

Learning with Incremental Iterative Regularization

Lorenzo Rosasco¹ and Silvia Villa²

¹ Università degli Studi di Genova
DIBRIS, Via Dodecaneso 35 – 16146 Genova

&

Massachusetts Institute of Technology and Istituto Italiano di Tecnologia
Laboratory for Computational and Statistical Learning, Cambridge, MA 02139, USA

lrosasco@mit.edu

²Massachusetts Institute of Technology and Istituto Italiano di Tecnologia
Laboratory for Computational and Statistical Learning, Cambridge, MA 02139, USA

silvia.villa@iit.it

June 16, 2015

Abstract

Within a statistical learning setting, we propose and study an iterative regularization algorithm for least squares defined by an incremental gradient method. In particular, we show that, if all other parameters are fixed a priori, the number of passes over the data (epochs) acts as a regularization parameter, and prove strong universal consistency, i.e. almost sure convergence of the risk, as well as sharp finite sample bounds for the iterates. Our results are a step towards understanding the effect of multiple epochs in stochastic gradient techniques in machine learning and rely on integrating statistical and optimization results.

1 Introduction

Machine learning applications often require efficient statistical procedures to process potentially massive amount of high dimensional data. Motivated by such applications, the broad objective of our study is deriving learning procedures with optimal statistical properties, and, at the same time, computational complexities proportional to the *generalization* properties allowed by the data, rather than their raw amount [5]. In this paper, we focus on iterative regularization as a viable approach towards this goal. The key observation behind these techniques is that iterative optimization schemes applied to scattered, noisy data exhibit a self-regularizing property, in the sense that early termination (early-stop) of the iterative process has a regularizing effect [19, 22]. Indeed, iterative regularization algorithms are classical in inverse problems [14], and have been recently considered in machine learning [6, 32, 2, 4, 8, 24], where they have been proved to achieve optimal learning bounds, matching those of variational regularization schemes such as Tikhonov [7, 29].

In this paper, we consider an iterative regularization algorithm for the square loss, based on a recursive procedure updating the solution after processing one training set point at each iteration. Methods of the latter form, often broadly referred to as online learning algorithms, have become standard in the processing of large data-sets, because of their often low iteration cost and good practical performance. Theoretical studies for this class of algorithms have been developed within different frameworks. In composite optimization [17], in stochastic optimization [18, 27], in online learning, a.k.a. sequential prediction [10], and finally, in statistical learning [9]. The latter is the setting of interest in this paper, where we aim at developing an analysis keeping into account simultaneously both statistical and computational aspects. To place our contribution in context, it is useful to emphasize the role of regularization and different ways in which it can be incorporated in online learning algorithms. The key idea of regularization is that controlling the *complexity* of a solution can help avoiding overfitting and ensure stability to achieve improved results [31]. Classically, regularization is achieved perturbing (penalizing) the objective function with some suitable functional, or replacing the original risk minimization problem by a constrained problem obtained restricting the space of possible solutions [31]. Model selection is then performed to determine the amount of regularization suitable for the data at hand. More recently, there has been an interest in alternative, possibly more efficient, ways to incorporate regularization. We mention in particular [33, 1] (see also [30]) where there is no explicit regularization by penalization, and the step-size of an iterative procedure is shown to act as a regularization parameter. Here, for each fixed step-size, each data point is processed once, but multiple passes are indeed needed to perform model selection (that is pick the best step-size). We also mention [20] where an interesting adaptive approach is proposed, which seemingly avoid model selection under certain assumptions.

In this paper, we consider a different regularization strategy, which is widely used in practice. Namely, we consider no explicit penalization, fix the step size a priori, and analyze the effect of the number of passes over the data, which becomes the only free parameter to avoid overfitting, i.e. regularize. The associated regularization strategy, that we dub *incremental iterative regularization*, is hence based on early stopping. The latter is a well known "trick", for example in training large neural networks [16], and is known to perform very well in practice [15]. Our goal is to provide a theoretical understanding of the generalization property of the above heuristic for incremental/online techniques by grounding it in solid theoretical terms. Towards this end, we develop a theoretical analysis considering the behavior of both the excess risk and the iterates themselves. For the latter we obtain sharp finite sample bounds matching those for Tikhonov regularization in the same setting. Finite sample bounds for the excess risk can then be easily derived, albeit possibly suboptimal. Our results are developed in a capacity independent setting [11, 28], that is under no conditions on the covering or entropy numbers [28]. In this sense our analysis is worst case and dimension free. To the best of our knowledge the analysis in the paper is the first theoretical study of regularization by early stopping in incremental/online algorithms, and thus a first step towards understanding the effect of multiple epochs of stochastic gradient for risk minimization.

The rest of the paper is organized as follows. In Section 2 we describe the setting and the main assumptions, and in Section 3 we state the main results, discuss them and provide the main elements of the proof, which is deferred to the supplementary material. In Section 4 we present some experimental results on real and synthetic datasets.

Notation We denote by $\mathbb{R}_+ = [0, +\infty[$, $\mathbb{R}_{++} =]0, +\infty[$, and $\mathbb{N}^* = \mathbb{N} \setminus \{0\}$. Given a normed space \mathcal{B} and linear operators $(A_i)_{1 \leq i \leq m}$, $A_i: \mathcal{B} \rightarrow \mathcal{B}$ for every i , their composition $A_m \circ \dots \circ A_1$ will be denoted as $\prod_{i=1}^m A_i$. By convention, if $j > m$, we set $\prod_{i=j}^m A_i = I$, where I is the identity of \mathcal{B} . The operator norm will be denoted by $\|\cdot\|$ and the Hilbert-Schmidt norm by $\|\cdot\|_{HS}$. Also, if $j > m$, we

set $\sum_{i=j}^m A_i = 0$.

2 Setting and Assumptions

We first describe the setting we consider and then introduce and discuss the main assumptions that will hold throughout the paper. We essentially follow the framework proposed in [12, 25] and further developed in a series of follow up works [7, 2, 26, 8]. Unlike these papers where a reproducing kernel Hilbert space (RKHS) setting is considered (see Section A), here we develop an equivalent formulation within an abstract Hilbert space. This latter formulation is close to the setting of functional regression [23] and reduces to standard linear regression if \mathcal{H} is finite dimensional, see Section A.

Let \mathcal{H} be a separable Hilbert space with inner product and norm denoted by $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and $\|\cdot\|_{\mathcal{H}}$. Let (X, Y) be a pair of random variables on a probability space $(\Omega, \mathfrak{G}, \mathbb{P})$, with values in \mathcal{H} and \mathbb{R} , respectively. Denote by ρ the distribution of (X, Y) , by $\rho_{\mathcal{H}}$ the marginal measure on \mathcal{H} , and by $\rho(\cdot|x)$ the conditional measure on \mathbb{R} given $x \in \mathcal{H}$. Considering the square loss function, the problem under study is the minimization of the risk,

$$\inf_{w \in \mathcal{H}} \mathcal{E}(w), \quad \mathcal{E}(w) = \int_{\mathcal{H} \times \mathbb{R}} (\langle w, x \rangle_{\mathcal{H}} - y)^2 d\rho(x, y), \quad (1)$$

provided the distribution ρ is fixed but known only through a *training set* $\mathbf{z} = \{(x_1, y_1), \dots, (x_n, y_n)\}$, that is a realization of $n \in \mathbb{N}^*$ independent identical copies of (X, Y) . In the following, we measure the quality of an approximate solution $\hat{w} \in \mathcal{H}$ (an estimator) controlling the excess risk

$$\mathcal{E}(\hat{w}) - \inf_{\mathcal{H}} \mathcal{E}.$$

If the set of solutions of Problem (1) is non empty, that is $\mathcal{O} = \operatorname{argmin}_{\mathcal{H}} \mathcal{E} \neq \emptyset$, we also consider

$$\left\| \hat{w} - w^\dagger \right\|_{\mathcal{H}}, \quad \text{where } w^\dagger = \operatorname{argmin}_{w \in \mathcal{O}} \|w\|_{\mathcal{H}}. \quad (2)$$

More precisely we are interested in deriving almost sure convergence results and finite sample bounds on the above error measures. This requires making some assumptions that we discuss next. We make throughout the following basic assumption.

Assumption 1 There exist $M \in]0, +\infty[$ and $\kappa \in]0, +\infty[$ such that $|y| \leq M$ ρ -almost surely, and $\|x\|_{\mathcal{H}}^2 \leq \kappa$ $\rho_{\mathcal{H}}$ -almost surely.

The above assumption is fairly standard. The boundness assumption on the output is satisfied in classification, see Section A, and can be easily relaxed, see e.g. [7]. The boundness assumption on the input can also be relaxed, but the resulting analysis is more involved. We omit these developments for the sake of clarity. It is well known that (see e.g. [13]), under Assumption 1, the risk is a convex and continuous functional on $L^2(\mathcal{H}, \rho_{\mathcal{H}})$, the space of square-integrable functions with norm $\|f\|_{\rho}^2 = \int_{\mathcal{H} \times \mathbb{R}} |f(x)|^2 d\rho(x, y)$. The minimizer of the risk on $L^2(\mathcal{H}, \rho_{\mathcal{H}})$ is the regression function $f_{\rho}(x) = \int y d\rho(y|x)$ for $\rho_{\mathcal{H}}$ -almost every $x \in \mathcal{H}$. By considering Problem (1) we are restricting the search for a solution to linear functions. Note that, since \mathcal{H} is in general infinite dimensional, the minimum in (1) might not be achieved. Indeed, bounds on the above error measures depend on if, and how well, the regression function can be linearly approximated. The following assumption quantifies in a precise way such a requirement.

Assumption 2 Consider the space $\mathcal{L}_\rho = \{f : \mathcal{H} \rightarrow \mathbb{R} \mid \exists w \in \mathcal{H} \text{ with } f = \langle w, \cdot \rangle_{\rho_{\mathcal{H}}} \text{ a. s.}\}$, and let $\overline{\mathcal{L}}_\rho$ be its closure in $L^2(\mathcal{H}, \rho_{\mathcal{H}})$. Moreover, consider the operator

$$L : L^2(\mathcal{H}, \rho_{\mathcal{H}}) \rightarrow L^2(\mathcal{H}, \rho_{\mathcal{H}}), \quad Lf(x) = \int \langle x, x' \rangle f(x') d\rho(x'), \quad \forall f \in L^2(\mathcal{H}, \rho_{\mathcal{H}}). \quad (3)$$

Define $g_\rho = \operatorname{argmin}_{g \in \overline{\mathcal{L}}_\rho} \|f_\rho - g\|_\rho$. Let $r \in [0, +\infty[$, and assume that

$$(\exists g \in L^2(\mathcal{H}, \rho_{\mathcal{H}})) \quad \text{such that} \quad g_\rho = L^r g. \quad (4)$$

The above assumption is standard [28]. Since its statement is somewhat technical, and we provide a more general formulation in a Hilbert space with respect to the usual RKHS setting, we further comment on its interpretation. We begin noting that \mathcal{L}_ρ is the space of linear functions indexed by \mathcal{H} and is a proper subspace of $L^2(\mathcal{H}, \rho_{\mathcal{H}})$, if Assumption 1 holds. Moreover, under the same assumption, it is easy to see that the operator L is linear, self-adjoint, positive definite and trace class, hence compact, so that its fractional power in (3) is well defined. It can be shown fairly easily that the space \mathcal{L}_ρ can be characterized in terms of the operator L , namely

$$\mathcal{L}_\rho = L^{1/2} (L^2(\mathcal{H}, \rho_{\mathcal{H}})). \quad (5)$$

This last observation allows to provide an interpretation of Condition (4). Indeed, given (5), for $r = 1/2$, Condition (4) states that g_ρ belongs to \mathcal{L}_ρ , rather than its closure. In this case, Problem 1 has at least one solution, and the set \mathcal{O} in (2) is not empty. Vice versa, if $\mathcal{O} \neq \emptyset$ then $g_\rho \in \mathcal{L}_\rho$, and w^\dagger is well-defined. If $r > 1/2$ the condition is stronger than for $r = 1/2$, since the images of $L^2(\mathcal{H}, \rho_{\mathcal{H}})$ with respect to L^r are nested subspaces for increasing r ¹.

2.1 Iterative Incremental Regularized Learning

The learning algorithm we consider is defined by the following iteration.

Let $w_0 \in \mathcal{H}$ and $\gamma \in \mathbb{R}_{++}$. Consider the sequence $(\hat{w}_t)_{t \in \mathbb{N}}$ generated through the following procedure: given $t \in \mathbb{N}$ and $\hat{w}_t \in \mathcal{H}$, define

$$\hat{w}_{t+1} = \hat{u}_t^n, \quad (6)$$

where \hat{u}_t^n is obtained at the end of one cycle, namely as the last step of the recursion

$$\hat{u}_t^0 = \hat{w}_t; \quad \hat{u}_t^i = \hat{u}_t^{i-1} - \frac{\gamma}{n} (\langle \hat{u}_t^{i-1}, x_i \rangle_{\mathcal{H}} - y_i) x_i, \quad i = 1, \dots, n. \quad (7)$$

Each cycle, called an epoch, corresponds to one pass over data. The above iteration can be seen as the incremental gradient method [3, 17] for the minimization of the empirical risk corresponding to \mathbf{z} , that is the functional

$$\hat{\mathcal{E}}(w) = \frac{1}{n} \sum_{i=1}^n (\langle w, x_i \rangle_{\mathcal{H}} - y_i)^2. \quad (8)$$

¹ If $r < 1/2$ then the regression function does not have a best linear approximation since $g_\rho \notin \mathcal{L}_\rho$, and in particular, for $r = 0$ we are not making any assumption. Intuitively, for $0 < r < 1/2$, the condition quantifies *how far* g_ρ is from \mathcal{L}_ρ , that is to be well approximated by a linear function.

(see also Section B.2). Indeed, there is a vast literature on how the iterations (6), (7) can be used to minimize the empirical risk [3, 17]. Unlike these studies in this paper we are interested in how the iterations (6), (7) can be used to approximately minimize the risk \mathcal{E} . The key idea is that while \hat{w}_t is close to a minimizer of the empirical risk when t is sufficiently large, a good approximate solution of Problem (1) can be found by terminating the iterations earlier (early stopping). The analysis in the next few sections grounds theoretically this latter intuition.

3 Early stopping for incremental iterative regularization

In this section we present and discuss the main results of the paper, together with a sketch of the proof. The complete proofs can be found in Appendix B. We first present convergence results and then finite sample bounds for the norm and the excess risk.

Theorem 3.1 *In the setting of Section 2, let Assumption 1 hold. Let $\gamma \in]0, \kappa^{-1}]$. Then the following hold:*

(i) *If we choose a stopping rule $t^* : \mathbb{N}^* \rightarrow \mathbb{N}^*$ such that*

$$\lim_{n \rightarrow +\infty} t^*(n) = +\infty \quad \text{and} \quad \lim_{n \rightarrow +\infty} \frac{t^*(n)^3 \log n}{n} = 0 \quad (9)$$

then

$$\lim_{n \rightarrow +\infty} \mathcal{E}(\hat{w}_{t^*(n)}) - \inf_{w \in \mathcal{H}} \mathcal{E}(w) = 0 \quad \mathbb{P}\text{-almost surely.} \quad (10)$$

(ii) *Suppose additionally that the set \mathcal{O} of minimizers of (1) is nonempty and let w^\dagger be defined as in (2). If we choose a stopping rule $t^* : \mathbb{N}^* \rightarrow \mathbb{N}^*$ satisfying the conditions in (9) then*

$$\|\hat{w}_{t^*(n)} - w^\dagger\|_{\mathcal{H}} \rightarrow 0 \quad \mathbb{P}\text{-almost surely.} \quad (11)$$

The above result shows that for an a priori fixed step-sized, consistency is achieved computing a suitable number $t^*(n)$ of iterations of algorithm (6)-(7) given n points. The number of required iterations tends to infinity as the number of available training points increases. Condition (9) can be interpreted as an early stopping rule, since it requires the number of epochs not to grow too fast. In particular, this excludes the choice $t^*(n) = 1$ for all $n \in \mathbb{N}^*$, namely considering only one pass over the data. Given the currently known results, the failure of a single pass does not appear to be surprising, considering the stepsize is fixed (not depending on n) and we are not averaging. In the following remark we make clear that, if we let the step size to depend on the length of one epoch, convergence is recovered also for one pass.

Remark 3.2 (Recovering Stochastic Gradient descent) Note that if in Theorem 3.1 we let γ to depend on n , than we can choose $t^*(n) = 1$. Indeed, choosing $\gamma(n) = \kappa^{-1}n^\alpha$, with $\alpha < 1/4$, which corresponds to a stochastic gradient method with step-size $\kappa^{-1}n^{\alpha-1}$, we can derive almost sure convergence of $\mathcal{E}(\hat{w}_1) - \inf_{\mathcal{H}} \mathcal{E}$ as $n \rightarrow +\infty$ relying on the same proof.

To derive finite sample bounds we need to impose additional conditions. We will see that the behavior of the bias of the estimator depends on the smoothness assumption (4). We are in position to state our main result, giving a finite sample bound.

Theorem 3.3 (Finite sample bounds in \mathcal{H}) *In the setting of Section 2, let $\gamma \in]0, \kappa^{-1}]$ for every $t \in \mathbb{N}$. Suppose that Assumption (2) is satisfied for some $r \in]1/2, +\infty[$. Then the set \mathcal{O} of minimizers of (1) is nonempty, and w^\dagger in (2) is well defined. Moreover, the following hold:*

(i) *There exists $c \in]0, +\infty[$ such that, for every $t \in \mathbb{N}^*$, with probability greater than $1 - \delta$,*

$$\|\hat{w}_t - w^\dagger\|_{\mathcal{H}} \leq \frac{32 \log(16/\delta)}{\sqrt{n}} \left(M\kappa^{-1/2} + 2M^2\kappa^{-1} + 3\|g\|_{\rho}\kappa^{r-3/2} \right) t \quad (12)$$

$$+ \left(\frac{r-1/2}{\gamma} \right)^{r-1/2} \|g\|_{\rho} t^{1/2-r}. \quad (13)$$

(ii) *For the stopping rule $t^* : \mathbb{N}^* \rightarrow \mathbb{N}^*$ defined by $t^*(n) = \lceil n^{\frac{1}{2r+1}} \rceil$, with probability greater than $1 - \delta$,*

$$\|\hat{w}_{t^*(n)} - w^\dagger\|_{\mathcal{H}} \leq \left[32 \log \left(\frac{16}{\delta} \right) \left(M\kappa^{-1/2} + 2M^2\kappa^{-1} + 3\|g\|_{\rho}\kappa^{r-3/2} \right) + \left(\frac{r-1/2}{\gamma} \right)^{r-1/2} \|g\|_{\rho} \right] n^{-\frac{r-1/2}{2r+1}}. \quad (14)$$

Note that the dependence on κ suggests that a big κ , which corresponds to a small γ , helps in decreasing the sample error, but increases the approximation error. Next we present the result for the excess risk. We consider only the attainable case, that is the case $r > 1/2$ in Assumption 2. The case $r \leq 1/2$ is deferred to Appendix A, since both the proof and the statement are conceptually similar to the attainable case.

Theorem 3.4 (Finite sample bounds for the risk – attainable case) *In the setting of Section 2, let Assumptions 1 holds, and let $\gamma \in]0, \kappa^{-1}]$. Let Assumption 2 be satisfied for some $r \in]1/2, +\infty[$. Then the following hold:*

(i) *For every $t \in \mathbb{N}^*$, with probability greater than $1 - \delta$,*

$$\mathcal{E}(\hat{w}_t) - \inf_{\mathcal{H}} \mathcal{E} \leq \frac{2(32 \log(16/\delta))^2}{n} \left[M + 2M^2\kappa^{-1/2} + 3\kappa^r \|g\|_{\rho} \right]^2 t^2 + 2 \left(\frac{r}{\gamma t} \right)^{2r} \|g\|_{\rho}^2 \quad (15)$$

(ii) *For the stopping rule $t^* : \mathbb{N}^* \rightarrow \mathbb{N}^*$ $t^*(n) = \lceil n^{\frac{1}{2(1+r)}} \rceil$, with probability greater than $1 - \delta$,*

$$\mathcal{E}(\hat{w}_{t^*(n)}) - \inf_{\mathcal{H}} \mathcal{E} \leq \left[8 \left(32 \log \frac{16}{\delta} \right)^2 \left(M + 2M^2\kappa^{-1/2} + 3\kappa^r \|g\|_{\rho} \right)^2 + 2 \left(\frac{r}{\gamma} \right)^{2r} \|g\|_{\rho}^2 \right] n^{-r/(r+1)} \quad (16)$$

Equations (12) and (15) arise from a form of bias-variance (sample-approximation) decomposition of the error. Choosing the number of epochs that optimize the bounds in (12) and (15) we derive a priori stopping rules and corresponding bounds (14) and (16). Again, these results confirm that the number of epochs acts as a regularization parameter and the best choices following from equations (12) and (15) suggest multiple passes over the data to be beneficial. In both cases, the stopping rule depends on the smoothness parameter r which is typically unknown, and hold-out cross validation is often used in practice. Following [8], it is possible to show that this procedure allows to adaptively achieve the same convergence rate as in (16).

3.1 Discussion

We discuss a few comments and comparisons. In Theorem 3.3, the obtained bound can be compared to known lower bounds, as well as to previous results for least squares algorithms obtained under Assumption 2. Minimax lower bounds and individual lower bounds [7, 29], suggest that, for $r > 1/2$, $O(n^{(r-1/2)/(2r+1)})$ is the optimal capacity-independent bound for the \mathcal{H} norm². In this sense, Theorem 3.3 provides sharp bounds on the iterates. Bounds can be improved only under stronger assumptions, e.g. on the covering numbers or on the eigenvalues of L [28]. This question is left for future work. The lower bounds for the excess risk [7, 29] are of the form $O(n^{-2r/(2r+1)})$ and in this case the results in Theorems A.1 and 3.4 are not sharp, and in principle could be improved. Our results can be contrasted with online learning algorithms mentioned in the introduction that use step-size as regularization parameter. Optimal capacity independent bounds are obtained in [33], see also [30] and indeed such results can be further improved considering capacity assumptions, see [1] and references therein.

Interestingly, our results can also be contrasted with non incremental iterative regularization approaches [6, 32, 2, 4, 8, 24]. Our current results show that incremental iterative regularization, with distribution independent step-size, behaves as a full gradient descent, at least in terms of iterates convergence. We believe that proving additional advantages of incremental regularization over the batch one is an interesting future research direction. Finally, we note that optimal capacity independent and dependent bounds are known for several least squares algorithms including Tikhonov regularization see e.g. [29] and references therein, as well as for a larger class of so called spectral filtering methods [2, 8]. These algorithms can be seen to be essentially equivalent from a statistical perspective but different from a computational point of view.

3.2 Elements of the proof

The proofs of the main results are based on a decomposition of the error to be estimated in two terms. The idea is to build an auxiliary sequence and to majorize the error with the sum of two quantities that can be interpreted as a sample and an approximation error, respectively. Bounds on these two terms are then provided. The main technical contribution of the paper is the sample error bound. The difficulty in proving this result is due to the fact that multiple passes over the data induce complex statistical dependences in the iterates.

Error decomposition. We consider an auxiliary iteration $(w_t)_{t \in \mathbb{N}}$ which is the expectation of the iterations (6) and (7), starting from $w_0 \in \mathcal{H}$ with step-size $\gamma \in \mathbb{R}_{++}$. More explicitly, given $w_t \in \mathcal{H}$, the considered iteration generates w_{t+1} according to

$$w_{t+1} = u_t^n, \tag{17}$$

where u_t^n is given by

$$u_t^0 = w_t; \quad u_t^i = u_t^{i-1} - \frac{\gamma}{n} \int_{\mathcal{H} \times \mathbb{R}} (\langle u_t^{i-1}, x \rangle_{\mathcal{H}} - y) x \, d\rho(x, y). \tag{18}$$

²In a recent manuscript, it has been proved that these are indeed minimax lower bounds (G. Blanchard, personal communication)

If we let $S: \mathcal{H} \rightarrow L^2(\mathcal{H}, \rho_{\mathcal{H}})$ be the linear map $w \mapsto \langle w, \cdot \rangle_{\mathcal{H}}$, which is bounded by $\sqrt{\kappa}$ under Assumption (1) holds, then it is well-known that

$$\begin{aligned} (\forall t \in \mathbb{N}) \quad \mathcal{E}(\hat{w}_t) - \inf_{\mathcal{H}} \mathcal{E} &= \|S\hat{w}_t - g_{\rho}\|_{\rho}^2 \leq 2\|S\hat{w}_t - Sw_t\|_{\rho}^2 + 2\|Sw_t - g_{\rho}\|_{\rho}^2 \\ &\leq 2\kappa\|\hat{w}_t - w_t\|_{\mathcal{H}}^2 + 2(\mathcal{E}(w_t) - \inf_{\mathcal{H}} \mathcal{E}). \end{aligned} \quad (19)$$

In this paper, we refer to the gap between the empirical and the expected iterates $\|\hat{w}_t - w_t\|_{\mathcal{H}}$ as the *sample error*, and to $\mathcal{A}(t, \gamma, n) = \mathcal{E}(w_t) - \inf_{\mathcal{H}} \mathcal{E}$ as the *approximation error*. Similarly, if w^{\dagger} (as defined in (2)) exists, using the triangle inequality, we obtain

$$\|\hat{w}_t - w^{\dagger}\|_{\mathcal{H}} \leq \|\hat{w}_t - w_t\|_{\mathcal{H}} + \|w_t - w^{\dagger}\|_{\mathcal{H}}. \quad (20)$$

Proof main steps. In the setting of Section 2, we summarize the key steps to derive a general bound for the sample error. Indeed, this is the main technical contribution of the paper. The proof of the behavior of the approximation error is more standard. The bound on the sample error is derived through many technical lemmas and uses concentration inequalities applied to martingales. Its complete derivation is reported in Section B.2. We underline that the crucial point is the probabilistic inequality in **STEP 5** below. We need to introduce the additional linear operators: $T: \mathcal{H} \rightarrow \mathcal{H}$: $T = S^*S$, ($\forall x \in \mathcal{X}$) $S_x: \mathcal{H} \rightarrow \mathbb{R}$: $S_x w = \langle w, x \rangle$, and $T_x: \mathcal{H} \rightarrow \mathcal{H}$: $T_x = S_x S_x^*$. Moreover, set $\hat{T} = \sum_{i=1}^n T_{x_i}/n$. We are now ready to state the main steps of the proof of the main results.

Sample error bound (STEP 1 to 5)

STEP 1 (see Proposition B.6): Find equivalent formulations for the sequences \hat{w}_t and w_t :

$$\begin{aligned} \hat{w}_{t+1} &= (I - \gamma\hat{T})\hat{w}_t + \gamma\left(\frac{1}{n} \sum_{j=1}^n S_{x_j}^* y_j\right) + \gamma^2 (\hat{A}\hat{w}_t - \hat{b}) \\ w_{t+1} &= (I - \gamma T)w_t + \gamma S^* g_{\rho} + \gamma^2 (Aw_t - b), \end{aligned}$$

with

$$\begin{aligned} \hat{A} &= \frac{1}{n^2} \sum_{k=2}^n \left[\prod_{i=k+1}^n \left(I - \frac{\gamma}{n} T_{x_i} \right) \right] T_{x_k} \sum_{j=1}^{k-1} T_{x_j}, & \hat{b} &= \frac{1}{n^2} \sum_{k=2}^n \left[\prod_{i=k+1}^n \left(I - \frac{\gamma}{n} T_{x_i} \right) \right] T_{x_k} \sum_{j=1}^{k-1} S_{x_j}^* y_j. \\ A &= \frac{1}{n^2} \sum_{k=2}^n \left[\prod_{i=k+1}^n \left(I - \frac{\gamma}{n} T \right) \right] T \sum_{j=1}^{k-1} T, & b &= \frac{1}{n^2} \sum_{k=2}^n \left[\prod_{i=k+1}^n \left(I - \frac{\gamma}{n} T \right) \right] T \sum_{j=1}^{k-1} S^* g_{\rho}. \end{aligned}$$

STEP 2 (see Lemma B.7): Use the formulation obtained in **STEP 1** to derive the following recursive inequality

$$\hat{w}_t - w_t = \left(I - \gamma\hat{T} + \gamma^2\hat{A} \right)^t (\hat{w}_0 - w_0) + \gamma \sum_{k=0}^{t-1} \left(I - \gamma\hat{T} + \gamma\hat{A} \right)^{t-k-1} \zeta_k \quad (21)$$

with

$$\zeta_k = (T - \hat{T})w_k + \gamma(\hat{A} - A)w_k + \left(\frac{1}{n} \sum_{i=1}^n \hat{S}_{x_i}^* y_i - S^* g_{\rho} \right) + \gamma(b - \hat{b}). \quad (22)$$

STEP 3 (see Lemmas B.8 and B.9): Initialize $\hat{w}_0 = w_0 = 0$, prove that $\|I - \gamma\hat{T} + \gamma\hat{A}\| \leq 1$ and derive from **STEP 2** that

$$\|\hat{w}_t - w_t\|_{\mathcal{H}} \leq \gamma(\|T - \hat{T}\| + \gamma\|\hat{A} - A\|) \sum_{k=0}^{t-1} \|w_k\|_{\mathcal{H}} + \gamma t \left(\left\| \frac{1}{n} \sum_{i=1}^n \hat{S}_{x_i}^* y_i - S^* g_{\rho} \right\| + \gamma\|b - \hat{b}\| \right).$$

STEP 4 (see Lemma B.10): Let Assumption 2 hold for some $r \in \mathbb{R}_+$ and $g \in L^2(\mathcal{H}, \rho_{\mathcal{H}})$. Prove that

$$(\forall t \in \mathcal{H}) \quad \|w_t\|_{\mathcal{H}} \leq \begin{cases} \max\{\kappa^{r-1/2}, (\gamma t)^{1/2-r}\} \|g\|_{\rho} & \text{if } r \in [0, 1/2[\\ \kappa^{r-1/2} \|g\|_{\rho} & \text{if } r \in [1/2, +\infty[\end{cases}$$

STEP 5 (see Lemma B.11 and Proposition B.12): Prove that with probability greater than $1 - \delta$ the following inequalities hold:

$$\begin{aligned} \|\hat{A} - A\|_{HS} &\leq \frac{32\kappa^2}{3\sqrt{n}} \log \frac{4}{\delta}, & \|\hat{b} - b\|_{\mathcal{H}} &\leq \frac{32\kappa M^2}{3\sqrt{n}} \log \frac{4}{\delta} \\ \left\| \frac{1}{n} \sum_{i=1}^n T_{x_i} - T \right\|_{HS} &\leq \frac{16\kappa}{3\sqrt{n}} \log \frac{2}{\delta}, & \left\| \frac{1}{n} \sum_{i=1}^n S_{x_i}^* y_i - S^* g_{\rho} \right\|_{\mathcal{H}} &\leq \frac{16\sqrt{\kappa} M}{3\sqrt{n}} \log \frac{2}{\delta} \end{aligned}$$

STEP 6 (approximation error bound, see Theorem B.14): Prove that, if Assumption 2 holds for some $r \in]0, +\infty[$, then

$$\mathcal{E}(w_t) - \inf_{\mathcal{H}} \mathcal{E} \leq \left(\frac{r}{\gamma t} \right)^{2r} \|g\|_{\rho}^2.$$

Moreover, if Assumption 2 holds with $r = 1/2$, then $\|w_t - w^{\dagger}\|_{\mathcal{H}} \rightarrow 0$, and if Assumption 2 holds for some $r \in]1/2, +\infty[$, then

$$\|w_t - w^{\dagger}\|_{\mathcal{H}} \leq \left(\frac{r - 1/2}{\gamma t} \right)^{r-1/2} \|g\|_{\rho}.$$

STEP 7: Plug the sample and approximation error bounds obtained in **STEP 1-5** and **STEP 6** in (19) and (20), respectively.

4 Experiments

Synthetic data. We consider a linear regression problem with random design in $\mathbb{R}^d \times \mathbb{R}$. The input points $(x_i)_{1 \leq i \leq n}$ are uniformly distributed in $[0, 1]$ and the output points are obtained as $y_i = \langle w^*, \Phi(x_i) \rangle + N_i$, where N_i is a gaussian noise with zero mean and standard deviation 1 and $\Phi = (\varphi_k)_{1 \leq k \leq d}$ is a dictionary of trigonometric functions whose k -th element is $\varphi_k(x) = \cos((k-1)x) + \sin((k-1)x)$. In Figure 1, we plot the test error for $d = 5$ (with $n = 80$ in (a) and 800 in (b)). The plots show that the number of the epochs acts as a regularization parameter, and that early stopping is beneficial to achieve a better test error. Moreover, according to our theoretical findings, the experimental results suggest that the number of performed epochs should increase if the number of available training points increases.

Real data. We tested the kernelized version of our algorithm (see Appendix A) on the cpuSmall³, Adult and Breast Cancer Wisconsin (Diagnostic)⁴ real-world datasets. We considered a subset of Adult, with $n = 1600$. The results are shown in Figure 2. A comparison of the test errors obtained with the kernelized version of the method proposed in this paper (Kernel Incremental Iterative Regularization (KIIR)), Kernel Iterative Regularization (KIR), that is the kernelized version of gradient descent, and Kernel Ridge Regression (KRR) is reported in Table 1. The results show that the proposed method achieve a test error comparable to that of KIR and KRR.

³Available at <http://www.cs.toronto.edu/~delve/data/comp-activ/desc.html>

⁴Adult and Breast Cancer Wisconsin (Diagnostic), UCI repository, 2013.

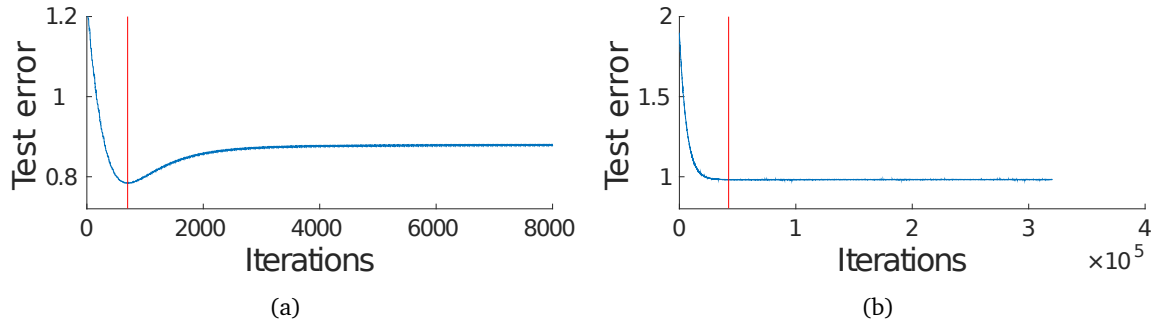


Figure 1: Evolution of the test error with the number of iterations. In (a), $n = 80$, and total number of iterations of IIR is 8000, corresponding to 100 epochs. In (b), $n = 800$ and the total number of epochs is 400. The best test error is obtained for 9 epochs in (a) and for 31 epochs in (b).

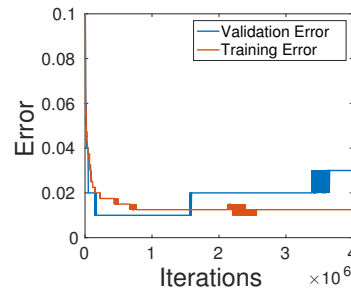


Figure 2: Training (orange) and validation (blue) classification errors obtained by KIIR on the Breast Cancer dataset as a function of the number of iterations over the training points. As can be seen by visual inspection, the test error increases after a certain number of iterations, while the training error is approximately decreasing with the number of iterations.

Table 1: Test error comparison on real datasets. Median values over 5 trials.

Dataset	n_{tr}	d	Error Measure	KIIR	KRR	KIR
cpuSmall	5243	12	RMSE	5.9125	3.6841	5.4665
Adult	1600	123	Class. Err.	0.167	0.164	0.154
Breast Cancer	400	30	Class. Err.	0.0118	0.0118	0.0237

References

- [1] F. Bach and A. Dieuleveut. Non-parametric stochastic approximation with large step sizes. *arXiv:1408.0361*, 2014.
- [2] F. Bauer, S. Pereverzev, and L. Rosasco. On regularization algorithms in learning theory. *Journal of complexity*, 23(1):52–72, 2007.
- [3] D. P. Bertsekas. A new class of incremental gradient methods for least squares problems. *SIAM J. Optim.*, 7(4):913–926, 1997.

- [4] G. Blanchard and N. Krämer. Optimal learning rates for kernel conjugate gradient regression. In *Advances in Neural Inf. Proc. Systems (NIPS)*, pages 226–234, 2010.
- [5] L. Bottou and O. Bousquet. The tradeoffs of large scale learning. In Suvrit Sra, Sebastian Nowozin, and Stephen J. Wright, editors, *Optimization for Machine Learning*, pages 351–368. MIT Press, 2011.
- [6] P. Buhlmann and B. Yu. Boosting with the l2 loss: Regression and classification. *Journal of the American Statistical Association*, 98:324–339, 2003.
- [7] A. Caponnetto and E. De Vito. Optimal rates for regularized least-squares algorithm. *Found. Comput. Math.*, 2006.
- [8] A. Caponnetto and Yuan Yao. Adaptive rates for regularization operators in learning theory. *Analysis and Applications*, 08, 2010.
- [9] N. Cesa-Bianchi, A. Conconi, and C. Gentile. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057, 2004.
- [10] N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.
- [11] F. Cucker and D. X. Zhou. *Learning Theory: An Approximation Theory Viewpoint*. Cambridge University Press, 2007.
- [12] E. De Vito, L. Rosasco, A. Caponnetto, U. De Giovannini, and F. Odone. Learning from examples as an inverse problem. *Journal of Machine Learning Research*, 6:883–904, 2005.
- [13] E. De Vito, L. Rosasco, A. Caponnetto, M. Piana, and A. Verri. Some properties of regularized kernel methods. *Journal of Machine Learning Research*, 5:1363–1390, 2004.
- [14] H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of inverse problems*. Kluwer, 1996.
- [15] P.-S. Huang, H. Avron, T. Sainath, V. Sindhvani, and B. Ramabhadran. Kernel methods match deep neural networks on timit. In *IEEE ICASSP*, 2014.
- [16] Y. LeCun, L. Bottou, G. Orr, and K. Muller. Efficient backprop. In G. Orr and Muller K., editors, *Neural Networks: Tricks of the trade*. Springer, 1998.
- [17] A. Nedic and D. P Bertsekas. Incremental subgradient methods for nondifferentiable optimization. *SIAM Journal on Optimization*, 12(1):109–138, 2001.
- [18] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.*, 19(4):1574–1609, 2008.
- [19] A. Nemirovskii. The regularization properties of adjoint gradient method in ill-posed problems. *USSR Computational Mathematics and Mathematical Physics*, 26(2):7–16, 1986.
- [20] F. Orabona. Simultaneous model selection and optimization through parameter-free stochastic learning. *NIPS Proceedings*, 2014.
- [21] I. Pinelis. Optimum bounds for the distributions of martingales in Banach spaces. *Ann. Probab.*, 22(4):1679–1706, 1994.

- [22] B. Polyak. *Introduction to Optimization*. Optimization Software, New York, 1987.
- [23] J. Ramsay and B. Silverman. *Functional Data Analysis*. Springer Series in Statistics. Springer-Verlag, New York, 2005.
- [24] G. Raskutti, M. Wainwright, and B. Yu. Early stopping for non-parametric regression: An optimal data-dependent stopping rule. In *49th Annual Allerton Conference*, pages 1318–1325. IEEE, 2011.
- [25] S. Smale and D. Zhou. Shannon sampling II: Connections to learning theory. *Applied and Computational Harmonic Analysis*, 19(3):285–302, November 2005.
- [26] S. Smale and D.-X. Zhou. Learning theory estimates via integral operators and their approximations. *Constr. Approx.*, 26(2):153–172, 2007.
- [27] N. Srebro, K. Sridharan, and A. Tewari. Optimistic rates for learning with a smooth loss. *arXiv:1009.3896*, 2012.
- [28] I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, 2008.
- [29] I. Steinwart, D. R. Hush, and C. Scovel. Optimal rates for regularized least squares regression. In *COLT*, 2009.
- [30] P. Tarrès and Y. Yao. Online learning as stochastic approximation of regularization paths: optimality and almost-sure convergence. *IEEE Trans. Inform. Theory*, 60(9):5716–5735, 2014.
- [31] V.N. Vapnik. *Statistical learning theory*, volume 2. Wiley, New York, 1998.
- [32] Y. Yao, L. Rosasco, and A. Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007.
- [33] Y. Ying and M. Pontil. Online gradient descent learning algorithms. *Foundations of Computational Mathematics*, 8(5):561–596, 2008.

A Some Special Cases of Interest

Finally, before proving our main results, we illustrate and discuss a few special instances of the considered setting and related quantities.

Linear and Functional Regression In classical linear regression, data are described by the following model

$$y_i = w_*^T x_i + \delta_i, \quad i = 1, \dots, n$$

where $\delta_i, i = 1, \dots, n$, are i.i.d. sample from a normal distribution and $w_*, x_1, \dots, x_n \in \mathbb{R}^d, d \in \mathbb{N}^*$. In fixed design regression, the inputs x_1, \dots, x_n are assumed to be fixed, while in random design regression they are random sample according to some fixed unknown distribution [28]. It is easy to see that this latter setting is a special case of the framework in the paper (indeed the analysis in the paper can be also adapted with minor modifications to the fixed design setting). The regression model can be further complicated assuming the function of interest to be non linear (while we might still restrict the search of a solution to linear estimators). This can be dealt with for example considering kernel methods as we discuss below. Another special case of the setting in the paper is that of functional regression, where the input points are assumed to be infinite dimensional objects, for example curves, and they are formally described as functions in a Hilbert space. Clearly also this example is subsumed as a special case of our setting.

Learning with Kernels The setting in the paper reduces to nonparametric learning in RKHS as a special case. Let $\Xi \times \mathbb{R}$ be a probability space with distribution μ , that be can seen as the input/output space. The goal is then to minimize the risk, that, considering the square loss function, is given by

$$\mathcal{E}(f) = \int (y - f(\xi))^2 d\mu(\xi, y) \quad (23)$$

and is well defined for all measurable functions. A common way to build an estimator is to consider a symmetric kernel $K : \Xi \times \Xi \rightarrow \mathbb{R}$ which is positive definite, that is for which the matrix with entries $K(\xi_i, \xi_j), i, j = 1 \dots n$, is positive semidefinite for all in $\xi_1, \dots, \xi_n \in \Xi, n \in \mathbb{N}^*$. Such a kernel defines a unique Hilbert space of function \mathcal{H}_K with inner product $\langle \cdot, \cdot \rangle_K$ and such that for all $\xi \in \Xi, K_\xi(\cdot) = K(\xi, \cdot) \in \mathcal{H}_K$ and the following reproducing property holds for all $f \in \mathcal{H}_K, f(\xi) = \langle f, K_\xi \rangle_K$. To see how this setting is subsumed by the one in the paper, it is useful to introduce the (feature) map $\Phi : \Xi \rightarrow \mathcal{H}_K$, where $\Phi(\xi) = K_\xi$, for $\xi \in \Xi$ and further consider $\bar{\Phi} : \Xi \times \mathbb{R} \rightarrow \mathcal{H}_K \times \mathbb{R}$, where $\bar{\Phi}(\xi, y) = (K_\xi, y)$, for $\xi \in \Xi$ and $y \in \mathbb{R}$. Assuming the kernel to be measurable, we can view $\bar{\Phi}$ as a random variable. If we denote its distribution on $\mathcal{H}_K \times \mathbb{R}$ by $\mu_{\bar{\Phi}}$, then we can then let $\mathcal{H} = \mathcal{H}_K$ and $\rho = \mu_{\bar{\Phi}}$. It is known that the functions in a RKHS a measurable provided that the kernel is measurable [28], hence if we consider the risk of a function $f \in \mathcal{H}_K$ we have

$$\int_{\Xi \times \mathbb{R}} (y - f(\xi))^2 d\mu(\xi, y) = \int_{\Xi \times \mathbb{R}} (y - \langle f, K_\xi \rangle_K)^2 d\mu(\xi, y) = \int_{\mathcal{H} \times \mathbb{R}} (y - \langle f, x \rangle)^2 d\rho(x, y),$$

where we made the change of variables $(K_\xi, y) = (x, y)$. As is well known in machine learning, we can view learning a function using a kernel as learning a linear function in suitable Hilbert space.

Integral and Covariance Operators. The operator L can be seen as an integral operator associated to a linear kernel and is closely related to the covariance operator, or rather the second moment operator defined by ρ . This connections allows to interpret Assumption 2 in terms of the principal components.

To see this note that, under Assumption 1 it is easy to see that S is bounded, and its adjoint is given by

$$S^* : L^2(\mathcal{H}, \rho_{\mathcal{H}}) \rightarrow \mathcal{H}, \quad S^* w = \int x f(x) d\rho_{\mathcal{H}}(x), \quad \forall f \in L^2(\mathcal{H}, \rho_{\mathcal{H}}).$$

Then a straightforward calculation shows that $L = SS^*$. Moreover we can define $T : L^2(\mathcal{H}, \rho_{\mathcal{H}}) \rightarrow \mathcal{H}$ as $T = S^*S$ and check that

$$Tw = \int x \otimes x d\rho_{\mathcal{H}}(x), \quad \forall w \in \mathcal{H},$$

where $x \otimes x = \langle x, \cdot \rangle x$, for all $x \in \mathcal{H}$. The operator T is the second moment operator associated to ρ and its eigenfunctions are the principal components. Under Assumption 1, the operators T, L are linear, positive, self-adjoint and trace class, S, S^* are bounded and Hilbert Schmidt, hence compact. The operators T, L have the same non zero eigenvalues $(\sigma_j)_j$ which are the square of the singular values of S . If we denote by $(v_j)_j$ the eigenfunctions of T , the eigenfunctions of L can be chosen to be $(u_j)_j$ with $u_j(x) = \sigma_j^{-1} \langle v_j, x \rangle_{\mathcal{H}}$, $\rho_{\mathcal{H}}$ -almost surely. This latter observation allows an interpretation of Condition (4). By considering higher fractional power we are essentially assuming that the regression function can be linearly approximated and its approximation can be effectively represented considering the principal components associated to large eigenvalues.

Binary Classification The results in the paper can be directly applied to binary classification. Indeed, in this setting the outputs are binary valued i.e. $\{-1, 1\}$ and the goal is to learn a classifier $c : \mathcal{H} \rightarrow \{-1, 1\}$ with small misclassification risk

$$R(c) = \mathbb{P}(c(X) \neq Y). \tag{24}$$

The above risk is minimized by the so called Bayes decision rule defined by $b_{\rho}(x) = \text{sign}(2\rho(1|x) - 1)$, $\rho_{\mathcal{H}}$ - almost surely, and where for $a \in \mathbb{R}$, $\text{sign}(a) = 1$, if $a \geq 1$ and $\text{sign}(a) = -1$ otherwise. A relaxation approach is usually considered to learn a classification rule, which is based on replacing the risk R with convex error functional defined over real valued functions, e.g. considering (23). A classification rule is then obtained by taking the sign.

So called comparison results quantify the cost of the relaxation. In particular, it is known that the following inequality related R and \mathcal{E} defined in (24), (23) respectively,

$$R(\text{sign}(f)) - R(b_{\rho}) \leq \sqrt{\mathcal{E}(f) - \mathcal{E}(f_{\rho})}$$

for all measurable functions f . The latter inequality allows to derive excess misclassification risk and can be improved under additional assumption. We refer to [32], for further details in this direction.

B Proofs

In this appendix we prove the main results. The proof is quite long, and will be given relying on a series of lemmas.

B.1 Preliminary Results

We collect very general results that will be applied to our setting.

Let \mathcal{B} be a normed space. For every $r \in \mathbb{N}$, let $A_r : \mathcal{B} \rightarrow \mathcal{B}$ be a linear operator, let $(B_r)_{r \in \mathbb{N}}$ be a sequence in \mathcal{B} , and define the sequence $(X_r)_{r \in \mathbb{N}}$ in \mathcal{B} recursively as

$$X_{r+1} = A_r X_r + B_r. \quad (25)$$

We repeatedly use the following well-known equality, which is valid for every $r \in \mathbb{N}^*$ and for every integer $s \leq r$,

$$X_r = \left(\prod_{i=s}^{r-1} A_i \right) X_s + \sum_{k=s}^{r-1} \left(\prod_{i=k+1}^{r-1} A_i \right) B_k. \quad (26)$$

We next state an auxiliary lemma, establishing the minimizing property of the gradient descent iteration also when the infimum is not attained. Despite the result is a basic property of a very classical algorithm, we were not able to find the proof of this fact. Convergence properties of gradient descent are usually studied in two settings: for differentiable functions (not necessarily convex) and for convex functions. In the first case, the typical results do not assume existence of a minimizer and establish convergence to zero of the gradient of the function [22]. In the convex setting, the minimizing property is established assuming the existence of a minimizer [22].

Lemma B.1 *Let \mathcal{H} be a Hilbert space, and $F : \mathcal{H} \rightarrow \mathbb{R}$ be a convex and differentiable function with β -Lipschitz continuous gradient. Let $v_0 \in \mathcal{H}$, let $(\eta)_{k \in \mathbb{N}}$ be such that, for every $k \in \mathbb{N}$, $\eta_k \in]0, 2/\beta[$ and define, for every $k \in \mathbb{N}$, $v_{k+1} = v_k - \eta_k \nabla F(v_k)$. Then*

$$(\forall u \in \mathcal{H}) \quad F(v_k) - F(u) \leq \frac{\|u - v_0\|_{\mathcal{H}}}{2 \sum_{j=0}^k \eta_j} \quad (27)$$

In particular, if $\sum_{k \in \mathbb{N}} \eta_k = +\infty$, $F(v_k) \rightarrow \inf F$.

Proof. Since F is convex and differentiable,

$$\begin{aligned} (\forall k \in \mathbb{N})(\forall u \in \mathcal{H}) \quad F(u) - F(v_k) &\geq \langle \nabla F(v_k), u - v_k \rangle_{\mathcal{H}} \\ &= \eta_k^{-1} \langle v_k - v_{k+1}, u - v_k \rangle_{\mathcal{H}}. \end{aligned} \quad (28)$$

Therefore,

$$\begin{aligned} (\forall k \in \mathbb{N}) \quad 2\eta_k(F(u) - F(v_k)) &\geq -2 \langle v_{k+1} - v_k, u - v_k \rangle_{\mathcal{H}} \\ &= \|u - v_{k+1}\|_{\mathcal{H}}^2 - \|v_{k+1} - v_k\|_{\mathcal{H}}^2 - \|u - v_k\|_{\mathcal{H}}^2 \\ &= \eta_k^2 \|\nabla F(v_k)\|_{\mathcal{H}}^2 + \|u - v_{k+1}\|_{\mathcal{H}}^2 - \|u - v_k\|_{\mathcal{H}}^2 \end{aligned} \quad (29)$$

Let $t \in \mathbb{N}$ and define $\sigma_t = \sum_{k=0}^t \eta_k$. Summing (29) for $k = 0, \dots, t$ we obtain

$$2\sigma_t F(u) - 2 \sum_{k=0}^t \eta_k F(v_k) \geq \sum_{k=0}^t \eta_k^2 \|\nabla F(v_k)\|_{\mathcal{H}}^2 + \|u - v_{t+1}\|_{\mathcal{H}}^2 - \|u - v_0\|_{\mathcal{H}}^2. \quad (30)$$

Using the Lipschitz continuity of the gradient of F (see [22, Equation (15) p.6]),

$$(\forall k \in \mathbb{N}) \quad F(v_k) - F(v_{k+1}) \geq \eta_k \left(1 - \frac{\eta_k \beta}{2} \right) \|\nabla F(v_k)\|_{\mathcal{H}}^2$$

Therefore,

$$(\forall k \in \mathbb{N}) \quad \sigma_k F(v_k) - \sigma_{k+1} F(v_{k+1}) + \eta_{k+1} F(v_{k+1}) \geq \sigma_k \eta_k \left(1 - \frac{\eta_k \beta}{2}\right) \|\nabla F(v_k)\|_{\mathcal{H}}^2 \quad (31)$$

Summing (31) for $k = 0, \dots, t-1$ we get, for every $t \in \mathbb{N}$

$$-\sigma_t F(v_t) + \sum_{k=0}^t \eta_k F(v_k) \geq \sum_{k=0}^{t-1} \sigma_k \eta_k \left(1 - \frac{\eta_k \beta}{2}\right) \|\nabla F(v_k)\|_{\mathcal{H}}^2 \quad (32)$$

Adding (32) to (30) we get, for every $u \in \mathcal{H}$

$$2\sigma_t(F(u) - F(v_t)) \geq \sum_{k=0}^t \eta_k^2 \|\nabla F(v_k)\|_{\mathcal{H}} + \|u - v_{t+1}\|_{\mathcal{H}}^2 - \|u - v_0\|_{\mathcal{H}}^2 + 2 \sum_{k=0}^{t-1} \sigma_k \eta_k \left(1 - \frac{\eta_k \beta}{2}\right) \|\nabla F(v_k)\|_{\mathcal{H}}^2$$

and hence,

$$(\forall u \in \mathcal{H}) \quad F(v_t) - F(u) \leq \frac{\|u - v_0\|_{\mathcal{H}}^2}{2\sigma_t}.$$

□

We next recall a probabilistic inequality for martingales [21, Theorem 3.4] (see also [30, Lemma A.1 and Corollaries A.2 and A.3]).

Theorem B.2 *Let $(\xi_i, \mathcal{F}_i)_{1 \leq i \leq n}$ be an adapted family of random vectors taking values in a Hilbert space with norm $\|\cdot\|$, such that $\mathbb{E}[\xi_i | \mathcal{F}_{i-1}] = 0$ a.s. Assume that there exist $M \in \mathbb{R}_{++}$ such that $\|\xi_i\| \leq M$. Then, for every $\delta \in]0, 1[$ the following holds*

$$\mathbb{P} \left(\left\{ \sup_{1 \leq j \leq n} \left\| \frac{1}{n} \sum_{i=1}^j \xi_i \right\| \leq \frac{8M}{3\sqrt{n}} \log \frac{2}{\delta} \right\} \right) \geq 1 - \delta.$$

B.2 Proof of STEP 1

Here we first introduce a recursive expression which is satisfied by the sequence $(\hat{w}_t)_{t \in \mathbb{N}}$, that allows to interpret the incremental gradient iteration as a gradient descent iteration with errors. To do so, we start by introducing some further notation and then show that the iteration presented in (6)-(7) results from the application of the incremental gradient method to the empirical risk.

Let $x \in \mathcal{H}$ and define $S_x : \mathcal{H} \rightarrow \mathbb{R}$, by setting $S_x w = \langle w, x \rangle_{\mathcal{H}}$. Then S_x is a bounded linear operator and $\|S_x\| \leq \|x\|_{\mathcal{H}}$. Using this linear operator, and the operator S introduced in Assumption 2, Problems (1) and (8) can be expressed as convex quadratic minimization problems. The empirical risk can be written as

$$\hat{\mathcal{E}}(w) = \frac{1}{n} \sum_{i=1}^n (S_{x_i} w - y_i)^2, \quad (33)$$

and recalling Assumption 2 and (19), we have

$$\mathcal{E}(w) = \int_{\mathcal{H} \times \mathbb{R}} (Sw(x) - y)^2 d\rho(x, y) = \|Sw - g_\rho\|_\rho^2 + \inf_{\mathcal{H}} \mathcal{E}. \quad (34)$$

Define the operators

$$T: \mathcal{H} \rightarrow \mathcal{H}: T = S^*S, \quad \text{and} \quad (\forall x \in \mathcal{X}) \quad T_x: \mathcal{H} \rightarrow \mathcal{H}: T_x = S_x S_x^*. \quad (35)$$

Then, computing the gradients of \mathcal{E} and $\hat{\mathcal{E}}$ respectively, (6)-(7) can be rewritten as

$$\hat{u}_t^0 = \hat{w}_t; \quad \hat{u}_t^i = \hat{u}_t^{i-1} - \frac{\gamma}{n}(T_{x_i} \hat{u}_t^{i-1} - S_{x_i}^* y_i), \quad i = 1, \dots, n \quad (36)$$

and (17)-(18) can be expressed as

$$u_t^0 = w_t; \quad u_t^i = u_t^{i-1} - \frac{\gamma}{n}(T u_t^{i-1} - S^* g_\rho). \quad (37)$$

It is apparent from (36) that the considered iteration is derived from the application of the incremental gradient algorithm to the empirical error (see [3, 17]). At the same time (37) shows that iteration (17)-(18) can be seen as the result of applying the incremental gradient descent algorithm to the expected loss, which clearly, for fixed n , can be written as

$$w \mapsto \frac{1}{n} \sum_{i=1}^n \mathcal{E}(w).$$

Lemma B.3 *Let $t \in \mathbb{N}$, and let w_t be defined as in (17)-(18), with $w_0 = 0$. Let $\eta = \gamma/n$. Then*

$$w_t = \eta \sum_{j=0}^{nt-1} (I - \eta T)^{nt-j-1} S^* g_\rho. \quad (38)$$

Proof. Let, for every $k \in \mathbb{N}$,

$$v_{k+1} = (I - \eta T)v_k + \eta S^* g_\rho.$$

Then, by (37), $w_t = v_{nt}$, and (38) follows. \square

In other words, the statement of Lemma B.3 states that the t -th epoch of the incremental gradient descent iteration in (17)-(18). coincides with n steps of gradient descent with stepsize γ/n .

Next, we relate the iteration (6)-(7) to the gradient descent iteration on the empirical error. These will be used in the error analysis and provide some useful comparison between these two methods. Hereafter, \hat{T} is the operator $(\sum_{i=1}^n T_{x_i})/n$.

The following lemma provides an alternative expression for the composition of linear operators.

Lemma B.4 *Let $n \in \mathbb{N}^*$, let $(T_i)_{1 \leq i \leq n}$ be a family of linear operators from \mathcal{H} to \mathcal{H} , and let $(w_i)_{1 \leq i \leq n} \subset \mathcal{H}^n$. Then*

$$\prod_{i=1}^n (I - T_i) = I - \sum_{j=1}^n T_j + \sum_{k=2}^n \left(\prod_{i=k+1}^n (I - T_i) \right) T_k \sum_{j=1}^{k-1} T_j \quad (39)$$

and

$$\sum_{i=1}^n \left(\prod_{k=i+1}^n (I - T_k) \right) w_i = \sum_{i=1}^n w_i - \sum_{k=2}^n \left(\prod_{i=k+1}^n (I - T_i) \right) T_k \sum_{j=1}^{k-1} w_j \quad (40)$$

Proof. By induction. Equality (39) is trivially satisfied for $n = 1$. Suppose now that $n \geq 2$, and that (39) holds for $n - 1$. Then

$$\begin{aligned}
\prod_{i=1}^n (I - T_i) &= (I - T_n) \prod_{i=1}^{n-1} (I - T_i) \\
&= (I - T_n) \left(I - \sum_{j=1}^{n-1} T_j + \sum_{k=2}^{n-1} \left(\prod_{i=k+1}^{n-1} (I - T_i) \right) T_k \sum_{j=1}^{k-1} T_j \right) \\
&= I - \sum_{j=1}^n T_j + T_n \sum_{j=1}^{n-1} T_j + \sum_{k=2}^{n-1} \left(\prod_{i=k+1}^n (I - T_i) \right) T_k \sum_{j=1}^{k-1} T_j \\
&= I - \sum_{j=1}^n T_j + \sum_{k=2}^n \left(\prod_{i=k+1}^n (I - T_i) \right) T_k \sum_{j=1}^{k-1} T_j,
\end{aligned}$$

and the validity of (39) for every $n \in \mathbb{N}^*$ follows by induction.

Equality (40) is trivially satisfied for $n = 1$. Suppose now that $n \geq 2$ and that (39) holds for $n - 1$. Then

$$\begin{aligned}
\sum_{i=1}^n \left(\prod_{k=i+1}^n (I - T_k) \right) w_i &= \sum_{i=1}^{n-1} \left(\prod_{k=i+1}^n (I - T_k) \right) w_i + w_n \\
&= (I - T_n) \left(\sum_{i=1}^{n-1} w_i - \sum_{k=2}^{n-1} \left(\prod_{i=k+1}^{n-1} (I - T_i) \right) T_k \sum_{j=1}^{k-1} w_j \right) + w_n \\
&= \sum_{i=1}^n w_i - T_n \sum_{j=1}^{n-1} w_j - \sum_{k=2}^{n-1} \left(\prod_{i=k+1}^n (I - T_i) \right) T_k \sum_{j=1}^{k-1} w_j \\
&= \sum_{i=1}^n w_i - \sum_{k=2}^n \left(\prod_{i=k+1}^n (I - T_i) \right) T_k \sum_{j=1}^{k-1} w_j,
\end{aligned}$$

and the conclusion follows. \square The following lemma establishes an equivalent expression for the iterates (6)-(7).

Lemma B.5 *Let $t \in \mathbb{N}$. Then,*

$$\hat{w}_{t+1} = \prod_{i=1}^n \left(I - \frac{\gamma}{n} T_{x_i} \right) \hat{w}_t + \frac{\gamma}{n} \sum_{i=1}^n \prod_{k=i+1}^n \left(I - \frac{\gamma}{n} T_{x_k} \right) S_{x_i}^* y_i \quad (41)$$

Proof. For every $i \in \{1, \dots, n\}$, the update of \hat{v}_t^i in (7) can be equivalently written as in (36), and is of the form (25), with $A_r = I - (\gamma/n)T_{x_{r+1}}$, and $B_r = (\gamma/n)S_{x_{r+1}}^* y_{r+1}$, for every $r \in \{0, \dots, n-1\}$, and $X_0 = \hat{w}_t$. Equation (41) follows by writing (26) for $r = n-1$. \square

Proposition B.6 *For every $t \in \mathbb{N}$, the iteration (6)-(7) can be written as*

$$\hat{w}_{t+1} = (I - \gamma \hat{T}) \hat{w}_t + \gamma \left(\frac{1}{n} \sum_{j=1}^n S_{x_j}^* y_j \right) + \gamma^2 \left(\hat{A} \hat{w}_t - \hat{b} \right), \quad (42)$$

with

$$\hat{A} = \frac{1}{n^2} \sum_{k=2}^n \prod_{i=k+1}^n \left(I - \frac{\gamma}{n} T_{x_i} \right) T_{x_k} \sum_{j=1}^{k-1} T_{x_j}, \quad \hat{b} = \frac{1}{n^2} \sum_{k=2}^n \prod_{i=k+1}^n \left(I - \frac{\gamma}{n} T_{x_i} \right) T_{x_k} \sum_{j=1}^{k-1} S_{x_j}^* y_j. \quad (43)$$

The iteration (17)-(18) applied to \mathcal{E} can be expressed as

$$w_{t+1} = (I - \gamma T) w_t + \gamma S^* g_\rho + \gamma^2 (A w_t - b). \quad (44)$$

with

$$A = \frac{1}{n^2} \sum_{k=2}^n \left[\prod_{i=k+1}^n \left(I - \frac{\gamma}{n} T \right) \right] T \sum_{j=1}^{k-1} T, \quad b = \frac{1}{n^2} \sum_{k=2}^n \left[\prod_{i=k+1}^n \left(I - \frac{\gamma}{n} T \right) \right] T \sum_{j=1}^{k-1} S^* g_\rho. \quad (45)$$

Proof. Equations (42) and (43) follow from Lemma B.4 and Lemma B.5 applied with $(\forall i \in \{1, \dots, n\}) T_i = (\gamma/n) T_{x_i}$. Equations (44) and (45) follow from Lemma B.4 and Lemma B.5 applied with $(\forall i \in \{1, \dots, n\}) T_i = (\gamma/n) T$. \square

Note that, although not explicitly specified, the operator A and the element $b \in \mathcal{H}$ in Proposition B.6 depend on n . Equation (42) allows to compare the update resulting from one epoch of the iteration (6)-(7) with the one of a standard gradient descent on the empirical error with stepsize γ , which is given by

$$\hat{v}_{t+1} = \left(I - \frac{\gamma}{n} \sum_{j=1}^n T_{x_j} \right) \hat{v}_t + \frac{\gamma}{n} \sum_{j=1}^n S_{x_j}^* y_j \quad (46)$$

for an arbitrary $\hat{v}_0 \in \mathcal{H}$. As can be seen comparing (42) and (46), In particular, the incremental gradient descent can be interpreted as a perturbed gradient descent step, with perturbation

$$\hat{e}_t = \gamma^2 \left(\hat{A} \hat{w}_t - \hat{b} \right).$$

B.3 Proof of STEP 2

The following recursive expression is key to get sample bounds estimates, and is at the basis of Lemma B.9.

Lemma B.7 *In the setting of Section 2, let $t \in \mathbb{N}$ and let \hat{w}_t and w_t be defined as in (6)-(7) and w_t be defined as (17)-(18), respectively. Define \hat{A} and \hat{b} as in (43), and A and b as in (45). Then*

$$\hat{w}_t - w_t = \left(I - \gamma \hat{T} + \gamma^2 \hat{A} \right)^t (\hat{w}_0 - w_0) + \gamma \sum_{k=0}^{t-1} \left(I - \gamma \hat{T} + \gamma \hat{A} \right)^{t-k-1} \zeta_k \quad (47)$$

with

$$\zeta_k = (T - \hat{T}) w_k + \gamma (\hat{A} - A) w_k + \left(\frac{1}{n} \sum_{i=1}^n \hat{S}_{x_i}^* y_i - S^* g_\rho \right) + \gamma (b - \hat{b}). \quad (48)$$

Proof. We have

$$\hat{w}_{t+1} = (I - \gamma \hat{T} + \gamma^2 \hat{A}) \hat{w}_t + \frac{\gamma}{n} \sum_{i=1}^n \hat{S}_{x_i}^* y_i - \gamma^2 \hat{b}. \quad (49)$$

Adding and subtracting $(-\gamma\hat{T} + \gamma^2\hat{A})w_t$ it follows from (44) that

$$\begin{aligned}\hat{w}_{t+1} - w_{t+1} &= (I - \gamma\hat{T} + \gamma^2\hat{A})(\hat{w}_t - w_t) + \gamma(T - \hat{T})w_t + \gamma^2(\hat{A} - A)w_t \\ &\quad + \gamma\left(\frac{1}{n}\sum_{i=1}^n \hat{S}_{x_i}^* y_i - S^* g_\rho\right) + \gamma^2(b - \hat{b})\end{aligned}$$

Relying on equation (25) we get (53). \square

B.4 Proof of STEP 3

Next we provide a lemma to bound the norm of the operator appearing in (53), and acting on the random variable ζ_k in (54).

Lemma B.8 *In the setting of Section 2, let $\gamma \in]0, n\kappa^{-1}]$. Then*

$$\|I - \gamma\hat{T} + \gamma^2\hat{A}\| \leq 1. \quad (50)$$

Proof. It follows from Lemma B.4, equation (39) applied with $T_i = T_{x_i}$ and the definition of \hat{A}_k in (43) that

$$I - \gamma\hat{T} + \gamma^2\hat{A} = \prod_{i=1}^n \left(I - \frac{\gamma}{n} T_{x_i}\right). \quad (51)$$

Since $\|T_{x_i}\| \leq \kappa$ and by assumption $\gamma/n \leq \kappa^{-1}$, $\|I - (\gamma/n)T_{x_i}\| \leq 1$ and the statement follows. \square

We next provide a first inequality for the sample error.

Lemma B.9 *Let $t \in \mathbb{N}$ and let \hat{w}_t and w_t be defined as in (6)-(7) and w_t be defined as (17)-(18), respectively. Define \hat{A} and \hat{b} as in (43), and A and b as in (45). Then*

$$\|\hat{w}_t - w_t\|_{\mathcal{H}} \leq \gamma(\|T - \hat{T}\| + \gamma\|\hat{A} - A\|) \sum_{k=0}^{t-1} \|w_k\|_{\mathcal{H}} + \gamma t \left(\left\| \frac{1}{n} \sum_{i=1}^n \hat{S}_{x_i}^* y_i - S^* g_\rho \right\| + \gamma\|b - \hat{b}\| \right). \quad (52)$$

Proof. By Lemma B.7 we derive that

$$\hat{w}_t - w_t = \left(I - \gamma\hat{T} + \gamma^2\hat{A}\right)^t (\hat{w}_0 - w_0) + \gamma \sum_{k=0}^{t-1} \left(I - \gamma\hat{T} + \gamma\hat{A}\right)^{t-k-1} \zeta_k \quad (53)$$

with

$$\zeta_k = (T - \hat{T})w_k + \gamma(\hat{A} - A)w_k + \left(\frac{1}{n}\sum_{i=1}^n \hat{S}_{x_i}^* y_i - S^* g_\rho\right) + \gamma(b - \hat{b}). \quad (54)$$

Since $w_0 = \hat{w}_0$, we have

$$\|\hat{w}_t - w_t\|_{\mathcal{H}} \leq \gamma \sum_{k=0}^{t-1} \|I - \gamma\hat{T} + \gamma^2\hat{A}\|^{k+1} \|\zeta_k\| \quad (55)$$

By Lemma B.8, for every $j \in \mathbb{N}$, $\|I - \gamma\hat{T} + \gamma^2\hat{A}\| \leq 1$, and therefore

$$\|\hat{w}_t - w_t\|_{\mathcal{H}} \leq \gamma(\|T - \hat{T}\| + \gamma\|\hat{A} - A\|) \sum_{k=0}^{t-1} \|w_k\| + \left(\gamma\left\|\frac{1}{n}\sum_{i=1}^n \hat{S}_{x_i}^* y_i - S^* g_\rho\right\| + \gamma^2\|b - \hat{b}\|\right)t$$

\square

B.5 Proof of STEP 4

Here, we provide a bound for $\|w_t\|_{\mathcal{H}}$. The proof technique is similar to that of known results in inverse problems [14], although the bound obtained in (56) is novel.

Lemma B.10 *In the setting of Section 2, let Assumption 1 hold, let $n \in \mathbb{N}^*$, and let $\gamma \in]0, n\kappa^{-1}[$. Let $w_0 = 0$, and let $t \in \mathbb{N}$. Then the following hold:*

(i) *Let Assumption 2 hold with $r \in [0, 1/2]$. Then*

$$\|w_t\|_{\mathcal{H}} \leq \max\{\kappa^{r-1/2}, (\gamma t)^{1/2-r}\} \|g\|_{\rho}. \quad (56)$$

(ii) *Suppose that \mathcal{O} is nonempty. Then there exist w^\dagger as in (2) and $\beta \in \mathbb{R}_{++}$ such that*

$$\|w_{t+1}\|_{\mathcal{H}} \leq \beta.$$

(iii) *Let Assumption 2 hold with $r \in]1/2, +\infty[$. Then w^\dagger is well defined and*

$$\|w_{t+1}\|_{\mathcal{H}} \leq \kappa^{r-1/2} \|g\|_{\rho}. \quad (57)$$

Proof. Let $\epsilon \in]0, \kappa]$ and set $\eta = \gamma/n$. By Lemma B.3 and by the spectral theorem [14, equation (2.43)], we derive

$$w_{t+1} = \sum_{j=0}^{nt-1} S^* \eta \prod_{i=j+1}^{nt-1} (I - \eta L) g_{\rho}. \quad (58)$$

(i): By (58)

$$\begin{aligned} \|w_{t+1}\|_{\mathcal{H}} &\leq \eta \left\| S^* L^r \sum_{j=0}^{nt-1} (I - \eta L)^{nt-j+1} \right\| \|g\|_{\rho} \\ &\leq \sup_{\sigma \in [0, \kappa]} \sigma^{1/2+r} \left| \sum_{j=0}^{nt-1} \eta (1 - \eta \sigma)^{nt-j+1} \right| \|g\|_{\rho} \\ &= \max \left\{ \sup_{\sigma \in [0, \epsilon]} nt \eta \sigma^{1/2+r}, \sup_{\sigma \in [\epsilon, \kappa]} \sigma^{r-1/2} \left(1 - (1 - \eta \sigma)^{nt} \right) \right\} \|g\|_{\rho} \\ &\leq \psi(\epsilon) \|g\|_{\rho} \end{aligned} \quad (59)$$

where $(\forall \epsilon \in \mathbb{R}_+) \psi(\epsilon) = \max \left\{ \epsilon^{1/2+r} nt \eta, \epsilon^{r-1/2} \right\}$. Since ϵ is arbitrary,

$$\|w_{t+1}\|_{\mathcal{H}} \leq \inf_{\epsilon \in]0, \kappa]} \psi(\epsilon) \|g\|_{\rho}. \quad (60)$$

Now note that $\epsilon \in]0, +\infty] \mapsto nt \eta \epsilon^{1/2+r}$ is strictly increasing, and has limit equal to zero at zero. On the contrary, $\epsilon \in]0, +\infty] \mapsto \epsilon^{r-1/2}$ is strictly decreasing and $\lim_{\epsilon \rightarrow 0^+} \epsilon^{r-1/2} = +\infty$. Hence, there exists a unique point $\bar{\epsilon} \in]0, +\infty]$ such that

$$nt \eta \bar{\epsilon}^{1/2+r} = \bar{\epsilon}^{r-1/2} \quad \text{and} \quad \psi(\epsilon) = \begin{cases} \epsilon^{r-1/2} & \text{if } \epsilon \in]0, \bar{\epsilon}] \\ nt \eta \epsilon^{1/2+r} & \text{if } \epsilon \in [\bar{\epsilon}, +\infty[\end{cases}, \quad (61)$$

therefore $\bar{\epsilon}$ is the unique minimizer of ψ . Solving for $\bar{\epsilon}$ in (61), we get $\bar{\epsilon} = (nt\eta)^{-1}$. We derive, again from (61), that

$$\min_{\epsilon \in]0, \kappa]} \psi(\epsilon) = \max\{\kappa^{r-1/2}, (\gamma t)^{1/2-r}\}.$$

Finally, (60) yields

$$\|w_{t+1}\|_{\mathcal{H}} \leq (\gamma t)^{1/2-r} \|g\|_{\rho}.$$

(ii): First note that, by Fermat's rule, $S^*g_{\rho} = Tw^{\dagger}$. It follows from Lemma B.3 that

$$\begin{aligned} w_t - w^{\dagger} &= \left(\sum_{j=0}^{nt-1} \eta T (I - \eta T)^{nt-j+1} - I \right) w^{\dagger} \\ &= (I - \eta T)^{nt} w^{\dagger}. \end{aligned}$$

Let $(\sigma_m, h_m)_{m \in \mathbb{N}}$ be an eigensystem of T . Since $w^{\dagger} \in N(T)^{\perp}$ (see [14, Proposition 2.3]), it follows that $w^{\dagger} = \sum_{m \in \mathbb{N}} \langle w^{\dagger}, h_m \rangle h_m$. Therefore,

$$\|w_t - w^{\dagger}\|_{\mathcal{H}}^2 = \sum_{m \in \mathbb{N}} \left| (1 - \eta \sigma_m)^{nt} \langle w^{\dagger}, h_m \rangle \right|^2 \quad (62)$$

Since, for every $m \in \mathbb{N}$, each summand is bounded by $|\langle w^{\dagger}, h_m \rangle|^2$, and $\sum_{m \in \mathbb{N}} |\langle w^{\dagger}, h_m \rangle|^2$, the Dominated Convergence Theorem yields

$$\lim_{t \rightarrow +\infty} \|w_t - w^{\dagger}\|_{\mathcal{H}}^2 = \sum_{m \in \mathbb{N}} \lim_{t \rightarrow +\infty} \left| (1 - \eta \sigma_m)^{nt} \langle w^{\dagger}, h_m \rangle \right|^2 = 0. \quad (63)$$

Hence, the sequence $(\|w_t\|_{\mathcal{H}})_{t \in \mathbb{N}}$ is bounded.

(iii): Arguing as in the proof of (i), it follows from (58) and Assumption 2 that

$$\begin{aligned} \|w_{t+1}\|_{\mathcal{H}} &\leq \sup_{\sigma \in [0, \kappa]} \sigma^{r-1/2} \left(1 - (1 - \eta \sigma)^{nt} \right) \|g\|_{\rho} \\ &\leq \kappa^{r-1/2} \|g\|_{\rho} \end{aligned}$$

□

B.6 Proof of STEP 5

The main novel probabilistic estimates are given in the following proposition. This is the more involved part of the proof, where many tricks are needed in order to get a manageable expression. The proof is based on writing the terms $\hat{A} - A$ and $\hat{b} - b$ as a sum of martingales and then apply Theorem B.2 to derive concentration inequalities. We start with the following well-known lemma, which is a direct consequence of Theorem B.2 (see also [12]).

Lemma B.11 *In the Setting of Section 2, let Assumption 1 hold. For every $\delta \in]0, 1]$*

$$\mathbb{P} \left(\left\| \frac{1}{n} \sum_{i=1}^n T_{x_i} - T \right\|_{HS} \leq \frac{16\kappa}{3\sqrt{n}} \log \frac{2}{\delta} \right) \geq 1 - \delta, \quad (64)$$

and

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^n S_{x_i}^* y_i - S^* g_\rho\right\|_{\mathcal{H}} \leq \frac{16\sqrt{\kappa}M}{3\sqrt{n}} \log \frac{2}{\delta}\right) \geq 1 - \delta. \quad (65)$$

Proof. Equation 64 follows from Theorem B.2, since $(T_{x_i} - T)_{1 \leq i \leq n}$ is a family of i.i.d. random operators taking values in the space of Hilbert-Schmidt operators satisfying $\|T\|_{HS} \leq \kappa$ and $\|T_{x_i}\|_{HS} \leq \kappa$ (see also [12]). Equation 65 follows from Theorem B.2 applied to the i.i.d. random vectors $(S_{x_i}^* y_i - S^* f_\rho)_{1 \leq i \leq n}$ in \mathcal{H} whose norms are bounded by $2\kappa M$. \square

Proposition B.12 *In the Setting of Section 2, let Assumption 1 hold, let $\gamma \in]0, n\kappa^{-1}[$, and let $\delta \in]0, 1[$. Then*

$$\mathbb{P}\left(\|\hat{A} - A\|_{HS} \leq \frac{32\kappa^2}{3\sqrt{n}} \log \frac{4}{\delta}\right) \geq 1 - \delta, \quad (66)$$

and

$$\mathbb{P}\left(\|\hat{b} - b\|_{\mathcal{H}} \leq \frac{32\kappa M^2}{3\sqrt{n}} \log \frac{4}{\delta}\right) \geq 1 - \delta. \quad (67)$$

Proof. We first show a useful decomposition. Recall that

$$\hat{A} = \frac{1}{n} \sum_{j=2}^n \frac{1}{n} \prod_{i=j+1}^n (I - \gamma T_{x_i}) T_{x_j} \sum_{l=1}^{j-1} T_{x_l} \quad A = \frac{1}{n} \sum_{j=2}^n \frac{1}{n} \prod_{i=j+1}^n (I - \gamma T) T \sum_{l=1}^{j-1} T.$$

For every $j \in \{2, \dots, n\}$, set

$$\hat{B}_j = \left[\prod_{i=j+1}^n (I - \gamma T_{x_i}) \right] T_{x_j}, \quad B_j = \left[\prod_{i=j+1}^n (I - \gamma T) \right] T$$

we have

$$\begin{aligned} \hat{A} - A &= \frac{1}{n} \sum_{j=2}^n \hat{B}_j \left(\frac{1}{n} \sum_{l=1}^{j-1} T_{x_l} \right) - \frac{1}{n} \sum_{j=2}^n B_j \frac{j-1}{n} T \\ &= \frac{1}{n} \left[\sum_{j=2}^n \hat{B}_j \left(\frac{1}{n} \sum_{l=1}^{j-1} (T_{x_l} - T) \right) + Q_n T \right], \end{aligned} \quad (68)$$

with

$$Q_n = \sum_{j=2}^n (\hat{B}_j - B_j). \quad (69)$$

We next bound each term appearing in (68). By Lemma B.11, with probability greater than $1 - \delta$,

$$\sup_{j \in \{1, \dots, n\}} \left\| \frac{1}{n} \sum_{l=1}^{j-1} (T_{x_l} - T) \right\|_{HS} \leq \frac{16\kappa}{3\sqrt{n}} \log \frac{2}{\delta}. \quad (70)$$

On the other hand

$$\|\hat{B}_j\| \leq \prod_{i=j+1}^n \left\| I - \frac{\gamma}{n} T_{x_i} \right\| \|T_{x_j}\| \leq \kappa. \quad (71)$$

Note that $\sum_{j=2}^n \hat{B}_j \left(\frac{1}{n} \sum_{l=1}^{j-1} (T_{x_l} - T) \right)$ is Hilbert-Schmidt, for T_{x_l} and T are Hilbert-Schmidt operators, with $\|T_{x_l}\|_{HS} \leq \kappa$ and $\|T\|_{HS} \leq \kappa$, and the family of Hilbert-Schmidt operators is an ideal with respect to the composition in $L(\mathcal{H})$. Therefore, by (70) and (71),

$$\frac{1}{n} \left\| \sum_{j=2}^n \hat{B}_j \left(\frac{1}{n} \sum_{l=1}^{j-1} (T_{x_l} - T) \right) \right\|_{HS} \leq \frac{16\kappa^2}{3\sqrt{n}} \log \frac{2}{\delta} \quad (72)$$

holds with probability greater than $1 - \delta$, for any $\delta \in]0, 1[$. Next we write the quantity Q_n appearing in the second term in (68) as the sum of a martingale. For short, we set $\eta = \frac{\gamma}{n}$ and for all $j \in \{2, \dots, n\}$ we denote

$$\hat{\Pi}_j^n = \prod_{i=j+1}^n (I - \eta T_{x_i}), \quad \Pi_j^n = \prod_{i=j+1}^n (I - \eta T),$$

so that from the definition of Q_n in (69),

$$Q_n = \sum_{j=2}^n (\hat{\Pi}_j^n T_{x_j} - \Pi_j^n T).$$

We can derive a recursive update that determine a different expression for the quantity Q_n as follows. Let $s \in \{1, \dots, n-1\}$.

$$\begin{aligned} Q_{s+1} &= \sum_{j=2}^{s+1} (\hat{\Pi}_j^{s+1} T_{x_j} - \Pi_j^{s+1} T) \\ &= (T_{x_{s+1}} - T) + \sum_{j=2}^s (\hat{\Pi}_j^{s+1} T_{x_j} - \Pi_j^{s+1} T) \\ &= (T_{x_{s+1}} - T) + \sum_{j=2}^s ((I - \eta T_{x_{s+1}}) \hat{\Pi}_j^s T_{x_j} - (I - \eta T) \Pi_j^s T) \\ &= (T_{x_{s+1}} - T) + (I - \eta T_{x_{s+1}}) \sum_{j=2}^s (\hat{\Pi}_j^s T_{x_j} - \Pi_j^s T) + \eta (T - T_{x_{s+1}}) \sum_{j=2}^s \Pi_j^s T \\ &= (I - \eta T_{x_{s+1}}) Q_s + (T_{x_{s+1}} - T) \left(I - \eta \sum_{j=2}^n \Pi_j^s T \right). \end{aligned}$$

Applying equation (25), since $Q_1 = 0$, we get

$$Q_n = \sum_{l=1}^n \Theta_l \quad (73)$$

where, $\Theta_1 = 0$ and, for every $l \in \{2, \dots, n\}$,

$$\Theta_l = \prod_{i=l+1}^n \left(I - \frac{\gamma}{n} T_{x_i} \right) (T_{x_l} - T) \left(I - \frac{\gamma}{n} \sum_{j=2}^{l-1} \prod_{i=j+1}^{l-1} \left(I - \frac{\gamma}{n} T \right) T \right).$$

For every $l = 1, \dots, n$

$$\mathbb{E}[\Theta_l] = 0,$$

being T_{x_2}, \dots, T_{x_n} independent and $\mathbb{E}[(T_{x_l} - T)] = 0$. Moreover the conditional expectation

$$\mathbb{E}[\Theta_l \mid \Theta_{l+1}, \dots, \Theta_n] = 0,$$

since T_{x_l} is independent from $T_{x_{l+1}}, \dots, T_{x_n}$. Therefore the sequence $(\Theta_l)_{1 \leq l \leq n}$ is a martingale difference sequence. The operator Θ_l is Hilbert-Schmidt, since it is the composition of a Hilbert-Schmidt operator with a continuous one. Moreover, $\|\Theta_1\| = 0$. Next, since the operator T is compact and self-adjoint and $0 \leq \gamma/n \leq 1/\|T\|$, from the spectral mapping theorem, for every $l \in \{2, \dots, n\}$

$$\left\| I - \frac{\gamma}{n} \sum_{j=2}^{l-1} \prod_{i=j+1}^{l-1} \left(I - \frac{\gamma}{n} T \right) T \right\| = \sup_{x \in [0,1]} \left| 1 - \sum_{j=2}^{l-1} x(1-x)^{l-j-1} \right|.$$

We have

$$0 \leq \sum_{j=2}^{l-1} x(1-x)^{l-j-1} = \sum_{j=2}^{l-1} \left((1-x)^{l-j-1} - (1-x)^{l-j} \right) = 1 - (1-x)^{l-2} \leq 1$$

Therefore

$$\left\| I - \frac{\gamma}{n} \sum_{j=2}^{l-1} \prod_{i=j+1}^{l-1} \left(I - \frac{\gamma}{n} T \right) T \right\| \leq 1.$$

Using the last inequality, we derive

$$\begin{aligned} \|\Theta_l\|_{HS} &\leq \left\| \prod_{i=l+1}^n \left(I - \frac{\gamma}{n} T_{x_i} \right) \right\| \|T_{x_l} - T\|_{HS} \left\| I - \frac{\gamma}{n} \sum_{j=2}^{l-1} \prod_{i=j+1}^{l-1} \left(I - \frac{\gamma}{n} T \right) T \right\| \\ &\leq \|T_{x_l} - T\|_{HS} \\ &\leq 2\kappa. \end{aligned}$$

Then, Theorem B.2 applied to $(\Theta_l)_{i \leq l \leq n}$, yields

$$\left\| \frac{1}{n} \sum_{l=1}^n \Theta_l \right\|_{HS} \leq \frac{16\kappa}{3\sqrt{n}} \log \frac{2}{\delta} \quad (74)$$

with probability greater than $1 - \delta$. Therefore, with probability greater than $1 - \delta$

$$\frac{1}{n} \left\| Q_n T \right\|_{HS} \leq \frac{16\kappa^2}{3\sqrt{n}} \log \frac{2}{\delta}. \quad (75)$$

The statement then follows recalling the decomposition in (68), and summing (75) with (72). From the definition of \hat{b} and b in equations (42) and (45) respectively, we have

$$\hat{b} - b = \frac{1}{n^2} \sum_{j=2}^n \hat{B}_j \sum_{l=1}^{j-1} S_{x_l}^* y_l - \frac{1}{n^2} \sum_{j=2}^n B_j \sum_{l=1}^{j-1} S^* g_\rho, \quad (76)$$

and equation (67) follows reasoning as in the previous part of the proof. \square

B.7 Sample error

The proof of the bound on the sample error easily follows from the above results.

Theorem B.13 (Sample error) *Let Assumption 1 hold. Let $n \in \mathbb{N}^*$, suppose that $\gamma \in]0, n\kappa^{-1}]$, and let $\hat{w}_0 = w_0 = 0$. Let $\delta \in]0, 1[$, and, for every $t \in \mathbb{N}^*$, let \hat{w}_t and w_t be defined as in (6)-(7) and (17)-(18), respectively. Then the following hold:*

- (i) *Let Assumption 2 hold, for some $r \in [0, 1/2]$, and let $t \in \mathbb{N}^*$. Then, with probability greater than $1 - \delta$*

$$\begin{aligned} \|\hat{w}_t - w_t\|_{\mathcal{H}} \leq & \frac{\log(16/\delta)}{3\sqrt{n}} \left[(16\sqrt{\kappa}M + 32\kappa M^2\gamma)\gamma t \right. \\ & \left. + (16\kappa + 32\kappa^2\gamma)\|g\|_{\rho} \max \left\{ \kappa^{r-1/2}\gamma t, \left(\frac{1-2r}{3-2r} + \frac{2}{3-2r}t^{3/2-r} \right) \gamma^{3/2-r} \right\} \right]. \end{aligned} \quad (77)$$

- (ii) *Let $t \in \mathbb{N}^*$, and let Assumption 2 hold for some $r \in [1/2, +\infty]$. Then, with probability greater than $1 - \delta$*

$$\|\hat{w}_t - w_t\|_{\mathcal{H}} \leq \frac{\log(16/\delta)}{3\sqrt{n}} \left[16\sqrt{\kappa}M + 32\kappa M^2\gamma + (16\kappa + 32\kappa^2\gamma)\|g\|_{\rho}\kappa^{r-1/2} \right] \gamma t. \quad (78)$$

Proof. Substituting the bounds obtained in Lemma B.11 and Proposition B.12 with $\delta/4$ into (52), and applying Lemma B.9, we obtain yield that with probability bigger than $1 - \delta$

$$\|\hat{w}_t - w_t\|_{\mathcal{H}} \leq \frac{\log(16/\delta)}{3\sqrt{n}} \left(\gamma(16\kappa + 32\kappa^2\gamma) \sum_{k=0}^{t-1} \|w_k\|_{\mathcal{H}} + \gamma t (16\sqrt{\kappa}M + 32\kappa M^2\gamma) \right).$$

Statements (i) and (ii) directly follow from the bound on $\|w_k\|_{\mathcal{H}}$ obtained in Lemma B.10. \square

B.8 Proof of STEP 6 – approximation error

The proof of this result is similar to that of the approximation error bounds obtained in [32, 33], and uses spectral techniques, which are classical in linear inverse problems [14].

Theorem B.14 (Approximation error) *In the setting of Section 2, let Assumption 1 hold, let $n \in \mathbb{N}$, let $w_0 \in \mathcal{H}$, let $\gamma \in]0, n\kappa^{-1}[$ and let $(w_t)_{t \in \mathbb{N}}$ be defined as in (17)-(18). Then the following hold:*

- (i) *The approximation error $\mathcal{E}(w_t) - \inf_{\mathcal{H}} \mathcal{E} \rightarrow 0$.*
(ii) *Suppose that \mathcal{O} is nonempty. Then w^\dagger in (2) exists and $\|w_t - w^\dagger\|_{\mathcal{H}} \rightarrow 0$.*
(iii) *Let Assumption 2 hold for some $r \in]0, +\infty[$. Then*

$$\mathcal{E}(w_t) - \inf_{\mathcal{H}} \mathcal{E} \leq \left(\frac{r}{\gamma t} \right)^{2r} \|g\|_{\rho}^2.$$

(iv) Let Assumption 2 hold for some $r \in]1/2, +\infty[$. Then w^\dagger in (2) is well-defined and

$$\|w_t - w^\dagger\|_{\mathcal{H}} \leq \left(\frac{r-1/2}{\gamma t}\right)^{r-1/2} \|g\|_{\rho}.$$

Proof. (i): This is a direct consequence of Lemma B.1.

(ii): The proof is the same lines as that of Lemma B.10-(ii). See in particular (63).

(iii): It follows from Lemma B.3 that

$$\begin{aligned} \|Sv_{t+1} - g_{\rho}\|_{\rho} &= \left\| \left(\sum_{j=0}^{nt-1} L\eta(I - \eta L)^{nt-j+1} \right) - I \right\| g_{\rho} \Big\|_{\rho} \\ &= \sup_{\sigma \in [0, \|L\|]} \sigma^r (1 - \eta\sigma)^{nt} \|g\|_{\rho}. \end{aligned}$$

Note that, the last term is maximized at $\sigma = rn/(\gamma(r + nt + 1))$, hence for every $\sigma \in [0, +\infty[$,

$$\begin{aligned} \sigma^r \left(1 - \frac{\gamma}{n}\sigma\right)^{nt} &\leq \left(\frac{nr}{\gamma(r + nt)}\right)^r \left(1 - \frac{r}{r + nt}\right)^{nt} \\ &\leq \left(\frac{nr}{\gamma(r + nt)}\right)^r \\ &\leq \left(\frac{r}{t\gamma}\right)^r. \end{aligned}$$

Finally, the equality

$$\mathcal{E}(w_t) - \inf_{\mathcal{H}} \mathcal{E} = \|Sw_t - g_{\rho}\|_{\rho}^2,$$

yields the statement.

(iv): Since $L^{1/2}$ is a partial isometry between $L^2(\mathcal{H}, \rho_{\mathcal{H}})$ and $S(\mathcal{H})$ and $g_{\rho} = L^{1/2}(L^{r-1/2}g)$ by Assumption (2), it follows that $g_{\rho} \in S(\mathcal{H})$, and thus \mathcal{O} is nonempty, and w^\dagger is well defined. Moreover, since $S^*g_{\rho} = Tw^\dagger$, we also get that $Tw^\dagger = S^*Lr g = T^r S^*g$, implying that $w^\dagger = T^\dagger T^r S^*g$. It follows from Lemma B.3 and [14, Equation 2.24] that

$$\begin{aligned} \|w_t - w^\dagger\|_{\mathcal{H}} &\leq \left\| T^\dagger T^r \left(\sum_{j=0}^{nt-1} \eta T(I - \eta T)^{nt-j+1} - I \right) S^* \right\| \|g\|_{\rho} \\ &= \sup_{\sigma \in [0, \|L\|]} \sigma^{r-1/2} (1 - \eta\sigma)^{nt} \|g\|_{\rho} \\ &\leq \left(\frac{r-1/2}{t\gamma}\right)^{r-1/2} \|g\|_{\rho}, \end{aligned}$$

where the last inequality can be derived proceeding as in (iii). \square

B.9 STEP 7: proof of the main results

Let us denote by $\mathcal{S}(t, n, \delta)$ the right hand side of (77). Then, combining the sample error estimate with the error decomposition (19), we can immediately derive the following inequalities

$$\begin{aligned} \mathcal{E}(\hat{w}_t) - \inf_{\mathcal{H}} \mathcal{E} &\leq 2\kappa(\mathcal{S}(t, n, \delta))^2 + 2\mathcal{A}(t, \gamma, n) \\ \|\hat{w}_t - w^\dagger\|_{\mathcal{H}} &\leq \mathcal{S}(t, n, \delta) + \|w_t - w^\dagger\| \end{aligned}$$

with probability greater than $1 - \delta$. Note that analogous inequalities hold for the case $r > 1/2$ and with respect to the norm in \mathcal{H} . We are now ready to prove the Theorems stated in Section 3. The proof of universal consistency is a consequence of the sample and approximation error bounds, and of the application of Borel-Cantelli Lemma.

Proof of Theorem 3.1. (i): Recalling the error decomposition in (19), we have

$$\mathcal{E}(\hat{w}_t) - \inf_{\mathcal{H}} \mathcal{E} \leq 2\kappa \|\hat{w}_t - w_t\|_{\mathcal{H}}^2 + 2(\mathcal{E}(w_t) - \inf_{\mathcal{H}} \mathcal{E}).$$

Suppose that $t > 1$. Theorem B.13 applied with $r = 0$ yields that there exists $c \in \mathbb{R}_{++}$ such that, with probability greater than $1 - \delta$

$$\|\hat{w}_t - w_t\|_{\mathcal{H}} \leq c \frac{t^{3/2} \log(16/\delta)}{3\sqrt{n}} + \mathcal{E}(w_t) - \inf_{\mathcal{H}} \mathcal{E}. \quad (79)$$

Since $\sum_{k \in \mathbb{N}} \gamma = +\infty$ and $\gamma \leq \kappa^{-1} \leq n\kappa^{-1}$, by Theorem B.14(i), $\mathcal{A}(t) = \mathcal{E}(w_t) - \inf_{\mathcal{H}} \mathcal{E} \rightarrow 0$. Moreover, $\mathcal{A}(t^*(n)) \rightarrow 0$ since $t^*(n) \rightarrow +\infty$. Let $\eta \in \mathbb{R}_{++}$ and let $\alpha \in]1, +\infty[$. By (9), there exists $\bar{n} \in \mathbb{N}$ such that, for every $n \geq \bar{n}$, $\eta n / t^*(n)^3 > \alpha \log n$. Define

$$A_{n,\eta} = \left\{ \mathcal{E}(\hat{w}_{t^*(n)}) - \inf_{\mathcal{H}} \mathcal{E} \geq \mathcal{A}(t^*(n)) + c\eta \right\}. \quad (80)$$

By (79), for every $n \geq \bar{n}$, $\mathbb{P}(A_{n,\eta}) \leq 16 \exp(-\eta n / t^*(n)^3)^{1-\theta} \leq \exp(-\alpha \log n)$. Therefore,

$$\sum_{n \geq \bar{n}} \mathbb{P}(A_{n,\eta}) \leq \sum_{n \geq \bar{n}} n^{-\alpha} < +\infty,$$

hence the Borel-Cantelli lemma yields $\mathbb{P}(\bigcap_{k \geq \bar{n}} \bigcup_{n \geq k} A_{n,\eta}) = 0$, and almost sure convergence follows.

(ii): Since \mathcal{O} is nonempty, it follows that there $g_\rho \in S(\mathcal{H}) = L^{1/2}(\mathcal{H})$. Therefore, Assumption 2 is satisfied with $r = 1/2$. From Theorem B.13(ii), that there exists $c_1 \in \mathbb{R}_{++}$ such that, with probability greater than $1 - \delta$

$$\|\hat{w}_t - w_t\|_{\mathcal{H}} \leq \frac{c_1 t \log(16/\delta)}{3\sqrt{n}} \quad (81)$$

Moreover, since $\sum_{k \in \mathbb{N}} \gamma = +\infty$ and $\gamma \leq n\kappa^{-1}$, by Theorem B.14(ii), $\|w_t - w^\dagger\|_{\mathcal{H}} \rightarrow 0$. Reasoning as in (i), we obtain (11). \square

Proof of Theorem 3.3. (i): It follows from Theorem B.13(ii), that with probability greater than $1 - \delta$

$$\|\hat{w}_t - w_t\|_{\mathcal{H}} \leq \frac{32 \log(16/\delta)}{n} \left[M\kappa^{-1/2} + 2M^2\kappa^{-1} + 3\kappa^{r-1/2} \|g\|_\rho \right] t \quad (82)$$

Moreover, Theorem B.14(iv) yields

$$\|w_t - w^\dagger\|_{\mathcal{H}} \leq \left(\frac{r-1/2}{\gamma t} \right)^{r-1/2} \|g\|_\rho. \quad (83)$$

Inequality (12) follows by adding (82) with (83).

(ii): Let $\alpha \in]0, +\infty[$ and let $t^*(n) = \lceil n^\alpha \rceil$. Minimizing the right hand side in (12), we get

$$\alpha - 1/2 = \alpha(1/2 - r)$$

leading to the expression of $t^*(n)$. Now, let $n \in \mathbb{N}^*$ and $\beta \in [1, 2[$ be such that $n^\alpha \leq t^*(n) = \beta n^\alpha \leq n^\alpha + 1$. Then, by (12) we get

$$\|\hat{w}_t - w_t\|_{\mathcal{H}} \leq \beta^{1/2-r} 2r + 132 \log(16/\delta) \left[M\kappa^{-1/2} + 2M^2\kappa^{-1} + 3\kappa^{r-1/2}\|g\|_\rho \right] n^{\frac{1/2-r}{2r+1}} \quad (84)$$

and

$$\|w_t - w^\dagger\|_{\mathcal{H}} \leq \left(\frac{r-1/2}{\beta} \right)^{r-1/2} \|g\|_\rho n^{\frac{1/2-r}{2r+1}}. \quad (85)$$

Equation (14) follows recalling that $\beta \in [1, 2[$ in (i). \square

Proof of Theorem 3.4. (i): Recalling (19), it follows from Theorem B.13(i) and Theorem B.14(iii), that

$$\mathcal{E}(\hat{w}_t) - \inf_{\mathcal{H}} \mathcal{E} \leq 2 \frac{(32 \log(16/\delta))^2}{n} \left[M + 2M^2\kappa^{-1/2} + 3\kappa^r \|g\|_\rho \right]^2 t^2 + \left(\frac{r}{\gamma t} \right)^{2r} \|g\|_\rho^2 \quad (86)$$

(ii): As in the proof of Theorem 3.3(ii), set $t = \lceil n^\alpha \rceil$. Then, minimizing the right hand side in (86), we derive

$$2\alpha - 1 = -2\alpha r$$

which gives $\alpha = 1/(2r+2)$. Let $n \in \mathbb{N}^*$ and $\beta \in [1, 2[$ be such that $n^\alpha \leq t^*(n) = \beta n^\alpha \leq n^\alpha + 1$. Then, plugging the expression of $t^*(n)$ into (86), we get

$$\begin{aligned} \mathcal{E}(\hat{w}_{t^*(n)}) - \inf_{\mathcal{H}} \mathcal{E} &\leq 8 \frac{(32 \log(16/\delta))^2}{n} \left[M + 2M^2\kappa^{-1/2} + 3\kappa^r \|g\|_\rho \right]^2 n^{1/(r+1)} \\ &\quad + 2 \left(\frac{r}{\gamma} \right)^{2r} n^{-r/(r+1)} \|g\|_\rho^2. \end{aligned}$$

\square

A Non attainable case

Theorem A.1 (Finite sample bounds for the risk – non attainable case) *In the setting of Section 2, let Assumption 1 hold, and let $\gamma \in]0, \kappa^{-1}[$. Let Assumption 2 be satisfied for some $r \in]0, 1/2[$. Then the following hold:*

(i) *For every $t \in \mathbb{N}^*$, with probability greater than $1 - \delta$,*

$$\begin{aligned} \mathcal{E}(\hat{w}_t) - \inf_{\mathcal{H}} \mathcal{E} &\leq 8 \frac{(32 \log(16/\delta))^2}{n} \left[\left(M + 2M^2\kappa^{-1/2} + \frac{6\|g\|_\rho\kappa^{-1/2}}{3-2r} \right) t^{3/2-r} \right. \\ &\quad \left. + 3\kappa^r \frac{1-2r}{3-2r} \|g\|_\rho \right]^2 + 2 \left(\frac{r}{\gamma t} \right)^{2r} \|g\|_\rho^2 \end{aligned} \quad (87)$$

(ii) For the stopping rule $t^* : \mathbb{N}^* \rightarrow \mathbb{N}^*$

$$t^*(n) = \lceil n^{\frac{1}{3}} \rceil \quad (88)$$

with probability greater than $1 - \delta$,

$$\begin{aligned} \mathcal{E}(\hat{w}_{t^*(n)}) - \inf_{\mathcal{H}} \mathcal{E} \leq & \left[8 \left(32 \log \frac{16}{\delta} \right)^2 \left(M + 2M^2 \kappa^{-1/2} + \frac{6\|g\|_\rho}{3-2r} + \frac{3-6r}{3-2r} \kappa^r \|g\|_\rho \right)^2 \right. \\ & \left. + 2 \left(\frac{r}{\gamma} \right)^{2r} \|g\|_\rho^2 \right] n^{-2r/3} \quad (89) \end{aligned}$$

As for the attainable case, equation (87) arises from a form of bias-variance (sample-approximation) decomposition of the error. Choosing the number of epochs that optimize the bounds in (87), we derive a priori stopping rules (88) and corresponding bound (89). Again, these results confirm that the number of epochs acts as a regularization parameter and the best choice follows from equation (87).

Proof of Theorem A.1. (i): This follows from Theorem B.13(i) and Theorem B.14(iii).

(ii): It follows plugging the expression of $t^*(n)$ into the inequality in (i). \square