\rho_{\mathcal{X}})$ 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 (see Koltchinskii and Beznosova, 2005). With respect to this latter work, our analysis uses the explicit characterization of the kernel ridge regression estimator in terms of linear operators on $\mathcal{H}$ (see Caponnetto and De Vito, 2007). 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 $\mathbb{E}\left(y|x\right)$).

Finally note that the exponent of λ could be reduced by using a refined analysis under additional regularity assumption of $\rho_{\mathcal{X}}$ and $\mathbb{E}\left(y|x\right)$ (as *source condition* and *intrinsic dimension* from Caponnetto and De Vito, 2007), but it is beyond the scope of this paper.

D.2. Proofs

Here we prove that Kernel Ridge Regression achieves exponential classification rates under assumptions **(A1)**, **(A4)**. In particular by Lemma 12 we bound $\left\|\widehat{g}_\lambda - g_\lambda\right\|_{\mathcal{H}}$ in high probability and then we use Lemma 1 that gives exponential classfication rates when $\left\|\widehat{g}_\lambda - g_\lambda\right\|_{\mathcal{H}}$ is small enough in high probability.

Proof of Lemma 12 Denote by $\widehat{\Sigma}_\lambda$ the operator $\widehat{\Sigma} + \lambda I$ and with Σ_λ the operator $\Sigma + \lambda I$. We have

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

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

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

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

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

■

Proof of Theorem 13 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 $u_n = \|\frac{1}{n}\sum_{i=1}^n (y_i K_{x_i} - \mathbb{E}\left[yK_x\right])\|_{\mathcal{H}}$ and $v_n = \|\Sigma - \widehat{\Sigma}\|_{\mathrm{op}}$. For u_n we can apply Pinelis inequality (Thm. 3.5 Pinelis, 1994), since $(x_i, y_i)_{i=1}^n$ are sampled independently according to the probability ρ and that $y_i K_{x_i} - \mathbb{E}\left[yK_x\right]$ is zero mean. Since

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

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 $\left\|\cdot\right\|_{HS}$ the Hilbert-Schmidt norm and recall that $\left\|\cdot\right\| \leq \left\|\cdot\right\|_{HS}$. To bound v_n we apply again the Pinelis inequality (see also Rosasco et al.,

2010) 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 $\rho_{\mathcal{X}}$ and that $\mathbb{E}\left[K_{x_i} \otimes K_{x_i}\right] = \Sigma$. In particular we apply it with $D = 1$ and $b_*^2 = \frac{4R^4}{n}$, so

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

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. ■

Appendix E. 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 E.2 we prove that specific settings naturally satisfy **(A4)**.

E.1. From $g_* \in \mathcal{H}$ to (A4)

Here we assume that there exists $g_* \in \mathcal{H}$ such that $g_*(x) = \mathbb{E}\left(y|x\right)$ a.e. on the support of $\rho_{\mathcal{X}}$. 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 $\left\|g_\lambda - g_*\right\|_{\mathcal{H}}$ 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 14 *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

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

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 15 *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

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

then $g_\lambda = \Sigma_\lambda^{-1}\mathbb{E}\left[yK_x\right] = \Sigma_\lambda^{-1}\Sigma g_*$. 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

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

Now divide the series in two parts

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

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 2 By Lemma 15 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 14. 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)**. ■

E.2. Examples

In this subsection we first introduce some notation and basic results about Sobolev spaces, then we prove Prop. 3 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

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

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

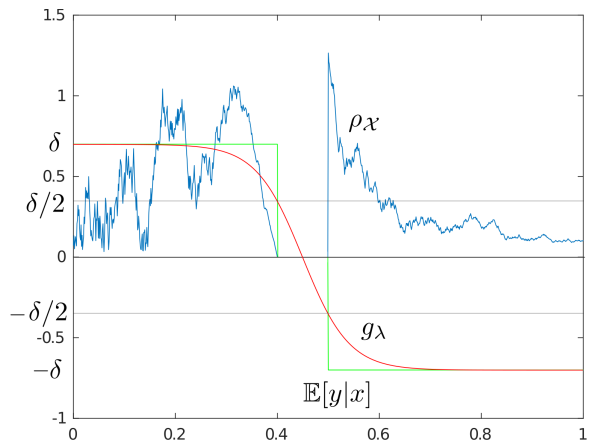

Figure 4: Pictorial representation of a model in 1D satisfying Example 1, ($p = 0.15$). Blue: $\rho_{\mathcal{X}}$, green: $\mathbb{E}\,(y|x)$, red: g_λ.

Proof 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 $W^{s,2}$. ■

Proof of Proposition 3 Since we are under **(A5)**, we can apply Prop. 16 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 (see Adams and Fournier, 2003), 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. 2. ■

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. 16. So we can apply Prop. 3, that given g_+^*, g_-^* guarantees that **(A4)** is satisfied. See an example in Figure 4. ■

Appendix F. Preliminaries for Stochastic Gradient Descent

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

F.1. Proof of the optimality condition on g_*

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

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

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 $\bar{\mathcal{H}}^{L_2}$ 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 $g_* - \tilde{g}_* = \mathcal{P}_{\mathcal{H}^\perp}(g_*) \in \mathcal{F}^\perp$.

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

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

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

F.2. Proof of Lemma 4: reformulation of SGD as noisy recursion

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

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

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

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

Appendix G. Proof of stochastic gradient descent results

Let us recall for the Appendix the SGD recursion defined in Eq. (3):

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

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,

$$\eta_n = M(n,n) \eta_{n-1} + \gamma_n \varepsilon_n \tag{10}$$

$$\eta_n = M(n,1) \eta_0 + \sum_{k=1}^n \gamma_k M(n, k+1) \varepsilon_k, \tag{11}$$

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

G.1. Non-averaged SGD - Proof of Theorem 5

In this section, we define the two 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:

$$\eta_n = \underbrace{M(n,1)\eta_0}_{\textbf{Biais term}} + \underbrace{W_n}_{\textbf{Noise term}} \quad , \tag{12}$$

- 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.

G.1.1. GENERAL RESULT FOR ALL (γ_n)

As $W_n = \sum_{k=1}^{n} \gamma_k Z_k$, we want to apply Corollary 11 of section B to $(\gamma_k Z_k)_{k\in\mathbb{N}}$ that is why we need the following lemma:

Lemma 17 *We have the following bounds:*

$$\sup_{k\leqslant n} \|\gamma_k Z_k\| \leqslant c^{1/2}\zeta_n, \ \textit{and} \tag{13}$$

$$\sum_{k=1}^{n} \mathbb{E}\left[\|\gamma_k Z_k\|^2|\mathcal{F}_{k-1}\right] \leqslant \operatorname{Tr} C\beta_n, \tag{14}$$

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

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

Second,

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

Hence,

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

■

Proposition 18 *We have the following inequality: for $t > 0, n \geqslant 1$,*

$$\begin{aligned} \|\eta_n\| &\leqslant \alpha_n \|\eta_0\| + V_n, \quad \textit{with} & (15) \\ \mathbb{P}\left(V_n \geqslant t\right) &\leqslant 2\exp\left(-\frac{t^2}{2(\operatorname{Tr} C\beta_n + c^{1/2}\zeta_n t/3)}\right). & (16) \end{aligned}$$

Proof We just need to apply Lemma 17 and Corollary 11 to the martingale W_n and $V_n = \|W_n\|$ for all n. ■

G.1.2. 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 Bach and Moulines (2011).

Lemma 19 *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) $\alpha_n \leqslant \exp\left(-\dfrac{\gamma\lambda}{2^{1-\alpha}(1-\alpha)}n^{1-\alpha}\right)$, (ii) $\beta_n \leqslant \dfrac{2^{\alpha+1}\gamma}{\lambda n^\alpha}$, (iii) $\zeta_n \leqslant \dfrac{\gamma}{n^\alpha}$.

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 $m \in [\![1, n]\!]$,

$$\beta_n = \sum_{k=1}^{n} \gamma_k^2 \prod_{i=k+1}^{n} (1-\gamma_i\lambda)^2 = \sum_{k=1}^{m} \gamma_k^2 \prod_{i=k+1}^{n} (1-\gamma_i\lambda)^2 + \sum_{k=m+1}^{n} \gamma_k^2 \prod_{i=k+1}^{n} (1-\gamma_i\lambda)^2$$

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

By taking $\gamma_n = \dfrac{\gamma}{n^\alpha}$ and $m = \lfloor \dfrac{n}{2} \rfloor$, 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 5. 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 20 (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. (3). 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*
$$\mathbb{P}\left(V_n \geqslant t\right) \leqslant 2\exp\left(-\frac{t^2}{4^{3/2}(\operatorname{Tr} C)\gamma^2/((1-2\gamma\lambda)n^{\gamma\lambda}) + 4tc^{1/2}\gamma/3}\cdot n^{\gamma\lambda}\right);$$
- *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*
$$\mathbb{P}\left(V_n \geqslant t\right) \leqslant 2\exp\left(-\frac{t^2}{2\gamma(\operatorname{Tr} C/\lambda + tc^{1/2}/3)}\right);$$
- *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*
$$\mathbb{P}\left(V_n \geqslant t\right) \leqslant 2\exp\left(-\frac{t^2}{\gamma(2^{\alpha+2}\operatorname{Tr} C/\lambda + 2c^{1/2}t/3)}\cdot n^\alpha\right).$$

Proof [Proof of Theorem 5] We apply Proposition 18, and the bound found on α_n, β_n and ζ_n in Lemma 19 to get the results. ■

G.2. Averaged SGD for the variance term ($\eta_0 = 0$) - Proof of Theorem 6

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 10 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 $(A^{1/2}\bar{Z}_k)_k$.

5. See Appendix Section G Lemma 19 for more details.

Lemma 21 *Assuming (**H1**), (**H2**), (**H3**),(**H4**), (**H5**), we have the following bounds for* $\bar{Z}_k = \sum_{i=k}^{n} M(i,k+1)\varepsilon_k$*:*

$$\sup_{k\leqslant n} \|A^{1/2}\bar{Z}_k\| \leqslant \frac{c^{1/2}\|A\|_{op}^{1/2}}{\gamma\lambda} \tag{17}$$

$$\sum_{k=1}^{n} \mathbb{E}\left[\|A^{1/2}\bar{Z}_k\|^2|\mathcal{F}_{k-1}\right] \leqslant n\frac{1}{\gamma^2}\frac{1}{1-\gamma/2\gamma_0}\operatorname{Tr}\left(AH^{-2}\cdot C\right). \tag{18}$$

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, $\|M(i,k+1)\|_{\text{op}} \leqslant (1-\gamma\lambda)^{i-k}$ and,

$$\|\bar{Z}_k\| \leqslant \|\varepsilon_k\|\sum_{i=k}^{n}\|M(i,k+1)\|_{\text{op}} \leqslant c^{1/2}\sum_{i=k}^{n}(1-\gamma\lambda)^{i-k} \leqslant \frac{c^{1/2}}{\gamma\lambda}$$

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}.

$$\mathbb{E}\left[\|A^{1/2}\bar{Z}_k\|^2|\mathcal{F}_{k-1}\right] \leqslant 2\sum_{i=k}^{n}\operatorname{Tr}\left(A\left(\gamma H\right)^{-1}\mathbb{E}\left[M(i,k+1)CM(i,k+1)^*\right]\right),$$

- **Step 2:** we then upperbound this sum to a telescopic one involving H^{-2} to finally show:

$$\mathbb{E}\left[\|A^{1/2}\bar{Z}_k\|^2|\mathcal{F}_{k-1}\right] \leqslant \frac{1}{\gamma^2}\frac{1}{1-\gamma/2\gamma_0}\operatorname{Tr}\left(AH^{-2}C\right).$$

Step 1: We write,

$$\begin{aligned}
\mathbb{E}\left[\|A^{1/2}\bar{Z}_k\|^2|\mathcal{F}_{k-1}\right] &= \mathbb{E}\left[\sum_{k\leqslant i,j\leqslant n}\left\langle A^{1/2}M(i,k+1)\varepsilon_k, A^{1/2}M(j,k+1)\varepsilon_k\right\rangle|\mathcal{F}_{k-1}\right] \\
&= \mathbb{E}\left[\sum_{k\leqslant i,j\leqslant n}\left\langle M(i,k+1)\varepsilon_k, AM(j,k+1)\varepsilon_k\right\rangle|\mathcal{F}_{k-1}\right] \\
&= \sum_{k\leqslant i,j\leqslant n}\mathbb{E}\left[\operatorname{Tr}\left(M(i,k+1)^*AM(j,k+1)\cdot\varepsilon_k\otimes\varepsilon_k\right)\right] \\
&= \sum_{k\leqslant i,j\leqslant n}\operatorname{Tr}\left(\mathbb{E}\left[M(i,k+1)^*AM(j,k+1)\right]\cdot\mathbb{E}\left[\varepsilon_k\otimes\varepsilon_k\right]\right).
\end{aligned}$$

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),

$$\sum_{k\leqslant i,j\leqslant n} \mathrm{Tr}\left(\mathbb{E}\left[{M_{k+1}^i}^* A M_{k+1}^j\right]\cdot C\right) = \sum_{i=k}^{n} \mathrm{Tr}\left(\mathbb{E}\left[{M_{k+1}^i}^* A M_{k+1}^i\right]\cdot C\right) + 2\sum_{k\leqslant i<j\leqslant n} \mathrm{Tr}\left(\mathbb{E}\left[{M_{k+1}^i}^* A M_{k+1}^j\right]\cdot C\right).$$

Moreover,

$$\begin{aligned}
&\sum_{k\leqslant i<j\leqslant n} \mathrm{Tr}\left(\mathbb{E}\left[{M_{k+1}^i}^* A M_{k+1}^j\right]\cdot C\right)\\
=\ &\sum_{k\leqslant i<j\leqslant n} \mathrm{Tr}\left(\mathbb{E}\left[{M_{k+1}^i}^* A\left(I-\gamma H\right)^{j-i} M_{k+1}^i\right]\cdot C\right)\\
=\ &\sum_{i=k}^{n} \mathrm{Tr}\left(\mathbb{E}\left[{M_{k+1}^i}^* A \sum_{j=i+1}^{n}\left(I-\gamma H\right)^{j-i} M_{k+1}^i\right]\cdot C\right)\\
=\ &\sum_{i=k}^{n} \mathrm{Tr}\left(\mathbb{E}\left[{M_{k+1}^i}^* A\left[\left(I-\gamma H\right)\left(I-\left(I-\gamma H\right)^{n-i}\right)\left(\gamma H\right)^{-1}\right] M_{k+1}^i\right]\cdot C\right)\\
\leqslant\ &\sum_{i=k}^{n} \mathrm{Tr}\left(\mathbb{E}\left[{M_{k+1}^i}^* A\left[\left(\gamma H\right)^{-1}-I\right] M_{k+1}^i\right]\cdot C\right)\\
\leqslant\ &\sum_{i=k}^{n} \mathrm{Tr}\left(\mathbb{E}\left[{M_{k+1}^i}^* A\left(\gamma H\right)^{-1} M_{k+1}^i\right]\cdot C\right) - \sum_{i=k}^{n} \mathrm{Tr}\left(\mathbb{E}\left[{M_{k+1}^i}^* A M_{k+1}^i\right]\cdot C\right).
\end{aligned}$$

Hence,

$$\begin{aligned}
&\sum_{k\leqslant i,j\leqslant n} \mathrm{Tr}\left(\mathbb{E}\left[{M_{k+1}^i}^* A M_{k+1}^j\right]\cdot C\right)\\
&\quad= \sum_{i=k}^{n} \mathrm{Tr}\left(\mathbb{E}\left[{M_{k+1}^i}^* A M_{k+1}^i\right]\cdot C\right) + 2\sum_{k\leqslant i<j\leqslant n} \mathrm{Tr}\left(\mathbb{E}\left[{M_{k+1}^i}^* A M_{k+1}^j\right]\cdot C\right)\\
&\quad\leqslant 2\sum_{i=k}^{n} \mathrm{Tr}\left(\mathbb{E}\left[{M_{k+1}^i}^* A\left(\gamma H\right)^{-1} M_{k+1}^i\right]\cdot C\right) - \sum_{i=k}^{n} \mathrm{Tr}\left(\mathbb{E}\left[{M_{k+1}^i}^* A M_{k+1}^i\right]\cdot C\right)\\
&\quad\leqslant 2\sum_{i=k}^{n} \mathrm{Tr}\left(\mathbb{E}\left[{M_{k+1}^i}^* A\left(\gamma H\right)^{-1} M_{k+1}^i\right]\cdot C\right)\\
&\quad\leqslant 2\sum_{i=k}^{n} \mathrm{Tr}\left(A\left(\gamma H\right)^{-1}\mathbb{E}\left[M_{k+1}^i C {M_{k+1}^i}^*\right]\right)
\end{aligned}$$

This concludes step 1.

Step 2: Let us now try to bound $\sum_{i=k}^{n} \operatorname{Tr}\left(A\left(\gamma H\right)^{-1}\mathbb{E}\left[M_{k+1}^{i} C {M_{k+1}^{i}}^{*}\right]\right)$. 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,

$$\begin{aligned}
\operatorname{Tr}\left(A\left(\gamma H\right)^{-1}\mathbb{E}\left[M_{k+1}^{i+1} C\left(\gamma H\right)^{-1} {M_{k+1}^{i+1}}^{*}\right]\right) &= \operatorname{Tr}\left(A\left(\gamma H\right)^{-1}\mathbb{E}\left[M_{k+1}^{i} C\left(\gamma H\right)^{-1} {M_{k+1}^{i}}^{*}\right]\right) \\
&\quad - 2\operatorname{Tr}\left(A\left(\gamma H\right)^{-1}\mathbb{E}\left[M_{k+1}^{i} C {M_{k+1}^{i}}^{*}\right]\right) \\
&\quad + \gamma\operatorname{Tr}\left(A\left(\gamma H\right)^{-1}\mathbb{E}\left[M_{k+1}^{i}\mathbb{E}\left[H_{i+1}CH^{-1}H_{i+1}\right] {M_{k+1}^{i}}^{*}\right]\right),
\end{aligned}$$

And as $\mathbb{E}\Big[H_k C H^{-1} H_k\Big] \preccurlyeq \gamma_0^{-1} C$ we have,

$$\gamma\operatorname{Tr}\left(A\left(\gamma H\right)^{-1}\mathbb{E}\left[M_{k+1}^{i}\mathbb{E}\left[H_{i+1}CH^{-1}H_{i+1}\right] {M_{k+1}^{i}}^{*}\right]\right) \leqslant \gamma/\gamma_0 \operatorname{Tr}\left(A\left(\gamma H\right)^{-1}\mathbb{E}\left[M_{k+1}^{i} C {M_{k+1}^{i}}^{*}\right]\right),$$

thus,

$$\begin{aligned}
\operatorname{Tr}\left(A\left(\gamma H\right)^{-1}\mathbb{E}\left[M_{k+1}^{i+1} C\left(\gamma H\right)^{-1} {M_{k+1}^{i+1}}^{*}\right]\right) &\leqslant \operatorname{Tr}\left(A\left(\gamma H\right)^{-1}\mathbb{E}\left[M_{k+1}^{i} C\left(\gamma H\right)^{-1} {M_{k+1}^{i}}^{*}\right]\right) \\
&- \ 2\operatorname{Tr}\left(A\left(\gamma H\right)^{-1}\mathbb{E}\left[M_{k+1}^{i} C {M_{k+1}^{i}}^{*}\right]\right) \\
&+ \ \gamma/\gamma_0\operatorname{Tr}\left(A\left(\gamma H\right)^{-1}\mathbb{E}\left[M_{k+1}^{i} C {M_{k+1}^{i}}^{*}\right]\right)
\end{aligned}$$

$$\begin{aligned}
&\operatorname{Tr}\left(A\left(\gamma H\right)^{-1}\mathbb{E}\left[M_{k+1}^{i} C {M_{k+1}^{i}}^{*}\right]\right) \\
&\leqslant \frac{1}{2-\frac{\gamma}{\gamma_0}}\left(\operatorname{Tr}\left(A\left(\gamma H\right)^{-1}\mathbb{E}\left[M_{k+1}^{i} C\left(\gamma H\right)^{-1} {M_{k+1}^{i}}^{*}\right]\right) - \operatorname{Tr}\left(A\left(\gamma H\right)^{-1}\mathbb{E}\left[M_{k+1}^{i+1} C\left(\gamma H\right)^{-1} {M_{k+1}^{i+1}}^{*}\right]\right)\right).
\end{aligned}$$

If we take all the calculations from the beginning,

$$\begin{aligned}
\mathbb{E}\left[\|A^{1/2}\bar{Z}_k\|^2|\mathcal{F}_{k-1}\right] &\leqslant \sum_{k\leqslant i,j\leqslant n}\operatorname{Tr}\left(\mathbb{E}\left[{M_{k+1}^{i}}^{*} A M_{k+1}^{j}\right]\cdot C\right) \\
&\leqslant 2\sum_{i=k}^{n}\operatorname{Tr}\left(A\left(\gamma H\right)^{-1}\mathbb{E}\left[M_{k+1}^{i} C {M_{k+1}^{i}}^{*}\right]\right) \\
&\leqslant \frac{2}{2-\gamma/\gamma_0}\sum_{i=k}^{n}\operatorname{Tr}\left(A\left(\gamma H\right)^{-1}\mathbb{E}\left[M_{k+1}^{i} C\left(\gamma H\right)^{-1} {M_{k+1}^{i}}^{*}\right]\right) \\
&\qquad - \operatorname{Tr}\left(A\left(\gamma H\right)^{-1}\mathbb{E}\left[M_{k+1}^{i+1} C\left(\gamma H\right)^{-1} {M_{k+1}^{i+1}}^{*}\right]\right) \\
&\leqslant \frac{2}{2-\gamma/\gamma_0}\operatorname{Tr}\left(A\left(\gamma H\right)^{-1}\mathbb{E}\left[M_{k+1}^{k} C\left(\gamma H\right)^{-1} {M_{k+1}^{k}}^{*}\right]\right) \\
&\leqslant \frac{1}{\gamma^2}\frac{1}{1-\gamma/2\gamma_0}\operatorname{Tr}\left(AH^{-2}\cdot C\right),
\end{aligned}$$

which concludes the proof if we sum this inequality from 1 to n.

■

We can now prove Theorem 6:

Proof [Proof of Theorem 6] We apply Corollary 11 to the sequence $\left(\frac{\gamma}{n+1}A^{1/2}Z_k\right)_{k\leqslant n}$ thanks to Lemma 21. We have:

$$\begin{aligned}
\sup_{k\leqslant n}\|\frac{\gamma}{n+1}A^{1/2}Z_k\| &\leqslant \frac{c^{1/2}\|A^{1/2}\|}{(n+1)\lambda} \\
\sum_{k=1}^{n}\mathbb{E}\left[\|\frac{\gamma}{n+1}A^{1/2}Z_k\|^2|\mathcal{F}_{k-1}\right] &\leqslant \frac{1}{n+1}\frac{1}{1-\gamma/2\gamma_0}\operatorname{Tr}\left(AH^{-2}\cdot C\right),
\end{aligned}$$

so that,

$$\mathbb{P}\left(\left\|A^{1/2}\bar{\eta}_n\right\|\geqslant t\right)=\mathbb{P}\left(\left\|\sum_{k=1}^{n}\frac{\gamma}{n+1}A^{1/2}Z_k\right\|\geqslant t\right)\leqslant 2\exp\left(-\frac{t^2}{2\left(\frac{\operatorname{Tr}(AH^{-2}\cdot C)}{(n+1)(1-\gamma/2\gamma_0)}+\frac{c^{1/2}\|A^{1/2}\|t}{3\lambda(n+1)}\right)}\right)$$

$$\mathbb{P}\left(\left\|A^{1/2}\bar{\eta}_n\right\|\geqslant t\right)\leqslant 2\exp\left(-\frac{(n+1)t^2}{\frac{2\operatorname{Tr}(AH^{-2}\cdot C)}{(1-\gamma/2\gamma_0)}+\frac{2\|A^{1/2}\|c^{1/2}t}{3\lambda}}\right).$$

■

G.3. Tail-averaged SGD - Proof of Corollary 7

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 6 for n and $n/2$, as $\bar{\eta}_n^{\text{tail}}=2\bar{\eta}_n-\bar{\eta}_{n/2}$.

Proof [Proof of Corollary 7]

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,

$$\begin{aligned}
\mathbb{P}(L_n \geqslant t) &\leqslant 2\left[\exp\left(-\frac{(n+1)t^2}{4E_t}\right) + \exp\left(-\frac{(n+1)t^2}{2E_t}\right)\right] \\
&\leqslant 4\exp\left(-\frac{(n+1)t^2}{4E_t}\right).
\end{aligned}$$

■

Appendix H. 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 (3):

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

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 22 (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:*

$$\begin{array}{rcl} \|\varepsilon_n\| & \leqslant & R(1 + 2\|\tilde{g}_* - g_\lambda\|_{L_\infty}) = c^{1/2} \\ \mathbb{E}\varepsilon_n \otimes \varepsilon_n & \preccurlyeq & 2\left(1 + \|\tilde{g}_* - g_\lambda\|_\infty^2\right)\Sigma = C \\ \mathbb{E}\|\varepsilon_n\|^2 & \leqslant & 2\left(1 + \|\tilde{g}_* - g_\lambda\|_\infty^2\right)\operatorname{Tr}\Sigma = \operatorname{Tr} C. \end{array}$$

- ***(H4)*** *We have:*

$$\mathbb{E}\Big[H_k C H^{-1} H_k\Big] \quad \preccurlyeq \quad \left(R^2 + 2\lambda\right) C = \gamma_0^{-1} C \; .$$

Proof **(H1)**, **(H2)** are obviously satisfied.
Let us show **(H3)**:

$$\begin{array}{rcl} \|\varepsilon_n\| & = & \|\xi_n K_{x_n} + (\tilde{g}_*(x_n) - g_\lambda(x_n))K_{x_n} - \mathbb{E}\left[(\tilde{g}_*(x_n) - g_\lambda(x_n))K_{x_n}\right]\| \\ & \leqslant & (|\xi_n| + |\tilde{g}_*(x_n) - g_\lambda(x_n)|)\|K_{x_n}\| + \mathbb{E}\left[|\tilde{g}_*(x_n) - g_\lambda(x_n)|\|K_{x_n}\|\right] \\ & \leqslant & (1 + \|\tilde{g}_* - g_\lambda\|_\infty)R + \|\tilde{g}_* - g_\lambda\|_\infty R \\ & = & R(1 + 2\|\tilde{g}_* - g_\lambda\|_\infty) \end{array}$$

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,

$$\begin{aligned} &\mathbb{E}[((\tilde{g}_*(x_n) - g_\lambda(x_n))K_{x_n} - \mathbb{E}\left[(\tilde{g}_*(x_n) - g_\lambda(x_n)K_{x_n}\right]) \\ &\qquad\qquad \otimes ((\tilde{g}_*(x_n) - g_\lambda(x_n))K_{x_n} - \mathbb{E}\left[(\tilde{g}_*(x_n) - g_\lambda(x_n))K_{x_n}\right])] \\ &= \mathbb{E}\left[(\tilde{g}_*(x_n) - g_\lambda(x_n))^2(x_n)K_{x_n} \otimes K_{x_n}\right] - \mathbb{E}\left[(\tilde{g}_*(x_n) - g_\lambda(x_n))K_{x_n}\right] \\ &\qquad\qquad \otimes \mathbb{E}\left[(\tilde{g}_*(x_n) - g_\lambda(x_n))K_{x_n}\right] \\ &\preccurlyeq \mathbb{E}\left[(\tilde{g}_*(x_n) - g_\lambda(x_n))^2(x_n)K_{x_n} \otimes K_{x_n}\right] \\ &\preccurlyeq \|\tilde{g}_* - g_\lambda\|_\infty^2 \Sigma. \end{aligned}$$

So that,

$$\mathbb{E}\varepsilon_n \otimes \varepsilon_n \preccurlyeq 2\left(1 + \|\tilde{g}_* - g_\lambda\|_\infty^2\right)\Sigma$$

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 $x \in \mathcal{H}$, $\langle x, (a+b)\otimes(a+b)x\rangle = (\langle a+b, x\rangle)^2 = (\langle a, x\rangle + \langle b, x\rangle)^2 \leqslant 2\langle a, x\rangle^2 + 2\langle b, x\rangle^2 = 2\langle x, (a\otimes a)x\rangle + 2\langle x, (b\otimes b)x\rangle$.

Finally $\mathbb{E}\varepsilon_n \otimes \varepsilon_n \preccurlyeq 2\left(1 + \|\tilde{g}_* - g_\lambda\|_\infty^2\right)\Sigma$, we have $\operatorname{Tr}\mathbb{E}\varepsilon_n \otimes \varepsilon_n \leqslant 2\left(1 + \|\tilde{g}_* - g_\lambda\|_\infty^2\right)\operatorname{Tr}\Sigma$, thus

$$\operatorname{Tr}\mathbb{E}\varepsilon_n \otimes \varepsilon_n = \mathbb{E}\operatorname{Tr}\varepsilon_n \otimes \varepsilon_n = \mathbb{E}\|\varepsilon_n\|^2 \leqslant 2\left(1 + \|\tilde{g}_* - g_\lambda\|_\infty^2\right)\operatorname{Tr}\Sigma.$$

To conclude the proof of this lemma, let us show **(H4)**. We have:

$$\mathbb{E}\Big[(K_{x_k} \otimes K_{x_k} + \lambda I)\Sigma(\Sigma + \lambda I)^{-1}(K_{x_k} \otimes K_{x_k} + \lambda I)\Big] = \mathbb{E}\Big[K_{x_k} \otimes K_{x_k}\Sigma(\Sigma + \lambda I)^{-1}K_{x_k} \otimes K_{x_k}\Big] + \lambda\Sigma\Sigma(\Sigma + \lambda I)^{-1} + \lambda\Sigma$$

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 $\mathbb{E}\Big[(K_{x_k} \otimes K_{x_k} + \lambda I)\Sigma(\Sigma + \lambda I)^{-1}(K_{x_k} \otimes K_{x_k} + \lambda I)\Big] \preccurlyeq R^2\Sigma + \lambda\Sigma + \lambda\Sigma = (R^2 + 2\lambda)\Sigma$.

■

H.1. SGD with decreasing step-size: proof of Theorem 8

Proof [Proof of Theorem 8]

Let us apply Theorem 5 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 22). Let δ correspond to the one of **(A4)**. We have for $t = \delta/(4R), n \geqslant 1$:

$$\|g_n - g_\lambda\|_{\mathcal{H}} \leqslant \exp\left(-\frac{\gamma\lambda}{1-\alpha}\left((n+1)^{1-\alpha} - 1\right)\right)\|g_0 - g_\lambda\|_{\mathcal{H}} + \|W_n\|_{\mathcal{H}}, \text{ a.s, with}$$

$$\mathbb{P}\left(\|W_n\|_{\mathcal{H}} \geqslant \delta/(4R)\right) \leqslant 2\exp\left(-\frac{\delta^2}{C_R}n^\alpha\right), \quad C_R = \gamma(2^{\alpha+6}R^2\operatorname{Tr}C/\lambda + 8Rc^{1/2}\delta/3).$$

Then if n is such that $\exp\left(-\frac{\gamma\lambda}{1-\alpha}\left((n+1)^{1-\alpha} - 1\right)\right) \leqslant \dfrac{\delta}{5R\|g_0 - g_\lambda\|_{\mathcal{H}}}$,

$$\begin{aligned}\|g_n - g_\lambda\|_{\mathcal{H}} &\leqslant \frac{\delta}{5R} + \frac{\delta}{4R}, \text{ with probability } 1 - 2\exp\left(-\frac{\delta^2}{C_R}n^\alpha\right),\\ \|g_n - g_\lambda\|_{\mathcal{H}} &< \frac{\delta}{2R}, \text{ with probability } 1 - 2\exp\left(-\frac{\delta^2}{C_R}n^\alpha\right).\end{aligned}$$

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

$$\begin{aligned}&C_R = \gamma(2^{\alpha+6}R^2\operatorname{Tr}C/\lambda + 8Rc^{1/2}\delta/3)\\ &C_R == \gamma\left(\frac{2^{\alpha+7}R^2\operatorname{Tr}\Sigma\left(1 + \|\tilde{g}_* - g_\lambda\|_\infty^2\right)}{\lambda} + \frac{8R^2\delta(1 + 2\|\tilde{g}_* - g_\lambda\|_\infty)}{3}\right).\end{aligned}$$

■

H.2. Tail averaged SGD with constant step-size: proof of Theorem 9

Proof [Proof of Theorem 9]

Let us apply Corollary 7 to $g_n - g_\lambda$. We assume **(A2)**, **(A3)** and $A = I$, such that **(H1)**, **(H2)**, **(H3)**,**(H4)**, **(H5)** are verified (Lemma 22). 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 $(1-\gamma\lambda)^{n/2} \leqslant \dfrac{\delta}{5R\|g_0 - g_\lambda\|_{\mathcal{H}}}$,

$$\begin{aligned} \left\|\bar{g}_n^{\text{tail}} - g_\lambda\right\|_{\mathcal{H}} &\leqslant \frac{\delta}{5R} + \frac{\delta}{4R}, \text{ with probability } 1 - 4\exp\left(-(n+1)\delta^2/(64R^2 E_{\delta/(4R)})\right), \\ \left\|\bar{g}_n^{\text{tail}} - g_\lambda\right\|_{\mathcal{H}} &< \frac{\delta}{2R}, \text{ with probability } 1 - 4\exp\left(-(n+1)\delta^2/(64R^2 E_{\delta/(4R)})\right). \end{aligned}$$

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

$$\begin{aligned} K_R = 64R^2 E_{\delta/(4R)} &= 64R^2\left(4\operatorname{Tr}(H^{-2}C) + \frac{2c^{1/2}}{3\lambda}\cdot\frac{\delta}{4R}\right) \\ &= 512R^2\left(1 + \|\tilde{g}_* - g_\lambda\|_\infty^2\right)\operatorname{Tr}((\Sigma+\lambda I)^{-2}\Sigma) + \frac{32\delta R^2(1+2\|\tilde{g}_* - g_\lambda\|_\infty)}{3\lambda}. \end{aligned}$$

■

Appendix I. Extension of Corollary 7 and Theorem 9 for the full averaged case.

I.1. Extension of Corollary 7 for the full averaged case.

Let us recall the SGD abstract recursion defined in Eq. (3) 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 6 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

$$w_n = (I - \gamma \tilde{H}_n) w_{n-1} + \gamma \tilde{\varepsilon}_n. \tag{19}$$

Thus, w_n follows the same recursion as Eq.(3) with $(\tilde{H}_n, \tilde{\varepsilon}_n)$. We thus have the following corollary:

Corollary 23 *Assume that the sequence (w_n) defined in Eq. (19) 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 6 to the sequence (w_n) defined in Eq. (19). ∎

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 24 *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. (3). We have for $t > 0, n \geqslant 1$:*

$$\left\|A^{1/2}\bar{\eta}_n\right\|_{\mathcal{H}} \leqslant \frac{\left\|A^{1/2}\right\| \|\eta_0\|_{\mathcal{H}}}{(n+1)\gamma\lambda} + L_n \quad \text{, with} \tag{20}$$

$$\mathbb{P}(L_n \geqslant t) \leqslant 4\exp\left(-\frac{(n+1)t^2}{\max(E_t, \tilde{E}_t)}\right). \tag{21}$$

Proof [Proof of Theorem 24] 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,

$$\mathbb{P}(L_n \geqslant t) \quad \leqslant \quad 4\exp\left(-\frac{(n+1)t^2}{\max(E_t, \tilde{E}_t)}\right).$$

■

I.2. Extension of Theorem 9 for the full averaged case.

Same situation here, we want to apply full averaged SGD instead of the tail-averaged technique.

Theorem 25 *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. (2), as soon as: $n \geqslant \dfrac{5R\|g_0 - g_\lambda\|_{\mathcal{H}}}{\lambda\gamma\delta}$, then*

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

and in particular

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

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 [Proof of Theorem 25]

We want to apply Theorem 24 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 22. 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 26 (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:*

$$\mathbb{E}\left[\tilde{H}_k\tilde{C}H^{-1}\tilde{H}_k\right] \quad \preccurlyeq \quad \left(R^2+\lambda\right)\tilde{C} = \tilde{\gamma}_0^{-1}\tilde{C}\,.$$

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)** :

$$\begin{aligned}\mathbb{E}\Big[\tilde{H}_k \tilde{C} H^{-1} \tilde{H}_k\Big] = H\tilde{C} = R^2 \|g_0 - g_\lambda\|_{\mathcal{H}} (\Sigma^2 + \lambda \Sigma) \quad &\preccurlyeq \quad R^2 \|g_0 - g_\lambda\|_{\mathcal{H}} (\|\Sigma\|_{\text{op}} + \lambda)\Sigma \\ &\preccurlyeq \quad \left(R^2 + \lambda\right) \tilde{C} = \tilde{\gamma_0}^{-1} \tilde{C}.\end{aligned}$$

■

Let us apply now Theorem 24 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, ε_n) and $(\tilde{H}_n, \tilde{\varepsilon}_n)$) (Lemma 22,26). Let δ correspond to the one of Assumption 4. We have for $t = \delta/(4R), n \geqslant 1$:

$$\begin{aligned}\|\bar{g}_n - g_\lambda\|_{\mathcal{H}} \quad &\leqslant \quad \frac{\|g_0 - g_\lambda\|_{\mathcal{H}}}{(n+1)\gamma\lambda} + L_n \quad \text{, with} \\ \mathbb{P}(L_n \geqslant t) \quad &\leqslant \quad 4\exp\left(-\frac{(n+1)t^2}{\max(E_t, \tilde{E}_t)}\right).\end{aligned}$$

Then as soon as $\dfrac{1}{(n+1)\lambda\gamma} \leqslant \dfrac{\delta}{5R\|g_0 - g_\lambda\|_{\mathcal{H}}}$,

$$\begin{aligned}\|\bar{g}_n - g_\lambda\|_{\mathcal{H}} \quad &\leqslant \quad \frac{\delta}{5R} + \frac{\delta}{4R}, \quad \text{with probability } 1 - 4\exp\left(-\frac{(n+1)\delta^2}{16R^2 \max(E_{\delta/4R}, \tilde{E}_{\delta/4R})}\right), \\ \|\bar{g}_n - g_\lambda\|_{\mathcal{H}} \quad &< \quad \frac{\delta}{2R}, \quad \text{with probability } 1 - 4\exp\left(-\frac{(n+1)\delta^2}{16R^2 \max(E_{\delta/4R}, \tilde{E}_{\delta/4R})}\right).\end{aligned}$$

Now assume **(A1)**, **(A4)**, we now only have to apply Lemma 1 to the estimator $\bar{g}_n$ with the probability $q = 4\exp\left(-\frac{(n+1)\delta^2}{16R^2\max(E_{\delta/4R}, \tilde{E}_{\delta/4R})}\right)$. And,

$$K_R^{-1} = 16R^2\max(E_{\delta/4R}, \tilde{E}_{\delta/4R})$$
$$= \max\begin{cases} 128R^2\left(1+\|\tilde{g}_* - g_\lambda\|_\infty^2\right)\operatorname{Tr}((\Sigma+\lambda I)^{-2}\Sigma) + \frac{8R^2(1+2\|\tilde{g}_* - g_\lambda\|_\infty)}{3\lambda} \\ 64R^4\|g_0 - g_\lambda\|_{\mathcal{H}}\operatorname{Tr}((\Sigma+\lambda I)^{-2}\Sigma) + \frac{16R^4\|g_0 - g_\lambda\|_{\mathcal{H}}}{3\lambda}. \end{cases}$$

■

Appendix J. 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 9 with the new assumptions. Note that it could also be shown for the full average sequence $\bar{g}_n$.

Corollary 27 (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. (2), and* $\bar{g}_n^{tail} = \frac{1}{\lfloor n/2 \rfloor}\sum_{i=\lfloor n/2\rfloor}^n g_i$, *as soon as* $n \geqslant \frac{2}{\gamma\lambda}\ln\left(\frac{5R\|g_0 - g_\lambda\|_{\mathcal{H}}}{\delta}\right)$, *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 Audibert and Tsybakov (2007).

Let $\delta, \lambda > 0$, such that $\|g_* - g_\lambda\|_\infty \leqslant \delta$. Remark that $\forall j \in \mathbb{N}$,
$\mathbb{P}\left(\operatorname{sign}(g_*(X))g_\lambda(X) \leqslant 2^j\delta\right) \leqslant \mathbb{P}\left(|g_\lambda(X)| \leqslant 2^j\delta\right) \leqslant \mathbb{P}\left(|g_*(X)| \leqslant 2^{j+1}\delta\right) \leqslant 2^{\alpha j}\delta^\alpha$.

Note $A_0 = \{x \in \mathcal{X}|\ \operatorname{sign}(g_*)g_\lambda \leqslant \delta\}$ and for $j \geqslant 1$, $A_j = \{x \in \mathcal{X}|\ 2^{j-1}\delta < \operatorname{sign}(g_*)g_\lambda \leqslant 2^j\delta\}$. Then,

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 $\gamma = \frac{r-1/2}{2r+1/\beta}$ (Caponnetto and De Vito, 2007).

$$\begin{aligned}
\mathbb{E}\left[R(\bar{g}_n^{\text{tail}}) - R^*\right] &= \sum_{j\in\mathbb{N}} \mathbb{E}\left[\left(R(\bar{g}_n^{\text{tail}}) - R^*\right)\mathbf{1}_{A_j}\right] \\
&= \mathbb{E}\left[\left(R(\bar{g}_n^{\text{tail}}) - R^*\right)\mathbf{1}_{\text{sign}(g_*)g_\lambda \leqslant \delta}\right] + \sum_{j\geqslant 1} \mathbb{E}\left[\left(R(\bar{g}_n^{\text{tail}}) - R^*\right)\mathbf{1}_{A_j}\right] \\
&\leqslant \mathbb{P}\left(\text{sign}(g_*(X))g_\lambda(X) \leqslant \delta\right) + \sum_{j\geqslant 1} \mathbb{E}\left[\left(R(\bar{g}_n^{\text{tail}}) - R^*\right)\mathbf{1}_{2^{j-1}\delta < \text{sign}(g_*(X))g_\lambda(X) \leqslant 2^j\delta}\right] \\
&\leqslant \delta^\alpha + \sum_{j\geqslant 1} \mathbb{E}_X\left[\mathbb{E}_{x_1,\dots,x_n}\left[\underbrace{\left(R(\bar{g}_n^{\text{tail}}) - R^*\right)\mathbf{1}_{2^{j-1}\delta < \text{sign}(g_*(X))g_\lambda(X)}}_{\text{Theorem 9}} | x_1,\dots,x_n\right]\right. \\
&\qquad\qquad \left. \cdot \mathbf{1}_{\text{sign}(g_*(X))g_\lambda(X) \leqslant 2^j\delta}\right] \\
&\leqslant \delta^\alpha + 4\sum_{j\geqslant 1} \mathbb{P}\left(\text{sign}(g_*(X))g_\lambda(X) \leqslant 2^j\delta\right)\exp\left(-(2^j\delta)^2 K_R(\delta)(n+1)\right) \\
&\leqslant \delta^\alpha + 4\delta^\alpha \sum_{j\geqslant 1} 2^{\alpha j}\exp\left(-(2^j\delta)^2 K_R(\delta)(n+1)\right),
\end{aligned}$$

and $K_R(\delta)^{-1} = 2^9 R^2\left(1 + \|\tilde{g}_* - g_\lambda\|_\infty^2\right)\text{Tr}(\Sigma(\Sigma+\lambda I)^{-2}) + \dfrac{32\delta R^2(1+2\|\tilde{g}_* - g_\lambda\|_\infty)}{3\lambda}$. 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 assumed **(A8)**, we chose $\lambda = \delta^{1/\gamma}$ such that $\|g_* - g_\lambda\|_\infty \leqslant \lambda^\gamma = \delta$. Second, **(A9)** implies (see Caponnetto and De Vito, 2007) that $\text{Tr}(\Sigma(\Sigma+\lambda I)^{-2}) \leqslant \dfrac{\beta}{(\beta-1)\lambda^{1+1/\beta}}$. For δ small enough, we have:

$$\begin{aligned}
K_R(\delta)^{-1} &\leqslant 2^{10}\frac{\beta R^2}{(\beta-1)\delta^{\frac{1+1/\beta}{\gamma}}} + 32\delta^{(\gamma-1)/\gamma}R^2 \\
K_R(\delta)^{-1} &\leqslant 2^{11}\frac{\beta R^2}{(\beta-1)}\cdot\delta^{-(\beta+1)/\beta\gamma}
\end{aligned}$$

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.

■