Statistical Optimality of Stochastic Gradient Descent on Hard Learning Problems through Multiple Passes

Loucas Pillaud-Vivien, Alessandro Rudi, Francis Bach

Statistical Optimality of Stochastic Gradient Descent on Hard Learning Problems through Multiple Passes

Loucas Pillaud-Vivien
INRIA - Ecole Normale Supérieure
PSL Research University
loucas.pillaud-vivien@inria.fr

Alessandro Rudi
INRIA - Ecole Normale Supérieure
PSL Research University
alessandro.rudi@inria.fr

Francis Bach
INRIA - Ecole Normale Supérieure
PSL Research University
francis.bach@inria.fr

Abstract

We consider stochastic gradient descent (SGD) for least-squares regression with potentially several passes over the data. While several passes have been widely reported to perform practically better in terms of predictive performance on unseen data, the existing theoretical analysis of SGD suggests that a single pass is statistically optimal. While this is true for low-dimensional easy problems, we show that for hard problems, multiple passes lead to statistically optimal predictions while single pass does not; we also show that in these hard models, the optimal number of passes over the data increases with sample size. In order to define the notion of hardness and show that our predictive performances are optimal, we consider potentially infinite-dimensional models and notions typically associated to kernel methods, namely, the decay of eigenvalues of the covariance matrix of the features and the complexity of the optimal predictor as measured through the covariance matrix. We illustrate our results on synthetic experiments with non-linear kernel methods and on a classical benchmark with a linear model.

1 Introduction

Stochastic gradient descent (SGD) and its multiple variants—averaged [1], accelerated [2], variance-reduced [3, 4, 5]—are the workhorses of large-scale machine learning, because (a) these methods looks at only a few observations before updating the corresponding model, and (b) they are known in theory and in practice to generalize well to unseen data [6].

Beyond the choice of step-size (often referred to as the learning rate), the number of passes to make on the data remains an important practical and theoretical issue. In the context of finite-dimensional models (least-squares regression or logistic regression), the theoretical answer has been known for many years: a single passes suffices for the optimal statistical performance [1, 7]. Worse, most of the theoretical work only apply to single pass algorithms, with some exceptions leading to analyses of multiple passes when the step-size is taken smaller than the best known setting [8, 9].

However, in practice, multiple passes are always performed as they empirically lead to better generalization (e.g., loss on unseen test data) [6]. But no analysis so far has been able to show that, given the appropriate step-size, multiple pass SGD was theoretically better than single pass SGD.

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

The main contribution of this paper is to show that for least-squares regression, while single pass averaged SGD is optimal for a certain class of "easy" problems, multiple passes are needed to reach optimal prediction performance on another class of "hard" problems.

In order to define and characterize these classes of problems, we need to use tools from infinite-dimensional models which are common in the analysis of kernel methods. De facto, our analysis will be done in infinite-dimensional feature spaces, and for finite-dimensional problems where the dimension far exceeds the number of samples, using these tools are the only way to obtain non-vacuous dimension-independent bounds. Thus, overall, our analysis applies both to finite-dimensional models with explicit features (parametric estimation), and to kernel methods (non-parametric estimation).

The two important quantities in the analysis are:

(a) The decay of eigenvalues of the covariance matrix Σ of the input features, so that the ordered eigenvalues λ_m decay as $O(m^{-\alpha})$; the parameter $\alpha \geqslant 1$ characterizes the size of the feature space, $\alpha = 1$ corresponding to the largest feature spaces and $\alpha = +\infty$ to finite-dimensional spaces. The decay will be measured through $\mathrm{tr}\Sigma^{1/\alpha} = \sum_m \lambda_m^{1/\alpha}$, which is small when the decay of eigenvalues is faster than $O(m^{-\alpha})$.

(b) The complexity of the optimal predictor θ_* as measured through the covariance matrix Σ, that is with coefficients $\langle e_m, \theta_* \rangle$ in the eigenbasis $(e_m)_m$ of the covariance matrix that decay so that $\langle \theta_*, \Sigma^{1-2r}\theta_* \rangle$ is small. The parameter $r \geqslant 0$ characterizes the difficulty of the learning problem: $r = 1/2$ corresponds to characterizing the complexity of the predictor through the squared norm $\|\theta_*\|^2$, and thus r close to zero corresponds to the hardest problems while r larger, and in particular $r \geqslant 1/2$, corresponds to simpler problems.

Dealing with non-parametric estimation provides a simple way to evaluate the optimality of learning procedures. Indeed, given problems with parameters r and α, the best prediction performance (averaged square loss on unseen data) is well known [10] and decay as $O(n^{\frac{-2r\alpha}{2r\alpha+1}})$, with $\alpha = +\infty$ leading to the usual parametric rate $O(n^{-1})$. For *easy problems*, that is for which $r \geqslant \frac{\alpha-1}{2\alpha}$, then it is known that most iterative algorithms achieve this optimal rate of convergence (but with various running-time complexities), such as exact regularized risk minimization [11], gradient descent on the empirical risk [12], or averaged stochastic gradient descent [13].

We show that for *hard problems*, that is for which $r \leqslant \frac{\alpha-1}{2\alpha}$ (see Example 1 for a typical hard problem), then multiple passes are superior to a single pass. More precisely, under additional assumptions detailed in Section 2 that will lead to a subset of the hard problems, with $\Theta(n^{(\alpha-1-2r\alpha)/(1+2r\alpha)})$ passes, we achieve the optimal statistical performance $O(n^{\frac{-2r\alpha}{2r\alpha+1}})$, while for all other hard problems, a single pass only achieves $O(n^{-2r})$. This is illustrated in Figure 1.

We thus get a number of passes that grows with the number of observations n and depends precisely on the quantities r and α. In synthetic experiments with kernel methods where α and r are known, these scalings are precisely observed. In experiments on parametric models with large dimensions, we also exhibit an increasing number of required passes when the number of observations increases.

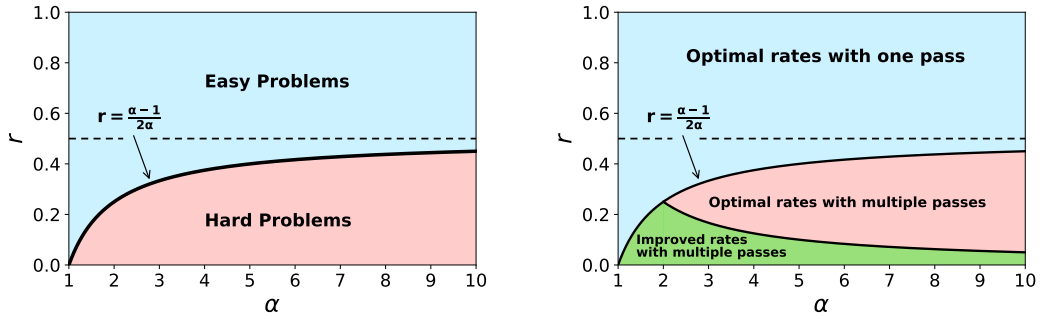

Figure 1 – (Left) easy and hard problems in the (α, r)-plane. (Right) different regions for which multiple passes improved known previous bounds (green region) or reaches optimality (red region).

2 Least-squares regression in finite dimension

We consider a joint distribution ρ on pairs of input/output $(x, y) \in \mathcal{X} \times \mathbb{R}$, where $\mathcal{X}$ is any input space, and we consider a feature map Φ from the input space $\mathcal{X}$ to a feature space $\mathcal{H}$, which we assume Euclidean in this section, so that all quantities are well-defined. In Section 4, we will extend all the notions to Hilbert spaces.

2.1 Main assumptions

We are considering predicting y as a linear function $f_\theta(x) = \langle \theta, \Phi(x) \rangle_{\mathcal{H}}$ of $\Phi(x)$, that is estimating $\theta \in \mathcal{H}$ such that $F(\theta) = \frac{1}{2}\mathbb{E}(y - \langle \theta, \Phi(x) \rangle_{\mathcal{H}})^2$ is as small as possible. Estimators will depend on n observations, with standard sampling assumptions:

(A1) *The n observations $(x_i, y_i) \in \mathcal{X} \times \mathbb{R}$, $i = 1, \dots, n$, are independent and identically distributed from the distribution ρ.*

Since $\mathcal{H}$ is finite-dimensional, F always has a (potentially non-unique) minimizer in $\mathcal{H}$ which we denote θ_*. We make the following standard boundedness assumptions:

(A2) *$\|\Phi(x)\| \leqslant R$ almost surely, $|y - \langle \theta_*, \Phi(x) \rangle_{\mathcal{H}}|$ is almost surely bounded by σ and $|y|$ is almost surely bounded by M.*

In order to obtain improved rates with multiple passes, and motivated by the equivalent previously used condition in reproducing kernel Hilbert spaces presented in Section 4, we make the following extra assumption (we denote by $\Sigma = \mathbb{E}[\Phi(x) \otimes_{\mathcal{H}} \Phi(x)]$ the (non-centered) covariance matrix).

(A3) *For $\mu \in [0, 1]$, there exists $\kappa_\mu \geqslant 0$ such that, almost surely, $\Phi(x) \otimes_{\mathcal{H}} \Phi(x) \preccurlyeq_{\mathcal{H}} \kappa_\mu^2 R^{2\mu} \Sigma^{1-\mu}$. Note that it can also be written as $\|\Sigma^{\mu/2 - 1/2} \Phi(x)\|_{\mathcal{H}} \leqslant \kappa_\mu R^\mu$.*

Assumption (A3) is always satisfied with any $\mu \in [0, 1]$, and has particular values for $\mu = 1$, with $\kappa_1 = 1$, and $\mu = 0$, where κ_0 has to be larger than the dimension of the space $\mathcal{H}$.

We will also introduce a parameter α that characterizes the decay of eigenvalues of Σ through the quantity $\mathrm{tr}\Sigma^{1/\alpha}$, as well as the difficulty of the learning problem through $\|\Sigma^{1/2-r}\theta_*\|_{\mathcal{H}}$, for $r \in [0, 1]$. In the finite-dimensional case, these quantities can always be defined and most often finite, but may be very large compared to sample size. In the following assumptions the quantities are assumed to be finite and small compared to n.

(A4) *There exists $\alpha > 1$ such that $\mathrm{tr}\ \Sigma^{1/\alpha} < \infty$.*

Assumption (A4) is often called the "capacity condition". First note that this assumption implies that the decreasing sequence of the eigenvalues of Σ, $(\lambda_m)_{m \geqslant 1}$, satisfies $\lambda_m = o\left(1/m^\alpha\right)$. Note that $\mathrm{tr}\Sigma^\mu \leqslant \kappa_\mu^2 R^{2\mu}$ and thus often we have $\mu \geqslant 1/\alpha$, and in the most favorable cases in Section 4, this bound will be achieved. We also assume:

(A5) *There exists $r \geqslant 0$, such that $\|\Sigma^{1/2-r}\theta_*\|_{\mathcal{H}} < \infty$.*

Assumption (A5) is often called the "source condition". Note also that for $r = 1/2$, this simply says that the optimal predictor has a small norm.

In the subsequent sections, we essentially assume that α, μ and r are chosen (by the theoretical analysis, not by the algorithm) so that all quantities R_μ, $\|\Sigma^{1/2-r}\theta_*\|_{\mathcal{H}}$ and $\mathrm{tr}\Sigma^{1/\alpha}$ are finite and small. As recalled in the introduction, these parameters are often used in the non-parametric literature to quantify the hardness of the learning problem (Figure 1).

We will use result with $O(\cdot)$ and $\Theta(\cdot)$ notations, which will all be independent of n and t (number of observations and number of iterations) but can depend on other finite constants. Explicit dependence on all parameters of the problem is given in proofs. More precisely, we will use the usual $O(\cdot)$ and $\Theta(\cdot)$ notations for sequences b_{nt} and a_{nt} that can depend on n and t, as $a_{nt} = O(b_{nt})$ if and only if, there exists $M > 0$ such that for all n, t, $a_{nt} \leqslant M b_{nt}$, and $a_{nt} = \Theta(b_{nt})$ if and only if, there exist $M, M' > 0$ such that for all n, t, $M' b_{nt} \leqslant a_{nt} \leqslant M b_{nt}$.

2.2 Related work

Given our assumptions above, several algorithms have been developed for obtaining low values of the expected excess risk $\mathbb{E}\big[F(\theta)\big] - F(\theta_*)$.

Regularized empirical risk minimization. Forming the empirical risk $\hat{F}(\theta)$, it minimizes $\hat{F}(\theta) + \lambda\|\theta\|_{\mathcal{H}}^2$, for appropriate values of λ. It is known that for easy problems where $r \geqslant \frac{\alpha-1}{2\alpha}$, it achieves the optimal rate of convergence $O(n^{\frac{-2r\alpha}{2r\alpha+1}})$ [11]. However, algorithmically, this requires to solve a linear system of size n times the dimension of $\mathcal{H}$. One could also use fast variance-reduced stochastic gradient algorithms such as SAG [3], SVRG [4] or SAGA [5], with a complexity proportional to the dimension of $\mathcal{H}$ times $n + R^2/\lambda$.

Early-stopped gradient descent on the empirical risk. Instead of solving the linear system directly, one can use gradient descent with early stopping [12, 14]. Similarly to the regularized empirical risk minimization case, a rate of $O(n^{-\frac{2r\alpha}{2r\alpha+1}})$ is achieved for the easy problems, where $r \geqslant \frac{\alpha-1}{2\alpha}$. Different iterative regularization techniques beyond batch gradient descent with early stopping have been considered, with computational complexities ranging from $O(n^{1+\frac{\alpha}{2r\alpha+1}})$ to $O(n^{1+\frac{\alpha}{4r\alpha+2}})$ times the dimension of $\mathcal{H}$ (or n in the kernel case in Section 4) for optimal predictions [12, 15, 16, 17, 14].

Stochastic gradient. The usual stochastic gradient recursion is iterating from $i = 1$ to n,

$$\theta_i = \theta_{i-1} + \gamma\big(y_i - \langle \theta_{i-1}, \Phi(x_i)\rangle_{\mathcal{H}}\big)\Phi(x_i),$$

with the averaged iterate $\bar{\theta}_n = \frac{1}{n}\sum_{i=1}^n \theta_i$. Starting from $\theta_0 = 0$, [18] shows that the expected excess performance $\mathbb{E}[F(\bar{\theta}_n)] - F(\theta_*)$ decomposes into a *variance* term that depends on the noise σ^2 in the prediction problem, and a *bias term*, that depends on the deviation $\theta_* - \theta_0 = \theta_*$ between the initialization and the optimal predictor. Their bound is, up to universal constants, $\frac{\sigma^2 \dim(\mathcal{H})}{n} + \frac{\|\theta_*\|_{\mathcal{H}}^2}{\gamma n}$.

Further, [13] considered the quantities α and r above to get the bound, up to constant factors:

$$\frac{\sigma^2 \mathrm{tr}\Sigma^{1/\alpha}(\gamma n)^{1/\alpha}}{n} + \frac{\|\Sigma^{1/2-r}\theta_*\|^2}{\gamma^{2r} n^{2r}}.$$

We recover the finite-dimensional bound for $\alpha = +\infty$ and $r = 1/2$. The bounds above are valid for all $\alpha \geqslant 1$ and all $r \in [0, 1]$, and the step-size γ is such that $\gamma R^2 \leqslant 1/4$, and thus we see a natural trade-off appearing for the step-size γ, between bias and variance.

When $r \geqslant \frac{\alpha-1}{2\alpha}$, then the optimal step-size minimizing the bound above is $\gamma \propto n^{\frac{-2\alpha\min\{r,1\}-1+\alpha}{2\alpha\min\{r,1\}+1}}$, and the obtained rate is optimal. Thus a single pass is optimal. However, when $r \leqslant \frac{\alpha-1}{2\alpha}$, the best step-size does not depend on n, and one can only achieve $O(n^{-2r})$.

Finally, in the same multiple pass set-up as ours, [9] has shown that for easy problems where $r \geqslant \frac{\alpha-1}{2\alpha}$ (and single-pass averaged SGD is already optimal) that multiple-pass non-averaged SGD is becoming optimal after a correct number of passes (while single-pass is not). Our proof principle of comparing to batch gradient is taken from [9], but we apply it to harder problems where $r \leqslant \frac{\alpha-1}{2\alpha}$. Moreover we consider the multi-pass averaged-SGD algorithm, instead of non-averaged SGD, and take explicitly into account the effect of Assumption (A3).

3 Averaged SGD with multiple passes

We consider the following algorithm, which is stochastic gradient descent with sampling with replacement with multiple passes over the data (we experiment in Section E of the Appendix with cycling over the data, with or without reshuffling between each pass).

- **Initialization**: $\theta_0 = \bar{\theta}_0 = 0$, t = maximal number of iterations, $\gamma = 1/(4R^2)$ = step-size
- **Iteration**: for $u = 1$ to t, sample $i(u)$ uniformly from $\{1, \dots, n\}$ and make the step

$$\theta_u = \theta_{u-1} + \gamma\big(y_{i(u)} - \langle \theta_{t-1}, \Phi(x_{i(u)})\rangle_{\mathcal{H}}\big)\Phi(x_{i(u)}) \quad \text{and} \quad \bar{\theta}_u = (1 - \tfrac{1}{u})\bar{\theta}_{u-1} + \tfrac{1}{u}\theta_u.$$

In this paper, following [18, 13], but as opposed to [19], we consider unregularized recursions. This removes a unnecessary regularization parameter (at the expense of harder proofs).

3.1 Convergence rate and optimal number of passes

Our main result is the following (see full proof in Appendix):

Theorem 1. *Let $n \in \mathbb{N}^*$ and $t \geqslant n$, under Assumptions (A1), (A2), (A3), (A4), (A5), (A6), with $\gamma = 1/(4R^2)$.*

- *For $\mu\alpha < 2r\alpha + 1 < \alpha$, if we take $t = \Theta(n^{\alpha/(2r\alpha+1)})$, we obtain the following rate:*

$$\mathbb{E}F(\bar{\theta}_t) - F(\theta_*) = O(n^{-2r\alpha/(2r\alpha+1)}).$$

- *For $\mu\alpha \geqslant 2r\alpha + 1$, if we take $t = \Theta(n^{1/\mu} (\log n)^{\frac{1}{\mu}})$, we obtain the following rate:*

$$\mathbb{E}F(\bar{\theta}_t) - F(\theta_*) \leqslant O(n^{-2r/\mu}).$$

Sketch of proof. The main difficulty in extending proofs from the single pass case [18, 13] is that as soon as an observation is processed twice, then statistical dependences are introduced and the proof does not go through. In a similar context, some authors have considered stability results [8], but the large step-sizes that we consider do not allow this technique. Rather, we follow [16, 9] and compare our multi-pass stochastic recursion θ_t to the batch gradient descent iterate η_t defined as $\eta_t = \eta_{t-1} + \frac{\gamma}{n}\sum_{i=1}^n \big(y_i - \langle \eta_{t-1}, \Phi(x_i)\rangle_{\mathcal{H}}\big)\Phi(x_i)$ with its averaged iterate $\bar{\eta}_t$. We thus need to study the predictive performance of $\bar{\eta}_t$ and the deviation $\bar{\theta}_t - \bar{\eta}_t$. It turns out that, given the data, the deviation $\theta_t - \eta_t$ satisfies an SGD recursion (with the respect to the randomness of the sampling with replacement). For a more detailed summary of the proof technique see Section B.

The novelty compared to [16, 9] is (a) to use refined results on averaged SGD for least-squares, in particular convergence in various norms for the deviation $\bar{\theta}_t - \bar{\eta}_t$ (see Section A), that can use our new Assumption (A3). Moreover, (b) we need to extend the convergence results for the batch gradient descent recursion from [14], also to take into account the new assumption (see Section D). These two results are interesting on their own.

Improved rates with multiple passes. We can draw the following conclusions:

- If $2\alpha r + 1 \geqslant \alpha$, that is, easy problems, it has been shown by [13] that a single pass with a smaller step-size than the one we propose here is optimal, and our result does not apply.
- If $\mu\alpha < 2r\alpha + 1 < \alpha$, then our proposed number of iterations is $t = \Theta(n^{\alpha/(2\alpha r+1)})$, which is now greater than n; the convergence rate is then $O(n^{\frac{-2r\alpha}{2r\alpha+1}})$, and, as we will see in Section 4.2, the predictive performance is then optimal when $\mu \leqslant 2r$.
- If $\mu\alpha \geqslant 2r\alpha + 1$, then with a number of iterations is $t = \Theta(n^{1/\mu})$, which is greater than n (thus several passes), with a convergence rate equal to $O(n^{-2r/\mu})$, which improves upon the best known rates of $O(n^{-2r})$. As we will see in Section 4.2, this is not optimal.

Note that these rates are theoretically only bounds on the optimal number of passes over the data, and one should be cautious when drawing conclusions; however our simulations on synthetic data, see Figure 2 in Section 5, confirm that our proposed scalings for the number of passes is observed in practice.

4 Application to kernel methods

In the section above, we have assumed that $\mathcal{H}$ was finite-dimensional, so that the optimal predictor $\theta_* \in \mathcal{H}$ was always defined. Note however, that our bounds that depends on α, r and μ are *independent of the dimension*, and hence, intuitively, following [19], should apply immediately to infinite-dimensional spaces.

We now first show in Section 4.1 how this intuition can be formalized and how using kernel methods provides a particularly interesting example. Moreover, this interpretation allows to characterize the statistical optimality of our results in Section 4.2.

4.1 Extension to Hilbert spaces, kernel methods and non-parametric estimation

Our main result in Theorem 1 extends directly to the case where $\mathcal{H}$ is an infinite-dimensional Hilbert space. In particular, given a feature map $\Phi : \mathcal{X} \to \mathcal{H}$, any vector $\theta \in \mathcal{H}$ is naturally associated to a function defined as $f_\theta(x) = \langle \theta, \Phi(x) \rangle_{\mathcal{H}}$. Algorithms can then be run with infinite-dimensional objects if the *kernel* $K(x', x) = \langle \Phi(x'), \Phi(x) \rangle_{\mathcal{H}}$ can be computed efficiently. This identification of elements θ of $\mathcal{H}$ with functions f_θ endows the various quantities we have introduced in the previous sections, with natural interpretations in terms of functions. The stochastic gradient descent described in Section 3 adapts instantly to this new framework as the iterates $(\theta_u)_{u \leqslant t}$ are linear combinations of feature vectors $\Phi(x_i)$, $i = 1, \dots, n$, and the algorithms can classically be "kernelized" [20, 13], with an overall running time complexity of $O(nt)$.

First note that Assumption (A3) is equivalent to, for all $x \in \mathcal{X}$ and $\theta \in \mathcal{H}$, $|f_\theta(x)|^2 \leqslant \kappa_\mu^2 R^{2\mu} \langle f_\theta, \Sigma^{1-\mu} f_\theta \rangle_{\mathcal{H}}$, that is, $\|g\|_{L_\infty}^2 \leqslant \kappa_\mu^2 R^{2\mu} \|\Sigma^{1/2-\mu/2} g\|_{\mathcal{H}}^2$ for any $g \in \mathcal{H}$ and also implies[1] $\|g\|_{L_\infty} \leqslant \kappa_\mu R^\mu \|g\|_{\mathcal{H}}^\mu \|g\|_{L_2}^{1-\mu}$, which are common assumptions in the context of kernel methods [22], essentially controlling in a more refined way the regularity of the whole space of functions associated to $\mathcal{H}$, with respect to the L^∞-norm, compared to the too crude inequality $\|g\|_{L^\infty} = \sup_x |\langle \Phi(x), g \rangle_{\mathcal{H}}| \leqslant \sup_x \|\Phi(x)\|_{\mathcal{H}} \|g\|_{\mathcal{H}} \leqslant R \|g\|_{\mathcal{H}}$.

The natural relation with functions allows to analyze effects that are crucial in the context of learning, but difficult to grasp in the finite-dimensional setting. Consider the following prototypical example of a hard learning problem,

Example 1 (Prototypical hard problem on simple Sobolev space)**.** *Let $\mathcal{X} = [0, 1]$, with x sampled uniformly on X and*

$$y = \mathit{sign}(x - 1/2) + \epsilon, \quad \Phi(x) = \{|k|^{-1} e^{2ik\pi x}\}_{k \in \mathbb{Z}^*}.$$

This corresponds to the kernel $K(x, y) = \sum_{k \in \mathbb{Z}^*} |k|^{-2} e^{2ik\pi(x-y)}$, which is well defined (and lead to the simplest Sobolev space). Note that for any $\theta \in \mathcal{H}$, which is here identified as the space of square-summable sequences $\ell^2(\mathbb{Z})$, we have $f_\theta(x) = \langle \theta, \Phi(x) \rangle_{\ell^2(\mathbb{Z})} = \sum_{k \in \mathbb{Z}^*} \frac{\theta_k}{|k|} e^{2ik\pi x}$. This means that for any estimator $\hat{\theta}$ given by the algorithm, $f_{\hat{\theta}}$ is at least once continuously differentiable, while the target function $\mathrm{sign}(\cdot - 1/2)$ is not even continuous. Hence, we are in a situation where θ_*, the minimizer of the excess risk, does not belong to $\mathcal{H}$. Indeed let represent $\mathrm{sign}(\cdot - 1/2)$ in $\mathcal{H}$, for almost all $x \in [0, 1]$, by its Fourier series $\mathrm{sign}(x - 1/2) = \sum_{k \in \mathbb{Z}_*} \alpha_k e^{2ik\pi x}$, with $|\alpha_k| \sim 1/k$, an informal reasoning would lead to $(\theta_*)_k = \alpha_k |k| \sim 1$, which is not square-summable and thus $\theta_* \notin \mathcal{H}$. For more details, see [23, 24].

This setting generalizes important properties that are valid for Sobolev spaces, as shown in the following example, where α, r, μ are characterized in terms of the smoothness of the functions in $\mathcal{H}$, the smoothness of f^* and the dimensionality of the input space $\mathcal{X}$.

Example 2 (Sobolev Spaces [25, 22, 26, 10])**.** *Let $\mathcal{X} \subseteq \mathbb{R}^d$, $d \in \mathbb{N}$, with $\rho_{\mathcal{X}}$ supported on $\mathcal{X}$, absolutely continous with the uniform distribution and such that $\rho_{\mathcal{X}}(x) \geqslant a > 0$ almost everywhere, for a given a. Assume that $f^*(x) = \mathbb{E}[y|x]$ is s-times differentiable, with $s > 0$. Choose a kernel, inducing Sobolev spaces of smoothness m with $m > d/2$, as the Matérn kernel*

$$K(x', x) = \|x' - x\|^{m - d/2} \mathcal{K}_{d/2-m}(\|x' - x\|),$$

where $\mathcal{K}_{d/2-m}$ is the modified Bessel function of the second kind. Then the assumptions are satisfied for any $\epsilon > 0$, with $\alpha = \frac{2m}{d}$, $\mu = \frac{d}{2m} + \epsilon$, $r = \frac{s}{2m}$.

In the following subsection we compare the rates obtained in Thm. 1, with known lower bounds under the same assumptions.

4.2 Minimax lower bounds

In this section we recall known lower bounds on the rates for classes of learning problems satisfying the conditions in Sect. 2.1. Interestingly, the comparison below shows that our results in Theorem 1

[1] Indeed, for any $g \in \mathcal{H}$, $\|\Sigma^{1/2-\mu/2} g\|_{\mathcal{H}} = \|\Sigma^{-\mu/2} g\|_{L_2} \leqslant \|\Sigma^{-1/2} g\|_{L_2}^\mu \|g\|_{L_2}^{1-\mu} = \|g\|_{\mathcal{H}}^\mu \|g\|_{L_2}^{1-\mu}$, where we used that for any $g \in \mathcal{H}$, any bounded operator A, $s \in [0, 1]$: $\|A^s g\|_{L_2} \leqslant \|Ag\|_{L_2}^s \|g\|_{L_2}^{1-s}$ (see [21]).

are optimal in the setting $2r \geqslant \mu$. While the optimality of SGD was known for the regime $\{2r\alpha+1 \geqslant \alpha \cap 2r \geqslant \mu\}$, here we extend the optimality to the new regime $\alpha \geqslant 2r\alpha + 1 \geqslant \mu\alpha$, covering essentially all the region $2r \geqslant \mu$, as it is possible to observe in Figure 1, where for clarity we plotted the best possible value for μ that is $\mu = 1/\alpha$ [10] (which is true for Sobolev spaces).

When $r \in (0, 1]$ is fixed, but there are no assumptions on α or μ, then the optimal minimax rate of convergence is $O(n^{-2r/(2r+1)})$, attained by regularized empirical risk minimization [11] and other spectral filters on the empirical covariance operator [27].

When $r \in (0, 1]$ and $\alpha \geqslant 1$ are fixed (but there are no constraints on μ), the optimal minimax rate of convergence $O(n^{\frac{-2r\alpha}{2r\alpha+1}})$ is attained when $r \geqslant \frac{\alpha-1}{2\alpha}$, with empirical risk minimization [14] or stochastic gradient descent [13].

When $r \geqslant \frac{\alpha-1}{2\alpha}$, the rate of convergence $O(n^{\frac{-2r\alpha}{2r\alpha+1}})$ is known to be a lower bound on the optimal minimax rate, but the best upper-bound so far is $O(n^{-2r})$ and is achieved by empirical risk minimization [14] or stochastic gradient descent [13], and the optimal rate is not known.

When $r \in (0, 1]$, $\alpha \geqslant 1$ and $\mu \in [1/\alpha, 1]$ are fixed, then the rate of convergence $O(n^{\frac{-\max\{\mu,2r\}\alpha}{2\max\{\mu,2r\}\alpha+1}})$ is known to be a lower bound on the optimal minimax rate [10]. This is attained by regularized empirical risk minimization when $2r \geqslant \mu$ [10], and *now by SGD with multiple passes*, and it is thus the optimal rate in this situation. When $2r < \mu$, the only known upper bound is $O(n^{-2\alpha r/(\mu\alpha+1)})$, and the optimal rate is not known.

5 Experiments

In our experiments, the main goal is to show that with more that one pass over the data, we can improve the accuracy of SGD when the problem is hard. We also want to highlight our dependence of the optimal number of passes (that is t/n) with respect to the number of observations n.

Synthetic experiments. Our main experiments are performed on artificial data following the setting in [21]. For this purpose, we take kernels K corresponding to splines of order q (see [24]) that fulfill Assumptions (A1) (A2) (A3) (A4) (A5) (A6). Indeed, let us consider the following function

$$\Lambda_q(x, z) = \sum_{k \in \mathbb{Z}} \frac{e^{2i\pi k(x-z)}}{|k|^q},$$

defined almost everywhere on $[0, 1]$, with $q \in \mathbb{R}$, and for which we have the interesting relationship: $\langle \Lambda_q(x, \cdot), \Lambda_{q'}(z, \cdot)\rangle_{L_2(d\rho_{\mathcal{X}})} = \Lambda_{q+q'}(x, z)$ for any $q, q' \in \mathbb{R}$. Our setting is the following:

- **Input distribution:** $\mathcal{X} = [0, 1]$ and $\rho_{\mathcal{X}}$ is the uniform distribution.
- **Kernel:** $\forall (x, z) \in [0, 1],\ K(x, z) = \Lambda_\alpha(x, z)$.
- **Target function:** $\forall x \in [0, 1],\ \theta_* = \Lambda_{r\alpha+\frac{1}{2}}(x, 0)$.
- **Output distribution :** $\rho(y|x)$ is a Gaussian with variance σ^2 and mean θ_*.

For this setting we can show that the learning problem satisfies Assumptions (A1) (A2) (A3) (A4) (A5) (A6) with $r,\ \alpha,$ and$\mu = 1/\alpha$. We take different values of these parameters to encounter all the different regimes of the problems shown in Figure 1.

For each n from 100 to 1000, we found the optimal number of steps $t_*(n)$ that minimizes the test error $F(\bar{\theta}_t) - F(\theta_*)$. Note that because of overfitting the test error increases for $t > t_*(n)$. In Figure 2, we show $t_*(n)$ with respect to n in log scale. As expected, for the easy problems (where $r \geqslant \frac{\alpha-1}{2\alpha}$, see top left and right plots), the slope of the plot is 1 as one pass over the data is enough: $t_*(n) = \Theta(n)$. But we see that for hard problems (where $r \leqslant \frac{\alpha-1}{2\alpha}$, see bottom left and right plots), we need more than one pass to achieve optimality as the optimal number of iterations is very close to $t_*(n) = \Theta\left(n^{\frac{\alpha}{2r\alpha+1}}\right)$. That matches the theoretical predictions of Theorem 1. We also notice in the plots that, the bigger $\frac{\alpha}{2r\alpha+1}$ the harder the problem is and the bigger the number of epochs we have to take. Note, that to reduce the noise on the estimation of $t_*(n)$, plots show an average over 100 replications.

To conclude, the experiments presented in the section correspond exactly to the theoretical setting of the article (sampling with replacement), however we present in Figures 4 and 5 of Section E of the Appendix results on the same datasets for two different ways of sampling the data: (a)*without replacement*: for which we select randomly the data points but never use twice the same point in one epoch, (b) *cycles*: for which we pick successively the data points in the same order. The obtained scalings relating number of iterations or passes to number of observations are the same.

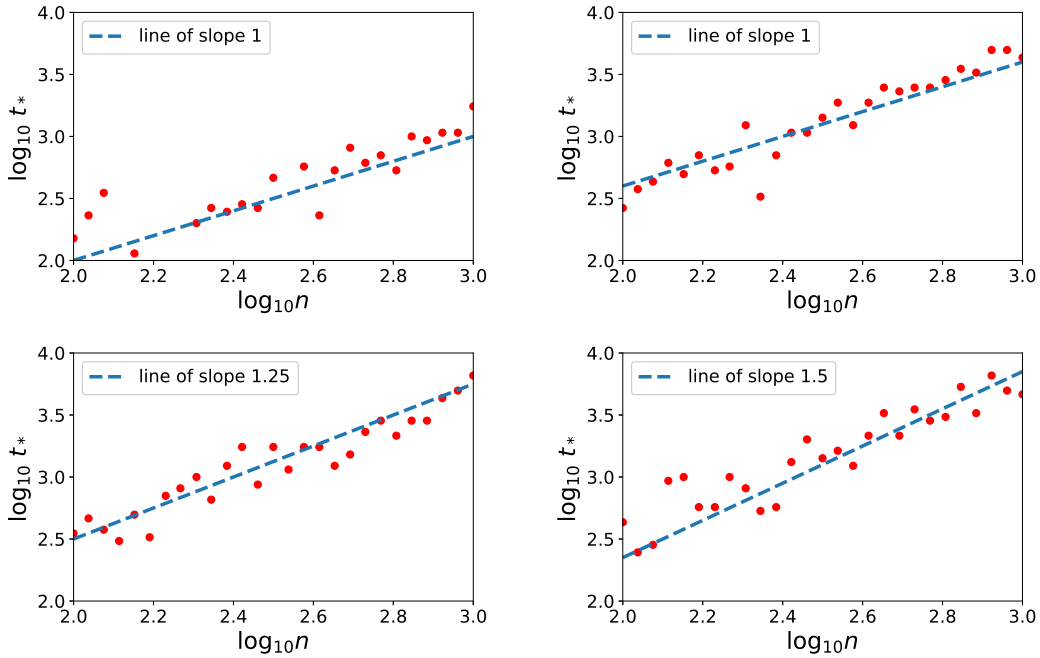

Figure 2 – The four plots represent each a different configuration on the (α, r) plan represented in Figure 1, for $r = 1/(2\alpha)$. **Top left** ($\alpha = 1.5$) and **right** ($\alpha = 2$) are two easy problems (Top right is the limiting case where $r = \frac{\alpha-1}{2\alpha}$) for which one pass over the data is optimal. **Bottom left** ($\alpha = 2.5$) and **right** ($\alpha = 3$) are two hard problems for which an increasing number of passes is required. The blue dotted line are the slopes predicted by the theoretical result in Theorem 1.

Linear model. To illustrate our result with some real data, we show how the optimal number of passes over the data increases with the number of samples. In Figure 3, we simply performed linear least-squares regression on the MNIST dataset and plotted the optimal number of passes over the data that leads to the smallest error on the test set. Evaluating α and r from Assumptions (A4) and (A5), we found $\alpha = 1.7$ and $r = 0.18$. As $r = 0.18 \leqslant \frac{\alpha-1}{2\alpha} \sim 0.2$, Theorem 1 indicates that this corresponds to a situation where only one pass on the data is not enough, confirming the behavior of Figure 3. This suggests that learning MNIST with linear regression is a *hard problem*.

6 Conclusion

In this paper, we have shown that for least-squares regression, in hard problems where single-pass SGD is not statistically optimal ($r < \frac{\alpha-1}{2\alpha}$), then multiple passes lead to statistical optimality with a number of passes that somewhat surprisingly needs to grow with sample size, with a convergence rate which is superior to previous analyses of stochastic gradient. Using a non-parametric estimation, we show that under certain conditions ($2r \geqslant \mu$), we attain statistical optimality.

Our work could be extended in several ways: (a) our experiments suggest that cycling over the data and cycling with random reshuffling perform similarly to sampling with replacement, it would be interesting to combine our theoretical analysis with work aiming at analyzing other sampling schemes [28, 29]. (b) Mini-batches could be also considered with a potentially interesting effects compared to the streaming setting. Also, (c) our analysis focuses on least-squares regression, an extension to all smooth loss functions would widen its applicability. Moreover, (d) providing optimal

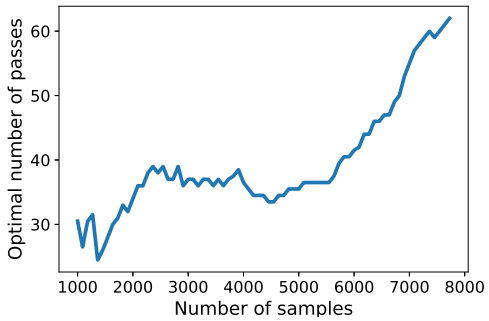

Figure 3 – For the MNIST data set, we show the optimal number of passes over the data with respect to the number of samples in the case of the linear regression.

efficient algorithms for the situation $2r < \mu$ is a clear open problem (for which the optimal rate is not known, even for non-efficient algorithms). Additionally, (e) in the context of classification, we could combine our analysis with [30] to study the potential discrepancies between training and testing losses and errors when considering high-dimensional models [31]. More generally, (f) we could explore the effect of our analysis for methods based on the least squares estimator in the context of structured prediction [32, 33, 34] and (non-linear) multitask learning [35]. Finally, (g) to reduce the computational complexity of the algorithm, while retaining the (optimal) statistical guarantees, we could combine multi-pass stochastic gradient descent, with approximation techniques like *random features* [36], extending the analysis of [37] to the more general setting considered in this paper.

Acknowledgements

We acknowledge support from the European Research Council (grant SEQUOIA 724063). We also thank Raphaël Berthier and Yann Labbé for their enlightening advices on this project.

References

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

[2] Guanghui Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1-2):365–397, 2012.

[3] Nicolas L. Roux, Mark Schmidt, and Francis Bach. A stochastic gradient method with an exponential convergence rate for finite training sets. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

[4] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, 2013.

[5] 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.

[6] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

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

[8] Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, 2016.

[9] Junhong Lin and Lorenzo Rosasco. Optimal rates for multi-pass stochastic gradient methods. *Journal of Machine Learning Research*, 18(97):1–47, 2017.

[10] Simon Fischer and Ingo Steinwart. Sobolev norm learning rates for regularized least-squares algorithm. Fakultät für Mathematik und Physik, Universität Stuttgart, 2017.

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

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

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

[14] Junhong Lin, Alessandro Rudi, Lorenzo Rosasco, and Volkan Cevher. Optimal rates for spectral algorithms with least-squares regression over hilbert spaces. *Applied and Computational Harmonic Analysis*, 2018.

[15] L Lo Gerfo, L Rosasco, F Odone, E De Vito, and A Verri. Spectral algorithms for supervised learning. *Neural Computation*, 20(7):1873–1897, 2008.

[16] Lorenzo Rosasco and Silvia Villa. Learning with incremental iterative regularization. In *Advances in Neural Information Processing Systems*, pages 1630–1638, 2015.

[17] Gilles Blanchard and Nicole Krämer. Convergence rates of kernel conjugate gradient for random design regression. *Analysis and Applications*, 14(06):763–794, 2016.

[18] Francis Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$. In *Advances in Neural Information Processing Systems (NIPS)*, pages 773–781, 2013.

[19] Aymeric Dieuleveut, Nicolas Flammarion, and Francis Bach. Harder, better, faster, stronger convergence rates for least-squares regression. *Journal of Machine Learning Research*, 18(1):3520–3570, 2017.

[20] Yiming Ying and Massimiliano Pontil. Online gradient descent learning algorithms. *Foundations of Computational Mathematics*, 8(5):561–596, Oct 2008.

[21] Alessandro Rudi and Lorenzo Rosasco. Generalization properties of learning with random features. In *Advances in Neural Information Processing Systems*, pages 3215–3225, 2017.

[22] Ingo Steinwart, Don R. Hush, and Clint Scovel. Optimal rates for regularized least squares regression. In *Proc. COLT*, 2009.

[23] R. A. Adams. *Sobolev spaces / Robert A. Adams*. Academic Press New York, 1975.

[24] G. Wahba. *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics, 1990.

[25] Holger Wendland. *Scattered data approximation*, volume 17. Cambridge university press, 2004.

[26] Francis Bach. On the equivalence between kernel quadrature rules and random feature expansions. *Journal of Machine Learning Research*, 18(21):1–38, 2017.

[27] Gilles Blanchard and Nicole Mücke. Optimal rates for regularization of statistical inverse learning problems. *Foundations of Computational Mathematics*, pages 1–43, 2017.

[28] Ohad Shamir. Without-replacement sampling for stochastic gradient methods. In *Advances in Neural Information Processing Systems 29*, pages 46–54, 2016.

[29] Mert Gürbüzbalaban, Asu Ozdaglar, and Pablo Parrilo. Why random reshuffling beats stochastic gradient descent. Technical Report 1510.08560, arXiv, 2015.

[30] Loucas Pillaud-Vivien, Alessandro Rudi, and Francis Bach. Exponential convergence of testing error for stochastic gradient methods. In *Proceedings of the 31st Conference On Learning Theory*, volume 75, pages 250–296, 2018.

[31] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

[32] Carlo Ciliberto, Lorenzo Rosasco, and Alessandro Rudi. A consistent regularization approach for structured prediction. In *Advances in Neural Information Processing Systems*, pages 4412–4420, 2016.

[33] Anton Osokin, Francis Bach, and Simon Lacoste-Julien. On structured prediction theory with calibrated convex surrogate losses. In *Advances in Neural Information Processing Systems*, pages 302–313, 2017.

[34] Carlo Ciliberto, Francis Bach, and Alessandro Rudi. Localized structured prediction. *arXiv preprint arXiv:1806.02402*, 2018.

[35] Carlo Ciliberto, Alessandro Rudi, Lorenzo Rosasco, and Massimiliano Pontil. Consistent multitask learning with nonlinear output relations. In *Advances in Neural Information Processing Systems*, pages 1986–1996, 2017.

[36] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, pages 1177–1184, 2008.

[37] Luigi Carratino, Alessandro Rudi, and Lorenzo Rosasco. Learning with sgd and random features. In *Advances in Neural Information Processing Systems 31*, pages 10213–10224, 2018.

[38] R. Aguech, E. Moulines, and P. Priouret. On a perturbation approach for the analysis of stochastic tracking algorithms. *SIAM J. Control and Optimization*, 39(3):872–899, 2000.

[39] Alessandro Rudi, Guillermo D Canas, and Lorenzo Rosasco. On the sample complexity of subspace learning. In *Advances in Neural Information Processing Systems*, pages 2067–2075, 2013.

[40] Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

[41] Alessandro Rudi, Luigi Carratino, and Lorenzo Rosasco. Falkon: An optimal large scale kernel method. In *Advances in Neural Information Processing Systems*, pages 3888–3898, 2017.

Appendix

The appendix in constructed as follows:

- We first present in Section A a new result for stochastic gradient recursions which generalizes the work of [18] and [13] to more general norms. This result could be used in other contexts.
- The proof technique for Theorem 1 is presented in Section B.
- In Section C we give a proof of the various lemmas needed in the first part of the proof of Theorem 1 (deviation between SGD and batch gradient descent).
- In Section D we provide new results for the analysis of batch gradient descent, which are adapted to our new (A3), and instrumental in proving Theorem 1 in Section B.
- Finally, in Section E we present experiments for different sampling techniques.

A A general result for the SGD variance term

Independently of the problem studied in this paper, we consider i.i.d. observations $(z_t, \xi_t) \in \mathcal{H} \times \mathcal{H}$ a Hilbert space, and the recursion started from $\mu_0 = 0$.

$$\mu_t = (I - \gamma z_t \otimes z_t)\, \mu_{t-1} + \gamma \xi_t \tag{1}$$

(this will applied with $z_t = \Phi(x_{i(t)})$). This corresponds to the variance term of SGD. We denote by $\bar{\mu}_t$ the averaged iterate $\bar{\mu}_t = \frac{1}{t} \sum_{i=1}^t \mu_i$.

The goal of the proposition below is to provide a bound on $\mathbb{E}\left[\left\|H^{u/2} \bar{\mu}_t\right\|^2\right]$ for $u \in [0, \frac{1}{\alpha} + 1]$, where $H = \mathbb{E}\left[z_t \otimes z_t\right]$ is such that $\operatorname{tr} H^{1/\alpha}$ is finite. Existing results only cover the case $u = 1$.

Proposition 1 (A general result for the SGD variance term)**.** *Let us consider the recursion in Eq. (1) started at $\mu_0 = 0$. Denote $\mathbb{E}\left[z_t \otimes z_t\right] = H$, assume that $\operatorname{tr} H^{1/\alpha}$ is finite, $\mathbb{E}\left[\xi_t\right] = 0$, $\mathbb{E}\left[(z_t \otimes z_t)^2\right] \preccurlyeq R^2 H$, $\mathbb{E}\left[\xi_t \otimes \xi_t\right] \preccurlyeq \sigma^2 H$ and $\gamma R^2 \leqslant 1/4$, then for $u \in [0, \frac{1}{\alpha} + 1]$:*

$$\mathbb{E}\left[\left\|H^{u/2} \bar{\mu}_t\right\|^2\right] \leqslant 4 \sigma^2 \gamma^{1-u}\, \frac{\gamma^{1/\alpha} \operatorname{tr} H^{1/\alpha}}{t^{u - 1/\alpha}}. \tag{2}$$

A.1 Proof principle

We follow closely the proof technique of [18], and prove Proposition 1 by showing it first for a "semi-stochastic" recursion, where $z_t \otimes z_t$ is replaced by its expectation (see Lemma 1). We will then compare our general recursion to the semi-stochastic one.

A.2 Semi-stochastic recursion

Lemma 1 (Semi-stochastic SGD)**.** *Let us consider the following recursion $\mu_t = (I - \gamma H)\, \mu_{t-1} + \gamma \xi_t$ started at $\mu_0 = 0$. Assume that $\operatorname{tr} H^{1/\alpha}$ is finite, $\mathbb{E}\left[\xi_t\right] = 0$, $\mathbb{E}\left[\xi_t \otimes \xi_t\right] \preccurlyeq \sigma^2 H$ and $\gamma H \preccurlyeq I$, then for $u \in [0, \frac{1}{\alpha} + 1]$:*

$$\mathbb{E}\left[\left\|H^{u/2} \bar{\mu}_t\right\|^2\right] \leqslant \sigma^2 \gamma^{1-u}\, \gamma^{1/\alpha} \operatorname{tr} H^{1/\alpha} t^{1/\alpha - u}. \tag{3}$$

Proof. For $t \geqslant 1$ and $u \in [0, \frac{1}{\alpha} + 1]$, using an explicit formula for μ_t and $\bar{\mu}_t$ (see [18] for details), we get:

$$\mu_t = (I - \gamma H)\,\mu_{t-1} + \gamma \xi_t = (I - \gamma H)^t\,\mu_0 + \gamma \sum_{k=1}^{t} (I - \gamma H)^{t-k}\,\xi_k$$

$$\bar{\mu}_t = \frac{1}{t}\sum_{u=1}^{t} \mu_u = \frac{\gamma}{t}\sum_{u=1}^{t}\sum_{k=1}^{u} (I - \gamma H)^{u-k}\,\xi_k = \frac{1}{t}\sum_{k=1}^{t} H^{-1}\left(I - (I - \gamma H)^{t-k+1}\right)\xi_k$$

$$\mathbb{E}\left[\left\|H^{u/2}\bar{\mu}_t\right\|^2\right] = \frac{1}{t^2}\mathbb{E}\sum_{k=1}^{t} \mathrm{tr}\left[\left(I - (I - \gamma H)^{t-k+1}\right)^2 H^{u-2}\xi_k \otimes \xi_k\right]$$

$$\leqslant \frac{\sigma^2}{t^2}\sum_{k=1}^{t} \mathrm{tr}\left[\left(I - (I - \gamma H)^{k}\right)^2 H^{u-1}\right] \text{ using } \mathbb{E}\left[\xi_t \otimes \xi_t\right] \preccurlyeq \sigma^2 H.$$

Now, let $(\lambda_i)_{i \in \mathbb{N}^*}$ be the non-increasing sequence of eigenvalues of the operator H. We obtain:

$$\mathbb{E}\left[\left\|H^{u/2}\bar{\mu}_t\right\|^2\right] \leqslant \frac{\sigma^2}{t^2}\sum_{k=1}^{t}\sum_{i=1}^{\infty}\left(I - (I - \gamma\lambda_i)^k\right)^2 \lambda_i^{u-1}.$$

We can now use a simple result[2] that for any $\rho \in [0,1]$, $k \geqslant 1$ and $u \in [0, \frac{1}{\alpha} + 1]$, we have : $(1 - (1-\rho)^k)^2 \leqslant (k\rho)^{1-u+1/\alpha}$, applied to $\rho = \gamma\lambda_i$. We get, by comparing sums to integrals:

$$\begin{aligned}
\mathbb{E}\left[\left\|H^{u/2}\bar{\mu}_t\right\|^2\right] &\leqslant \frac{\sigma^2}{t^2}\sum_{k=1}^{t}\sum_{i=1}^{\infty}\left(I - (I - \gamma\lambda_i)^k\right)^2 \lambda_i^{u-1} \\
&\leqslant \frac{\sigma^2}{t^2}\sum_{k=1}^{t}\sum_{i=1}^{\infty}(k\gamma\lambda_i)^{1-u+1/\alpha}\lambda_i^{u-1} \\
&\leqslant \frac{\sigma^2}{t^2}\gamma^{1-u+1/\alpha}\mathrm{tr}H^{1/\alpha}\sum_{k=1}^{t} k^{1-u+1/\alpha} \\
&\leqslant \frac{\sigma^2}{t^2}\gamma^{1-u+1/\alpha}\mathrm{tr}H^{1/\alpha}\int_1^t y^{1-u+1/\alpha}dy \\
&\leqslant \frac{\sigma^2}{t^2}\gamma^{1-u}\,\gamma^{1/\alpha}\mathrm{tr}H^{1/\alpha}\frac{t^{2-u+1/\alpha}}{2-u+1/\alpha} \\
&\leqslant \sigma^2\gamma^{1-u}\,\gamma^{1/\alpha}\mathrm{tr}H^{1/\alpha}t^{1/\alpha-u},
\end{aligned}$$

which shows the desired result. □

A.3 Relating the semi-stochastic recursion to the main recursion

Then, to relate the semi-stochastic recursion with the true one, we use an expansion in the powers of γ using recursively the perturbation idea from [38].

For $r \geqslant 0$, we define the sequence $(\mu_t^r)_{t \in \mathbb{N}}$, for $t \geqslant 1$,

$$\mu_t^r = (I - \gamma H)\mu_{t-1}^r + \gamma \Xi_t^r, \text{ with } \Xi_t^r = \begin{cases} (H - z_t \otimes z_t)\mu_{t-1}^{r-1} \text{ if } r \geqslant 1 \\ \Xi_t^0 = \xi_t \end{cases}. \tag{4}$$

We will show that $\mu_t \simeq \sum_{i=0}^{\infty}\mu_t^i$. To do so, notice that for $r \geqslant 0$, $\mu_t - \sum_{i=0}^{r}\mu_t^i$ follows the recursion:

$$\mu_t - \sum_{i=0}^{r}\mu_t^i = (I - z_t \otimes z_t)\left(\mu_{t-1} - \sum_{i=0}^{r}\mu_{t-1}^i\right) + \gamma\Xi_t^{r+1}, \tag{5}$$

so that by bounding the covariance operator we can apply a classical SGD result. This is the purpose of the following lemma.

[2] Indeed, adapting a similar result from [18], on the one hand, $1 - (1-\rho)^k \leqslant 1$ implying that $(1 - (1 - \rho)^k)^{1-1/\alpha+u} \leqslant 1$. On the other hand, $1 - (1-\gamma x)^k \leqslant \gamma k x$ implying that $(1 - (1-\rho)^k)^{1+1/\alpha-u} \leqslant (k\rho)^{1+1/\alpha-u}$. Thus by multiplying the two we get $(1 - (1-\rho)^k)^2 \leqslant (k\rho)^{1-u+1/\alpha}$.

Lemma 2 (Bound on covariance operator)**.** *For any $r \geqslant 0$, we have the following inequalities:*

$$\mathbb{E}\left[\Xi_t^r \otimes \Xi_t^r\right] \preccurlyeq \gamma^r R^{2r}\sigma^2 H \;\; \text{and} \;\; \mathbb{E}\left[\mu_t^r \otimes \mu_t^r\right] \preccurlyeq \gamma^{r+1} R^{2r}\sigma^2 I. \tag{6}$$

Proof. We propose a proof by induction on r. For $r = 0$, and $t \geqslant 0$, $\mathbb{E}\left[\Xi_t^0 \otimes \Xi_t^0\right] = \mathbb{E}\left[\xi_t \otimes \xi_t\right] \preccurlyeq \sigma^2 H$ by assumption. Moreover,

$$\mathbb{E}\left[\mu_t^0 \otimes \mu_t^0\right] = \gamma^2 \sum_{k=1}^{t-1} (I-\gamma H)^{t-k} \mathbb{E}\left[\Xi_t^0 \otimes \Xi_t^0\right] (I-\gamma H)^{t-k} \preccurlyeq \gamma^2\sigma^2 \sum_{k=1}^{t-1} (I-\gamma H)^{2(t-k)} H \preccurlyeq \gamma\sigma^2 I.$$

Then, for $r \geqslant 1$,

$$\begin{aligned}
\mathbb{E}\left[\Xi_t^{r+1} \otimes \Xi_t^{r+1}\right] &\preccurlyeq \mathbb{E}[(H - z_t \otimes z_t)\mu_{t-1}^r \otimes \mu_{t-1}^r (H - z_t \otimes z_t)] \\
&= \mathbb{E}[(H - z_t \otimes z_t)\mathbb{E}[\mu_{t-1}^r \otimes \mu_{t-1}^r](H - z_t \otimes z_t)] \\
&\preccurlyeq \gamma^{r+1} R^{2r}\sigma^2 \mathbb{E}[(H - z_t \otimes z_t)^2] \\
&\preccurlyeq \gamma^{r+1} R^{2r+2}\sigma^2 H.
\end{aligned}$$

And,

$$\begin{aligned}
\mathbb{E}\left[\mu_t^{r+1} \otimes \mu_t^{r+1}\right] &= \gamma^2 \sum_{k=1}^{t-1} (I-\gamma H)^{t-k} \mathbb{E}\left[\Xi_t^{r+1} \otimes \Xi_t^{r+1}\right] (I-\gamma H)^{t-k} \\
&\preccurlyeq \gamma^{r+3} R^{2r+2}\sigma^2 \sum_{k=1}^{t-1} (I-\gamma H)^{2(t-k)} H \preccurlyeq \gamma^{r+2} R^{2r+2}\sigma^2 I,
\end{aligned}$$

which thus shows the lemma by induction. □

To bound $\mu_t - \sum_{i=0}^r \mu_t^i$, we prove a very loose result for the average iterate, that will be sufficient for our purpose.

Lemma 3 (Bounding SGD recursion)**.** *Let us consider the following recursion $\mu_t = (I - \gamma z_t \otimes z_t)\,\mu_{t-1} + \gamma\xi_t$ starting at $\mu_0 = 0$. Assume that $\mathbb{E}[z_t \otimes z_t] = H$, $\mathbb{E}\left[\xi_t\right] = 0$, $\|x_t\|^2 \leqslant R^2$, $\mathbb{E}\left[\xi_t \otimes \xi_t\right] \preccurlyeq \sigma^2 H$ and $\gamma R^2 < I$, then for $u \in [0, \frac{1}{\alpha} + 1]$:*

$$\mathbb{E}\left[\left\|H^{u/2}\bar{\mu}_t\right\|^2\right] \leqslant \sigma^2\gamma^2 R^u \mathrm{tr} H\; t. \tag{7}$$

Proof. Let us define the operators for $j \leqslant i$: $M_j^i = (I - \gamma z_{i(i)} \otimes z_{i(i)}) \cdots (I - \gamma z_{i(j)} \otimes z_{i(j)})$ and $M_{i+1}^i = I$. Since $\mu_0 = 0$, note that we have we have, $\mu_i = \gamma \sum_{k=1}^i M_{k+1}^i \xi_k$. Hence, for $i \geqslant 1$,

$$\begin{aligned}
\mathbb{E}\left\|H^{u/2}\mu_i\right\|^2 &= \gamma^2 \mathbb{E} \sum_{k,j} \langle M_{j+1}^i \xi_j, H^u M_{k+1}^i \xi_k\rangle \\
&= \gamma^2 \mathbb{E} \sum_{k=1}^{i} \langle M_{k+1}^i \xi_k, H^u M_{k+1}^i \xi_k\rangle \\
&= \gamma^2 \mathrm{tr}\left(\mathbb{E}\left[\sum_{k=1}^{i} {M_{k+1}^i}^* H^u M_{k+1}^i \xi_k \otimes \xi_k\right]\right) \leqslant \sigma^2\gamma^2 \mathbb{E}\left[\sum_{k=1}^{i} \mathrm{tr}\left({M_{k+1}^i}^* H^u M_{k+1}^i H\right)\right] \\
&\leqslant \sigma^2\gamma^2 R^u i\; \mathrm{tr} H,
\end{aligned}$$

because $\mathrm{tr}\left({M_{k+1}^i}^* H^u M_{k+1}^i H\right) \leqslant R^u \mathrm{tr} H$. Then,

$$\begin{aligned}
\mathbb{E}\left\|H^{u/2}\bar{\mu}_t\right\|^2 &= \frac{1}{t^2} \sum_{i,j} \langle H^{u/2}\mu_i, H^{u/2}\mu_j\rangle \\
&\leqslant \frac{1}{t^2}\mathbb{E}\left(\sum_{i=1}^{t} \left\|H^{u/2}\mu_i\right\|\right)^2 \leqslant \frac{1}{t}\sum_{i=1}^{t} \mathbb{E}\left\|H^{u/2}\mu_i\right\|^2 \leqslant \sigma^2\gamma^2 R^u \mathrm{tr} H\; t,
\end{aligned}$$

which finishes the proof of Lemma 3. □

A.4 Final steps of the proof

We have now all the material to conclude. Indeed by the triangular inequality:

$$\left(\mathbb{E}\left\|H^{u/2}\bar{\mu}_t\right\|^2\right)^{1/2} \leqslant \sum_{i=1}^{r}\left(\underbrace{\mathbb{E}\left\|H^{u/2}\bar{\mu}_t^i\right\|^2}_{\text{Lemma 1}}\right)^{1/2} + \left(\underbrace{\mathbb{E}\left\|H^{u/2}\left(\bar{\mu}_t - \sum_{i=1}^{r}\bar{\mu}_t^i\right)\right\|^2}_{\text{Lemma 3}}\right)^{1/2}.$$

With Lemma 2, we have all the bounds on the covariance of the noise, so that:

$$\begin{aligned}\left(\mathbb{E}\left\|H^{u/2}\bar{\mu}_t\right\|^2\right)^{1/2} &\leqslant \sum_{i=1}^{r}\left(\gamma^i R^{2i}\sigma^2\gamma^{1-u}\,\gamma^{1/\alpha}\mathrm{tr}H^{1/\alpha}t^{1/\alpha-u}\right)^{1/2} + \left(\gamma^{r+2}R^{2r+u}\mathrm{tr}H\,t\right)^{1/2} \\ &\leqslant (\sigma^2\gamma^{1-u}\,\gamma^{1/\alpha}\mathrm{tr}H^{1/\alpha}t^{1/\alpha-u})^{1/2}\sum_{i=1}^{r}\left(\gamma R^2\right)^{i/2} + \left(\gamma^{r+2}R^{2r+u}\mathrm{tr}H\,t\right)^{1/2}.\end{aligned}$$

Now we make r go to infinity and we obtain:

$$\left(\mathbb{E}\left\|H^{u/2}\bar{\mu}_t\right\|^2\right)^{1/2} \leqslant (\sigma^2\gamma^{1-u}\,\gamma^{1/\alpha}\mathrm{tr}H^{1/\alpha}t^{1/\alpha-u})^{1/2}\frac{1}{1-\sqrt{\gamma R^2}} + \underbrace{\left(\gamma^{r+2}R^{2r+u}\mathrm{tr}H\,t\right)^{1/2}}_{\underset{r\to\infty}{\longrightarrow}0}$$

Hence with $\gamma R^2 \leqslant 1/4$,

$$\mathbb{E}\left\|H^{u/2}\bar{\mu}_t\right\|^2 \leqslant 4\sigma^2\gamma^{1-u}\,\gamma^{1/\alpha}\mathrm{tr}H^{1/\alpha}t^{1/\alpha-u},$$

which finishes to prove Proposition 1.

B Proof sketch for Theorem 1

We consider the batch gradient descent recursion, started from $\eta_0 = 0$, with the same step-size:

$$\eta_t = \eta_{t-1} + \frac{\gamma}{n}\sum_{i=1}^{n}\big(y_i - \langle\eta_{t-1}, \Phi(x_i)\rangle_{\mathcal{H}}\big)\Phi(x_i),$$

as well as its averaged version $\bar{\eta}_t = \frac{1}{t}\sum_{i=0}^{t}\eta_i$. We obtain a recursion for $\theta_t - \eta_t$, with the initialization $\theta_0 - \eta_0 = 0$, as follows:

$$\theta_t - \eta_t = \big[I - \Phi(x_{i(u)})\otimes_{\mathcal{H}}\Phi(x_{i(u)})\big](\theta_{t-1} - \eta_{t-1}) + \gamma\xi_t^1 + \gamma\xi_t^2,$$

with $\xi_t^1 = y_{i(u)}\Phi(x_{i(u)}) - \frac{1}{n}\sum_{i=1}^{n} y_i\Phi(x_i)$ and $\xi_t^2 = \big[\Phi(x_{i(u)})\otimes_{\mathcal{H}}\Phi(x_{i(u)}) - \frac{1}{n}\sum_{i=1}^{n}\Phi(x_i)\otimes_{\mathcal{H}}\Phi(x_i)\big]\eta_{t-1}$. We decompose the performance $F(\theta_t)$ in two parts, one analyzing the performance of batch gradient descent, one analyzing the deviation $\theta_t - \eta_t$, using

$$\mathbb{E}F(\bar{\theta}_t) - F(\theta_*) \leqslant 2\mathbb{E}\big[\|\Sigma^{1/2}(\theta_t - \eta_t)\|_{\mathcal{H}}^2\big] + 2\big[\mathbb{E}F(\bar{\eta}_t) - F(\theta_*)\big].$$

We denote by $\hat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^{n}\Phi(x_i)\otimes\Phi(x_i)$ the empirical second-order moment.

Deviation $\theta_t - \eta_t$. Denoting by $\mathcal{G}$ the σ-field generated by the data and by $\mathcal{F}_t$ the σ-field generated by $i(1), \dots, i(t)$, then, we have $\mathbb{E}(\xi_t^1|\mathcal{G}, \mathcal{F}_{t-1}) = \mathbb{E}(\xi_t^2|\mathcal{G}, \mathcal{F}_{t-1}) = 0$, thus we can apply results for averaged SGD (see Proposition 1 of the Appendix) to get the following lemma.

Lemma 4. *For any $t \geqslant 1$, if $\mathbb{E}\big[(\xi_t^1 + \xi_t^2)\otimes_{\mathcal{H}}(\xi_t^1 + \xi_t^2)|\mathcal{G}\big] \preccurlyeq \tau^2\hat{\Sigma}_n$, and $4\gamma R^2 = 1$, under Assumptions (A1), (A2), (A4),*

$$\mathbb{E}\big[\|\hat{\Sigma}_n^{1/2}(\bar{\theta}_t - \bar{\eta}_t)\|_{\mathcal{H}}^2|\mathcal{G}\big] \leqslant \frac{8\tau^2\gamma^{1/\alpha}\mathrm{tr}\,\hat{\Sigma}_n^{1/\alpha}}{t^{1-1/\alpha}}. \tag{8}$$

In order to obtain the bound, we need to bound τ^2 (which is dependent on $\mathcal{G}$) and go from a bound with the empirical covariance matrix $\hat{\Sigma}_n$ to bounds with the population covariance matrix Σ.

We have

$$\mathbb{E}\big[\xi_t^1 \otimes_{\mathcal{H}} \xi_t^1 | \mathcal{G}\big] \preccurlyeq_{\mathcal{H}} \mathbb{E}\big[y_{i(u)}^2 \Phi(x_{i(u)}) \otimes_{\mathcal{H}} \Phi(x_{i(u)}) | \mathcal{G}\big] \preccurlyeq_{\mathcal{H}} \|y\|_\infty^2 \hat{\Sigma}_n \preccurlyeq_{\mathcal{H}} (\sigma + \sup_{x \in \mathcal{X}} \langle \theta_*, \Phi(x) \rangle_{\mathcal{H}})^2 \hat{\Sigma}_n$$

$$\mathbb{E}\big[\xi_t^2 \otimes \xi_t^2 | \mathcal{G}\big] \preccurlyeq_{\mathcal{H}} \mathbb{E}\big[\langle \eta_{t-1}, \Phi(x_{i(u)}) \rangle^2 \Phi(x_{i(u)}) \otimes_{\mathcal{H}} \Phi(x_{i(u)}) | \mathcal{G}\big] \preccurlyeq_{\mathcal{H}} \sup_{t \in \{0,\dots,T-1\}} \sup_{x \in \mathcal{X}} \langle \eta_t, \Phi(x) \rangle_{\mathcal{H}})^2 \hat{\Sigma}_n$$

Therefore $\tau^2 = 2M^2 + 2 \sup_{t \in \{0,\dots,T-1\}} \sup_{x \in \mathcal{X}} \langle \eta_t, \Phi(x) \rangle_{\mathcal{H}}^2$ or using Assumption (A3) $\tau^2 = 2M^2 + 2 \sup_{t \in \{0,\dots,T-1\}} R^{2\mu} \kappa_\mu^2 \|\Sigma^{1/2-\mu/2} \eta_t\|_{\mathcal{H}}^2$.

In the proof, we rely on an event (that depend on $\mathcal{G}$) where $\hat{\Sigma}_n$ is close to Σ. This leads to the the following Lemma that bounds the deviation $\bar{\theta}_t - \bar{\eta}_t$.

Lemma 5. *For any $t \geqslant 1$, $4\gamma R^2 = 1$, under Assumptions (A1), (A2), (A4),*

$$\mathbb{E}\big[\|\Sigma^{1/2}(\bar{\theta}_t - \bar{\eta}_t)\|_{\mathcal{H}}^2\big] \leqslant 16 \tau_\infty^2 \left[R^{-2/\alpha} \operatorname{tr} \Sigma^{1/\alpha} t^{1/\alpha} \left(\frac{1}{t} + \left(\frac{4}{\mu} \frac{\log n}{n} \right)^{1/\mu} \right) + 1 \right]. \tag{9}$$

We make the following remark on the bound.

Remark 1. *Note that as defined in the proof τ_∞ may diverge in some cases as*

$$\tau_\infty^2 = \begin{cases} O(1) & \text{when } \mu \leqslant 2r, \\ O(n^{\mu - 2r}) & \text{when } 2r \leqslant \mu \leqslant 2r + 1/\alpha, \\ O(n^{1-2r/\mu}) & \text{when } \mu \geqslant 2r + 1/\alpha, \end{cases}$$

with $O(\cdot)$ are defined explicitly in the proof.

Convergence of batch gradient descent. The main result is summed up in the following lemma, with $t = O(n^{1/\mu})$ and $t \geqslant n$.

Lemma 6. *Let $t > 1$, under Assumptions (A1), (A2), (A3), (A4), (A5), (A6), when, with $4\gamma R^2 = 1$,*

$$t = \begin{cases} \Theta(n^{\alpha/(2r\alpha+1)}) & 2r\alpha + 1 > \mu\alpha \\ \Theta(n^{1/\mu} (\log n)^{\frac{1}{\mu}}) & 2r\alpha + 1 \leqslant \mu\alpha. \end{cases} \tag{10}$$

then,

$$\mathbb{E} F(\bar{\eta}_t) - F(\theta_*) \leqslant \begin{cases} O(n^{-2r\alpha/(2r\alpha+1)}) & 2r\alpha + 1 > \mu\alpha \\ O(n^{-2r/\mu}) & 2r\alpha + 1 \leqslant \mu\alpha \end{cases} \tag{11}$$

with $O(\cdot)$ are defined explicitly in the proof.

Remark 2. *In all cases, we can notice that the speed of convergence of Lemma 6 are slower that the ones in Lemma 5, hence, the convergence of the gradient descent controls the rates of convergence of the algorithm.*

C Bounding the deviation between SGD and batch gradient descent

In this section, following the proof sketch from Section B, we provide a bound on the deviation $\theta_t - \eta_t$. In all the following let us denote $\mu_t = \theta_t - \eta_t$ that deviation between the stochastic gradient descent recursion and the batch gradient descent recursion.

C.1 Proof of Lemma 5

We need to (a) go from $\hat{\Sigma}_n$ to Σ in the result of Lemma 4 and (b) to have a bound on τ. To prove this result we are going to need the two following lemmas:

Lemma 7. *Let $\lambda > 0$, $\delta \in (0, 1]$. Under Assumption (A3), when $n \geqslant 11(1 + \kappa_\mu^2 R^{2\mu} \gamma^\mu t^\mu) \log \frac{8R^2}{\lambda\delta}$, the following holds with probability $1 - \delta$,*

$$\left\| (\Sigma + \lambda I)^{1/2} (\hat{\Sigma}_n + \lambda I)^{-1/2} \right\|^2 \leqslant 2. \tag{12}$$

Proof. This Lemma is proven and stated lately in Lemma 14 in Section D.3. We recalled it here for the sake of clarity. □

Lemma 8. *Let $\lambda > 0$, $\delta \in (0,1]$. Under Assumption (A3), for $t = O\left(\frac{1}{n^{1/\mu}}\right)$ then the following holds with probability $1-\delta$,*

$$\tau^2 \leqslant \tau_\infty^2 \quad \text{and} \quad \tau_\infty^2 = \begin{cases} O(1), \text{ when } \mu \leqslant 2r, \\ O\left(n^{\mu-2r}\right), \text{ when } 2r \leqslant \mu \leqslant 2r + 1/\alpha, \\ O\left(n^{1-2r/\mu}\right) \text{ when } \mu \geqslant 2r + 1/\alpha, \end{cases} \tag{13}$$

where the $O(\cdot)$-notation depend only on the parameters of the problem (and is independent of n and t).

Proof. This Lemma is a direct implication of Corollary 2 in Section D.3. We recalled it here for the sake of clarity. □

Note that we can take $\lambda_n^\delta = \left(\frac{\log \frac{n}{\delta}}{n}\right)^{1/\mu}$ so that Lemma 7 result holds. Now we are ready to prove Lemma 5.

Proof of Lemma 5. Let A_{δ_a} be the set for which inequality (12) holds and let B_{δ_b} be the set for which inequality (13) holds. Note that $\mathbb{P}(A_{\delta_a}^c) = \delta_a$ and $\mathbb{P}(B_{\delta_b}^c) = \delta_b$. We use the following decomposition:

$$\mathbb{E}\left\|\Sigma^{1/2}\bar{\mu}_t\right\|^2 \leqslant \mathbb{E}\left[\left\|\Sigma^{1/2}\bar{\mu}_t\right\|^2 \mathbf{1}_{A_{\delta_a}\cap B_{\delta_b}}\right] + \mathbb{E}\left[\left\|\Sigma^{1/2}\bar{\mu}_t\right\|^2 \mathbf{1}_{A_{\delta_a}^c}\right] + \mathbb{E}\left[\left\|\Sigma^{1/2}\bar{\mu}_t\right\|^2 \mathbf{1}_{B_{\delta_b}^c}\right].$$

First, let us bound roughly $\|\bar{\mu}_t\|^2$.

First, for $i \geqslant 1$, $\|\mu_i\|^2 \leqslant \gamma^2\left(\sum_{i=1}^t \|\xi_i^1\| + \|\xi_i^2\|\right)^2 \leqslant 16R^2\gamma^2\tau^2t^2$, so that $\|\bar{\mu}_t\|^2 \leqslant \frac{1}{t}\sum_{i=1}^t \|\mu_i\|^2 \leqslant 16R^2\gamma^2\tau^2t^2$. We can bound similarly $\tau^2 \leqslant 4M^2\gamma^2R^4t^2$, so that $\|\bar{\mu}_t\|^2 \leqslant 64R^2M^2\gamma^4t^4$. Thus, for the second term:

$$\mathbb{E}\left[\left\|\Sigma^{1/2}\bar{\mu}_t\right\|^2 \mathbf{1}_{A_{\delta_a}^c}\right] \leqslant 64R^8M^2\gamma^4t^4\mathbb{E}\mathbf{1}_{A_{\delta_a}^c} \leqslant 64R^8M^2\gamma^4t^4\delta_a,$$

and for the third term:

$$\mathbb{E}\left[\left\|\Sigma^{1/2}\bar{\mu}_t\right\|^2 \mathbf{1}_{B_{\delta_b}^c}\right] \leqslant 64R^8M^2\gamma^4t^4\mathbb{E}\mathbf{1}_{B_{\delta_b}^c} \leqslant 64R^8M^2\gamma^4t^4\delta_b.$$

And on for the first term,

$$\begin{aligned}
\mathbb{E}\left[\left\|\Sigma^{1/2}\bar{\mu}_t\right\|^2 \mathbf{1}_{A_{\delta_a}\cap B_{\delta_b}}\right] &\leqslant \mathbb{E}\left[\left\|\Sigma^{1/2}(\Sigma+\lambda_n^\delta I)^{-1/2}\right\|^2 \left\|(\Sigma+\lambda_n^\delta I)^{1/2}(\hat{\Sigma}_n+\lambda_n^\delta I)^{-1/2}\right\|^2 \right. \\
&\qquad \left. \left\|(\hat{\Sigma}_n+\lambda_n^\delta I)^{1/2}\bar{\mu}_t\right\|^2 \mathbf{1}_{A_{\delta_a}\cap B_{\delta_b}} \mid \mathcal{G}\right] \\
&\leqslant 2\mathbb{E}\left[\left\|(\hat{\Sigma}_n+\lambda_n^\delta I)^{1/2}\bar{\mu}_t\right\|^2 \mid \mathcal{G}\right] \\
&= 2\mathbb{E}\left[\left\|\hat{\Sigma}_n^{1/2}\bar{\mu}_t\right\|^2 \mid \mathcal{G}\right] + 2\lambda_n^\delta\mathbb{E}\left[\|\bar{\mu}_t\|^2 \mid \mathcal{G}\right] \\
&\leqslant 16\tau_\infty^2 \frac{\gamma^{1/\alpha}\mathbb{E}\left[\operatorname{tr}\hat{\Sigma}_n^{1/\alpha}\right]}{t^{1-1/\alpha}} + 8\lambda_n^\delta\tau_\infty^2\,\gamma^{1/\alpha}\mathbb{E}\left[\operatorname{tr}\hat{\Sigma}_n^{1/\alpha}\right]t^{1/\alpha},
\end{aligned}$$

using Proposition 1 twice with $u=1$ for the left term and $u=1$ for the right one.

As $x \to x^{1/\alpha}$ is a concave function, we can apply Jensen's inequality to have :

$$\mathbb{E}\left[\operatorname{tr}(\hat{\Sigma}_n^{1/\alpha})\right] \leqslant \operatorname{tr}\Sigma^{1/\alpha},$$

so that:

$$\mathbb{E}\left[\left\|\Sigma^{1/2}\bar{\mu}_t\right\|^2 \mathbf{1}_{A_{\delta_a}\cap B_{\delta_b}}\right] \leqslant 16\tau_\infty^2 \frac{\gamma^{1/\alpha}\mathrm{tr}\,\Sigma^{1/\alpha}}{t^{1-1/\alpha}} + 8\lambda_n^\delta \tau_\infty^2 \gamma\, \gamma^{1/\alpha}\mathrm{tr}\,\Sigma^{1/\alpha} t^{1/\alpha}$$
$$\leqslant 16\tau_\infty^2 \gamma^{1/\alpha}\mathrm{tr}\,\Sigma^{1/\alpha} t^{1/\alpha}\left(\frac{1}{t} + \lambda_n^\delta\right).$$

Now, we take $\delta_a = \delta_b = \frac{\tau_\infty^2}{4M^2R^8\gamma^4 t^4}$ and this concludes the proof of Lemma 5, with the bound:

$$\mathbb{E}\left\|\Sigma^{1/2}\bar{\mu}_t\right\|^2 \leqslant 16\tau_\infty^2\gamma^{1/\alpha}\mathrm{tr}\,\Sigma^{1/\alpha}t^{1/\alpha}\left(\frac{1}{t} + \left(\frac{2+2\log M + 4\log(\gamma R^2) + 4\log t}{n}\right)^{1/\mu}\right).$$

□

D Convergence of batch gradient descent

In this section we prove the convergence of averaged batch gradient descent to the target function. In particular, since the proof technique is valid for the wider class of algorithms known as spectral filters [15, 14], we will do the proof for a generic spectral filter (in Lemma 9, Sect. D.1 we prove that averaged batch gradient descent is a spectral filter).

In Section D.1 we provide the required notation and additional definitions. In Section D.2, in particular in Theorem D.2 we perform an analytical decomposition of the excess risk of the averaged batch gradient descent, in terms of basic quantities that will be controlled in expectation (or probability) in the next sections. In Section D.3 the various quantites obtained by the analytical decomposition are controlled, in particular, Corollary 2 controls the L^∞ norm of the averaged batch gradient descent algorithm. Finally in Section D.4, the main result, Theorem 3 controlling in expectation of the excess risk of the averaged batch gradient descent estimator is provided. In Corollary 3, a version of the result of Theorem 3 is given, with explicit rates for the regularization parameters and of the excess risk.

D.1 Notations

In this subsection, we study the convergence of batch gradient descent. For the sake of clarity we consider the RKHS framework (which includes the finite-dimensional case). We will thus consider elements of $\mathcal{H}$ that are naturally embedded in $L_2(d\rho_\mathcal{X})$ by the operator S from $\mathcal{H}$ to $L_2(d\rho_\mathcal{X})$ and such that: $(Sg)(x) = \langle g, K_x\rangle$, where we have $\Phi(x) = K_x = K(\cdot, x)$ where $K : \mathcal{X}\to\mathcal{X}\to\mathbb{R}$ is the kernel. We recall the recursion for η_t in the case of an RKHS feature space with kernel K:

$$\eta_t = \eta_{t-1} + \frac{\gamma}{n}\sum_{i=1}^n \left(y_i - \langle \eta_{t-1}, K_{x_i}\rangle_\mathcal{H}\right)K_{x_i},$$

Let us begin with some notations. In the following we will often use the letter g to denote vectors of $\mathcal{H}$, hence, Sg will denote functions of $L_2(d\rho_\mathcal{X})$. We also define the following operators (we may also use their adjoints, denoted with a $*$):

- The operator $\hat{S}_n$ from $\mathcal{H}$ to $\mathbb{R}^n$, $\hat{S}_n g = \frac{1}{\sqrt{n}}(g(x_1), \dots g(x_n))$.
- The operators from $\mathcal{H}$ to $\mathcal{H}$, Σ and $\hat{\Sigma}_n$, defined respectively as $\Sigma = \mathbb{E}\left[K_x\otimes K_x\right] = \int_\mathcal{X} K_x\otimes K_x d\rho_\mathcal{X}$ and $\hat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^n K_{x_i}\otimes K_{x_i}$. Note that Σ is the covariance operator.
- The operator $\mathcal{L} : L_2(d\rho_\mathcal{X}) \to L_2(d\rho_\mathcal{X})$ is defined by

$$(\mathcal{L}f)(x) = \int_\mathcal{X} K(x,z)f(z)d\rho_\mathcal{X}(x), \quad \forall f\in L_2(d\rho_\mathcal{X}).$$

  Moreover denote by $\mathcal{N}(\lambda)$ the so called *effective dimension* of the learning problem, that is defined as

$$\mathcal{N}(\lambda) = \mathrm{tr}(\mathcal{L}(\mathcal{L}+\lambda I)^{-1}),$$

  for $\lambda > 0$. Recall that by Assumption (A4), there exists $\alpha \geqslant 1$ and $Q > 0$ such that

$$\mathcal{N}(\lambda) \leqslant Q\lambda^{-1/\alpha}, \quad \forall\lambda > 0.$$

  We can take $Q = \mathrm{tr}\Sigma^{1/\alpha}$.

- $P : L_2(d\rho_{\mathcal{X}}) \to L_2(d\rho_{\mathcal{X}})$ projection operator on $\mathcal{H}$ for the $L_2(d\rho_{\mathcal{X}})$ norm s.t. $\text{ran}P = \text{ran}S$.

Denote by f_ρ the function so that $f_\rho(x) = \mathbb{E}[y|x] \in L_2(d\rho_{\mathcal{X}})$ the minimizer of the expected risk, defined by $F(f) = \int_{X\times\mathbb{R}}(f(x)-y)^2 d\rho(x,y)$.

Remark 3 (On Assumption (A5))**.** *With the notation above, we express assumption (A5), more formally, w.r.t. Hilbert spaces with infinite dimensions, as follows. There exists $r \in [0,1]$ and $\phi \in L_2(d\rho_{\mathcal{X}})$, such that*

$$Pf_\rho = \mathcal{L}^r \phi.$$

(A6) *Let $q \in [1,\infty]$ be such that $\|f_\rho - Pf_\rho\|_{L^{2q}(\mathcal{X},\rho_{\mathcal{X}})} < \infty$.*

The assumption above is always true for $q = 1$, moreover when the kernel is universal it is true even for $q = \infty$. Moreover if $r \geqslant 1/2$ then it is true for $q = \infty$. Note that we make the calculation in this Appendix for a general $q \in [1,\infty]$, but we presented the results for $q = \infty$ in the main paper. The following proposition relates the excess risk to a certain norm.

Proposition 2. *When $\widehat{g} \in \mathcal{H}$,*

$$F(\widehat{g}) - \inf_{g\in\mathcal{H}} F(g) = \|S\widehat{g} - Pf_\rho\|^2_{L_2(d\rho_{\mathcal{X}})}.$$

We introduce the following function $g_\lambda \in \mathcal{H}$ that will be useful in the rest of the paper $g_\lambda = (\Sigma + \lambda I)^{-1} S^* f_\rho$.

We introduce the estimators of the form, for $\lambda > 0$,

$$\widehat{g}_\lambda = q_\lambda(\hat{\Sigma}_n)\hat{S}_n^*\hat{y},$$

where $q_\lambda : \mathbb{R}_+ \to \mathbb{R}_+$ is a function called *filter*, that essentially approximates x^{-1} with the approximation controlled by λ. Denote moreover with r_λ the function $r_\lambda(x) = 1 - xq_\lambda(x)$. The following definition precises the form of the filters we want to analyze. We then prove in Lemma 9 that our estimator corresponds to such a filter.

Definition 1 (Spectral filters)**.** *Let $q_\lambda : \mathbb{R}_+ \to \mathbb{R}_+$ be a function parametrized by $\lambda > 0$. q_λ is called a* filter *when there exists $c_q > 0$ for which*

$$\lambda q_\lambda(x) \leqslant c_q, \quad r_\lambda(x)x^u \leqslant c_q\lambda^u, \quad \forall x > 0, \lambda > 0, u \in [0,1].$$

We now justify that we study estimators of the form $\widehat{g}_\lambda = q_\lambda(\hat{\Sigma}_n)\hat{S}_n^*\hat{y}$ with the following lemma. Indeed, we show that the average of batch gradient descent can be represented as a filter estimator, $\widehat{g}_\lambda$, for $\lambda = 1/(\gamma t)$.

Lemma 9. *For $t > 1$, $\lambda = 1/(\gamma t)$, $\bar{\eta}_t = \widehat{g}_\lambda$, with respect to the filter, $q^\eta(x) = \left(1 - \frac{1-(1-\gamma x)^t}{\gamma t x}\right)\frac{1}{x}$.*

Proof. Indeed, for $t > 1$,

$$\begin{aligned}
\eta_t &= \eta_{t-1} + \frac{\gamma}{n}\sum_{i=1}^n \left(y_i - \langle \eta_{t-1}, K_{x_i}\rangle_{\mathcal{H}}\right)K_{x_i} \\
&= \eta_{t-1} + \gamma(\hat{S}_n^*\hat{y} - \hat{\Sigma}_n\eta_{t-1}) \\
&= (I - \gamma\hat{\Sigma}_n)\eta_{t-1} + \gamma\hat{S}_n^*\hat{y} \\
&= \gamma\sum_{k=0}^{t-1}(I-\gamma\hat{\Sigma}_n)^k\hat{S}_n^*\hat{y} = \left[I - (I-\gamma\hat{\Sigma}_n)^t\right]\hat{\Sigma}_n^{-1}\hat{S}_n^*\hat{y},
\end{aligned}$$

leading to

$$\bar{\eta}_t = \frac{1}{t}\sum_{i=0}^t \eta_i = q^\eta\left(\hat{\Sigma}_n\right)\hat{S}_n^*\hat{y}.$$

Now, we prove that q has the properties of a filter. First, for $t > 1$, $\frac{1}{\gamma t}q^\eta(x) = \left(1 - \frac{1-(1-\gamma x)^t}{\gamma t x}\right)\frac{1}{\gamma t x}$ is a decreasing function so that $\frac{1}{\gamma t}q^\eta(x) \leqslant \frac{1}{\gamma t}q^\eta(0) \leqslant 1$. Second for $u \in [0,1]$, $x^u(1 - xq^\eta(x)) = \frac{1-(1-\gamma x)^t}{\gamma t x}x^u$. As used in Section A.2, $1-(1-\gamma x)^t \leqslant (\gamma t x)^{1-u}$, so that, $r^\eta(x)x^u \leqslant \frac{(\gamma t x)^{1-u}}{\gamma t x}x^u = \frac{1}{(\gamma t)^u}$, this concludes the proof that q^η is indeed a filter. □

D.2 Analytical decomposition

Lemma 10. *Let $\lambda > 0$ and $s \in (0, 1/2]$. Under Assumption (A5) (see Rem. 3), the following holds*

$$\|\mathcal{L}^{-s} S(\widehat{g}_\lambda - g_\lambda)\|_{L_2(d\rho_\mathcal{X})} \leqslant 2\lambda^{-s}\beta^2 c_q \|\Sigma_\lambda^{-1/2}(\hat{S}_n^*\hat{y} - \hat{\Sigma}_n g_\lambda)\|_\mathcal{H} + 2\beta c_q \|\phi\|_{L_2(d\rho_\mathcal{X})}\lambda^{r-s},$$

where $\beta := \|\Sigma_\lambda^{1/2}\widehat{\Sigma}_{n\lambda}^{-1/2}\|$.

Proof. By Prop. 2, we can characterize the excess risk of $\widehat{g}_\lambda$ in terms of the $L_2(d\rho_\mathcal{X})$ squared norm of $S\widehat{g}_\lambda - Pf_\rho$. In this paper, simplifying the analysis of [14], we perform the following decomposition

$$\begin{aligned}\mathcal{L}^{-s} S(\widehat{g}_\lambda - g_\lambda) = \mathcal{L}^{-s} S\widehat{g}_\lambda - \mathcal{L}^{-s} S q_\lambda(\hat{\Sigma}_n)\hat{\Sigma}_n g_\lambda \\ + \ \mathcal{L}^{-s} S q_\lambda(\hat{\Sigma}_n)\hat{\Sigma}_n g_\lambda - \mathcal{L}^{-s} S g_\lambda.\end{aligned}$$

Upper bound for the first term. By using the definition of $\widehat{g}_\lambda$ and multiplying and dividing by $\Sigma_\lambda^{1/2}$, we have that

$$\begin{aligned}\mathcal{L}^{-s} S\widehat{g}_\lambda - \mathcal{L}^{-s} S q_\lambda(\hat{\Sigma}_n)\hat{\Sigma}_n g_\lambda &= \mathcal{L}^{-s} S q_\lambda(\hat{\Sigma}_n)(\hat{S}_n^*\hat{y} - \hat{\Sigma}_n g_\lambda) \\ &= \mathcal{L}^{-s} S q_\lambda(\hat{\Sigma}_n)\Sigma_\lambda^{1/2}\ \Sigma_\lambda^{-1/2}(\hat{S}_n^*\hat{y} - \hat{\Sigma}_n g_\lambda),\end{aligned}$$

from which

$$\|\mathcal{L}^{-s} S(\widehat{g}_\lambda - q_\lambda(\hat{\Sigma}_n)\hat{\Sigma}_n g_\lambda)\|_{L_2(d\rho_\mathcal{X})} \leqslant \|\mathcal{L}^{-s} S q_\lambda(\hat{\Sigma}_n)\Sigma_\lambda^{1/2}\|\ \|\Sigma_\lambda^{-1/2}(\hat{S}_n^*\hat{y} - \hat{\Sigma}_n g_\lambda)\|_\mathcal{H}.$$

Upper bound for the second term. By definition of $r_\lambda(x) = 1 - xq_\lambda(x)$ and $g_\lambda = \Sigma_\lambda^{-1} S^* f_\rho$,

$$\begin{aligned}\mathcal{L}^{-s} S q_\lambda(\hat{\Sigma}_n)\hat{\Sigma}_n g_\lambda - \mathcal{L}^{-s} S g_\lambda &= \mathcal{L}^{-s} S(q_\lambda(\hat{\Sigma}_n)\hat{\Sigma}_n - I)g_\lambda \\ &= -\mathcal{L}^{-s} S r_\lambda(\hat{\Sigma}_n)\ \Sigma_\lambda^{-(1/2-r)}\ \Sigma_\lambda^{-1/2-r} S^*\mathcal{L}^r\ \phi,\end{aligned}$$

where in the last step we used the fact that $S^* f_\rho = S^* P f_\rho = S^*\mathcal{L}^r\phi$, by Asm. (A5) (see Rem. 3). Then

$$\begin{aligned}\|\mathcal{L}^{-s} S(q_\lambda(\hat{\Sigma}_n)\hat{\Sigma}_n - I)g_\lambda)\|_{L_2(d\rho_\mathcal{X})} &\leqslant \|\mathcal{L}^{-s} S r_\lambda(\hat{\Sigma}_n)\|\|\Sigma_\lambda^{-(1/2-r)}\|\|\Sigma_\lambda^{-1/2-r} S^*\mathcal{L}^r\|\|\phi\|_{L_2(d\rho_\mathcal{X})} \\ &\leqslant \lambda^{-(1/2-r)}\|\mathcal{L}^{-s} S r_\lambda(\hat{\Sigma}_n)\|\|\phi\|_{L_2(d\rho_\mathcal{X})},\end{aligned}$$

where the last step is due to the fact that $\|\Sigma_\lambda^{-(1/2-r)}\| \leqslant \lambda^{-(1/2-r)}$ and that $S^*\mathcal{L}^{2r}S = S^*(SS^*)^{2r}S = (S^*S)^{2r}S^*S = \Sigma^{1+2r}$ from which

$$\|\Sigma_\lambda^{-1/2-r} S^*\mathcal{L}^r\|^2 = \|\Sigma_\lambda^{-1/2-r} S^*\mathcal{L}^{2r} S\Sigma_\lambda^{-1/2-r}\| = \|\Sigma_\lambda^{-1/2-r}\Sigma^{1+2r}\Sigma_\lambda^{-1/2-r}\| \leqslant 1. \tag{14}$$

Additional decompositions. We further bound $\|\mathcal{L}^{-s} S r_\lambda(\hat{\Sigma}_n)\|$ and $\|\mathcal{L}^{-s} S q_\lambda(\hat{\Sigma}_n)\Sigma_\lambda^{1/2}\|$. For the first, by the identity $\mathcal{L}^{-s} S r_\lambda(\hat{\Sigma}_n) = \mathcal{L}^{-s} S\widehat{\Sigma}_{n\lambda}^{-1/2}\widehat{\Sigma}_{n\lambda}^{1/2} r_\lambda(\hat{\Sigma}_n)$, we have

$$\|\mathcal{L}^{-s} S r_\lambda(\hat{\Sigma}_n)\| = \|\mathcal{L}^{-s} S\widehat{\Sigma}_{n\lambda}^{-1/2}\|\|\widehat{\Sigma}_{n\lambda}^{1/2} r_\lambda(\hat{\Sigma}_n)\|,$$

where

$$\|\widehat{\Sigma}_{n\lambda}^{1/2} r_\lambda(\hat{\Sigma}_n)\| = \sup_{\sigma\in\sigma(\hat{\Sigma}_n)} (\sigma+\lambda)^{1/2} r_\lambda(\sigma) \leqslant \sup_{\sigma\geqslant 0}(\sigma+\lambda)^{1/2} r_\lambda(\sigma) \leqslant 2c_q\lambda^{1/2}.$$

Similarly, by using the identity

$$\mathcal{L}^{-s} S q_\lambda(\hat{\Sigma}_n)\Sigma_\lambda^{1/2} = \mathcal{L}^{-s} S\widehat{\Sigma}_{n\lambda}^{-1/2}\ \widehat{\Sigma}_{n\lambda}^{1/2} q_\lambda(\hat{\Sigma}_n)\widehat{\Sigma}_{n\lambda}^{1/2}\ \widehat{\Sigma}_{n\lambda}^{-1/2}\Sigma_\lambda^{1/2},$$

we have

$$\|\mathcal{L}^{-s} S q_\lambda(\hat{\Sigma}_n)\Sigma_\lambda^{1/2}\| = \|\mathcal{L}^{-s} S\widehat{\Sigma}_{n\lambda}^{-1/2}\|\ \|\widehat{\Sigma}_{n\lambda}^{1/2} q_\lambda(\hat{\Sigma}_n)\widehat{\Sigma}_{n\lambda}^{1/2}\|\ \|\widehat{\Sigma}_{n\lambda}^{-1/2}\Sigma_\lambda^{1/2}\|.$$

Finally note that

$$\|\mathcal{L}^{-s} S\widehat{\Sigma}_{n\lambda}^{-1/2}\| \leqslant \|\mathcal{L}^{-s} S\Sigma_\lambda^{-1/2+s}\|\|\Sigma_\lambda^{-s}\|\|\Sigma_\lambda^{1/2}\widehat{\Sigma}_{n\lambda}^{-1/2}\|,$$

and $\|\mathcal{L}^{-s} S\Sigma_\lambda^{-1/2+s}\| \leqslant 1$, $\|\Sigma_\lambda^{-s}\| \leqslant \lambda^{-s}$, and moreover

$$\|\widehat{\Sigma}_{n\lambda}^{1/2} q_\lambda(\hat{\Sigma}_n)\widehat{\Sigma}_{n\lambda}^{1/2}\| = \sup_{\sigma\in\sigma(\hat{\Sigma}_n)} (\sigma+\lambda) q_\lambda(\sigma) \leqslant \sup_{\sigma\geqslant 0}(\sigma+\lambda) q_\lambda(\sigma) \leqslant 2c_q,$$

so, in conclusion

$$\|\mathcal{L}^{-s} S r_\lambda(\hat{\Sigma}_n)\| \leqslant 2c_q\lambda^{1/2-s}\beta, \quad \|\mathcal{L}^{-s} S q_\lambda(\hat{\Sigma}_n)\Sigma_\lambda^{1/2}\| \leqslant 2c_q\lambda^{-s}\beta^2.$$

The final result is obtained by gathering the upper bounds for the three terms above and the additional terms of this last section. □

Lemma 11. *Let $\lambda > 0$ and $s \in (0, \min(r, 1/2)]$. Under Assumption (A5) (see Rem. 3), the following holds*

$$\|\mathcal{L}^{-s}(S\widehat{g}_\lambda - Pf_\rho)\|_{L_2(d\rho_\mathcal{X})} \leqslant \lambda^{r-s}\|\phi\|_{L_2(d\rho_\mathcal{X})}.$$

Proof. Since $S\Sigma_\lambda^{-1}S^* = \mathcal{L}\mathcal{L}_\lambda^{-1} = I - \lambda\mathcal{L}_\lambda^{-1}$, we have

$$\begin{aligned}\mathcal{L}^{-s}(Sg_\lambda - Pf_\rho) &= \mathcal{L}^{-s}(S\Sigma_\lambda^{-1}S^*f_\rho - Pf_\rho) = \mathcal{L}^{-s}(S\Sigma_\lambda^{-1}S^*Pf_\rho - Pf_\rho)\\ &= \mathcal{L}^{-s}(S\Sigma_\lambda^{-1}S^* - I)Pf_\rho = \mathcal{L}^{-s}(S\Sigma_\lambda^{-1}S^* - I)\mathcal{L}^r\ \phi\\ &= -\lambda\mathcal{L}^{-s}\mathcal{L}_\lambda^{-1}L^r\phi = -\lambda^{r-s}\ \lambda^{1-r+s}\mathcal{L}_\lambda^{-(1-r+s)}\ \mathcal{L}_\lambda^{-(r-s)}\mathcal{L}^{r-s}\ \phi,\end{aligned}$$

from which

$$\begin{aligned}\|\mathcal{L}^{-s}(Sg_\lambda - Pf_\rho)\|_{L_2(d\rho_\mathcal{X})} &\leqslant \lambda^{r-s}\|\lambda^{1-r+s}\mathcal{L}_\lambda^{-(1-r+s)}\|\|\mathcal{L}_\lambda^{-(r-s)}\mathcal{L}^{r-s}\|\ \|\phi\|_{L_2(d\rho_\mathcal{X})}\\ &\leqslant \lambda^{r-s}\|\phi\|_{L_2(d\rho_\mathcal{X})}.\end{aligned}$$

□

Theorem 2. *Let $\lambda > 0$ and $s \in (0, \min(r, 1/2)]$. Under Assumption (A5) (see Rem. 3), the following holds*

$$\|\mathcal{L}^{-s}(S\widehat{g}_\lambda - Pf_\rho)\|_{L_2(d\rho_\mathcal{X})} \leqslant 2\lambda^{-s}\beta^2 c_q\|\Sigma_\lambda^{-1/2}(\hat{S}_n^*\hat{y} - \hat{\Sigma}_n g_\lambda)\|_\mathcal{H} + \left(1 + \beta 2c_q\|\phi\|_{L_2(d\rho_\mathcal{X})}\right)\lambda^{r-s}$$

where $\beta := \|\Sigma_\lambda^{1/2}\widehat{\Sigma}_{n\lambda}^{-1/2}\|$.

Proof. By Prop. 2, we can characterize the excess risk of $\widehat{g}_\lambda$ in terms of the $L_2(d\rho_\mathcal{X})$ squared norm of $S\widehat{g}_\lambda - Pf_\rho$. In this paper, simplifying the analysis of [14], we perform the following decomposition

$$\begin{aligned}\mathcal{L}^{-s}(S\widehat{g}_\lambda - Pf_\rho) = \mathcal{L}^{-s}S\widehat{g}_\lambda - \mathcal{L}^{-s}Sg_\lambda\\ +\ \mathcal{L}^{-s}(Sg_\lambda - Pf_\rho).\end{aligned}$$

The first term is bounded by Lemma 10, the second is bounded by Lemma 11. □

D.3 Probabilistic bounds

In this section denote by $\mathcal{N}_\infty(\lambda)$, the quantity

$$\mathcal{N}_\infty(\lambda) = \sup_{x\in S}\|\Sigma_\lambda^{-1/2}K_x\|_\mathcal{H}^2,$$

where $S \subseteq \mathcal{X}$ is the support of the probability measure $\rho_\mathcal{X}$.

Lemma 12. *Under Asm. (A3), we have that for any $g \in \mathcal{H}$*

$$\sup_{x\in supp(\rho_\mathcal{X})}|g(x)| \leqslant \kappa_\mu R^\mu\|\Sigma^{1/2(1-\mu)}g\|_\mathcal{H} = \kappa_\mu R^\mu\|\mathcal{L}^{-\mu/2}Sg\|_{L_2(d\rho_\mathcal{X})}.$$

Proof. Note that, Asm. (A3) is equivalent to

$$\|\Sigma^{-1/2(1-\mu)}K_x\| \leqslant \kappa_\mu R^\mu,$$

for all x in the support of $\rho_\mathcal{X}$. Then we have, for any x in the support of $\rho_\mathcal{X}$,

$$\begin{aligned}|g(x)| = \langle g, K_x\rangle_\mathcal{H} &= \left\langle \Sigma^{1/2(1-\mu)}g, \Sigma^{-1/2(1-\mu)}K_x\right\rangle_\mathcal{H}\\ &\leqslant \|\Sigma^{1/2(1-\mu)}g\|_\mathcal{H}\|\Sigma^{-1/2(1-\mu)}K_x\| \leqslant \kappa_\mu R^\mu\|\Sigma^{1/2(1-\mu)}g\|_\mathcal{H}.\end{aligned}$$

Now note that, since $\Sigma^{1-\mu} = S^*\mathcal{L}^{-\mu}S$, we have

$$\|\Sigma^{1/2(1-\mu)}g\|_\mathcal{H}^2 = \left\langle g, \Sigma^{1-\mu}g\right\rangle_\mathcal{H} = \left\langle \mathcal{L}^{-\mu/2}Sg, \mathcal{L}^{-\mu/2}Sg\right\rangle_{L_2(d\rho_\mathcal{X})}.$$

□

Lemma 13. *Under Assumption (A3), we have*

$$\mathcal{N}_\infty(\lambda) \leqslant \kappa_\mu^2 R^{2\mu}\lambda^{-\mu}.$$

Proof. First denote with $f_{\lambda,u} \in \mathcal{H}$ the function $\Sigma_\lambda^{-1/2} u$ for any $u \in \mathcal{H}$ and $\lambda > 0$. Note that

$$\|f_{\lambda,u}\|_{\mathcal{H}} = \|\Sigma_\lambda^{-1/2} u\|_{\mathcal{H}} \leqslant \|\Sigma_\lambda^{-1/2}\| \|u\|_{\mathcal{H}} \leqslant \lambda^{-1/2} \|u\|_{\mathcal{H}}.$$

Moreover, since for any $g \in \mathcal{H}$ the identity $\|g\|_{L_2(d\rho_{\mathcal{X}})} = \|Sg\|_{\mathcal{H}}$, we have

$$\|f_{\lambda,u}\|_{L_2(d\rho_{\mathcal{X}})} = \|S\Sigma_\lambda^{-1/2} u\|_{\mathcal{H}} \leqslant \|S\Sigma_\lambda^{-1/2}\| \|u\|_{\mathcal{H}} \leqslant \|u\|_{\mathcal{H}}.$$

Now denote with $B(\mathcal{H})$ the unit ball in $\mathcal{H}$, by applying Asm. (A3) to $f_{\lambda,u}$ we have that

$$\begin{aligned}
\mathcal{N}_\infty(\lambda) &= \sup_{x \in S} \|\Sigma_\lambda^{-1/2} K_x\|^2 = \sup_{x \in S, u \in B(\mathcal{H})} \left\langle u, \Sigma_\lambda^{-1/2} K_x \right\rangle_{\mathcal{H}}^2 \\
&= \sup_{x \in S, u \in B(\mathcal{H})} \langle f_{\lambda,u}, K_x \rangle_{\mathcal{H}}^2 = \sup_{u \in B(\mathcal{H})} \sup_{x \in S} |f_{\lambda,u}(x)|^2 \\
&\leqslant \kappa_\mu^2 R^{2\mu} \sup_{u \in B(\mathcal{H})} \|f_{\lambda,u}\|_{\mathcal{H}}^{2\mu} \|f_{\lambda,u}\|_{L_2(d\rho_{\mathcal{X}})}^{2-2\mu} \\
&\leqslant \kappa_\mu^2 R^{2\mu} \lambda^{-\mu} \sup_{u \in B(\mathcal{H})} \|u\|_{\mathcal{H}}^2 \leqslant \kappa_\mu^2 R^{2\mu} \lambda^{-\mu}.
\end{aligned}$$

□

Lemma 14. *Let $\lambda > 0$, $\delta \in (0, 1]$ and $n \in \mathbb{N}$. Under Assumption (A3), we have that, when*

$$n \geqslant 11(1 + \kappa_\mu^2 R^{2\mu} \lambda^{-\mu}) \log \frac{8R^2}{\lambda\delta},$$

then the following holds with probability $1 - \delta$,

$$\|\Sigma_\lambda^{1/2} \widehat{\Sigma}_{n\lambda}^{-1/2}\|^2 \leqslant 2.$$

Proof. This result is a refinement of the one in [39] and is based on non-commutative Bernstein inequalities for random matrices [40]. By Prop. 8 in [21], we have that

$$\|\Sigma_\lambda^{1/2} \widehat{\Sigma}_{n\lambda}^{-1/2}\|^2 \leqslant (1 - t)^{-1}, \quad t := \|\Sigma_\lambda^{-1/2} (\Sigma - \hat{\Sigma}_n) \Sigma_\lambda^{-1/2}\|.$$

When $0 < \lambda \leqslant \|\Sigma\|$, by Prop. 6 of [21] (see also [41] Lemma 9 for more refined constants), we have that the following holds with probability at least $1 - \delta$,

$$t \leqslant \frac{2\eta(1 + \mathcal{N}_\infty(\lambda))}{3n} + \sqrt{\frac{2\eta \mathcal{N}_\infty(\lambda)}{n}},$$

with $\eta = \log \frac{8R^2}{\lambda\delta}$. Finally, by selecting $n \geqslant 11(1 + \kappa_\mu^2 R^{2\mu} \lambda^{-\mu})\eta$, we have that $t \leqslant 1/2$ and so $\|\Sigma_\lambda^{1/2} \widehat{\Sigma}_{n\lambda}^{-1/2}\|^2 \leqslant (1 - t)^{-1} \leqslant 2$, with probability $1 - \delta$.

To conclude note that when $\lambda \geqslant \|\Sigma\|$, we have

$$\|\Sigma_\lambda^{1/2} \widehat{\Sigma}_{n\lambda}^{-1/2}\|^2 \leqslant \|\Sigma + \lambda I\| \|(\hat{\Sigma}_n + \lambda I)^{-1}\| \leqslant \frac{\|\Sigma\| + \lambda}{\lambda} = 1 + \frac{\|\Sigma\|}{\lambda} \leqslant 2.$$

□

Lemma 15. *Under Assumption (A3), (A4), (A5) (see Rem. 3), (A6) we have*

1. *Let $\lambda > 0$, $n \in \mathbb{N}$, the following holds*

$$\mathbb{E}[\|\Sigma_\lambda^{-1/2} (\hat{S}_n^* \hat{y} - \hat{\Sigma}_n g_\lambda)\|_{\mathcal{H}}^2] \leqslant \|\phi\|_{L_2(d\rho_{\mathcal{X}})}^2 \lambda^{2r} + \frac{2\kappa_\mu^2 R^{2\mu} \lambda^{-(\mu-2r)}}{n} + \frac{4\kappa_\mu^2 R^{2\mu} A Q \lambda^{-\frac{q+\mu\alpha}{q\alpha+\alpha}}}{n},$$

where $A := \|f_\rho - P f_\rho\|_{L^{2q}(\mathcal{X}, \rho_{\mathcal{X}})}^{2-2/(q+1)}$.

2. *Let $\delta \in (0,1]$, under the same assumptions, the following holds with probability at least $1-\delta$*

$$\|\Sigma_\lambda^{-1/2}(\hat{S}_n^*\hat{y} - \hat{\Sigma}_n g_\lambda)\|_{\mathcal{H}} \leqslant c_0\lambda^r + \frac{4(c_1\lambda^{-\frac{\mu}{2}} + c_2\lambda^{-r-\mu})\log\frac{2}{\delta}}{n} + \sqrt{\frac{16\kappa_\mu^2 R^{2\mu}(\lambda^{-(\mu-2r)} + 2AQ\lambda^{-\frac{q+\mu\alpha}{q\alpha+\alpha}})\log\frac{2}{\delta}}{n}},$$

with $c_0 = \|\phi\|_{L_2(d\rho_{\mathcal{X}})}$, $c_1 = \kappa_\mu R^\mu M + \kappa_\mu^2 R^{2\mu}(2R)^{2r-\mu}\|\phi\|_{L_2(d\rho_{\mathcal{X}})}$, $c_2 = \kappa_\mu^2 R^{2\mu}\|\phi\|_{L_2(d\rho_{\mathcal{X}})}$

Proof. First denote with ζ_i the random variable

$$\zeta_i = (y_i - g_\lambda(x_i))\Sigma_\lambda^{-1/2}K_{x_i}.$$

In particular note that, by using the definitions of $\hat{S}_n$, $\hat{y}$ and $\hat{\Sigma}_n$, we have

$$\Sigma_\lambda^{-1/2}(\hat{S}_n^*\hat{y} - \hat{\Sigma}_n g_\lambda) = \Sigma_\lambda^{-1/2}(\frac{1}{n}\sum_{i=1}^n K_{x_i}y_i - \frac{1}{n}(K_{x_i}\otimes K_{x_i})g_\lambda) = \frac{1}{n}\sum_{i=1}^n \zeta_i.$$

I So, by noting that ζ_i are independent and identically distributed, we have

$$\begin{aligned}\mathbb{E}[\|\Sigma_\lambda^{-1/2}(\hat{S}_n^*\hat{y} - \hat{\Sigma}_n g_\lambda)\|_{\mathcal{H}}^2] &= \mathbb{E}[\|\frac{1}{n}\sum_{i=1}^n \zeta_i\|_{\mathcal{H}}^2] = \frac{1}{n^2}\sum_{i,j=1}^n \mathbb{E}[\langle\zeta_i,\zeta_j\rangle_{\mathcal{H}}] \\ &= \frac{1}{n}\mathbb{E}[\|\zeta_1\|_{\mathcal{H}}^2] + \frac{n-1}{n}\|\mathbb{E}[\zeta_1]\|_{\mathcal{H}}^2.\end{aligned}$$

Now note that

$$\mathbb{E}[\zeta_1] = \Sigma_\lambda^{-1/2}(\mathbb{E}[K_{x_1}y_1] - \mathbb{E}[K_{x_1}\otimes K_{x_1}]g_\lambda) = \Sigma_\lambda^{-1/2}(S^*f_\rho - \Sigma g_\lambda).$$

In particular, by the fact that $S^*f_\rho = Pf_\rho$, $Pf_\rho = \mathcal{L}^r\phi$ and $\Sigma g_\lambda = \Sigma\Sigma_\lambda^{-1}S^*f_\rho$ and $\Sigma\Sigma_\lambda^{-1} = I - \lambda\Sigma_\lambda^{-1}$, we have

$$\Sigma_\lambda^{-1/2}(S^*f_\rho - \Sigma g_\lambda) = \lambda\Sigma_\lambda^{-3/2}S^*f_\rho = \lambda^r\ \lambda^{1-r}\Sigma_\lambda^{-(1-r)}\ \Sigma_\lambda^{-1/2-r}S^*\mathcal{L}^r\ \phi.$$

So, since $\|\Sigma_\lambda^{-1/2-r}S^*\mathcal{L}^r\| \leqslant 1$, as proven in Eq. 14, then

$$\|\mathbb{E}[\zeta_1]\|_{\mathcal{H}} \leqslant \lambda^r\|\lambda^{1-r}\Sigma_\lambda^{-(1-r)}\|\ \|\Sigma_\lambda^{-1/2-r}S^*\mathcal{L}^r\|\ \|\phi\|_{L_2(d\rho_{\mathcal{X}})} \leqslant \lambda^r\|\phi\|_{L_2(d\rho_{\mathcal{X}})} := Z.$$

Morever

$$\begin{aligned}\mathbb{E}[\|\zeta_1\|_{\mathcal{H}}^2] &= \mathbb{E}[\|\Sigma_\lambda^{-1/2}K_{x_1}\|_{\mathcal{H}}^2(y_1 - g_\lambda(x_1))^2] = \mathbb{E}_{x_1}\mathbb{E}_{y_1|x_1}[\|\Sigma_\lambda^{-1/2}K_{x_1}\|_{\mathcal{H}}^2(y_1 - g_\lambda(x_1))^2] \\ &= \mathbb{E}_{x_1}[\|\Sigma_\lambda^{-1/2}K_{x_1}\|_{\mathcal{H}}^2(f_\rho(x_1) - g_\lambda(x_1))^2].\end{aligned}$$

Moreover we have

$$\begin{aligned}\mathbb{E}[\|\zeta_1\|_{\mathcal{H}}^2] &= \mathbb{E}_x[\|\Sigma_\lambda^{-1/2}K_x\|_{\mathcal{H}}^2(f_\rho(x) - g_\lambda(x))^2] \\ &= \mathbb{E}_x[\|\Sigma_\lambda^{-1/2}K_x\|_{\mathcal{H}}^2((f_\rho(x) - (Pf_\rho)(x)) + ((Pf_\rho)(x) - g_\lambda(x)))^2] \\ &\leqslant 2\mathbb{E}_x[\|\Sigma_\lambda^{-1/2}K_x\|_{\mathcal{H}}^2(f_\rho(x) - (Pf_\rho)(x))^2] + 2\mathbb{E}_x[\|\Sigma_\lambda^{-1/2}K_x\|_{\mathcal{H}}^2((Pf_\rho)(x) - g_\lambda(x))^2].\end{aligned}$$

Now since $\mathbb{E}[AB] \leqslant (\operatorname{ess\,sup} A)\mathbb{E}[B]$, for any two random variables A, B, we have

$$\begin{aligned}\mathbb{E}_x[\|\Sigma_\lambda^{-1/2}K_x\|_{\mathcal{H}}^2((Pf_\rho)(x) - g_\lambda(x))^2] &\leqslant \mathcal{N}_\infty(\lambda)\mathbb{E}_x[((Pf_\rho)(x) - g_\lambda(x))^2] \\ &= \mathcal{N}_\infty(\lambda)\|Pf_\rho - Sg_\lambda\|_{L_2(d\rho_{\mathcal{X}})}^2 \\ &\leqslant \kappa_\mu^2 R^{2\mu}\lambda^{-(\mu-2r)},\end{aligned}$$

where in the last step we bounded $\mathcal{N}_\infty(\lambda)$ via Lemma 13 and $\|Pf_\rho - Sg_\lambda\|_{L_2(d\rho_{\mathcal{X}})}^2$, via Lemma. 11 applied with $s = 0$. Finally, denoting by $a(x) = \|\Sigma_\lambda^{-1/2}K_x\|_{\mathcal{H}}^2$ and $b(x) = (f_\rho(x) - (Pf_\rho)(x))^2$

and noting that by Markov inequality we have $\mathbb{E}_x[\mathbf{1}_{\{b(x)>t\}}] = \rho_{\mathcal{X}}(\{b(x) > t\}) = \rho_{\mathcal{X}}(\{b(x)^q > t^q\}) \leqslant \mathbb{E}_x[b(x)^q]t^{-q}$, for any $t > 0$. Then for any $t > 0$ the following holds

$$\begin{aligned}\mathbb{E}_x[a(x)b(x)] &= \mathbb{E}_x[a(x)b(x)\mathbf{1}_{\{b(x)\leqslant t\}}] + \mathbb{E}_x[a(x)b(x)\mathbf{1}_{\{b(x)>t\}}] \\ &\leqslant t\mathbb{E}_x[a(x)] + N_\infty(\lambda)\mathbb{E}_x[b(x)\mathbf{1}_{\{b(x)>t\}}] \\ &\leqslant tN(\lambda) + N_\infty(\lambda)\mathbb{E}_x[b(x)^q]t^{-q}.\end{aligned}$$

By minimizing the quantity above in t, we obtain

$$\begin{aligned}\mathbb{E}_x[\|\Sigma_\lambda^{-1/2}K_x\|_{\mathcal{H}}^2(f_\rho(x) - (Pf_\rho)(x))^2] &\leqslant 2\|f_\rho - Pf_\rho\|_{L^q(X,\rho_{\mathcal{X}})}^{\frac{q}{q+1}}\mathcal{N}(\lambda)^{\frac{q}{q+1}}\mathcal{N}_\infty(\lambda)^{\frac{1}{q+1}} \\ &\leqslant 2\kappa_\mu^2 R^{2\mu}AQ\lambda^{-\frac{q+\mu\alpha}{q\alpha+\alpha}}.\end{aligned}$$

So finally

$$\mathbb{E}[\|\zeta_1\|_{\mathcal{H}}^2] \leqslant 2\kappa_\mu^2 R^{2\mu}\lambda^{-(\mu-2r)} + 4\kappa_\mu^2 R^{2\mu}AQ\lambda^{-\frac{q+\mu\alpha}{q\alpha+\alpha}} := W^2.$$

To conclude the proof, let us obtain the bound in high probability. We need to bound the higher moments of ζ_1. First note that

$$\mathbb{E}[\|\zeta_1 - \mathbb{E}[\zeta_1]\|_{\mathcal{H}}^p] \leqslant \mathbb{E}[\|\zeta_1 - \zeta_2\|_{\mathcal{H}}^p] \leqslant 2^{p-1}\mathbb{E}[\|\zeta_1|_{\mathcal{H}}^p + \|\zeta_2\|_{\mathcal{H}}^p] \leqslant 2^p\mathbb{E}[\|\zeta_1|_{\mathcal{H}}^p].$$

Moreover, denoting by $S \subseteq \mathcal{X}$ the support of $\rho_{\mathcal{X}}$ and recalling that y is bounded in $[-M, M]$, the following bound holds almost surely

$$\begin{aligned}\|\zeta_1\| &\leqslant \sup_{x\in S}\|\Sigma_\lambda^{-1/2}K_x\|(M + |g_\lambda(x)|) \leqslant (\sup_{x\in S}\|\Sigma_\lambda^{-1/2}K_x\|)(M + \sup_{x\in S}|g_\lambda(x)|) \\ &\leqslant \kappa_\mu R^\mu\lambda^{-\mu/2}(M + \kappa_\mu R^\mu\|\Sigma^{1/2(1-\mu)}g_\lambda\|_{\mathcal{H}}).\end{aligned}$$

where in the last step we applied Lemma 13 and Lemma 12. In particular, by definition of g_λ, the fact that $S^*f_\rho = S^*Pf_\rho$, that $Pf_\rho = \mathcal{L}^r\phi$ and that $\|\Sigma_\lambda^{-(1/2+r)}S^*\mathcal{L}^r\| \leqslant 1$ as proven in Eq. 14, we have

$$\begin{aligned}\|\Sigma^{1/2(1-\mu)}g_\lambda\|_{\mathcal{H}} &= \|\Sigma^{1/2(1-\mu)}\Sigma_\lambda^{-1}S^*\mathcal{L}^r\phi\|_{\mathcal{H}} \\ &\leqslant \|\Sigma^{1/2(1-\mu)}\Sigma^{-1/2(1-\mu)}\|\|\Sigma_\lambda^{-(\mu/2-r)}\|\|\Sigma_\lambda^{-(1/2+r)}S^*\mathcal{L}^r\|\|\phi\|_{L_2(d\rho_{\mathcal{X}})} \\ &\leqslant \|\Sigma_\lambda^{r-\mu/2}\|\|\phi\|_{L_2(d\rho_{\mathcal{X}})}.\end{aligned}$$

Finally note that if $r \leqslant \mu/2$ then $\|\Sigma_\lambda^{r-\mu/2}\| \leqslant \lambda^{-(\mu/2-r)}$, if $r \geqslant \mu/2$ then

$$\|\Sigma_\lambda^{r-\mu/2}\| = (\|C\| + \lambda)^{r-\mu/2} \leqslant (2\|C\|)^{r-\mu/2} \leqslant (2R)^{2r-\mu}.$$

So in particular

$$\|\Sigma_\lambda^{r-\mu/2}\| \leqslant (2R)^{2r-\mu} + \lambda^{-(\mu/2-r)}.$$

Then the following holds almost surely

$$\|\zeta_1\| \leqslant (\kappa_\mu R^\mu M + \kappa_\mu^2 R^{2\mu}(2R)^{2r-\mu}\|\phi\|_{L_2(d\rho_{\mathcal{X}})})\lambda^{-\mu/2} + \kappa_\mu^2 R^{2\mu}\|\phi\|_{L_2(d\rho_{\mathcal{X}})}\lambda^{r-\mu} := V.$$

So finally

$$\mathbb{E}[\|\zeta_1 - \mathbb{E}[\zeta_1]\|_{\mathcal{H}}^p] \leqslant 2^p\mathbb{E}[\|\zeta_1\|_{\mathcal{H}}^p] \leqslant \frac{p!}{2}(2V)^{p-2}(4W^2).$$

By applying Pinelis inequality, the following holds with probability $1 - \delta$

$$\|\frac{1}{n}\sum_{i=1}^n(\zeta_i - \mathbb{E}[\zeta_i])\|_{\mathcal{H}} \leqslant \frac{4V\log\frac{2}{\delta}}{n} + \sqrt{\frac{8W\log\frac{2}{\delta}}{n}}.$$

So with the same probability

$$\|\frac{1}{n}\sum_{i=1}^n\zeta_i\|_{\mathcal{H}} \leqslant \|\frac{1}{n}\sum_{i=1}^n(\zeta_i - \mathbb{E}[\zeta_i])\|_{\mathcal{H}} + \|\mathbb{E}[\zeta_1]\|_{\mathcal{H}} \leqslant Z + \frac{4V\log\frac{2}{\delta}}{n} + \sqrt{\frac{8W\log\frac{2}{\delta}}{n}}.$$

□

Lemma 16. *Let $\lambda > 0$, $n \in \mathbb{N}$ and $s \in (0, 1/2]$. Let $\delta \in (0, 1]$. Under Assumption (A3), (A4), (A5) (see Rem. 3), (A6), when*

$$n \geqslant 11(1 + \kappa_\mu^2 R^{2\mu} \lambda^{-\mu}) \log \frac{16R^2}{\lambda\delta},$$

then the following holds with probability $1 - \delta$,

$$\|\mathcal{L}^{-s} S(\widehat{g}_\lambda - g_\lambda)\|_{L_2(d\rho_\mathcal{X})} \leq c_0 \lambda^{r-s} + \frac{(c_1 \lambda^{-\frac{\mu}{2}-s} + c_2 \lambda^{r-\mu-s}) \log \frac{4}{\delta}}{n} + \sqrt{\frac{(c_3 \lambda^{-(\mu+2s-2r)} + c_4 \lambda^{-\frac{q+\mu\alpha}{q\alpha+\alpha}-2s}) \log \frac{4}{\delta}}{n}}.$$

with $c_0 = 7c_q\|\phi\|_{L_2(d\rho_\mathcal{X})}$, $c_1 = 16c_q(\kappa_\mu R^\mu M + \kappa_\mu^2 R^{2\mu}(2R)^{2r-\mu}\|\phi\|_{L_2(d\rho_\mathcal{X})})$, $c_2 = 16c_q\kappa_\mu^2 R^{2\mu}\|\phi\|_{L_2(d\rho_\mathcal{X})}$, $c_3 = 64\kappa_\mu^2 R^{2\mu} c_q^2$, $c_4 = 128\kappa_\mu^2 R^{2\mu} A Q c_q^2$.

Proof. Let $\tau = \delta/2$, the result is obtained by combining Lemma 10, with Lemma 15 with probability τ, and Lemma 14, with probability τ and then taking the intersection bound of the two events. □

Corollary 1. *Let $\lambda > 0$, $n \in \mathbb{N}$ and $s \in (0, 1/2]$. Let $\delta \in (0, 1]$. Under the assumptions of Lemma 16, when*

$$n \geqslant 11(1 + \kappa_\mu^2 R^{2\mu} \lambda^{-\mu}) \log \frac{16R^2}{\lambda\delta},$$

then the following holds with probability $1 - \delta$,

$$\|\mathcal{L}^{-s} S\widehat{g}_\lambda\|_{L_2(d\rho_\mathcal{X})} \leq R^{2r-2s} + (1 + c_0)\lambda^{r-s} + \frac{(c_1 \lambda^{-\frac{\mu}{2}-s} + c_2 \lambda^{r-\mu-s}) \log \frac{4}{\delta}}{n} + \sqrt{\frac{(c_3 \lambda^{-(\mu+2s-2r)} + c_4 \lambda^{-\frac{q+\mu\alpha}{q\alpha+\alpha}-2s}) \log \frac{4}{\delta}}{n}} +$$

with the same constants $c_0, \ldots, c_4$ as in Lemma 16.

Proof. First note that

$$\|\mathcal{L}^{-s} S\widehat{g}_\lambda\|_{L_2(d\rho_\mathcal{X})} \leqslant \|\mathcal{L}^{-s} S(\widehat{g}_\lambda - g_\lambda)\|_{L_2(d\rho_\mathcal{X})} + \|\mathcal{L}^{-s} S g_\lambda\|_{L_2(d\rho_\mathcal{X})}.$$

The first term on the right hand side is controlled by Lemma 16, for the second, by using the definition of g_λ and Asm. (A5) (see Rem. 3), we have

$$\begin{aligned}\|\mathcal{L}^{-s} S g_\lambda\|_{L_2(d\rho_\mathcal{X})} &\leqslant \|\mathcal{L}^{-s} S \Sigma_\lambda^{-1/2+s}\| \|\Sigma_\lambda^{-(s-r)}\| \|\Sigma_\lambda^{-1/2-r} S^* \mathcal{L}^r\| \|\phi\|_{L_2(d\rho_\mathcal{X})} \\ &\leqslant \|\Sigma_\lambda^{r-s}\| \|\phi\|_{L_2(d\rho_\mathcal{X})},\end{aligned}$$

where $\|\Sigma_\lambda^{-1/2-r} S^* \mathcal{L}^r\| \leqslant 1$ by Eq. 14 and analogously $\|\mathcal{L}^{-s} S \Sigma_\lambda^{-1/2+s}\| \leqslant 1$. Note that if $s \geqslant r$ then $\|\Sigma_\lambda^{r-s}\| \leqslant \lambda^{-(s-r)}$. If $s < r$, we have

$$\|\Sigma_\lambda^{r-s}\| = (\|\Sigma\| + \lambda)^{r-s} \leqslant \|C\|^{r-s} + \lambda^{r-s} \leqslant R^{2r-2s} + \lambda^{r-s}.$$

So finally $\|\Sigma_\lambda^{r-s}\| \leqslant R^{2r-2s} + \lambda^{r-s}$. □

Corollary 2. *Let $\lambda > 0$, $n \in \mathbb{N}$ and $s \in (0, 1/2]$. Let $\delta \in (0, 1]$. Under Assumption (A3), (A4), (A5) (see Rem. 3), (A6), when*

$$n \geqslant 11(1 + \kappa_\mu^2 R^{2\mu} \lambda^{-\mu}) \log \frac{16R^2}{\lambda\delta},$$

then the following holds with probability $1 - \delta$,

$$\sup_{x \in \mathcal{X}} |\widehat{g}_\lambda(x)| \leq \kappa_\mu R^\mu R^{2r-2s} + \kappa_\mu R^\mu (1 + c_0)\lambda^{r-\mu/2} + \kappa_\mu R^\mu \frac{(c_1 \lambda^{-\mu} + c_2 \lambda^{r-3/2\mu}) \log \frac{4}{\delta}}{n} + \kappa_\mu R^\mu \sqrt{\frac{(c_3 \lambda^{-(2\mu-2r)} + \kappa_\mu R^\mu c_4 \lambda^{-\frac{q+\mu\alpha}{q\alpha+\alpha}-\mu}) \log \frac{4}{\delta}}{n}}.$$

with the same constants $c_0, \ldots, c_4$ in Lemma 16.

Proof. The proof is obtained by applying Lemma 12 on $\widehat{g}_\lambda$ and then Corollary 1. □

D.4 Main Result

Theorem 3. *Let $\lambda > 0$, $n \in \mathbb{N}$ and $s \in (0, \min(r, 1/2)]$. Under Assumption (A3), (A4), (A5) (see Rem. 3), (A6), when*

$$n \geqslant 11(1 + \kappa_\mu^2 R^{2\mu}\lambda^{-\mu}) \log \frac{c_0}{\lambda^{3+4r-4s}},$$

then

$$\mathbb{E}[\|\mathcal{L}^{-s}(S\widehat{g}_\lambda - Pf_\rho)\|^2_{L_2(d\rho_\mathcal{X})}] \leqslant c_1 \frac{\lambda^{-(\mu+2s-2r)}}{n} + c_2 \frac{\lambda^{-\frac{q+\mu\alpha}{q\alpha+\alpha}-2s}}{n} + c_3\lambda^{2r-2s},$$

where $m_4 = M^4$, $c_0 = 32R^{4-4s}m_4 + 32R^{8-8r-8s}\|\phi\|^4_{L_2(d\rho_\mathcal{X})}$, $c_1 = 16c_q^2\kappa_\mu^2 R^{2\mu}$, $c_2 = 32c_q^2\kappa_\mu^2 R^{2\mu}AQ$, $c_3 = 3 + 8c_q^2\|\phi\|^2_{L_2(d\rho_\mathcal{X})}$.

Proof. Denote by $R(\widehat{g}_\lambda)$, the expected risk $R(\widehat{g}_\lambda) = \mathcal{E}(\widehat{g}_\lambda) - \inf_{g\in\mathcal{H}} \mathcal{E}(g)$. First, note that by Prop. 2, we have

$$R_s(\widehat{g}_\lambda) = \|\mathcal{L}^{-s}(S\widehat{g}_\lambda - Pf_\rho)\|^2_{L_2(d\rho_\mathcal{X})}.$$

Denote by E the event such that β as defined in Thm. 2, satisfies $\beta \leqslant 2$. Then we have

$$\mathbb{E}[R_s(\widehat{g}_\lambda)] = \mathbb{E}[R_s(\widehat{g}_\lambda)\mathbf{1}_E] + \mathbb{E}[R(\widehat{g}_\lambda)\mathbf{1}_{E^c}].$$

For the first term, by Thm. 2 and Lemma 15, we have

$$\begin{aligned}
\mathbb{E}[R_s(\widehat{g}_\lambda)\mathbf{1}_E] &\leqslant \mathbb{E}[\Big(2\lambda^{-2s}\beta^4 c_q^2\|\Sigma_\lambda^{-1/2}(\hat{S}_n^*\hat{y} - \hat{\Sigma}_n g_\lambda)\|^2_\mathcal{H} \\
&\qquad + 2\left(1 + \beta^2 2c_q^2\|\phi\|^2_{L_2(d\rho_\mathcal{X})}\right)\lambda^{2r-2s}\Big)\mathbf{1}_E] \\
&\leqslant 8\lambda^{-2s}c_q^2\mathbb{E}[\|\Sigma_\lambda^{-1/2}(\hat{S}_n^*\hat{y} - \hat{\Sigma}_n g_\lambda)\|^2_\mathcal{H}] + 2\left(1 + 4c_q^2\|\phi\|^2_{L_2(d\rho_\mathcal{X})}\right)\lambda^{2r-2s} \\
&\leqslant \frac{16c_q^2\kappa_\mu^2 R^{2\mu}\lambda^{-\mu+2r-2s}}{n} + \frac{32c_q^2\kappa_\mu^2 R^{2\mu}AQ\lambda^{-\frac{q+\mu\alpha}{q\alpha+\alpha}-2s}}{n} + \left(2 + 8c_q^2\|\phi\|^2_{L_2(d\rho_\mathcal{X})}\right)\lambda^{2r-2s}.
\end{aligned}$$

For the second term, since $\widehat{\Sigma}_{n\lambda}^{1/2} q_\lambda(\hat{\Sigma}_n)\widehat{\Sigma}_{n\lambda}^{1/2} = \hat{\Sigma}_{n\lambda} q_\lambda(\hat{\Sigma}_n) \leqslant \sup_{\sigma>0}(\sigma+\lambda)q_\lambda(\sigma) \leqslant c_q$ by definition of filters, and that $Pf_\rho = L^r\phi$, we have

$$\begin{aligned}
R_s(\widehat{g}_\lambda)^{1/2} &\leqslant \|\mathcal{L}^{-s}S\widehat{g}_\lambda\|_{L_2(d\rho_\mathcal{X})} + \|\mathcal{L}^{-s}Pf_\rho\|_{L_2(d\rho_\mathcal{X})} \\
&\leqslant \|\mathcal{L}^{-s}S\|\|\widehat{\Sigma}_{n\lambda}^{-1/2}\|\|\widehat{\Sigma}_{n\lambda}^{1/2}q_\lambda(\hat{\Sigma}_n)\widehat{\Sigma}_{n\lambda}^{1/2}\|\|\widehat{\Sigma}_{n\lambda}^{-1/2}\hat{S}_n^*\|\|\hat{y}\| + \|\mathcal{L}^{-s}\mathcal{L}^r\|\|\phi\|_{L_2(d\rho_\mathcal{X})} \\
&\leqslant R^{1/2-s}\lambda^{-1/2}\|\hat{y}\| + R^{2r-2s}\|\phi\|_{L_2(d\rho_\mathcal{X})} \\
&\leqslant \lambda^{-1/2}(R^{1/2-s}(n^{-1}\sum_{i=1}^n y_i) + R^{1+2r-2s}\|\phi\|_{L_2(d\rho_\mathcal{X})}),
\end{aligned}$$

where the last step is due to the fact that $1 \leqslant \lambda^{-1/2}\|\mathcal{L}\|^{1/2}$ since λ satisfies $0 < \lambda \leqslant \|\Sigma\| = \|\mathcal{L}\| \leqslant R^2$. Denote with δ the quantity $\delta = \lambda^{2+4r-4s}/c_0$. Since $\mathbb{E}[\mathbf{1}_E^c]$ corresponds to the probability of the event E^c, and, by Lemma 14, we have that E^c holds with probability at most δ since $n \geqslant 11(1 + \kappa_\mu^2 R^{2\mu}\lambda^{-\mu}) \log \frac{8R^2}{\lambda\delta}$, then we have that

$$\begin{aligned}
\mathbb{E}[R(\widehat{g}_\lambda)\mathbf{1}_{E^c}] &\leqslant \mathbb{E}[\|S\widehat{g}_\lambda\|^2_{L_2(d\rho_\mathcal{X})}\mathbf{1}_{E^c}] \leqslant \sqrt{\mathbb{E}[\|S\widehat{g}_\lambda\|^4_{L_2(d\rho_\mathcal{X})}]}\sqrt{\mathbb{E}[\mathbf{1}_{E^c}]} \\
&\leqslant \sqrt{\frac{4R^{2-4s}n^{-2}(\sum_{i,j=1}^n \mathbb{E}[y_i^2 y_j^2]) + 4R^{4-8r-8s}\|\phi\|^4_{L_2(d\rho_\mathcal{X})}}{\lambda^2}}\sqrt{\delta} \\
&\leqslant \frac{\sqrt{\delta}}{\lambda}\sqrt{4R^{2-4s}m_4 + 4R^{4-8r-8s}\|\phi\|^4_{L_2(d\rho_\mathcal{X})}} \\
&= \frac{\sqrt{\delta c_0/(8R^2)}}{\lambda} \leqslant \lambda^{2r-2s}.
\end{aligned}$$

□

Corollary 3. *Let $\lambda > 0$ and $n \in \mathbb{N}$ and $s = 0$. Under Assumption (A3), (A4), (A5) (see Rem. 3), (A6), when*

$$\lambda \;=\; B_1 \begin{cases} n^{-\alpha/\left(2r\alpha+1+\frac{\mu\alpha-1}{q+1}\right)} & 2r\alpha+1+\frac{\mu\alpha-1}{q+1} > \mu\alpha \\ n^{-1/\mu}\,(\log B_2 n)^{\frac{1}{\mu}} & 2r\alpha+1+\frac{\mu\alpha-1}{q+1} \leqslant \mu\alpha. \end{cases} \tag{15}$$

then,

$$\mathbb{E}\,\mathcal{E}(\widehat{g}_\lambda) - \inf_{g\in\mathcal{H}} \mathcal{E}(g) \leqslant B_3 \begin{cases} n^{-2r\alpha/\left(2r\alpha+1+\frac{\mu\alpha-1}{q+1}\right)} & 2r\alpha+1+\frac{\mu\alpha-1}{q+1} > \mu\alpha \\ n^{-2r/\mu} & 2r\alpha+1+\frac{\mu\alpha-1}{q+1} \leqslant \mu\alpha \end{cases} \tag{16}$$

where $B_2 = 3 \vee (32R^6 m_4)^{\frac{\mu}{3+4r}} B_1^{-\mu}$ and B_1 defined explicitly in the proof.

Proof. The proof of this corollary is a direct application of Thm. 3. In the rest of the proof we find the constants to guarantee that the condition relating n, λ in the theorem is always satisfied. Indeed to guarantee the applicability of Thm. 3, we need to be sure that $n \geqslant 11(1+\kappa_\mu^2 R^{2\mu}\lambda^{-\mu})\log\frac{32R^6 m_4}{\lambda^{3+4r}}$. This is satisfied when both the following conditions hold $n \geqslant 22\log\frac{32R^6 m_4}{\lambda^{3+4r}}$ and $n \geqslant 2\kappa_\mu^2 R^{2\mu}\lambda^{-\mu}\log\frac{32R^6 m_4}{\lambda^{3+4r}}$. To study the last two conditions, we recall that for $A, B, s, q > 0$ we have that $An^{-s}\log(Bn^q)$ satisfy

$$An^{-s}\log(Bn^q) = \frac{qAB^{s/q}}{s}\frac{\log B^{s/q}n^s}{B^{s/q}n^s} \leqslant \frac{qAB^{s/q}}{es},$$

for any $n > 0$, since $\frac{\log x}{x} \leqslant \frac{1}{e}$ for any $x > 0$. Now we define explicitly B_1, let $\tau = \alpha/\left(2r\alpha+1+\frac{\mu\alpha-1}{q+1}\right)$, we have

$$B_1 = \left(\frac{22(3+4r)}{e\mu}(32R^6 m_4)^{\frac{\mu}{3+4r}}\right)^{\frac{1}{\mu}} \vee \tag{17}$$

$$\vee \begin{cases} \left(\frac{2M(3+4r)}{e(1/\tau-\mu)}(32R^6 m_4)^{\frac{1/\tau-\mu}{3+4r}}\right)^{\tau} & 2r\alpha+1+\frac{\mu\alpha-1}{q+1} > \mu\alpha \\ \left(\frac{2M(3+4r)}{\mu}\right)^{\frac{1}{\mu}} & 2r\alpha+1+\frac{\mu\alpha-1}{q+1} \leqslant \mu\alpha \end{cases}. \tag{18}$$

For the first condition, we use the fact that λ is always larger than $B_1 n^{-1/\mu}$, so we have

$$\frac{22}{n}\log\frac{32R^6 m_4}{\lambda^{3+4r}} \leqslant \frac{22}{n}\log\frac{32R^6 m_4 n^{(3+4r)/\mu}}{B_1^{3+4r}} \leqslant \frac{22(3+4r)(32R^6 m_4)^{\mu/(3+4r)}}{e\mu B_1^\mu} \leqslant 1.$$

For the second inequality, when $2r\alpha+1+\frac{\mu\alpha-1}{q+1} \geqslant \mu\alpha$, we have $\lambda = B_1 n^{-\tau}$, so

$$\begin{aligned}\frac{2\kappa_\mu^2 R^{2\mu}}{n}\lambda^{-\mu}\log\frac{32R^6 m_4}{\lambda^{3+4r}} &\leqslant \frac{2\kappa_\mu^2 R^{2\mu}}{B_1^\mu n^{1-\mu\tau}}\log\frac{32R^6 m_4 n^{(3+4r)\tau}}{B_1^{3+4r}} \\ &\leqslant \frac{2\kappa_\mu^2 R^{2\mu}(3+4r)\tau}{e(1-\mu\tau)}\frac{(32R^6 m_4)^{\frac{1/\tau-\mu}{3+4r}}}{B_1^{1/\tau}} \leqslant 1.\end{aligned}$$

Finally, when $2r\alpha+1+\frac{\mu\alpha-1}{q+1} \geqslant \mu\alpha$, we have $\lambda = B_1 n^{-1/\mu}(\log B_2 n)^{1/\mu}$. So since $\log(B_2 n) > 1$, we have

$$\frac{2\kappa_\mu^2 R^{2\mu}}{n}\log\frac{32R^6 m_4}{\lambda^{3+4r}} \leqslant \frac{2\kappa_\mu^2 R^{2\mu}}{B_1^\mu}\frac{\log\frac{32R^6 m_4 n^{(3+4r)/\mu}}{B_1^{3+4r}}}{\log(B_2 n)} = \frac{2\kappa_\mu^2 R^{2\mu}(3+4r)}{\mu B_1^\mu}\frac{\log\frac{(32R^6 m_4)^{\mu/(3+4r)}n}{B_1^\mu}}{\log(B_2 n)} \leqslant 1.$$

So by selecting λ as in Eq. 15, we guarantee that the condition required by Thm. 3 is satisfied.

Finally the constant B_3 is obtained by

$$B_3 = c_1\max(1,w)^{-(\mu+2s-2r)} + c_2\max(1,w)^{-\frac{q+\mu\alpha}{q\alpha+\alpha}-2s} + c_3\max(1,w)^{2r-2s},$$

with $w = B_1\log(1+B_2)$ and c_1, c_2, c_3 as in Thm. 3. □

E Experiments with different sampling

We present here the results for two different types of sampling, which seem to be more stable, perform better and are widely used in practice :
Without replacement (Figure 4): for which we select randomly the data points but never use two times over the same point in one epoch.
Cycles (Figure 5): for which we pick successively the data points in the same order.

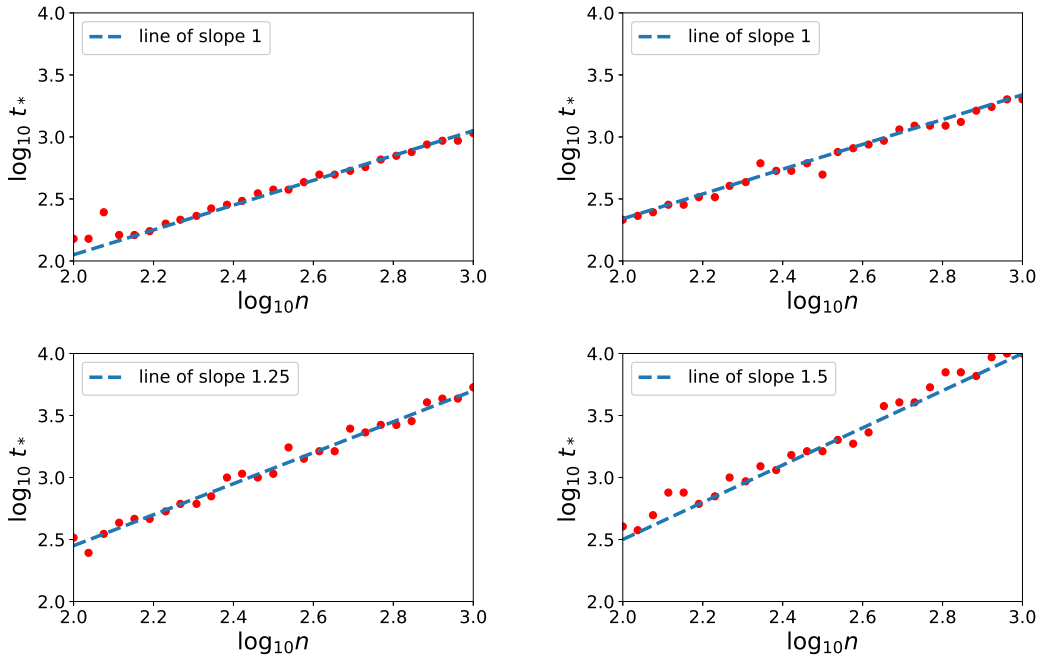

Figure 4 – The sampling is performed by **cycling over the data** The four plots represent each a different configuration on the (α, r) plan represented in Figure 1, for $r = 1/(2\alpha)$. **Top left** ($\alpha = 1.5$) and **right** ($\alpha = 2$) are two easy problems (Top right is the limiting case where $r = \frac{\alpha-1}{2\alpha}$) for which one pass over the data is optimal. **Bottom left** ($\alpha = 2.5$) and **right** ($\alpha = 3$) are two hard problems for which an increasing number of passes is recquired. The blue dotted line are the slopes predicted by the theoretical result in Theorem 1.

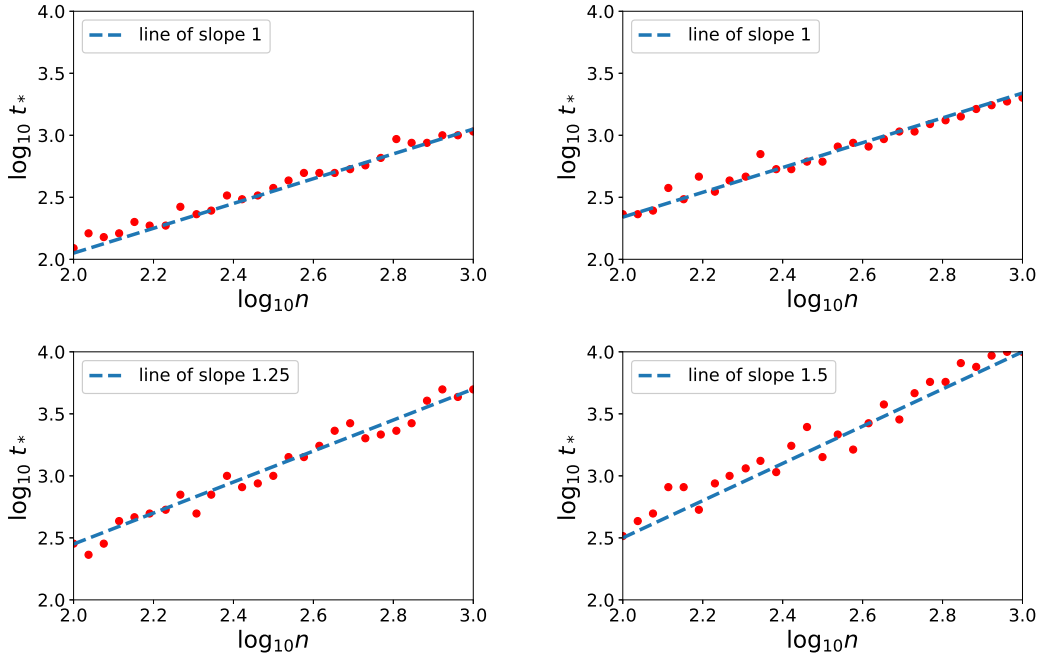

Figure 5 – The sampling is performed **without replacement**. The four plots represent each a different configuration on the (α, r) plan represented in Figure 1, for $r = 1/(2\alpha)$. **Top left** ($\alpha = 1.5$) and **right** ($\alpha = 2$) are two easy problems (Top right is the limiting case where $r = \frac{\alpha-1}{2\alpha}$) for which one pass over the data is optimal. **Bottom left** ($\alpha = 2.5$) and **right** ($\alpha = 3$) are two hard problems for which an increasing number of passes is recquired. The blue dotted line are the slopes predicted by the theoretical result in Theorem 1.