Journal of Machine Learning Research 17 (2016) 1-38 Submitted 3/15; Revised 3/16; Published 5/16

Iterative Regularization for Learning with Convex Loss Functions

Junhong Lin JHLIN5@HOTMAIL.COM
Department of Mathematics
City University of Hong Kong
Kowloon, Hong Kong, China

Lorenzo Rosasco LROSASCO@MIT.EDU
DIBRIS, Università di Genova
Via Dodecaneso, 35 — 16146 Genova, Italy
Laboratory for Computational and Statistical Learning
Istituto Italiano di Tecnologia and Massachusetts Institute of Technology
Bldg. 46-5155, 77 Massachusetts Avenue, Cambridge, MA 02139, USA

Ding-Xuan Zhou MAZHOU@CITYU.EDU.HK
Department of Mathematics
City University of Hong Kong
Kowloon, Hong Kong, China

Editor: Leon Bottou

Abstract

We consider the problem of supervised learning with convex loss functions and propose a new form of iterative regularization based on the subgradient method. Unlike other regularization approaches, in iterative regularization no constraint or penalization is considered, and generalization is achieved by (early) stopping an empirical iteration. We consider a nonparametric setting, in the framework of reproducing kernel Hilbert spaces, and prove consistency and finite sample bounds on the excess risk under general regularity conditions. Our study provides a new class of efficient regularized learning algorithms and gives insights on the interplay between statistics and optimization in machine learning.

1. Introduction

Availability of large high-dimensional data sets has motivated an interest in the interplay between statistics and optimization, towards developing new and more efficient learning solutions (Bousquet and Bottou, 2008). Indeed, while much theoretical work has been classically devoted to study statistical properties of estimators defined by variational schemes (for example Empirical Risk Minimization (Vapnik, 1998) or Tikhonov regularization (Tikhonov and Arsenin, 1977)), and to the computational properties of optimization procedures to solve the corresponding minimization problems (see for example Sra et al., 2011), much less work has considered the integration of statistical and optimization aspects, see for example Chandrasekaran and Jordan (2013); Wainwright (2014); Orabona (2014).

With the latter objective in mind, in this paper, we focus on iterative regularization. This class of methods, originated in a series of work in the mid-eighties (Nemirovskii, 1986;

Polyak, 1987), is based on the observation that early termination of an iterative optimization scheme applied to empirical data has a regularization effect. A critical implication of this fact is that the number of iterations serves as a regularization parameter, hence linking modeling and computational aspects: computational resources are directly linked to the generalization properties in the data, rather than their raw amount. Further, iterative regularization algorithms have a built-in "warm restart" property which allows to compute automatically a whole sequence of solutions corresponding to different levels of regularization (the regularization path). This latter property is especially relevant to efficiently determine the appropriate regularization level via model selection.

Iterative regularization techniques are well known in solving inverse problems, where several variants have been studied, see Engl et al. (1996); Kaltenbacher et al. (2008) and references therein. In machine learning, iterative regularization is often simply referred to as early stopping and is a well known "trick", for example in training neural networks (LeCun et al., 1998). Theoretical studies of iterative regularization in machine learning have mostly focused on the least square loss function (Buhlmann and Yu, 2003; Yao et al., 2007; Bauer et al., 2007; Blanchard and Nicole, 2010; Raskutti et al., 2014). Indeed, it is in this latter case that the connection to inverse problems can be made precise (De Vito et al., 2005). Interestingly, early stopping with the square loss has been shown to be related to boosting (Buhlmann and Yu, 2003) and also to be a special case of a large class of regularization approaches based on spectral filtering (Gerfo et al., 2008; Bauer et al., 2007). The regularizing effect of early stopping for loss functions other than the square loss has hardly been studied. Indeed, to the best of our knowledge the only papers considering related ideas are Bartlett and Traskin (2007); Bickel et al. (2006); Jiang (2004); Zhang and Yu (2005), where early stopping is studied in the context of boosting algorithms.

This paper is a different step towards understanding how early stopping can be employed with general convex loss functions. Within a statistical learning setting, we consider convex loss functions and propose a new form of iterative regularization based on the subgradient method, or the gradient descent if the loss is smooth. The resulting algorithms provide iterative regularization alternatives to Support Vector Machines or regularized logistic regression, and have built in the property of computing the whole regularization path. Our primary contribution in this paper is theoretical. By integrating optimization and statistical results, we establish consistency and non-asymptotic bounds quantifying the generalization properties of the proposed method under standard regularity assumptions. Interestingly, our study shows that considering the last iterate leads to essentially the same results as considering averaging, or selecting of the "best" iterate, as typically done in subgradient methods (Boyd and Vandenberghe, 2004). From a technical point of view, considering a general convex loss requires different error decompositions than the square loss. Moreover, operator theoretic techniques and matrix concentration inequalities need to be replaced by convex analysis and empirical process results. The error decomposition we consider, accounts for the contribution of both optimization and statistics to the error, and could be useful also for other methods.

The rest of the paper is organized as follows. We begin in Section 2 by briefly recalling the supervised learning problem, and then introduce our learning algorithm, discussing its numerical realization. In Section 3, after discussing the assumptions that underlie our analysis, we present and discuss our main theorems and illustrate the general error decom-

position which is composed of three error terms: computational, sample and approximation error. In Section 4, we will estimate the computational error, while in Section 5, we develop sample error bounds, and finally prove our main results.

2. Learning Algorithm

After briefly recalling the supervised learning problem, we introduce the algorithm we propose and give some comments on its numerical realization.

2.1 Problem Statement

In this paper we consider the problem of supervised learning. Let X be a separable metric space, $Y \subseteq \mathbb{R}$ and let ρ be a Borel probability measure on $Z = X \times Y$. Moreover, let $V : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ be a so-called loss function, measuring the *local error* $V(y, f(x))$ for $(x, y) \in Z$ and $f : X \to \mathbb{R}$. The *generalization error (or expected risk)* $\mathcal{E} = \mathcal{E}^V$ associated with V is given by

$$\mathcal{E}(f) = \int_Z V(y, f(x)) d\rho,$$

and is well defined for any measurable loss function V and measurable function f. We assume throughout that there exists a function f_ρ^V that minimizes the expected error $\mathcal{E}(f)$ among all measurable functions $f : X \to Y$. Roughly speaking, the goal of learning is to find an approximation of f_ρ^V when the measure ρ is known only through a sample $\mathbf{z} = \{z_i = (x_i, y_i)\}_{i=1}^m$ of size $m \in \mathbb{N}$ independently and identically drawn according to ρ. More precisely, given $\mathbf{z}$ the goal is to design a computational procedure to efficiently estimate a function $f_\mathbf{z}$, an estimator, for which it is possible to derive an explicitly probabilistic bound on the excess expected risk

$$\mathcal{E}(f_\mathbf{z}) - \mathcal{E}(f_\rho^V).$$

We end this section with a remark and an example.

Remark 1 *For several loss functions, it is possible to show that f_ρ^V exists, see the example below. However, as will be seen in the following, the search for an estimator in practice is often restricted to some hypothesis space $\mathcal{H}$ of measurable functions. In this case one should replace $\mathcal{E}(f_\rho^V)$ by $\inf_{f \in \mathcal{H}} \mathcal{E}(f)$. Interestingly, examples of hypothesis spaces are known for which $\mathcal{E}(f_\rho^V) = \inf_{f \in \mathcal{H}} \mathcal{E}(f)$, namely universal hypothesis spaces (Steinwart and Christmann, 2008). In the following, we consider $\mathcal{E}(f_\rho^V)$, with the understanding that it should be replaced by the infimum over $\mathcal{H}$, if the latter is not universal.*

Example 1 *The most classical example of loss function is probably the square loss $V(y, a) = (y - a)^2$, $y, a \in \mathbb{R}$. In this case, f_ρ^V is the regression function, defined at every point as the expectation of the conditional distribution of y given x (Cucker and Zhou, 2007; Steinwart and Christmann, 2008). Further examples include the absolute value loss $V(y, a) = |y - a|$ for which f_ρ^V is the median of the conditional distribution and more generally p-loss functions $V(y, a) = |y - a|^p$, $p \in \mathbb{N}$ [1]. Vapnik's ϵ-insensitive loss $V(y, a) = \max\{|y - a| - \epsilon, 0\}$, $\epsilon > 0$ and its generalizations $V(y, a) = \max\{|y - a|^p - \epsilon, 0\}$, $\epsilon > 0, p > 1$ provide yet other*

1. We denote the set of positive integers by $\mathbb{N}$.

examples. For classification, i.e., $Y = \{\pm 1\}$, other examples of loss functions used include the hinge loss $V(y,a) = \max\{1 - ya, 0\}$, the logistic loss $V(y,a) = \log(1 + e^{-ya})$ and the exponential loss $V(y,a) = e^{-ya}$. For all these examples f_ρ^V can be characterized, see for example Steinwart and Christmann (2008), and its measurability is easy to check.

2.2 Learning via Subgradient Methods with Early Stopping

To present the proposed learning algorithm we need a few preliminary definitions. Consider a reproducing kernel $K : X \times X \to \mathbb{R}$, that is a symmetric function, such that the matrix $(K(u_i, u_j))_{i,j=1}^\ell$ is positive semidefinite for any finite set of points $\{u_i\}_{i=1}^\ell$ in X. Recall that a reproducing kernel K defines a reproducing kernel Hilbert space (RKHS) $(\mathcal{H}_K, \|\cdot\|_K)$ as the completion of the linear span of the set $\{K_x(\cdot) := K(x,\cdot) : x \in X\}$ with respect to the inner product $\langle K_x, K_u\rangle_K := K(x,u), x, u \in X$ (Aronszajn, 1950). Moreover, assume the loss function V to be measurable and convex in its second argument, so that the corresponding left derivative V'_- exists and is non-decreasing at every point.

For a stepsize sequence $\{\eta_t > 0\}$, a stopping iteration $T > 2$ and an initial value $f_1 = 0$, we consider the iteration

$$f_{t+1} = f_t - \eta_t \frac{1}{m} \sum_{j=1}^m V'_-(y_j, f_t(x_j)) K_{x_j}, \qquad t = 1, \ldots, T. \tag{1}$$

The above iteration corresponds to the subgradient method (Bertsekas, 1999; Boyd et al., 2003) for minimizing the *empirical error* $\mathcal{E}_{\mathbf{z}} = \mathcal{E}_{\mathbf{z}}^V$ with respect to the loss V, which is given by

$$\mathcal{E}_{\mathbf{z}}(f) = \frac{1}{m} \sum_{j=1}^m V(y_j, f(x_j)).$$

Indeed, it is easy to see that $\frac{1}{m}\sum_{j=1}^m V'_-(y_j, f(x_j)) K_{x_j} \in \partial \mathcal{E}_{\mathbf{z}}(f)$, the subgradient of the empirical risk for $f \in \mathcal{H}_K$. In the special case where the loss function is smooth, then (1) reduces to the gradient descent algorithm. Since the subgradient method is not a descent algorithm, rather then the last iterate, the so-called Cesáro mean is often considered, corresponding, for $T \in \mathbb{N}$, to the following weighted average

$$a_T = \sum_{t=1}^T \omega_t f_t, \qquad \omega_t = \frac{\eta_t}{\sum_{t=1}^T \eta_t}, \quad t = 1, \ldots, T. \tag{2}$$

Alternatively, the *best iterate* is also often considered, which is defined for $T \in \mathbb{N}$ by

$$b_T = \operatorname*{arg\,min}_{f_t : t=1,\cdots,T} \mathcal{E}_{\mathbf{z}}(f_t). \tag{3}$$

In what follows, we will consider the learning algorithms obtained with these different choices.

We note that, classical results (Bertsekas, 1999; Boyd et al., 2003; Boyd and Vandenberghe, 2004) on the subgradient method focus on how the iteration (1) can be used to minimize $\mathcal{E}_{\mathbf{z}}$. Different to these studies, in the following we are interested in showing how iteration (1) can be used to define a *statistical estimator*, hence a learning algorithm to minimize the expected risk $\mathcal{E}$, rather than the empirical risk $\mathcal{E}_{\mathbf{z}}$. We add two remarks.

Remark 2 (Early Stopping SVM and Kernel Perceptron) *If we consider the hinge loss function in* (1)*, the corresponding algorithm is closely related to a batch (kernel) version of the perceptron (Rosenblatt, 1962; Aizerman et al., 1964), where an entire pass over the data is done before updating the solution. Such an algorithm can also be seen as an early stopping version of Support Vector Machines (Cortes and Vapnik, 1995). Interestingly, in this case the whole regularization path is computed incrementally albeit sparsity could be lost.*

Remark 3 (Multiple Passes SGD) *In practice stochastic/incremental approaches are often used. The latter correspond to considering the iteration*

$$f_{t+1} = f_t - \eta_t V'_-(y_{j_t}, f_t(x_{j_t}))K_{x_{j_t}}, \qquad t = 1, \dots, T.$$

given some initialization. Compared to (1) *in the above expression the "batch" gradient is replaced by a point-wise gradient. The sequence $(j_t)_t$ defines the order in which points are visited and can be stochastic. The obtained iteration is a form of stochastic gradient/subgradient method. When $T > n$ the algorithm visits the point multiple times. Each full pass over the data is called an epoch, or a cycle, and the obtained iteration corresponds to a form of incremental gradient/subgradient. The analysis in the paper can be modified to account for these iterations. However, to keep the paper self-contained we defer such an analysis to a future paper.*

2.3 Numerical Realization

The simplest case to derive a numerical procedure from Algorithm 1 is when $X = \mathbb{R}^d$ for some $d \in \mathbb{N}$ and K is the associated inner product. In this case it is straightforward to see that $f_{t+1}(x) = w_{t+1}^\top x$ for all $x \in X$, with $w_1 = 0_{d\times 1} \in \mathbb{R}^d$ and

$$w_{t+1} = w_t - \eta_t \frac{1}{m} \sum_{j=1}^{m} V'_-(y_j, w_t^\top x_j) x_j, \qquad t = 1, \dots, T.$$

Here, $w_t \in \mathbb{R}^d$ for all t. Beyond the linear kernel, it can be easily seen that given a finite *dictionary*

$$\{\phi_i : X \to \mathbb{R}, i = 1, \dots, p\}, \qquad p \in \mathbb{N},$$

one can consider the kernel $K(x, x') = \sum_{i=1}^p \phi_i(x')\phi_i(x)$. In this case, it holds $f_{t+1}(x) = \sum_{i=1}^p w_{t+1}^i \phi_i(x) = w_{t+1}^\top \Phi(x)$, $\Phi(x) = (\phi_1(x), \dots, \phi_p(x))^\top$ for all $x \in X$, with $w_1 = 0_{p\times 1} \in \mathbb{R}^p$ and

$$w_{t+1} = w_t - \eta_t \frac{1}{m} \sum_{j=1}^{m} V'_-(y_j, w_t^\top \Phi(x_j))\Phi(x_j), \qquad t = 1, \dots, T.$$

Finally, for a general kernel it is easy to prove by induction that $f_{t+1}(x) = \sum_{j=1}^m c_{t+1}^j K(x, x_j)$ for all $x \in X$, with

$$c_{t+1} = c_t - \eta_t \frac{1}{m} g_t, \qquad t = 1, \dots, T,$$

for $c_1 = 0_{m\times 1} \in \mathbb{R}^m$ and $g_t \in \mathbb{R}^m$ with its i-th component $g_t^i = V'_-(y_i, \sum_{j=1}^m c_t^j K(x_i, x_j)), \forall i = 1, \cdots, m$. Here, $c_t = (c_t^1, \cdots, c_t^m)^\top$ for $t \in \mathbb{N}$. Indeed, the base case is straightforward to

check and moreover by the inductive hypothesis

$$f_{t+1} = \sum_{j=1}^{m} c_t^j K_{x_j} - \eta_t \frac{1}{m} \sum_{j=1}^{m} V'_-(y_j, f_t(x_j)) K_{x_j} = \sum_{j=1}^{m} K_{x_j} \left(c_t^j - \eta_t \frac{1}{m} V'_-(y_j, f_t(x_j)) \right).$$

3. Main Results with Discussions

After presenting our main assumptions, in this section we state and discuss our main results.

3.1 Assumptions

Our results will be stated under several conditions on the triplet (ρ, V, K), that we describe and comment next. We begin with a basic assumption.

Assumption 1 *We assume the kernel to be bounded, that is* $\kappa = \sup_{x \in X} \sqrt{K(x,x)} < \infty$. *Moreover* $\|f_\rho^V\|_\infty < \infty$ *and* $|V|_0 := \sup_{y \in Y} V(y,0) < \infty$. *Furthermore, we consider the following growth condition for the left derivative* $V'_-(y,\cdot)$. *For some* $q \geq 0$ *and constant* $c_q > 0$, *it holds,*

$$\left|V'_-(y,a)\right| \leq c_q(1 + |a|^q), \qquad \forall a \in \mathbb{R}, y \in Y. \tag{4}$$

The boundedness conditions on K, f_ρ^V and V are fairly common (Cucker and Zhou, 2007; Steinwart and Christmann, 2008). They could probably be weakened by considering a more involved analysis which is outside the scope of this paper. Interestingly, the *growth* condition on the left derivative of V is weaker than assuming the loss, or its gradient, to be Lipschitz in its second entry which is standard both in learning theory (Cucker and Zhou, 2007; Steinwart and Christmann, 2008) and in optimization (Boyd and Vandenberghe, 2004). We note that the growth condition (4) is implied by the requirement for the loss function to be Nemitski when Y is bounded, as introduced in De Vito et al. (2004) (see also Steinwart and Christmann, 2008). This latter condition, which is satisfied by most loss functions, is natural to provide variational characterizations of the learning problem.

The second assumption refines the above boundedness condition by considering a variance-expectation bound which quantifies the noise (level) in the measure ρ with respect to *balls* in the RKHS $B_R = \{f \in \mathcal{H}_K : \|f\|_K \leq R\}$, $R > 0$ (Cucker and Zhou, 2007; Steinwart and Christmann, 2008).

Assumption 2 *We assume that there exists an exponent* $\tau \in [0,1]$ *and a positive constant* c_τ *such that for any* $R \geq 1$ *and* $f \in B_R$, *we have*

$$\int_{\mathcal{Z}} \left\{ \left(V(y, f(x)) - V(y, f_\rho^V(x)) \right)^2 \right\} d\rho \leq c_\tau R^{2+q-\tau} \left\{ \mathcal{E}(f) - \mathcal{E}(f_\rho^V) \right\}^\tau. \tag{5}$$

Assumption 2 always holds true for the square loss with $q = \tau = 1$, the hinge loss with $q = \tau = 0$, and more generally for Lipschitz loss functions with $\tau = 0$ and c_τ depending on $\|f_\rho^V\|_\infty$. In classification, the above condition can be related to the so-called Tsybakov margin condition. The latter quantifies the intuition that a classification problem is hard if the conditional probability of y given x is close to $1/2$ for many input points. More precisely if we denote by $\rho(y|x)$ the conditional probability for all $(x,y) \in Z$ and by ρ_X the marginal

probability on X, then we say that ρ satisfies the Tsybakov margin condition with exponent s if there exists a constant $C > 0$ such that for all $\delta > 0$

$$\rho_X(\{x \in X \ : \ |\rho(1|x) - \frac{1}{2}| \leq \delta\}) \leq (C\delta)^s.$$

Interestingly, under the Tsybakov margin condition, Assumption 2 holds for the hinge loss with $\tau = \frac{s}{s+1}$ and c_τ depending only on C.

The third condition is about the decay of the approximation error (Smale and Zhou, 2003).

Assumption 3 *Let $\lambda > 0$ and f_λ be a minimizer of:*

$$f_\lambda := \arg\min_{f \in \mathcal{H}_K} \mathcal{E}(f) + \lambda \|f\|_K^2. \tag{6}$$

The approximation error associated with the triplet (ρ, V, K) is defined by

$$\mathcal{D}(\lambda) = \mathcal{E}(f_\lambda) - \mathcal{E}(f_\rho^V) + \lambda \|f_\lambda\|_K^2. \tag{7}$$

We assume that for some $\beta \in (0, 1]$ and $c_\beta > 0$, the approximation error satisfies

$$\mathcal{D}(\lambda) \leq c_\beta \lambda^\beta, \qquad \forall \ \lambda > 0. \tag{8}$$

The above assumption is standard when analyzing regularized empirical risk minimization schemes and is related to the definition of interpolation spaces by means of K- functional (Cucker and Zhou, 2007). Interestingly, we will see in the following that it is also important when analyzing the approximation properties of the subgradient algorithm (1).

Finally, the last condition characterizes the *capacity* of a RKHS $\mathcal{H}_K$ in terms of empirical covering numbers, and plays an essential role in sample error estimates. Recall that for a subset $\mathcal{G}$ of a metric space (H, d), the covering number $\mathcal{N}(\mathcal{G}, \epsilon, d)$ is defined by

$$\mathcal{N}(\mathcal{G}, \epsilon, d) = \inf \left\{ l \in \mathbb{N} : \exists f_1, f_2, \cdots, f_l \subset H \text{ such that } \mathcal{G} \subset \bigcup_{i=1}^{\ell} \{f \in \mathcal{G} : d(f, f_i) \leq \epsilon\} \right\}.$$

Assumption 4 *Let $\mathcal{G}$ be a set of functions on X. The metric $d_{2,\mathbf{z}}$ is defined on $\mathcal{G}$ by*

$$d_{2,\mathbf{z}}(f, g) = \left\{ \frac{1}{m} \sum_{i=1}^{m} (f(z_i) - g(z_i))^2 \right\}^{1/2}, \quad f, g \in \mathcal{G}.$$

We assume that for some $\zeta \in (0, 2)$, $c_\zeta > 0$, the covering number of the unit ball B_1 in $\mathcal{H}_K$ with respect to $d_{2,\mathbf{z}}$ satisfies

$$\mathbb{E}_{\mathbf{z}} \left[\log \mathcal{N} \left(B_1, \epsilon, d_{2,\mathbf{z}} \right) \right] \leq c_\zeta \left(\frac{1}{\epsilon} \right)^\zeta, \qquad \forall \ \epsilon > 0. \tag{9}$$

The smaller ζ, the more stringent is the capacity assumption. As ζ approaches 2 we are essentially considering a capacity independent scenario, that is an arbitrary RKHS. In what follows, we will briefly comment on the connection between the above assumption and

other related assumptions. Recall that capacity of the RKHS may be measured by various concepts: covering numbers of balls B_R in $\mathcal{H}_K$, (dyadic) entropy numbers and decay of the eigenvalues of the integral operator $L_K : L^2_\rho \to L^2_\rho$ given by $L_K(f) = \int_X f(x)K_x d\rho_X(x)$, where $L^2_\rho = \{f : X \to \mathbb{R} \ : \ \int |f(x)|^2 d\rho_X(x) < \infty\}$. For a subset $\mathcal{G}$ of a metric space (H, d), the n-th entropy number is defined by

$$e_n(\mathcal{G}, d) = \inf \left\{ \varepsilon > 0 : \exists f_1, f_2, \cdots, f_{2^{n-1}} \text{ such that } \mathcal{G} \subset \bigcup_{i=1}^{2^{n-1}} \{f \in \mathcal{G} : d(f, f_i) \leq \varepsilon\} \right\}.$$

First, note that the covering and entropy numbers are equivalent (see for example Steinwart and Christmann, 2008, Lemma 6.21). Indeed, for $\zeta > 0$, the covering number $\mathcal{N}(\mathcal{G}, \epsilon, d)$ satisfies

$$\log \mathcal{N}(\mathcal{G}, \epsilon, d) \leq a_\zeta \left(\frac{1}{\epsilon}\right)^\zeta, \qquad \forall \epsilon > 0,$$

for some $a_\zeta > 0$, if and only if the entropy number $e_n(\mathcal{G}, d)$ satisfies

$$e_n(\mathcal{G}, d) \leq a'_\zeta \left(\frac{1}{n}\right)^{\frac{1}{\zeta}},$$

for some $a'_\zeta > 0$. Second, it is shown in Steinwart (2009) that if the eigenvalues of the integral operator L_K satisfy

$$\lambda_n \leq \tilde{a}_\zeta \left(\frac{1}{n}\right)^{\frac{2}{\zeta}}, \qquad n \geq 1,$$

for some constants $\tilde{a}_\zeta \geq 1$ and $\zeta \in (0, 2)$, then the expectation of the random entropy number $\mathbb{E}_{\mathbf{z}}[e_n(B_1, d_{2,\mathbf{z}})]$ satisfies

$$\mathbb{E}_{\mathbf{z}}[e_n(B_1, d_{2,\mathbf{z}})] \leq a_\zeta \left(\frac{1}{n}\right)^{\frac{1}{\zeta}}, \qquad n \geq 1,$$

for some constant a_ζ. Hence, using the equivalence of covering and entropy numbers, $\mathbb{E}_{\mathbf{z}}[\log \mathcal{N}(B_1, \epsilon, d_{2,\mathbf{z}}))]$ can be estimated from the eigenvalue decay of the integral operator L_K. Finally, since $d_{2,\mathbf{z}}(f, g) \leq \|f - g\|_\infty$, one has that for any $\epsilon > 0$, $\mathcal{N}(B_1, \epsilon, d_{2,\mathbf{z}})$ is bounded by $\mathcal{N}(B_1, \epsilon, \|\cdot\|_\infty)$, the uniform covering number of B_1 under the metric $\|\cdot\|_\infty$. Thus, the covering number $\mathcal{N}(B_1, \epsilon, d_{2,\mathbf{z}})$ can be also estimated given the uniform smoothness of the kernel (Zhou, 2003).

3.2 Finite Sample Bounds for General Convex Loss Functions

Our main results, in Theorems 4, 7 and 8, provide general stopping rules and corresponding upper bounds involving all the parameters defining the problem. These results are then illustrated and discussed in a series of corollaries considering special cases that allow for simpler statements, see in particular Corollary 6 in this subsection, Corollaries 9 and 10 in Subsection 3.3, and Theorems 11 and 12 in Subsection 3.4.

The following is our main result providing a general finite sample bound for the iterative regularization induced by the subgradient method for convex loss functions considering the last iterate. Here, $\lceil x \rceil$ denotes the smallest integer greater than or equal to $x \in \mathbb{R}$.

Theorem 4 *Assume (4) with $q \geq 0$, (5) with $\tau \in [0,1]$, (8) with $\beta \in (0,1]$ and (9) with $\zeta \in (0,2)$. Let $\eta_t = \eta_1 t^{-\theta}$ with $\frac{q}{q+1} < \theta < 1$ and η_1 satisfying*

$$0 < \eta_1 \leq \min\left\{ \frac{\sqrt{1-\theta}}{\sqrt{2}c_q(\kappa+1)^{q+1}}, \frac{1-\theta}{4|V|_0} \right\}. \tag{10}$$

If $T = \lceil m^\gamma \rceil$, then for any $0 < \delta < 1$, with confidence $1-\delta$, we have

$$\mathcal{E}(f_T) - \mathcal{E}(f_\rho^V) \leq \begin{cases} \widetilde{C} m^{-\alpha} \log \frac{2}{\delta}, & \text{when } \theta > \frac{q+1}{q+2}, \\ \widetilde{C} m^{-\alpha} \log m \log \frac{2}{\delta}, & \text{when } \theta \leq \frac{q+1}{q+2}, \end{cases}$$

where the power indices γ and α are defined as

$$\gamma = \begin{cases} \frac{2}{1-\theta} \frac{1}{(1+2\beta)(2-\tau+\zeta\tau/2)+q(1+\zeta/2)}, & \text{when } \theta \geq \frac{q+1}{q+2}, \\ \frac{2}{1-\theta} \frac{1}{\left(1+\frac{2\beta(\theta(1+q)-q)}{1-\theta}\right)(2-\tau+\zeta\tau/2)+q(1+\zeta/2)}, & \text{when } \theta < \frac{q+1}{q+2}, \end{cases} \tag{11}$$

$$\alpha = \begin{cases} \frac{\beta}{\beta(2-\tau+\zeta\tau/2)+\left\{\frac{2-\tau+\zeta\tau/2}{2}+\frac{q(1+\zeta/2)}{2}\right\}}, & \text{when } \theta \geq \frac{q+1}{q+2}, \\ \frac{\beta}{\beta(2-\tau+\zeta\tau/2)+\frac{1-\theta}{\theta(1+q)-q}\left\{\frac{2-\tau+\zeta\tau/2}{2}+\frac{q(1+\zeta/2)}{2}\right\}}, & \text{when } \theta < \frac{q+1}{q+2}, \end{cases} \tag{12}$$

and $\widetilde{C}$ is a positive constant independent of m or δ (given explicitly in the proof).

The proof is deferred to Section 5 and is based on a novel error decomposition, discussed in Section 3.6, integrating statistical and optimization aspects. We begin illustrating the above result for Lipschitz loss functions, that is considering $q = 0$, as follows.

Corollary 5 *Assume (4) with $q = 0$, (9) with $\zeta \in (0,2)$ and (8) with $\beta \in (0,1]$. Let $\eta_t = \eta_1 t^{-\theta}$ with $0 < \theta < 1$ and η_1 satisfying $0 < \eta_1 \leq \min\left\{ \frac{\sqrt{1-\theta}}{\sqrt{2}c_q(\kappa+1)}, \frac{1-\theta}{4|V|_0} \right\}$. If $T = \lceil m^\gamma \rceil$, then for any $0 < \delta < 1$, with confidence $1-\delta$, we have*

$$\mathcal{E}(f_T) - \mathcal{E}(f_\rho^V) \leq \begin{cases} \widetilde{C} m^{-\alpha} \log \frac{2}{\delta}, & \text{when } \theta > \frac{1}{2}, \\ \widetilde{C} m^{-\alpha} \log m \log \frac{2}{\delta}, & \text{when } \theta \leq \frac{1}{2}, \end{cases}$$

where the power indices γ and α are defined as

$$\gamma = \begin{cases} \frac{2}{(1-\theta)(2\beta+1)(2-\tau+\zeta\tau/2)}, & \text{when } \theta \geq \frac{1}{2}, \\ \frac{2}{(1-\theta+2\beta\theta)(2-\tau+\zeta\tau/2)}, & \text{when } \theta < \frac{1}{2}, \end{cases}$$

$$\alpha = \begin{cases} \frac{2\beta}{(2\beta+1)(2-\tau+\zeta\tau/2)}, & \text{when } \theta \geq \frac{1}{2}, \\ \frac{2\theta\beta}{(1-\theta+2\beta\theta)(2-\tau+\zeta\tau/2)}, & \text{when } \theta < \frac{1}{2}, \end{cases}$$

and $\widetilde{C}$ is a positive constant independent of m or δ.

For Lipschitz loss functions, Assumption 2 always holds true for $\tau = 0$. Also, if $f_\rho^V \in \mathcal{H}_K$, then Assumption 3 holds for $\beta = 1$ and $c_\beta \leq \|f_\rho^V\|_K^2$. In this case, γ and α from the above theorem are given by

$$\gamma = \max\left\{ \frac{1}{3(1-\theta)}, \frac{1}{1+\theta} \right\} \quad \text{and} \quad \alpha = \min\left\{ \frac{1}{3}, \frac{\theta}{1+\theta} \right\}.$$

Setting $\theta = 1/2$, we get the following result.

Corollary 6 *Assume (4) with $q = 0$, (9) with $\zeta \in (0, 2)$ and $f_\rho^V \in \mathcal{H}_K$. Let $\eta_t = \eta_1 t^{-1/2}$ with η_1 satisfying $0 < \eta_1 \leq \min\left\{\frac{1}{2c_q(\kappa+1)}, \frac{1}{8|V|_0}\right\}$. If $T = \lceil m^{2/3} \rceil$, then for any $0 < \delta < 1$, with confidence $1 - \delta$, we have*

$$\mathcal{E}(f_T) - \mathcal{E}(f_\rho^V) \leq \widetilde{C} m^{-1/3} \log m \log \frac{2}{\delta}.$$

The above results give finite sample bounds on the excess risk, provided that a suitable stopping rule is considered. While the stopping rule in the above theorems is distribution dependent, a data-driven stopping rule can be given by hold-out cross validation and adaptively achieves the same bounds. The proof of this latter result is straightforward using the techniques in Caponnetto and Yao (2010) and is omitted. The above bounds directly yield strong consistency (almost sure convergence) using standard arguments. Interestingly, our analysis suggests that a decaying stepsize needs to be chosen to achieve meaningful error bounds. The stepsize choice can influence both the early stopping rule and the error bounds. More precisely, if the stepsize decreases fast enough, i.e., $\theta \geq \frac{q+1}{q+2}$, the stopping rule depends on the decay speed but the error bound does not. In this case, the best possible choice for the early stopping rule is $\theta = \frac{q+1}{q+2}$, that is $\eta_t \sim 1/\sqrt{t}$ in the case of Lipschitz loss functions. With this choice, if for example we take the limit $\beta \to 1$, $\tau \to 0$, we have that the stopping rule scales as $O(m^{2/3})$ whereas the corresponding finite sample bounds are of order $O(m^{-1/3})$. A slower stepsize decay given by $\theta < \frac{q+1}{q+2}$ affects both the stopping rule and the error bounds, but these results are worse. A more detailed discussion of the obtained bounds in comparison to other learning algorithms is postponed to Section 3.5.

To see how the number of passes and the decaying rate θ of stepsize affects the performances of our algorithms, we carry out simple numerical simulations that complement the above result. In Fig. 1 we consider simulated data, i.e. simple binary classification problem where the input space is two dimensional. The training and test error as a function of the number of iterations are reported for different stepsize values. In Fig. 2 we consider a real benchmark data-set and again report the training and test error for different stepsize values. The same qualitative behavior can be observed in simulated and real data. The empirical error decreases as a function of the number of iterations while the expected (test) error as a minimum. The effect is more evident when the stepsize choice is more aggressive, that is for θ close to zero.

Next we discuss the behavior of different variants of the proposed algorithm. As mentioned before, in the subgradient method, when the goal is empirical risk minimization, the average or best iterates are often considered (see Equations (2), (3)). It is natural to ask what are the properties of the estimator obtained with these latter choices, that is when they are used as approximate minimizers of the expected, rather than the empirical risk. The following theorem provides an answer.

Theorem 7 *Under the assumptions of Theorem 4, if $T = \lceil m^\gamma \rceil$ and $g_T = a_T$ (or b_T) then for any $0 < \delta < 1$, with confidence $1 - \delta$, we have*

$$\mathcal{E}(g_T) - \mathcal{E}(f_\rho^V) \leq \begin{cases} \overline{C} m^{-\alpha} \log \frac{2}{\delta}, & \text{when } \theta \neq \frac{q+1}{q+2}, \\ \overline{C} m^{-\alpha} \log m \log \frac{2}{\delta}, & \text{when } \theta = \frac{q+1}{q+2}, \end{cases}$$

where the power indices γ and α are defined as in Theorem 4 and $\overline{C}$ is a positive constant independent of m or δ (can be given explicitly).

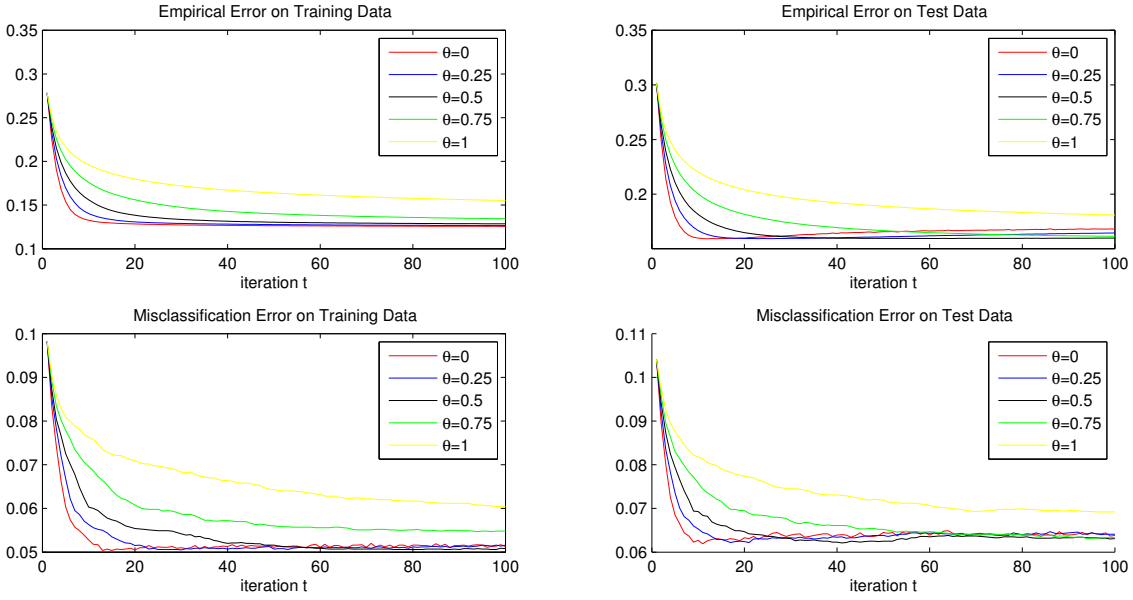

Figure 1: Performance of Algorithm (1) for the last iterates applied to synthetic data in binary classification with different θ, setting $\eta_1 = 1$, $V(y, f) = \max\{1 - yf, 0\}$ and $\mathcal{H}_K = \mathbb{R}^2$. Samples for two classes are drawn from bivariate Gaussian distributions. The parameters for the Gaussian distributions are $\mu_1 = [2, 0]^\top$, $\Sigma_1 = [5, 3; 3, 5]/2$ and $\mu_{-1} = -\mu_1$, $\Sigma_{-1} = \Sigma_1$. For each given θ, we run Algorithm (1) 100 times for 100 independent training data, and calculate the corresponding test errors for 100 independent test data. In each trial, both of the training data and the test data are of 100. The errors averaged over these 100 trials are depicted as the above.

The above result shows, perhaps surprisingly, that the behavior of the average or best iterates is essentially the same as the last iterate. Indeed, there is only a subtle difference between the upper bounds in Theorem 7 and those in Theorem 4, since the latter have an extra $\log m$ factor when $\theta < \frac{q+1}{q+2}$.

In the next section, we consider the case where loss is not only convex but also smooth.

3.3 Finite Sample Bounds for Smooth Loss Functions

In this section, we additionally assume that $V(y, \cdot)$ is differentiable and $V'(y, \cdot)$ is Lipschitz continuous with constant $L > 0$, i.e., for any $y \in Y$ and $a, b \in \mathbb{R}$,

$$|V'(y, b) - V'(y, a)| \le L|b - a|.$$

For the logistic loss in binary classification (see Example 1), it is easy to prove that both $V(y, \cdot)$ and $V'(y, \cdot)$ are Lipschitz continuous with $L = 1$, for all $y \in Y$. With the above smoothness assumption, we prove the following convergence result.

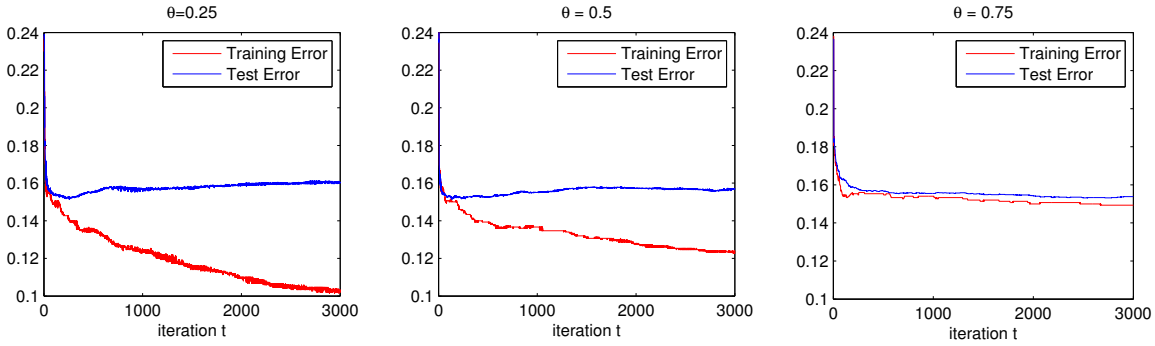

Figure 2: Misclassification errors of Algorithm (1) for the last iterates applied to Adult dataset with different values of θ, setting $V(y,f) = \max\{1-yf, 0\}$, $K(x,x') = \exp\{-\frac{\|x-x'\|_2^2}{2\sigma^2}\}$ and $m = 1500$. Here, σ is chosen as the median of the vector that consists of all Euclidean distances between training input vectors with different labels (Jaakkola et al., 1999). For each θ, η_1 is tuned using a holdout method.

Theorem 8 *Assume (4) with $q \geq 0$, (5) with $\tau \in [0,1]$, (8) with $\beta \in (0,1]$ and (9) with $\zeta \in (0,2)$. Assume that $V(y,\cdot)$ is differentiable and $V'(y,\cdot)$ is Lipschitz continuous with constant $L > 0$. Let $\eta_t = \eta_1 t^{-\theta}$ with $0 \leq \theta < 1$ and $0 < \eta_1 \leq \min(\frac{1-\theta}{2|V|_0}, (L\kappa^2)^{-1})$. If $T = \lceil m^\gamma \rceil$, then for any $0 < \delta < 1$, with confidence $1-\delta$, we have*

$$\mathcal{E}(f_T) - \mathcal{E}(f_\rho^V) \leq \widetilde{C} m^{-\alpha} \log \frac{2}{\delta},$$

where the power indices γ and α are defined as

$$\begin{aligned} \gamma &= \frac{2}{1-\theta} \frac{1}{(1+2\beta)(2-\tau+\zeta\tau/2) + q(1+\zeta/2)}, \\ \alpha &= \frac{\beta}{\beta(2-\tau+\zeta\tau/2) + \left\{\frac{2-\tau+\zeta\tau/2}{2} + \frac{q(1+\zeta/2)}{2}\right\}}, \end{aligned}$$

and $\widetilde{C}$ is a positive constant independent of m or δ.

The proof of this result will be given in Section 5. We can simplify the result by considering Lipschitz loss function ($q = 0$) and setting $\tau = 0$.

Corollary 9 *Under the assumptions of Theorem 8, let $q = 0$. If $T = \lceil m^{\frac{1}{(1-\theta)(2\beta+1)}} \rceil$, then for any $0 < \delta < 1$, with confidence $1-\delta$, we have*

$$\mathcal{E}(f_T) - \mathcal{E}(f_\rho^V) \leq \widetilde{C} m^{-\frac{\beta}{2\beta+1}} \log \frac{2}{\delta},$$

where $\widetilde{C}$ is a positive constant independent of m or δ.

The finite sample bound obtained above is essentially the same as the best possible bound obtained for general convex loss functions. However, the important difference is that for

smooth loss functions, a constant stepsize can be chosen and allows to considerably improve the stopping rule. Indeed, if for example we can consider the limit $\beta \to 1$, $\tau \to 0$, we have that the stopping is $O(m^{1/3})$, rather than $O(m^{2/3})$, whereas the corresponding finite sample bound is again $O(m^{-1/3})$.

A similar simplification can be done for the square loss. Here, as mentioned in Example 1, f_ρ^V is the regression function f_ρ, and there holds $\mathcal{E}(f) - \mathcal{E}(f_\rho) = \|f - f_\rho\|^2_{L^2_{\rho_X}}$. In this case, Assumption 2 holds true for $\tau = 1$ and $c_\tau = 1$, and condition (8) can be characterized by requiring that $f_\rho \in L_K^{\beta/2}(L^2_{\rho_X})$ (Smale and Zhou, 2003; Caponnetto and De Vito, 2007), where $L_K^{\beta/2}$ is the $\frac{\beta}{2}$-th power of the positive operator L_K.

Corollary 10 *Let $V(y,a) = (y-a)^2$ and $|y| \leq |V|_0 < \infty$ almost surely. Assume $f_\rho \in L_K^{\beta/2}(L^2_{\rho_X})$ with $\beta \in (0,1]$ and (9) with $\zeta \in (0,2)$. Let $\eta_t = \eta_1 t^{-\theta}$ with $0 \leq \theta < 1$ and $0 < \eta_1 \leq \min(\frac{1-\theta}{2|V|_0}, \kappa^{-2})$. If $T = \lceil m^{\frac{2}{(1-\theta)(\beta+1)(\zeta+2)}} \rceil$, then for any $0 < \delta < 1$, with confidence $1-\delta$, we have*

$$\|f_T - f_\rho\|^2_{L^2_{\rho_X}} \leq \widetilde{C} m^{-\frac{2\beta}{(\beta+1)(\zeta+2)}} \log \frac{2}{\delta}.$$

In particular, if $f_\rho \in \mathcal{H}_K$,

$$\|f_T - f_\rho\|^2_{L^2_{\rho_X}} \leq \widetilde{C} m^{-\frac{1}{\zeta+2}} \log \frac{2}{\delta}.$$

Before comparing our bounds with obtained with other algorithms we last specialize our results to a binary classification setting.

3.4 Iterative Regularization for Classification: Surrogate Loss Functions and Hinge Loss

We briefly discuss how the above results allow to derive error bounds in binary classification problems. In this latter case $Y = \{1, -1\}$ and a natural choice for the loss function is the misclassification loss given by

$$V(y, b(x)) = \Theta\left(-yb(x)\right) \tag{13}$$

for $b : X \to Y$ and $\Theta(a) = 1$, if $a \geq 0$, and $\Theta(a) = 0$ otherwise. The corresponding generalization error, denoted by $\mathcal{R}$, is called misclassification risk, since it can be shown to be the probability of the event $\{(x,y) \in Z : y \neq b(x)\}$. The minimizer of the misclassification error is the Bayes rule $b_\rho : X \to Y$ given by

$$b_\rho(x) = \begin{cases} 1, & \text{if the conditional probability } \rho(1|x) \geq 1/2, \\ -1, & \text{otherwise.} \end{cases}$$

The misclassification loss (13) is neither convex nor smooth and thus leads to computationally intractable problems. In practice, a convex (so-called *surrogate*) loss function is typically considered and a classifier is obtained by estimating a real function f and then taking its sign defined as

$$\operatorname{sign}(f)(x) = \begin{cases} 1, & \text{if } f(x) \geq 0, \\ -1, & \text{otherwise.} \end{cases}$$

The question arises of if, and how, error bounds on the excess risk $\mathcal{E}(f) - \mathcal{E}(f_\rho^V)$ yields results on $\mathcal{R}(\text{sign}f) - \mathcal{R}(b_\rho)$. Indeed, the so-called *comparison* results are known relating these different error measures, see for example Cucker and Zhou (2007); Steinwart and Christmann (2008) and references therein. We discuss in particular the case of the hinge loss function (see Example 1). In this case for all measurable functions f it holds that

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

Indeed, the hinge loss satisfies Assumption (4) with $q = 0$ and, under Tsybakov noise condition, Assumption (5). Misclassification error bound, for iterative regularization with the hinge loss, can then be obtained as a corollary of Theorem 4.

Theorem 11 *Let $Y = \{1, -1\}$ and V be the hinge loss. Let $0 < \epsilon < \frac{1}{3}$ and (8) be satisfied with $\beta \in (0, 1]$. Let $\eta_t = \eta_1 t^{-\theta}$ with $\theta > 1/2$ and $0 < \eta_1 \leq \min\left\{\frac{\sqrt{(1-\theta)}}{\sqrt{2}(\kappa+1)}, \frac{1-\theta}{4}\right\}$. If (9) is valid with $\zeta \in (0, 2)$ and $T = \lceil m^{\frac{1}{(1-\theta)(2\beta+1)}} \rceil$, then with confidence $1 - \delta$, we have*

$$\mathcal{R}\left(sign(f_T)\right) - \mathcal{R}(f_c) \leq \widetilde{C} m^{-\frac{\beta}{2\beta+1}} \log \frac{2}{\delta}. \tag{14}$$

In particular, if $\beta > \frac{1-3\epsilon}{1+6\epsilon}$ with $\epsilon \in (0, 1/3)$, then with confidence $1 - \delta$,

$$\mathcal{R}\left(sign(f_T)\right) - \mathcal{R}(f_c) \leq \widetilde{C} m^{\epsilon - \frac{1}{3}} \log \frac{2}{\delta}.$$

The proof of the above result is given in Section 5, and comments on the obtained rates are given in the next section.

We end noting that, as illustrated by the next result where the stopping rule is kept fixed while the stepsize is chosen in a distribution dependent way. This observation is made precise by the following result.

Theorem 12 *Let $Y = \{1, -1\}$ and V be the hinge loss. Let $0 < \epsilon < \frac{1}{3}$ and (8) is satisfied with $1 > \beta > \frac{4-3\epsilon}{4+6\epsilon}$. Let $\eta_t = \eta_1 t^{-\theta}$ with $\theta = \frac{4\beta-1+3\epsilon(2\beta+1)}{(2\beta+1)(2+3\epsilon)}$ and $0 < \eta_1 \leq \min\left\{\frac{\sqrt{2(1-\theta)}}{\kappa+1}, \frac{1-\theta}{4}\right\}$. If (9) is valid with $\zeta \in (0, 2)$ and $T = \lceil m^{\frac{2}{3}+\epsilon} \rceil$, then with confidence $1 - \delta$, we have*

$$\mathcal{R}\left(sign(f_T)\right) - \mathcal{R}(f_c) \leq \widetilde{C} m^{\frac{\epsilon}{4} - \frac{1}{3}} \log \frac{2}{\delta}.$$

3.5 Comparison with Other Learning Algorithms

As mentioned before iterative regularization has nice computational properties. The algorithm reduces to a simple first order method with low iteration cost and allows to easily compute the estimators corresponding to different regularization level (the regularization path), a crucial fact since model selection needs to be performed. In this view, the proposed procedure can be compared with standard approaches for example based on considering Support Vector Machines (SVM) or online variants such as Pegasos. In the former case, in principle a quadratic programming problem need to be solved for each regularization

parameter values. Our approach can be compared to more sophisticated approaches to compute the full SVM regularization path. In the latter case, the main difference is that in iterative regularization the early stopping rule is explicitly linked to the regularization level and in practice can be chosen by cross validation.

It is natural to compare the obtained statistical bounds with those for other learning algorithms. For general convex loss functions, the methods for which sharp bounds are available, are penalized empirical risk minimization (Tikhonov regularization), i.e.

$$f_{\mathbf{z},\lambda} = \arg\min_{f\in\mathcal{H}_K} \left\{\mathcal{E}_{\mathbf{z}}(f) + \lambda\|f\|_K^2\right\}, \quad \lambda > 0,$$

see for example Cucker and Zhou (2007); Steinwart and Christmann (2008) and references therein. The best error bounds for Tikhonov regularization with Lipschitz loss functions, see for example Steinwart and Christmann (2008, Chapter 7), are of order $O(m^{-\alpha'})$ with

$$\alpha' = \min\left\{\frac{2\beta}{\beta+1}, \frac{\beta}{(2-\zeta/2-\tau+\tau\zeta/2)\beta+\zeta/2}\right\},$$

which reduces to

$$\alpha' = \frac{\beta}{\beta+1}$$

in the capacity independent limit ($\zeta \to 2$). From Corollary 5 for Lipschitz loss functions, we see that the bounds we obtain are of order $O(m^{-\alpha})$ with the exponent

$$\alpha = \frac{2\beta}{(2\beta+1)(2-\tau+\zeta\tau/2)},$$

reducing to

$$\alpha = \frac{\beta}{2\beta+1}$$

in the capacity independent limit. Hence, the obtained bounds are worse than the best ones available for Tikhonov regularization. However, the analysis of the latter does not take into account the optimization error and it is still an open question whether the best rate is preserved when such an error is incorporated. At this point we believe this gap to be a byproduct of our analysis rather than a fundamental fact, and addressing this point should be a subject of further work. Moreover, we note that our analysis allows to derive error bound for all Nemitski loss functions rather than only Lipschiz loss functions.

Beyond Tikhnov regularization, we can compare with the online regularization scheme for the hinge loss. The online learning algorithms with a regularization sequence $\{\lambda_t > 0\}_t$ defined by

$$f_{t+1} = \begin{cases} (1-\eta_t\lambda_t)f_t, & \text{if } y_t f_t(x_t) > 1, \\ (1-\eta_t\lambda_t)f_t + \eta_t y_t K_{x_t}, & \text{if } y_t f_t(x_t) \le 1. \end{cases} \tag{15}$$

were studied in Ying and Zhou (2006); Ye and Zhou (2007). Our results improve the results in Ying and Zhou (2006); Ye and Zhou (2007) in two aspects. The bound obtained in Ying and Zhou (2006) is of the form $O(T^{\epsilon-\frac{1}{4}})$ while the bound in Theorem 12 is of type $O(T^{\frac{9}{8}\epsilon-\frac{1}{2}})$ by substituting the expression $m^{\frac{2}{3}+\epsilon}$ for T. Moreover, our results are with high probability and promptly yield almost sure convergence whereas the results in Ying and

Zhou (2006) are only in expectation. We note that, interestingly, sharp bounds for Lipschitz loss functions are derived in Orabona (2014), although the obtained results do not take into account the capacity and variance assumptions that could lead to large improvements.

We next compare with the previous results on iterative regularization. The main results available thus far have been obtained for the square loss, for which bounds have been first derived for gradient descent in Buhlmann and Yu (2003), but only for a fixed design regression setting, and in Yao et al. (2007) for a general statistical learning setting. While the bounds in Yao et al. (2007) are suboptimal, they have later been improved in Bauer et al. (2007); Caponnetto and Yao (2010); Raskutti et al. (2014). Interestingly, sharp error bounds have also been proved for iterative regularization induced by other, potentially faster, iterative techniques, including incremental gradient (Rosasco and Villa, 2014), conjugate gradient (Blanchard and Nicole, 2010) and the so-called ν-method (Bauer et al., 2007; Caponnetto and Yao, 2010), an accelerated gradient descent technique related to Chebyshev method (Engl et al., 1996). The best obtained bounds are of order $O(m^{-\frac{2\beta}{2\beta+\zeta}})$ and can be shown to be optimal since they match the corresponding minimax lower bound (Caponnetto and De Vito, 2007). The bound obtained in Corollary 10 holds for all smooth Nemitski loss functions but is of order $O(m^{-\frac{2\beta}{(2+\zeta)(\beta+1)}})$, which is worse. In the capacity independent limit, the best available bound we obtain is of order $O(m^{-\frac{\beta}{2(\beta+1)}})$, whereas the optimal bound is of order $O(m^{-\frac{\beta}{\beta+1}})$. Also, in this case, the reason for the gap appears to be of technical reason and should be further studied.

Finally, before giving the detailed proofs, in the next subsection, we discuss the general error decomposition underlying our approach, which highlights the interplay between statistics and optimization and could be also useful in other contexts.

3.6 Error Decomposition

Theorems 4 and 8 rely on a natural error decomposition that we derive next. The goal is to estimate the excess risk $\mathcal{E}(f_T) - \mathcal{E}(f_\rho^V)$, and the starting point is to split the error by introducing a *reference* function $f_* \in \mathcal{H}_K$,

$$\mathcal{E}(f_T) - \mathcal{E}(f_\rho^V) = \mathcal{E}(f_T) - \mathcal{E}(f_*) + \mathcal{E}(f_*) - \mathcal{E}(f_\rho^V). \tag{16}$$

The above equation can be further developed by considering

$$\mathcal{E}(f_T) - \mathcal{E}(f_\rho^V) = (\mathcal{E}_{\mathbf{z}}(f_T) - \mathcal{E}_{\mathbf{z}}(f_*)) + (\mathcal{E}(f_T) - \mathcal{E}_{\mathbf{z}}(f_T) + \mathcal{E}_{\mathbf{z}}(f_*) - \mathcal{E}(f_*)) + \left(\mathcal{E}(f_*) - \mathcal{E}(f_\rho^V)\right). \tag{17}$$

Inspection of the above expression provides several insights. The first term is a computational error related to optimization. It quantifies the discrepancy between the empirical errors of the iterate defined by the subgradient method and that of the reference function. The second term is a sample error and can be studied using empirical process theory, provided that a bound on the norm of the iterates (and of the reference function) is available. Indeed, to get a sharper concentration estimate *recentering* can be considered (Cucker and Zhou, 2007; Steinwart and Christmann, 2008)

$$\left\{\left(\mathcal{E}(f_T) - \mathcal{E}(f_\rho^V)\right) - \left(\mathcal{E}_{\mathbf{z}}(f_T) - \mathcal{E}_{\mathbf{z}}(f_\rho^V)\right)\right\} + \left\{\left(\mathcal{E}_{\mathbf{z}}(f_*) - \mathcal{E}_{\mathbf{z}}(f_\rho^V)\right) - \left(\mathcal{E}(f_*) - \mathcal{E}(f_\rho^V)\right)\right\}.$$

Note that the second addend can be negative so that we effectively only need to control

$$\mathcal{F}_{\mathbf{z}}(f_*) = \max\left\{(\mathcal{E}_{\mathbf{z}}(f_*) - \mathcal{E}_{\mathbf{z}}(f_\rho^V)) - (\mathcal{E}(f_*) - \mathcal{E}(f_\rho^V)),\ 0\right\}. \tag{18}$$

Finally the last term suggests that a natural choice for the reference function is an *almost minimizer* of the expected risk, having bounded norm, and for which the approximation level can be quantified. While there is a certain degree of freedom in the latter choice, in the following we will consider $f_* = f_\lambda$, the minimizer of (6). With this latter choice we can control

$$\mathcal{A}(f_*) = \mathcal{E}(f_*) - \mathcal{E}(f_\rho^V)$$

by $\mathcal{D}(\lambda)$ given in Assumption 3.

Collecting some of the above observations, we have the following lemma.

Lemma 13 *For $f_* \in \mathcal{H}_K$, we have*

$$\begin{aligned} &\mathcal{E}(f_T) - \mathcal{E}(f_\rho^V) \\ &\leq \left\{\left(\mathcal{E}(f_T) - \mathcal{E}(f_\rho^V)\right) - \left(\mathcal{E}_{\mathbf{z}}(f_T) - \mathcal{E}_{\mathbf{z}}(f_\rho^V)\right) + \mathcal{F}_{\mathbf{z}}(f_*)\right\} + (\mathcal{E}_{\mathbf{z}}(f_T) - \mathcal{E}_{\mathbf{z}}(f_*)) + \mathcal{A}(f_*). \end{aligned} \tag{19}$$

In the next sections, we proceed estimating the various error terms in the above error decomposition. We will first deal with the computational error, the analysis of which is the main technical contribution of the paper and then proceed to consider the sample and approximation error terms. The best stopping criterion and corresponding rates are derived by suitably balancing these different error terms.

4. Computational Error

In this section, we will bound the iterates and estimate the computational error from Lemma 13.

4.1 Bounds on Iterates

We introduce the following key lemma, which will be used several times in our analysis.

Lemma 14 *For any fixed $f \in \mathcal{H}_K$ and $t = 1, \ldots, T$,*

$$\|f_{t+1} - f\|_K^2 \leq \|f_t - f\|_K^2 + \eta_t^2 G_t^2 + 2\eta_t[\mathcal{E}_{\mathbf{z}}(f) - \mathcal{E}_{\mathbf{z}}(f_t)], \tag{20}$$

where

$$G_t^2 = \left\| \frac{1}{m} \sum_{j=1}^m V_-'(y_j, f_t(x_j)) K_{x_j} \right\|_K^2 \leq c_q^2 (\kappa + 1)^{2q+2} \max\left\{1, \|f_t\|_K^{2q}\right\}. \tag{21}$$

Proof Computing inner product $\langle f_{t+1} - f, f_{t+1} - f\rangle_K$ with f_{t+1} given by (1) yields

$$\|f_{t+1} - f\|_K^2 = \|f_t - f\|_K^2 + \eta_t^2 G_t^2 + \frac{2\eta_t}{m} \sum_{j=1}^m V_-'(y_j, f_t(x_j)) \left\langle K_{x_j}, f - f_t \right\rangle_K.$$

Using the reproducing property

$$f(x) = \langle f, K_x \rangle_K, \qquad \forall f \in \mathcal{H}_K, x \in X, \tag{22}$$

and Assumption 1, we get

$$\|f\|_\infty \leq \kappa \|f\|_K, \qquad \forall f \in \mathcal{H}_K, \tag{23}$$

and

$$\|f_{t+1} - f\|_K^2 = \|f_t - f\|_K^2 + \eta_t^2 G_t^2 + \frac{2\eta_t}{m} \sum_{j=1}^m V'_-(y_j, f_t(x_j))(f(x_j) - f_t(x_j)). \tag{24}$$

Since $V(y_j, \cdot)$ is a convex function, we have

$$V'_-(y_j, a)(b - a) \leq V(y_j, b) - V(y_j, a), \qquad \forall a, b \in \mathbb{R}.$$

Using this expression to (24) gives

$$\|f_{t+1} - f\|_K^2 \leq \|f_t - f\|_K^2 + \eta_t^2 G_t^2 + \frac{2\eta_t}{m} \sum_{j=1}^m \left[V(y_j, f(x_j)) - V(y_j, f_t(x_j)) \right],$$

where the last term is exactly $2\eta_t[\mathcal{E}_{\mathbf{z}}(f) - \mathcal{E}_{\mathbf{z}}(f_t)]$.

By (4), (23), and the observation $\|K_{x_j}\|_K = \sqrt{K(x_j, x_j)} \leq \kappa$, we find

$$\begin{aligned} G_t &= \left\| \frac{1}{m} \sum_{j=1}^m V'_-(y_j, f_t(x_j)) K_{x_j} \right\|_K \leq \frac{\kappa}{m} \sum_{j=1}^m \left| V'_-(y_j, f_t(x_j)) \right| \\ &\leq \frac{\kappa}{m} \sum_{j=1}^m c_q (1 + |f_t(x_j)|^q) \leq \kappa c_q (1 + \kappa^q \|f_t\|_K^q), \end{aligned}$$

and the desired bound follows. ■

Using the above lemma, we can bound the iterated sequence as follows.

Lemma 15 *Let $0 \leq \theta < 1$ satisfying $\theta \geq \frac{q}{q+1}$ and $\eta_t = \eta_1 t^{-\theta}$ with η_1 satisfying (10). Then for $t = 1, \ldots, T$,*

$$\|f_{t+1}\|_K \leq t^{\frac{1-\theta}{2}}. \tag{25}$$

Proof We prove our statement by induction. Taking $f = 0$ in Lemma 14, we know that

$$\|f_{t+1}\|_K^2 \leq \|f_t\|_K^2 + \eta_t^2 G_t^2 + 2\eta_t[\mathcal{E}_{\mathbf{z}}(0) - \mathcal{E}_{\mathbf{z}}(f_t)] \leq \|f_t\|_K^2 + \eta_t^2 G_t^2 + 2\eta_t |V|_0.$$

This verifies (25) for the case $t = 1$ since $f_1 = 0$ and $\eta_1^2 c_q^2 (\kappa + 1)^{2q+2} + 2\eta_1 |V|_0 \leq 1$. Assume $\|f_t\|_K \leq (t-1)^{\frac{1-\theta}{2}}$ with $t \geq 2$. Then

$$G_t^2 \leq c_q^2 (\kappa + 1)^{2q+2} (t-1)^{(1-\theta)q}.$$

Hence,

$$\begin{aligned}\|f_{t+1}\|_K^2 &\le (t-1)^{1-\theta} + \eta_1^2 t^{-2\theta} c_q^2 (\kappa+1)^{2q+2} t^{(1-\theta)q} + 2\eta_1 t^{-\theta}|V|_0 \\ &\le t^{1-\theta}\left\{\left(1-\frac{1}{t}\right)^{1-\theta} + \frac{\eta_1^2 c_q^2(\kappa+1)^{2q+2}}{t^{(q+1)\theta+1-q}} + \frac{2\eta_1|V|_0}{t}\right\}.\end{aligned}$$

Since $\left(1-\frac{1}{t}\right)^{1-\theta} \le 1-\frac{1-\theta}{t}$ and the condition $\theta \ge \frac{q}{q+1}$ implies $(q+1)\theta+1-q \ge 1$, we have

$$\|f_{t+1}\|_K^2 \le t^{1-\theta}\left\{1-\frac{1-\theta}{t} + \frac{\eta_1^2 c_q^2(\kappa+1)^{2q+2}}{t} + \frac{2\eta_1|V|_0}{t}\right\}.$$

Finally we use the restriction (10) for η_1 and find $\|f_{t+1}\|_K^2 \le t^{1-\theta}$. This completes the induction procedure and proves our conclusion. ■

By taking $f = f_t$ in (20), we see the following estimate for $\|f_{t+1} - f_t\|_K$ from Lemmas 14 and 15.

Corollary 16 *Under the assumptions of Lemma 15, we have for $t = 1, \ldots, T$,*

$$\|f_{t+1} - f_t\|_K \le \eta_1 c_q (\kappa+1)^{q+1} t^{\frac{(1-\theta)q}{2}-\theta}. \tag{26}$$

Observe from the restriction $\theta \ge \frac{q}{q+1}$ in Lemma 15 that the power index in (26) satisfies $\frac{(1-\theta)q}{2} - \theta \le -\frac{q}{2(q+1)} \le 0$.

4.2 Computational Error for the Last Iterate

In this subsection, we estimate the computational error $\mathcal{E}_{\mathbf{z}}(f_T) - \mathcal{E}_{\mathbf{z}}(f_*)$ for an arbitrary $f_* \in \mathcal{H}_K$. Some ideas for estimating the average error in our proof are taken from Boyd et al. (2003); Shamir and Zhang (2013).

Lemma 17 *Assume (4) with $q \ge 0$. Let $f_* \in \mathcal{H}_K$. If $\eta_t = \eta_1 t^{-\theta}$ with $0 < \theta < 1$ satisfying $\theta > \frac{q}{q+1}$ and η_1 satisfying (10), then we have*

$$\begin{aligned}\mathcal{E}_{\mathbf{z}}(f_T) - \mathcal{E}_{\mathbf{z}}(f_*) &\le \left(\frac{\|f_*\|_K^2}{2\eta_1} + \widetilde{C}_1\right)\Lambda_{T,\theta} \\ &+ \frac{T^\theta}{2\eta_1}\sum_{k=1}^{T-1}\frac{1}{k+1}\left[\frac{1}{k}\sum_{t=T-k+1}^{T} 2\eta_t - 2\eta_{T-k}\right]\{\mathcal{E}_{\mathbf{z}}(f_{T-k}) - \mathcal{E}_{\mathbf{z}}(f_*)\},\end{aligned} \tag{27}$$

where $\Lambda_{T,\theta}$ is defined by

$$\Lambda_{T,\theta} = \begin{cases} \frac{1}{(q+2)\theta-(q+1)}T^{-(1-\theta)}, & \text{when } \theta > \frac{q+1}{q+2}, \\ (\log T)T^{-(1-\theta)}, & \text{when } \theta = \frac{q+1}{q+2}, \\ \frac{1}{(q+1)-(q+2)\theta}(\log T)T^{-(\theta(1+q)-q)}, & \text{when } \theta < \frac{q+1}{q+2}, \end{cases} \tag{28}$$

and $\widetilde{C}_1$ is a positive constant depending on q, κ, θ (independent of T, m or f_* and given explicitly in the proof).

Proof Lemma 14 plays a key role in our proof. In particular, we shall apply the following equivalent form of inequality (20) from Lemma 14 several times with various choices of $f \in \mathcal{H}_K$:

$$2\eta_t \left[\mathcal{E}_{\mathbf{z}}(f_t) - \mathcal{E}_{\mathbf{z}}(f)\right] \leq \left\{\|f_t - f\|_K^2 - \|f_{t+1} - f\|_K^2\right\} + \eta_t^2 G_t^2. \tag{29}$$

Step 1: Error decomposition. Decompose the weighted empirical error $2\eta_T \mathcal{E}_{\mathbf{z}}(f_T)$ as

$$\begin{aligned}
2\eta_T \mathcal{E}_{\mathbf{z}}(f_T) &= \frac{1}{2}\left\{2\eta_T \mathcal{E}_{\mathbf{z}}(f_T) + 2\eta_{T-1}\mathcal{E}_{\mathbf{z}}(f_{T-1})\right\} \\
&\quad + \frac{1}{2} 2\eta_T \left\{\mathcal{E}_{\mathbf{z}}(f_T) - \mathcal{E}_{\mathbf{z}}(f_{T-1})\right\} + \frac{1}{2}\left\{2\eta_T - 2\eta_{T-1}\right\}\mathcal{E}_{\mathbf{z}}(f_{T-1}) \\
&= \frac{1}{3}\left\{2\eta_T \mathcal{E}_{\mathbf{z}}(f_T) + 2\eta_{T-1}\mathcal{E}_{\mathbf{z}}(f_{T-1}) + 2\eta_{T-2}\mathcal{E}_{\mathbf{z}}(f_{T-2})\right\} \\
&\quad + \frac{1}{2\times 3}\left\{2\eta_T\left[\mathcal{E}_{\mathbf{z}}(f_T) - \mathcal{E}_{\mathbf{z}}(f_{T-2})\right] + 2\eta_{T-1}\left[\mathcal{E}_{\mathbf{z}}(f_{T-1}) - \mathcal{E}_{\mathbf{z}}(f_{T-2})\right]\right\} \\
&\quad + \frac{1}{2} 2\eta_T \left\{\mathcal{E}_{\mathbf{z}}(f_T) - \mathcal{E}_{\mathbf{z}}(f_{T-1})\right\} + \frac{1}{2}\left\{2\eta_T - 2\eta_{T-1}\right\}\mathcal{E}_{\mathbf{z}}(f_{T-1}) \\
&\quad + \frac{1}{2\times 3}\left\{\left[2\eta_T - 2\eta_{T-2}\right] + \left[2\eta_{T-1} - 2\eta_{T-2}\right]\right\}\mathcal{E}_{\mathbf{z}}(f_{T-2}).
\end{aligned}$$

Repeating the above process by means of the decomposition

$$\begin{aligned}
&\frac{1}{k}\sum_{j=0}^{k-1} 2\eta_{T-j}\mathcal{E}_{\mathbf{z}}(f_{T-j}) = \frac{1}{k+1}\sum_{j=0}^{k} 2\eta_{T-j}\mathcal{E}_{\mathbf{z}}(f_{T-j}) \\
&+ \frac{1}{k(k+1)}\sum_{j=0}^{k-1} 2\eta_{T-j}\left\{\mathcal{E}_{\mathbf{z}}(f_{T-j}) - \mathcal{E}_{\mathbf{z}}(f_{T-k})\right\} + \frac{1}{k(k+1)}\sum_{j=0}^{k-1}\left\{2\eta_{T-j} - 2\eta_{T-k}\right\}\mathcal{E}_{\mathbf{z}}(f_{T-k})
\end{aligned}$$

with $k = 3, \ldots, T-1$, we know that

$$\begin{aligned}
2\eta_T \mathcal{E}_{\mathbf{z}}(f_T) &= \frac{1}{T}\sum_{j=0}^{T-1} 2\eta_{T-j}\mathcal{E}_{\mathbf{z}}(f_{T-j}) + \sum_{k=1}^{T-1}\frac{1}{k(k+1)}\sum_{j=0}^{k-1} 2\eta_{T-j}\left\{\mathcal{E}_{\mathbf{z}}(f_{T-j}) - \mathcal{E}_{\mathbf{z}}(f_{T-k})\right\} \\
&\quad + \sum_{k=1}^{T-1}\frac{1}{k(k+1)}\sum_{j=0}^{k-1}\left\{2\eta_{T-j} - 2\eta_{T-k}\right\}\mathcal{E}_{\mathbf{z}}(f_{T-k}).
\end{aligned}$$

Applying the same process to the sequence $\{2\eta_t \mathcal{E}_{\mathbf{z}}(f_*)\}_{t=1}^T$ yields

$$2\eta_T \mathcal{E}_{\mathbf{z}}(f_*) = \frac{1}{T}\sum_{j=0}^{T-1} 2\eta_{T-j}\mathcal{E}_{\mathbf{z}}(f_*) + \sum_{k=1}^{T-1}\frac{1}{k(k+1)}\sum_{j=0}^{k-1}\left\{2\eta_{T-j} - 2\eta_{T-k}\right\}\mathcal{E}_{\mathbf{z}}(f_*).$$

Hence the following error decomposition holds true:

$$\begin{aligned} 2\eta_T \left\{\mathcal{E}_{\mathbf{z}}(f_T) - \mathcal{E}_{\mathbf{z}}(f_*)\right\} &= \frac{1}{T}\sum_{t=1}^{T} 2\eta_t \left\{\mathcal{E}_{\mathbf{z}}(f_t) - \mathcal{E}_{\mathbf{z}}(f_*)\right\} \\ &+ \sum_{k=1}^{T-1} \frac{1}{k(k+1)} \sum_{t=T-k+1}^{T} 2\eta_t \left\{\mathcal{E}_{\mathbf{z}}(f_t) - \mathcal{E}_{\mathbf{z}}(f_{T-k})\right\} \\ &+ \sum_{k=1}^{T-1} \frac{1}{k+1} \left[\frac{1}{k}\sum_{t=T-k+1}^{T} 2\eta_t - 2\eta_{T-k}\right] \left\{\mathcal{E}_{\mathbf{z}}(f_{T-k}) - \mathcal{E}_{\mathbf{z}}(f_*)\right\}. \end{aligned} \tag{30}$$

Step 2: Average error in the first term of (30). Choosing $f = f_*$ in (29) and taking summation over $t = 1, \ldots, T$ together with (21) and Lemma 15 yield

$$\begin{aligned} \sum_{t=1}^{T} 2\eta_t \left\{\mathcal{E}_{\mathbf{z}}(f_t) - \mathcal{E}_{\mathbf{z}}(f_*)\right\} &\leq \|f_1 - f_*\|_K^2 - \|f_{T+1} - f_*\|_K^2 + \sum_{t=1}^{T} \eta_t^2 G_t^2 \\ &\leq \|f_*\|_K^2 + \sum_{t=1}^{T} \eta_1^2 c_q^2 (\kappa+1)^{2q+2} t^{q(1-\theta)-2\theta}. \end{aligned}$$

Since $1 > \theta > \frac{q}{q+1}$, we find $-2 < q(1-\theta) - 2\theta < 0$. Moreover, $q(1-\theta) - 2\theta < -1$ if and only if $\theta > \frac{q+1}{q+2}$. The following bound for the first term of (30) then follows

$$\begin{aligned} &\frac{1}{T}\sum_{t=1}^{T} 2\eta_t \left\{\mathcal{E}_{\mathbf{z}}(f_t) - \mathcal{E}_{\mathbf{z}}(f_*)\right\} \\ &\leq \begin{cases} \left(\|f_*\|_K^2 + C_{q,\kappa}\frac{(2+q)\theta - q}{(2+q)\theta - q - 1}\right) T^{-1}, & \text{when } \theta > \frac{q+1}{q+2}, \\ \left(\|f_*\|_K^2 + 2C_{q,\kappa}\right)(\log T) T^{-1}, & \text{when } \theta = \frac{q+1}{q+2}, \\ \left(\|f_*\|_K^2 + C_{q,\kappa}\frac{2}{q+1-(2+q)\theta}\right) T^{q-(2+q)\theta}, & \text{when } \theta < \frac{q+1}{q+2}, \end{cases} \end{aligned}$$

where $C_{q,\kappa}$ is the constant given by

$$C_{q,\kappa} = \eta_1^2 c_q^2 (\kappa+1)^{2q+2}.$$

Step 3: Moving average error in the second term of (30). Let $k \in \{1, \ldots, T-1\}$. Choosing $f = f_{T-k}$ in (29) and taking summation over $t = T-k+1, \ldots, T$ yield

$$\sum_{t=T-k+1}^{T} 2\eta_t \left\{\mathcal{E}_{\mathbf{z}}(f_t) - \mathcal{E}_{\mathbf{z}}(f_{T-k})\right\} \leq \|f_{T-k+1} - f_{T-k}\|_K^2 + \sum_{t=T-k+1}^{T} \eta_t^2 G_t^2$$

By Corollary 16,

$$\|f_{T-k+1} - f_{T-k}\|_K^2 \leq \eta_1^2 c_q^2 (\kappa+1)^{2(q+1)} (T-k)^{(1-\theta)q-2\theta}.$$

This bound is the term with $t = T-k+1$ of the following estimate which is a consequence of Lemma 15

$$\sum_{t=T-k+1}^{T} \eta_t^2 G_t^2 \leq \sum_{t=T-k+1}^{T} \eta_1^2 c_q^2 (\kappa+1)^{2q+2} t^{q(1-\theta)-2\theta}.$$

Hence

$$\sum_{t=T-k+1}^{T} 2\eta_t \left\{\mathcal{E}_{\mathbf{z}}(f_t) - \mathcal{E}_{\mathbf{z}}(f_{T-k})\right\} \leq C_{q,\kappa} \sum_{t=T-k}^{T} t^{q(1-\theta)-2\theta} = C_{q,\kappa} \sum_{t=T-k}^{T} t^{-q^*},$$

where we denote $q^* = 2\theta - q(1-\theta)$. We know that $0 < q^* < 2$ and $q^* = 1$ when $\theta = \frac{q+1}{q+2}$. So

$$\sum_{t=T-k+1}^{T} t^{-q^*} \leq \int_{T-k}^{T} x^{-q^*} dx \leq \begin{cases} \frac{T^{1-q^*} - (T-k)^{1-q^*}}{1-q^*}, & \text{when } \theta \neq \frac{q+1}{q+2}, \\ \log \frac{T}{T-k}, & \text{when } \theta = \frac{q+1}{q+2}. \end{cases}$$

When $\theta < \frac{q+1}{q+2}$, we have $q^* < 1$ and for $k \leq \frac{T}{2}$, we see from the mean value theorem that

$$\frac{T^{1-q^*} - (T-k)^{1-q^*}}{1-q^*} = T^{1-q^*} \frac{1 - (1-\frac{k}{T})^{1-q^*}}{1-q^*} \leq T^{1-q^*} \frac{(1-q^*)(1-\frac{k}{T})^{-q^*} \frac{k}{T}}{1-q^*}$$

which is exactly $(T-k)^{-q^*} k$. Thus,

$$\sum_{t=T-k}^{T} t^{-q^*} \leq (T-k)^{-q^*} k + (T-k)^{-q^*} \leq 2k(T-k)^{-q^*} \leq 2 \cdot 2^{q^*} T^{-q^*} k.$$

For $k \geq \frac{T}{2}$,

$$\sum_{t=T-k}^{T} t^{-q^*} \leq \frac{T^{1-q^*} - (T-k)^{1-q^*}}{1-q^*} + (T-k)^{-q^*} \leq \frac{T^{1-q^*}}{1-q^*}.$$

It follows that

$$\begin{aligned}
&\sum_{k=1}^{T-1} \frac{1}{k(k+1)} \sum_{t=T-k+1}^{T} 2\eta_t \left\{\mathcal{E}_{\mathbf{z}}(f_t) - \mathcal{E}_{\mathbf{z}}(f_{T-k})\right\} \\
&\leq C_{q,\kappa} \sum_{k \leq T/2} \frac{1}{k(k+1)} \sum_{t=T-k}^{T} t^{-q^*} + C_{q,\kappa} \sum_{T-1 \geq k > T/2} \frac{1}{k(k+1)} \sum_{t=T-k}^{T} t^{-q^*} \\
&\leq 2C_{q,\kappa} \sum_{k \leq T/2} \frac{1}{k(k+1)} 2^{q^*} T^{-q^*} k + 2C_{q,\kappa} \sum_{T-1 \geq k > T/2} \frac{1}{k(k+1)} \frac{T^{1-q^*}}{1-q^*} \\
&\leq 2C_{q,\kappa} \left(2^{q^*} + \frac{2}{1-q^*}\right) (\log T) T^{q(1-\theta)-2\theta}.
\end{aligned}$$

When $\theta = \frac{q+1}{q+2}$, we see from the mean value theorem that

$$\log \frac{T}{T-k} = -\log\left(1 - \frac{k}{T}\right) \leq \frac{k}{T} \frac{1}{1-\frac{k}{T}} = \frac{k}{T-k}.$$

It follows that

$$\begin{aligned}
&\sum_{k=1}^{T-1}\frac{1}{k(k+1)}\sum_{t=T-k+1}^{T}2\eta_t\left\{\mathcal{E}_{\mathbf{z}}(f_t)-\mathcal{E}_{\mathbf{z}}(f_{T-k})\right\}\\
&\le C_{q,\kappa}\sum_{k=1}^{T-1}\frac{1}{(T-k)k}=C_{q,\kappa}\frac{1}{T}\sum_{k=1}^{T-1}\left\{\frac{1}{k}+\frac{1}{T-k}\right\}\\
&\le 4C_{q,\kappa}\frac{\log T}{T}.
\end{aligned}$$

When $\theta>\frac{q+1}{q+2}$, we have $q^*>1$ and for $k\le\frac{T}{2}$,

$$\frac{T^{1-q^*}-(T-k)^{1-q^*}}{1-q^*}=T^{1-q^*}\frac{(1-\frac{k}{T})^{1-q^*}-1}{q^*-1}\le 2^{q^*}T^{-q^*}k.$$

For $k>\frac{T}{2}$, $\sum_{t=T-k}^{T}t^{-q^*}\le(k+1)2^{q^*}T^{-q^*}$. Then,

$$\begin{aligned}
&\sum_{k=1}^{T-1}\frac{1}{k(k+1)}\sum_{t=T-k+1}^{T}2\eta_t\left\{\mathcal{E}_{\mathbf{z}}(f_t)-\mathcal{E}_{\mathbf{z}}(f_{T-k})\right\}\\
&\le 2^{q^*+1}C_{q,\kappa}T^{-q^*}\sum_{k=1}^{T-1}\frac{1}{k+1}\le 2^{q^*+1}C_{q,\kappa}T^{-q^*}\log T\\
&\le 2^{q^*+1}C_{q,\kappa}\frac{1}{\mathrm{e}(q^*-1)}T^{-1}.
\end{aligned}$$

Thus the second term of (30) can also be bounded as

$$\begin{aligned}
&\sum_{k=1}^{T-1}\frac{1}{k(k+1)}\sum_{t=T-k+1}^{T}2\eta_t\left\{\mathcal{E}_{\mathbf{z}}(f_t)-\mathcal{E}_{\mathbf{z}}(f_{T-k})\right\}\\
&\le\begin{cases}
\frac{2^{q^*+1}C_{q,\kappa}}{\mathrm{e}(q^*-1)}T^{-1}, & \text{when } \theta>\frac{q+1}{q+2},\\
4C_{q,\kappa}(\log T)T^{-1}, & \text{when } \theta=\frac{q+1}{q+2},\\
2C_{q,\kappa}\left(2^{q^*}+\frac{2}{1-q^*}\right)(\log T)T^{q-(2+q)\theta}, & \text{when } \theta<\frac{q+1}{q+2}.
\end{cases}
\end{aligned}$$

Putting all the above estimates for the first two terms into (30), we see that the desired bound (27) holds true with the constant $\widetilde{C}_1$ given by

$$\widetilde{C}_1=\begin{cases}
\eta_1c_q^2(\kappa+1)^{2q+2}\left((2+q)\theta-q+2^{(2+q)\theta-q}\right), & \text{when } \theta>\frac{q+1}{q+2},\\
6\eta_1c_q^2(\kappa+1)^{2q+2}, & \text{when } \theta=\frac{q+1}{q+2},\\
\eta_1c_q^2(\kappa+1)^{2q+2}\left(2^{(2+q)\theta-q}+3\right), & \text{when } \theta<\frac{q+1}{q+2}.
\end{cases}$$

The proof of Lemma 17 is complete. ■

Lemma 17 is useful and can be used in a stochastic convex optimization problem, other than learning. In what follows, we shall see that how it can be used in our specified learning problems. For notational simplicity, with $\widetilde{R}>0$ we denote

$$\mathcal{M}_{\mathbf{z}}(\widetilde{R})=\sup_{f\in B_{\widetilde{R}}}\max\left\{\mathcal{E}_{\mathbf{z}}(f_\rho^V)-\mathcal{E}_{\mathbf{z}}(f),\ 0\right\}. \tag{31}$$

Proposition 18 *Under the assumptions of Lemma 17, we have*

$$\mathcal{E}_{\mathbf{z}}(f_T) - \mathcal{E}_{\mathbf{z}}(f_*) \leq \frac{3}{1-\theta}\left\{\mathcal{M}_{\mathbf{z}}\left(T^{\frac{1-\theta}{2}}\right) + \mathcal{F}_{\mathbf{z}}(f_*) + \mathcal{A}(f_*)\right\} + \frac{\|f_*\|_K^2}{2\eta_1}\Lambda_{T,\theta} + \widetilde{C}_1\Lambda_{T,\theta}, \quad (32)$$

where $\Lambda_{T,\theta}$ and $\widetilde{C}_1$ are defined in Lemma 17.

Proof Note that by Lemma 17, we have (27). We only need to estimate the second term of (27) denoted as

$$J_{T,\mathbf{z}} := \frac{T^\theta}{2\eta_1}\sum_{k=1}^{T-1}\frac{1}{k+1}\left[2\eta_{T-k} - \frac{1}{k}\sum_{t=T-k+1}^{T}2\eta_t\right]\left\{\mathcal{E}_{\mathbf{z}}(f_*) - \mathcal{E}_{\mathbf{z}}(f_{T-k})\right\}.$$

Denote $\widetilde{R} = T^{\frac{1-\theta}{2}}$. Lemma 15 tells us that $f_k \in B_{\widetilde{R}}$ for each $k = 1, \cdots, T$. It follows that for $k = 1, \cdots, T-1$,

$$\begin{aligned}\mathcal{E}_{\mathbf{z}}(f_*) - \mathcal{E}_{\mathbf{z}}(f_{T-k}) &= \left\{\left(\mathcal{E}_{\mathbf{z}}(f_*) - \mathcal{E}_{\mathbf{z}}(f_\rho^V)\right) - \left(\mathcal{E}(f_*) - \mathcal{E}(f_\rho^V)\right)\right\} \\ &\quad + \left(\mathcal{E}(f_*) - \mathcal{E}(f_\rho^V)\right) + \mathcal{E}_{\mathbf{z}}(f_\rho^V) - \mathcal{E}_{\mathbf{z}}(f_{T-k}) \\ &\leq \mathcal{F}_{\mathbf{z}}(f_*) + \mathcal{A}(f_*) + \mathcal{M}_{\mathbf{z}}(\widetilde{R}).\end{aligned}$$

By the choice of the stepsizes, $2\eta_{T-k} - \frac{1}{k}\sum_{t=T-k+1}^{T}2\eta_t \geq 0$ for any $k \in \{1, \ldots, T-1\}$. Therefore, $J_{T,\mathbf{z}}$ can be bounded by

$$J_{T,\mathbf{z}} \leq \frac{T^\theta}{2\eta_1}\sum_{k=1}^{T-1}\frac{1}{k+1}\left[2\eta_{T-k} - \frac{1}{k}\sum_{t=T-k+1}^{T}2\eta_t\right]\left\{\mathcal{F}_{\mathbf{z}}(f_*) + \mathcal{A}(f_*) + \mathcal{M}_{\mathbf{z}}(\widetilde{R})\right\}.$$

Now we need to bound the above summation. Note that, for each k,

$$2\eta_{T-k} - \frac{1}{k}\sum_{t=T-k+1}^{T}2\eta_t = \frac{2\eta_1}{k}\sum_{t=T-k+1}^{T}\left((T-k)^{-\theta} - t^{-\theta}\right).$$

Applying the mean value theorem to the function $g(x) = -x^{-\theta}$ on $[T-k, t]$ with $t \in \{T-k+1, \ldots, T\}$, we find that for some $c \in [T-k, t]$,

$$(T-k)^{-\theta} - t^{-\theta} = g(t) - g(T-k) = (t-(T-k))g'(c) \leq (t-(T-k))\theta(T-k)^{-\theta-1}.$$

Hence

$$\begin{aligned}&\sum_{k=1}^{T-1}\frac{1}{k+1}\left[2\eta_{T-k} - \frac{1}{k}\sum_{t=T-k+1}^{T}2\eta_t\right] \\ &\leq 2\eta_1\theta\sum_{k<T/2}\frac{(T-k)^{-\theta-1}}{k(k+1)}\sum_{t=T-k+1}^{T}(t-T+k) + \sum_{k\geq T/2}\frac{1}{k+1}2\eta_{T-k} \\ &\leq 2\eta_1\theta\sum_{k<T/2}\frac{(T-k)^{-\theta-1}}{k(k+1)}\frac{k(k+1)}{2} + \sum_{k\geq T/2}\frac{2}{T}2\eta_{T-k} \\ &\leq \eta_1\theta\sum_{k<T/2}(T-k)^{-\theta-1} + \frac{4\eta_1}{T}\sum_{k\geq T/2}(T-k)^{-\theta} \leq \frac{6\eta_1}{1-\theta}T^{-\theta}.\end{aligned}$$

Thus

$$J_{T,\mathbf{z}} \leq \frac{3}{1-\theta}\left\{\mathcal{F}_{\mathbf{z}}(f_*) + \mathcal{A}(f_*) + \mathcal{M}_{\mathbf{z}}(\widetilde{R})\right\}.$$

Then the desired bound follows from Lemma 17. ■

4.3 Computational Errors for Weighted Average and Best Iterate

Lemma 19 *Under the assumptions of Lemma 17, let $g_T = a_T$ (or $g_T = b_T$). Then*

$$\mathcal{E}_{\mathbf{z}}(g_T) - \mathcal{E}_{\mathbf{z}}(f_*) \leq \left(\frac{2\|f_*\|_K^2}{\eta_1} + \overline{C}_1\right)\overline{\Lambda}_{T,\theta},$$

where $\overline{\Lambda}_{T,\theta}$ is given by

$$\overline{\Lambda}_{T,\theta} = \begin{cases} \frac{1}{(2+q)\theta-(q+1)}T^{-(1-\theta)}, & \text{when } \theta > \frac{q+1}{q+2}, \\ (\log T)T^{-(1-\theta)}, & \text{when } \theta = \frac{q+1}{q+2}, \\ \frac{1}{(q+1)-(2+q)\theta}T^{-(\theta(1+q)-q)}, & \text{when } \theta < \frac{q+1}{q+2}, \end{cases} \tag{33}$$

and $\overline{C}_1$ is a positive constant depending on q, κ and θ (independent of T, m or f_* and given explicitly in the proof.)

Note that there is a subtle difference between $\overline{\Lambda}_{T,\theta}$ and $\Lambda_{T,\theta}$ defined by (39), where the latter term has an extra $\log T$ for $\theta < \frac{q+1}{q+2}$.

Proof For any $u \in \mathbb{R}$, we have

$$\sum_{t=1}^{T}\eta_t(\mathcal{E}_{\mathbf{z}}(f_t) - u) \geq \left(\sum_{t=1}^{T}\eta_t\right)\min_{t=1,\ldots,T}\mathcal{E}_{\mathbf{z}}(f_t) - \left(\sum_{t=1}^{T}\eta_t\right)u,$$

and by convexity of $\mathcal{E}_{\mathbf{z}}$,

$$\mathcal{E}_{\mathbf{z}}(a_T) = \mathcal{E}_{\mathbf{z}}\left(\sum_{t=1}^{T}\omega_t f_t\right) \leq \sum_{t=1}^{T}\omega_t\mathcal{E}_{\mathbf{z}}(f_T) = \frac{1}{\sum_{t=1}^{T}\eta_t}\sum_{t=1}^{T}\eta_t\mathcal{E}_{\mathbf{z}}(f_t).$$

Therefor, we have

$$\mathcal{E}_{\mathbf{z}}(b_T) - u \leq \frac{1}{\sum_{t=1}^{T}\eta_t}\sum_{t=1}^{T}\eta_t(\mathcal{E}_{\mathbf{z}}(f_t) - u)$$

and

$$\mathcal{E}_{\mathbf{z}}(a_T) - u \leq \frac{1}{\sum_{t=1}^{T}\eta_t}\sum_{t=1}^{T}\eta_t(\mathcal{E}_{\mathbf{z}}(f_t) - u).$$

We thus get

$$\mathcal{E}_{\mathbf{z}}(g_T) - \mathcal{E}_{\mathbf{z}}(f^*) \leq \frac{1}{\sum_{t=1}^{T}\eta_t}\sum_{t=1}^{T}\eta_t(\mathcal{E}_{\mathbf{z}}(f_t) - \mathcal{E}_{\mathbf{z}}(f^*)). \tag{34}$$

Following Step 2 of the proof of Lemma 17, we have

$$\begin{aligned}
&\sum_{t=1}^{T} 2\eta_t \left\{\mathcal{E}_{\mathbf{z}}(f_t) - \mathcal{E}_{\mathbf{z}}(f_*)\right\} \\
&\leq \begin{cases} \left(\|f_*\|_K^2 + C_{q,\kappa}\frac{(2+q)\theta - q}{(2+q)\theta - q - 1}\right), & \text{when } \theta > \frac{q+1}{q+2}, \\ \left(\|f_*\|_K^2 + 2C_{q,\kappa}\right)(\log T), & \text{when } \theta = \frac{q+1}{q+2}, \\ \left(\|f_*\|_K^2 + C_{q,\kappa}\frac{2}{q+1-(2+q)\theta}\right) T^{(1+q)-(2+q)\theta}, & \text{when } \theta < \frac{q+1}{q+2}. \end{cases}
\end{aligned}$$

Introducing the above inequality into (34), and using $\sum_{t=1}^{T} \eta_t \geq \eta_1 \int_{t=1}^{T+1} u^{-\theta} du \geq \eta_1 T^{1-\theta}/\mathrm{e}$, we get our desired result with $\overline{C}_1$ given by

$$\overline{C}_1 = \begin{cases} 3\eta_1 c_q^2 (\kappa+1)^{2q+2}\left((2+q)\theta - q\right), & \text{when } \theta > \frac{q+1}{q+2}, \\ 3\eta_1 c_q^2 (\kappa+1)^{2q+2}, & \text{when } \theta = \frac{q+1}{q+2}, \\ 3\eta_1 c_q^2 (\kappa+1)^{2q+2}, & \text{when } \theta < \frac{q+1}{q+2}. \end{cases}$$

■

While the above proof is shorter and easier than the proof of Lemma 17, it is surprising that the computational error bounds for the last iterate and the average (or the best one) are roughly of the same order.

4.4 Iterate Bound and Computational Error for Smooth Loss Functions

The following result can be proved by using the fact that $V'(y,\cdot)$ is Lipschitz. Its proof is a simple modification to RKHS of that in Nesterov (2004, Theorem 2.1.14), where the cases of Euclidean spaces are studied. For completeness, we provide the proof in Appendix A.

Lemma 20 *Assume that $V(y,\cdot)$ is differentiable and $V'(y,\cdot)$ is Lipschitz continuous with constant $L > 0$. Let $0 < \eta_t \leq (L\kappa^2)^{-1}$ for all $t \in \mathbb{N}$. Then we have*

$$\mathcal{E}_{\mathbf{z}}(f_T) - \mathcal{E}_{\mathbf{z}}(f_*) \leq \frac{\|f_*\|_K^2}{\sum_{k=1}^{T} 2\eta_k}.$$

In particular, if $\eta_t = \eta_1 t^{-\theta}$ with $\theta \in [0,1)$ satisfying $\eta_1 \leq (L\kappa^2)^{-1}$, then

$$\mathcal{E}_{\mathbf{z}}(f_T) - \mathcal{E}_{\mathbf{z}}(f_*) \leq \frac{2\|f_*\|_K^2}{\eta_1} T^{\theta-1}.$$

Using the above lemma, we can bound the iterates as follows.

Lemma 21 *Under the assumptions of Lemma 20, we have for $t = 1, \cdots, T$,*

$$\|f_{t+1}\|_K \leq \sqrt{2|V|_0 \sum_{k=1}^{t} \eta_k}.$$

In particular, if $\eta_t = \eta_1 t^{-\theta}$ with $\theta \in [0,1)$ satisfying $\eta_1 \leq \frac{1-\theta}{2|V|_0}$, then

$$\|f_{t+1}\|_K \leq t^{\frac{1-\theta}{2}}.$$

Proof Choosing $f_* = 0$ in (46) (see Appendix A), we get for $k = 1, \cdots, t$,

$$\|f_{k+1}\|_K^2 \le \|f_k\|_K^2 + 2\eta_k(\mathcal{E}_{\mathbf{z}}(0) - \mathcal{E}_{\mathbf{z}}(f_{k+1})) \le \|f_k\|_K^2 + 2\eta_k|V|_0.$$

Applying this relationship iteratively for $k = t, \cdots, 1$, with $f_1 = 0$, we get

$$\|f_{t+1}\|_K^2 \le 2|V|_0 \sum_{k=1}^{t} \eta_k,$$

which leads to the first conclusion. The second inequality can be proved by noting that

$$\sum_{k=1}^{t} \eta_k = \eta_1 \sum_{k=1}^{t} k^{-\theta} \le \eta_1 \left(1 + \frac{t^{1-\theta} - 1}{1-\theta}\right) \le \eta_1 \frac{t^{1-\theta}}{1-\theta}.$$

■

5. Sample Error and Finite Sample Bounds

In this subsection, we will estimate sample errors and then prove our main results.

5.1 Sample Error

We first bound the sample error term $\mathcal{F}_{\mathbf{z}}(f_*)$ for some fixed $f_* \in \mathcal{H}_K$ as follows. This is done by applying Bernstein inequality, see Appendix B for the proof.

Lemma 22 *Assume conditions (4) and (5) hold. For any $f_* \in \mathcal{H}_K$ with $\|f_*\|_K \le R$, where $R \ge 1$, with confidence at least $1 - \frac{\delta}{2}$,*

$$\mathcal{F}_{\mathbf{z}}(f_*) \le (C_1' + 2\sqrt{c_\tau}) \log \frac{2}{\delta} \max \left\{ \frac{R^{q+1}}{m}, \left(\frac{R^{2+q-\tau}}{m}\right)^{\frac{1}{2-\tau}}, \mathcal{A}(f_*) \right\}, \tag{35}$$

where C_1' is a positive constant independent of T, m, δ, given explicitly in the proof.

We next bound the empirical process over the ball $B_{\widetilde{R}}$ for some $\widetilde{R} > 0$ in the following lemma. It is essentially contained in Wu et al. (2007). We provide a short proof in Appendix B for the sake of completeness.

Lemma 23 *Assume (4) with $q \ge 0$, (5) with $\tau \in [0, 1]$, (8) with $\beta \in (0, 1]$ and (9) with $\zeta \in (0, 2)$. Let $\widetilde{R} > 1$. Then with confidence at least $1 - \frac{\delta}{2}$, there holds for every $g \in B_{\widetilde{R}}$,*

$$\begin{aligned} &\left(\mathcal{E}(g) - \mathcal{E}(f_\rho^V)\right) - \left(\mathcal{E}_{\mathbf{z}}(g) - \mathcal{E}_{\mathbf{z}}(f_\rho^V)\right) \\ \le &\frac{1}{2}\left(\mathcal{E}(g) - \mathcal{E}(f_\rho^V)\right) + C_3' \log \frac{2}{\delta} \max \left\{ \left(\frac{\widetilde{R}^{\frac{q(2+\zeta)+(4-2\tau+\zeta\tau)}{2}}}{m}\right)^{\frac{2}{4-2\tau+\zeta\tau}}, \frac{\widetilde{R}^{q+1}}{m^{\frac{2}{2+\zeta}}}, \left(\frac{\widetilde{R}^{2+q-\tau}}{m}\right)^{\frac{1}{2-\tau}} \right\}, \end{aligned} \tag{36}$$

and

$$\mathcal{M}_{\mathbf{z}}(\widetilde{R}) \leq C_3' \log\frac{2}{\delta} \max\left\{ \left(\frac{\widetilde{R}^{\frac{q(2+\zeta)+(4-2\tau+\zeta\tau)}{2}}}{m} \right)^{\frac{2}{4-2\tau+\zeta\tau}}, \ \frac{\widetilde{R}^{q+1}}{m^{\frac{2}{2+\zeta}}}, \ \left(\frac{\widetilde{R}^{2+q-\tau}}{m} \right)^{\frac{1}{2-\tau}} \right\}. \tag{37}$$

Here C_3' is a positive constant independent of T, m, δ, given explicitly in the proof.

5.2 Deriving the Finite Sample Bounds

We have the following result, which will be used for the proof of Theorem 4.

Proposition 24 *Assume (4) with $q \geq 0$, (5) with $\tau \in [0,1]$, (8) with $\beta \in (0,1]$ and (9) with $\zeta \in (0,2)$. Let $\eta_t = \eta_1 t^{-\theta}$ with $0 < \theta < 1$ satisfying $\theta > \frac{q}{q+1}$ and η_1 satisfying (10). Let $f_* \in \mathcal{H}_K$ be such that $\|f_*\|_K \leq R$, where $R \geq 1$. If $1 \leq R \leq T^{\frac{1-\theta}{2}}$ and $T^{\frac{q(1-\theta)}{2}} m^{-\frac{2}{2+\zeta}} \leq 1$, then with confidence $1-\delta$, we have*

$$\mathcal{E}(f_T) - \mathcal{E}(f_\rho^V) \leq \widetilde{C}_3 \log\frac{2}{\delta} \max\left\{ \left(\frac{T^{\frac{(1-\theta)(q(2+\zeta)+(4-2\tau+\zeta\tau))}{4}}}{m} \right)^{\frac{2}{4-2\tau+\zeta\tau}}, \ R^2\Lambda_T, \ \mathcal{A}(f_*) \right\}, \tag{38}$$

where Λ_T is defined by

$$\Lambda_T = \begin{cases} T^{-(1-\theta)}, & \text{when } \theta > \frac{q+1}{q+2}, \\ (\log T) T^{-(1-\theta)}, & \text{when } \theta = \frac{q+1}{q+2}, \\ (\log T) T^{-(\theta(1+q)-q)}, & \text{when } \theta < \frac{q+1}{q+2}, \end{cases} \tag{39}$$

and $\widetilde{C}_3$ is a positive constant independent of T, m, δ, given explicitly in the proof.

Proof Recall Lemma 13. Let $\tilde{R} = T^{\frac{1-\theta}{2}}$. Introducing with (32), we have

$$\begin{aligned} \mathcal{E}(f_T) - \mathcal{E}(f_\rho^V) \leq & \left\{ \left(\mathcal{E}(f_T) - \mathcal{E}(f_\rho^V) \right) - \left(\mathcal{E}_{\mathbf{z}}(f_T) - \mathcal{E}_{\mathbf{z}}(f_\rho^V) \right) \right\} + \frac{3}{1-\theta} \mathcal{M}_{\mathbf{z}}\left(\tilde{R}\right) \\ & + \frac{4-\theta}{1-\theta} \left(\mathcal{F}_{\mathbf{z}}(f_*) + \mathcal{A}(f_*) \right) + \frac{R^2}{2\eta_1} \Lambda_{T,\theta} + \widetilde{C}_1 \Lambda_{T,\theta}. \end{aligned}$$

Applying Lemmas 22 and 23 with $g = f_T$, with $R \in [1, \widetilde{R}]$ and the notation Λ_T defined by (39), we know that with confidence at least $1-\delta$,

$$\begin{aligned} \mathcal{E}(f_T) - \mathcal{E}(f_\rho^V) \ \leq \ & C_4' \log\frac{2}{\delta} \max\left\{ \left(\frac{\widetilde{R}^{\frac{q(2+\zeta)+(4-2\tau+\zeta\tau)}{2}}}{m} \right)^{\frac{2}{4-2\tau+\zeta\tau}}, \ \frac{\widetilde{R}^{q+1}}{m^{\frac{2}{2+\zeta}}}, \right. \\ & \left. \frac{\widetilde{R}^{\frac{2+q-\tau}{2-\tau}}}{m^{\frac{1}{2-\tau}}}, \ R^2\Lambda_T, \ \mathcal{A}(f_*) \right\} + \frac{1}{2} \left(\mathcal{E}(f_T) - \mathcal{E}(f_\rho^V) \right), \end{aligned} \tag{40}$$

where C_4' is the constant given by

$$C_4' = \frac{4-\theta}{1-\theta} C_3' + \frac{4-\theta}{1-\theta} \left(1 + C_1' + 2\sqrt{c_\tau} \right) + \left(\frac{1}{2\eta_1} + \widetilde{C}_1 \right) c_{\theta,q}.$$

Here $c_{\theta,q}$ is given by

$$c_{\theta,q} = \begin{cases} \frac{1}{|(q+2)\theta-(q+1)|}, & \text{when } \theta \neq \frac{q+1}{q+2}, \\ 1, & \text{when } \theta = \frac{q+1}{q+2}. \end{cases}$$

Since $\widetilde{R}^q m^{-2/(2+\zeta)} \leq 1$, $\tau \in [0,1]$ and $\zeta \in (0,2)$, one finds

$$\left(\frac{\widetilde{R}^{\frac{q(2+\zeta)+(4-2\tau+\zeta\tau)}{2}}}{m}\right)^{\frac{2}{4-2\tau+\zeta\tau}} \cdot \frac{m^{\frac{2}{2+\zeta}}}{\widetilde{R}^{q+1}} = \left\{\frac{\widetilde{R}^q}{m^{\frac{2}{2+\zeta}}}\right\}^{\frac{-(1-\tau)(2-\zeta)}{4-2\tau+\zeta\tau}} \geq 1,$$

and

$$\left(\frac{\widetilde{R}^{\frac{q(2+\zeta)+(4-2\tau+\zeta\tau)}{2}}}{m}\right)^{\frac{2}{4-2\tau+\zeta\tau}} \cdot \frac{m^{\frac{1}{2-\tau}}}{\widetilde{R}^{\frac{2+q-\tau}{2-\tau}}} = \left(\widetilde{R}^{2q(1-\tau)} m^{\tau}\right)^{\frac{\zeta}{(2-\tau)(4-2\tau+\zeta\tau)}} \geq 1.$$

Subtracting $\frac{1}{2}\left(\mathcal{E}(f_T) - \mathcal{E}(f_\rho^V)\right)$ from both sides of (40), and setting $\widetilde{C}_3 = 2C_4'$, we get the desired results. ■

Now we are in a position to prove the probabilistic upper bounds stated in Theorem 4.

Proof of Theorem 4 We use Proposition 24 with $f_* = f_\lambda$ to prove our result. Define a power index $\tilde{\theta}$ as

$$\tilde{\theta} = \begin{cases} 1-\theta, & \text{when } \theta \geq \frac{q+1}{q+2}, \\ \theta(1+q)-q, & \text{when } \theta < \frac{q+1}{q+2}. \end{cases} \tag{41}$$

Comparing this with the definition (39) for Λ_T, we see that

$$\Lambda_T = \begin{cases} T^{-\tilde{\theta}}, & \text{when } \theta > \frac{q+1}{q+2}, \\ T^{-\tilde{\theta}} \log T, & \text{when } \theta \leq \frac{q+1}{q+2}. \end{cases}$$

From the definition of $\mathcal{D}(\lambda)$, we have

$$\mathcal{A}(f_\lambda) \leq \mathcal{D}(\lambda) \qquad \text{and} \qquad \lambda \|f_\lambda\|_K^2 \leq \mathcal{D}(\lambda), \tag{42}$$

which imply $\|f_\lambda\|_K \leq \sqrt{\mathcal{D}(\lambda)/\lambda} = R$. Balancing the orders of the last two terms of (38) by setting

$$\lambda = \Lambda_T, \tag{43}$$

we find that the last two terms of (38) can be bounded as

$$\max\left\{R^2\Lambda_T,\ \mathcal{A}(f_\lambda)\right\} \leq \mathcal{D}(\lambda) \leq c_\beta \lambda^\beta \leq \begin{cases} c_\beta T^{-\beta\tilde{\theta}}, & \text{when } \theta > \frac{q+1}{q+2}, \\ c_\beta T^{-\beta\tilde{\theta}} \log T, & \text{when } \theta \leq \frac{q+1}{q+2}. \end{cases}$$

Then we balance the above main part with the first term of (38) by setting

$$\left(\frac{T^{\frac{(1-\theta)(q(2+\zeta)+(4-2\tau+\zeta\tau))}{4}}}{m}\right)^{\frac{2}{4-2\tau+\zeta\tau}} = T^{-\beta\tilde{\theta}}.$$

This leads us to choose T to be the integer part of

$$\lceil m^\gamma \rceil, \text{ where } \gamma := \frac{2}{\left(\frac{1-\theta}{2} + \beta\tilde{\theta}\right)(4 - 2\tau + \zeta\tau) + \frac{q(2+\zeta)(1-\theta)}{2}}. \tag{44}$$

With this choice, the main part of (38) can be bounded as

$$\begin{aligned}
&\max\left\{ \left(\frac{T^{\frac{(1-\theta)(q(2+\zeta)+(4-2\tau+\zeta\tau))}{4}}}{m} \right)^{\frac{2}{4-2\tau+\zeta\tau}}, \ R^2\Lambda_T, \ \mathcal{A}(f_*) \right\} \\
&\leq \begin{cases} 2c_\beta m^{-\beta\tilde{\theta}\gamma}, & \text{when } \theta > \frac{q+1}{q+2}, \\ 2\gamma c_\beta m^{-\beta\tilde{\theta}\gamma}\log m, & \text{when } \theta \leq \frac{q+1}{q+2}. \end{cases}
\end{aligned}$$

Noticing from the definition of $\tilde{\theta}$, one can easily prove that $\tilde{\theta} \leq 1-\theta$. Then $R/\sqrt{c_\beta} \leq \sqrt{\lambda^{\beta-1}} \leq \Lambda_T^{\frac{\beta-1}{2}} \leq T^{\frac{1-\theta}{2}}$ and the restriction for R in Theorem 24 is satisfied up to constants. The restriction $T^{\frac{q(1-\theta)}{2}} m^{-\frac{2}{2+\zeta}} \leq 1$ is also satisfied (up to a constant), because

$$T^{\frac{q(1-\theta)}{2}} \lesssim m^{\frac{q(1-\theta)\gamma}{2}} \lesssim m^{\frac{2}{2+\zeta}}.$$

Observe that $\gamma \leq \frac{2}{1-\theta}$. So by Proposition 24, with confidence $1-\delta$, we have

$$\mathcal{E}(f_T) - \mathcal{E}(f_\rho^V) \leq \begin{cases} 2c_\beta \widetilde{C}_3 m^{-\beta\tilde{\theta}\gamma}\log\frac{2}{\delta}, & \text{when } \theta > \frac{q+1}{q+2}, \\ \frac{4c_\beta\widetilde{C}_3}{1-\theta} m^{-\beta\tilde{\theta}\gamma}\log m\log\frac{2}{\delta}, & \text{when } \theta \leq \frac{q+1}{q+2}. \end{cases}$$

Observe that the power index $\beta\tilde{\theta}\gamma$ is

$$\beta\tilde{\theta}\gamma = \begin{cases} \frac{\beta}{\beta(2-\tau+\zeta\tau/2) + \left\{\frac{2-\tau+\zeta\tau/2}{2} + \frac{q(1+\zeta/2)}{2}\right\}}, & \text{when } \theta \geq \frac{q+1}{q+2}, \\ \frac{\beta}{\beta(2-\tau+\zeta\tau/2) + \frac{1-\theta}{\theta(1+q)-q}\left\{\frac{2-\tau+\zeta\tau/2}{2} + \frac{q(1+\zeta/2)}{2}\right\}}, & \text{when } \theta < \frac{q+1}{q+2}, \end{cases}$$

while the index γ can be expressed by (11). Then our desired learning rates are verified by setting the constant $\widetilde{C} = 2c_\beta\widetilde{C}_3$ when $\theta > \frac{q+1}{q+2}$ while $\widetilde{C} = \frac{4c_\beta\widetilde{C}_3}{1-\theta}$ when $\theta \leq \frac{q+1}{q+2}$. The proof of Theorem 4 is complete. ■

Proof of Theorem 7 We only sketch the proof for the case $g_T = a_T$. By applying Lemma 15, it is easy to prove that

$$\|a_T\|_K \leq T^{\frac{1-\theta}{2}}.$$

With the upper bound on a_T and Lemma 19, a similar argument as that for Theorem 4, one can prove the results. We omit the details. ■

Proof of Theorem 8 With Lemmas 20, 21, and a similar approach as that for Theorem 4, we can prove the convergence results for smooth loss functions. We omit the details. ■

Proof of Theorem 11 We use Theorem 4 to prove the results. The hinge loss satisfies (4) with $q = 0$ and $c_q = \frac{1}{2}$, $|V|_0 = 1$ and $\|f_\rho^V\|_\infty = 1$ where f_ρ^V is the Bayes rule f_c. Condition (5) is valid with $\tau = 0$ and $c_\tau = 1$. Since $\theta > 1/2$, by simple calculations, one finds that $\gamma = \frac{1}{(1-\theta)(2\beta+1)}$ and $\alpha = \frac{\beta}{2\beta+1}$.

Using the comparison theorem from Zhang (2004), we have

$$\mathcal{R}\left(\operatorname{sign}(f_T)\right) - \mathcal{R}(f_c) \leq \mathcal{E}(f_T) - \mathcal{E}(f_\rho^V).$$

So the desired probabilistic upper bound (14) for the hinge loss follows from the above inequality and Theorem 4.

It remains to prove the second part of the theorem. Since $0 < \epsilon < \frac{1}{3}$, the restriction $\beta > \frac{1-3\epsilon}{1+6\epsilon}$ for the approximation order tells us that

$$\alpha = \frac{\beta}{2\beta+1} = \frac{1}{2+1/\beta} \geq \frac{1}{3} - \epsilon.$$

The proof of Theorem 11 is complete. ■

Proof of Theorem 12 Since $0 < \epsilon < \frac{1}{3}$, the restriction $\beta > \frac{4-3\epsilon}{4+6\epsilon}$ for the approximation order tells us that the parameter θ satisfies $\frac{1}{2} < \theta < 1$ and the index

$$\gamma = \frac{1}{(1-\theta)(2\beta+1)} = \frac{2}{3} + \epsilon.$$

Finally we find that the index

$$\alpha = \frac{\beta}{2\beta+1} = \frac{1}{2+1/\beta} \geq \frac{1}{3} - \frac{\epsilon}{4}.$$

So the desired probabilistic upper bound follows from the first conclusion of Theorem 11. The proof of Theorem 12 is complete. ■

6. Conclusions

This paper proposes and studies iterative regularization approaches for learning with convex loss functions. More precisely, we study how regularization can be achieved by early stopping an empirical iteration induced by the subgradient method, or gradient descent in the case that the loss is also smooth. Finite sample bounds are established providing indications on how to suitably choose the stepsize and the stopping rule. Different to classical results on the subgradient method, we analyze the behavior of the last iterate showing that it has essentially the same properties of the average, to the best, iterate. These results provide a theoretical foundation for early stopping with convex losses.

Beyond the analysis in the paper our error decomposition provides an approach to incorporating statistics and optimization aspects in the analysis of learning algorithms. While a natural development will be to sharpen the bounds and perform extensive empirical tests, we hope the study in the paper can help deriving novel and faster algorithms, for example analyzing accelerations (Nesterov, 2004), or distributed approaches, within the framework we propose.

Acknowledgments

The work described in this paper is supported partially by the Research Grants Council of Hong Kong [Project No. CityU 104012] and by National Natural Science Foundation of China under Grant 11461161006. LR is supported by the FIRB project RBFR12M3AC and the Center for Minds, Brains and Machines (CBMM), funded by NSF STC award CCF-1231216. JL is now within LCSL, MIT & Istituto Italiano di Tecnologia. The authors would like to thank the referees and Dr. Yunlong Feng for their valuable comments.

References

A. Aizerman, E. M. Braverman and L. I. Rozoner. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 25:821–837, 1964.

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical society*, 68:337–404, 1950.

P. Bartlett and M. Traskin. Adaboost is consistent. *Journal of Machine Learning Research*, 8:2347–2368, 2007.

F. Bauer, S. Pereverzev and L. Rosasco. On regularization algorithms in learning theory. *Journal of Complexity*, 1:52–72, 2007.

D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, second edition, 1999.

P. J. Bickel, Y. Ritov and A. Zakai. Some theory for generalized boosting algorithms. *Journal of Machine Learning Research*, 7:705–732, 2006.

G. Blanchard and N. Krämer. Optimal learning rates for kernel conjugate gradient regression. *Advances in Neural Information Processing Systems*, 226–234, 2010.

O. Bousquet and L. Bottou. The tradeoffs of large scale learning. *Advances in Neural Information Processing Systems*, 161–168, 2008.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

S. Boyd, L. Xiao and A. Mutapcic. *Subgradient Methods*. Lecture notes of EE392o, Stanford University, Autumn Quarter, 2004 (2003).

P. Buhlmann and B. Yu. Boosting with the L_2 loss: regression and classification. *Journal of the American Statistical Association*, 462:324–339, 2003.

A. Caponnetto and E. De Vito. Optimal rates for regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7:331–368, 2007.

A. Caponnetto and Y. Yao. Cross-validation based adaptation for regularization operators in learning theory. *Analysis and Applications*, 8:161–183, 2010.

V. Chandrasekaran and M. I. Jordan. Computational and statistical tradeoffs via convex relaxation. *Proceedings of the National Academy of Sciences*, 110:E1181–E1190, 2013.

C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995.

F. Cucker and D.-X. Zhou. *Learning Theory: An Approximation Theory Viewpoint*. Cambridge University Press, 2007.

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.

E. De Vito, L. Rosasco, A. Caponnetto, U. Giovannini and F. Odone. Learning from examples as an inverse problem. *Journal of Machine Learning Research*, 6:1532–4435, 2005.

H. W. Engl, M. Hanke and A. Neubauer. *Regularization of Inverse Problems*. Kluwer, 1996.

L. Gerfo, L. Rosasco, F. Odone, E. De Vito and A. Verri. Spectral algorithms for supervised learning. *Neural Computation*, 20:1873–1897, 2008.

T. Hu, J. Fan, Q. Wu and D.-X. Zhou. Regularization schemes for minimum error entropy principle. *Analysis and Applications*, 13:437–455, 2015.

T. Jaakkola, M. Diekhaus and D. Haussler. Using the Fisher kernel method to detect remote protein homologies. *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*, 99:149–158, 1999.

W. Jiang. Process consistency for adaboost. *Annals of Statistics*, 32:13–29, 2004.

B. Kaltenbacher, A. Neubauer and O. Scherzer. *Iterative Regularization Methods for Nonlinear Ill-posed Problems*. Radon Series on Computational and Applied Mathematics, de Gruyter, Berlin, 2008.

Y. LeCun, L. Bottou, G. Orr and K. Muller. *Efficient Backprop*. Neural Networks: Tricks of the Trade, Springer, 1998.

A. Nemirovskii. The regularization properties of adjoint gradient method in ill-posed problems. *USSR Computational Mathematics and Mathematical Physics*, 26:7–16, 1986.

Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer, 2004.

F. Orabona. Simultaneous model selection and optimization through parameter-free stochastic learning. *Advances in Neural Information Processing Systems*, 1116–1124, 2014.

B. Polyak. *Introduction to Optimization*. Optimization Software, 1987.

G. Raskutti, M. J. Wainwright and B. Yu. Early stopping and non-parametric regression: an optimal data-dependent stopping rule. *Journal of Machine Learning Research*, 15:335–366, 2014.

L. Rosasco and S. Villa. Learning with incremental iterative regularization. *Advances in Neural Information Processing Systems*, 1621–1629, 2015.

F. Rosenblatt. *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Washington DC: Spartan Books, 1962.

O. Shamir and T. Zhang. Stochastic gradient descent for non-smooth optimization: convergence results and optimal averaging schemes. *International Conference on Machine Learning*, 71–79, 2013.

S. Smale and D.-X. Zhou. Estimating the approximation error in learning theory. *Analysis and Applications*, 1:17–41, 2003.

S. Sra, S. Nowozin and S. J. Wright. *Optimization for Machine Learning, Neural Information Processing Series*. MIT Press, 2011.

I. Steinwart. Oracle inequailities for support vector machines that are based on random entropy numbers. *Journal of Complexity*, 25:437–454, 2009.

I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, 2008.

A. N. Tikhonov and V. Y. Arsenin. *Solution of Ill-posed Problems*. Winston & Sons, 1977.

V. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, 1998.

M. J. Wainwright. Structured regularizers for high-dimensional problems: statistical and computational issues. *Annual Review of Statistics and Its Application*, 1:233–253, 2014.

Q. Wu, Y. Ying and D.-X. Zhou. Multi-kernel regularized classifiers. *Journal of Complexity*, 23:108–134, 2007.

Y. Yao, L. Rosasco and A. Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26:289–315, 2007.

G. B. Ye and D.-X. Zhou. Fully online classification by regularization. *Applied and Computational Harmonic Analysis*, 23:198–214, 2007.

Y. Ying and D.-X. Zhou. Online regularized classification algorithms. *IEEE Transaction on Information Theory*, 52:4775–4788, 2006.

T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, 32:56-85, 2004.

T. Zhang and B. Yu. Boosting with early stopping: convergence and consistency. *Annals of Statistics*, 33:1538–1579, 2005.

D.-X. Zhou. Capacity of reproducing kernel spaces in learning theory. *IEEE Transaction on Information Theory*, 49:1743–1752, 2003.

Appendix A. Proof of Lemma 20

Proof Since $V'(y,\cdot)$ is Lipschitz with constant L for any $y \in Y$, we have for any $a, b \in \mathbb{R}$,

$$V(y,b) \leq V(y,a) + V'(y,a)(b-a) + \frac{L}{2}(b-a)^2.$$

Choosing $y = y_j$, $b = f_{t+1}(x_j)$ and $a = f_t(x_j)$, according to the properties (22) and (23), we get for $j = 1, \cdots, m$ and $t \in \mathbb{N}$,

$$V(y_j, f_{t+1}(x_j)) \leq V(y_j, f_t(x_j)) + V'(y_j, f_t(x_j))\langle f_{t+1} - f_t, K_{x_j}\rangle_K + \frac{L\kappa^2}{2}\|f_{t+1} - f_t\|_K^2.$$

Summing up over $j = 1, \cdots, m$, with $g_t = \frac{1}{m}\sum_{j=1}^m V'(y_j, f_t(x_j))K_{x_j}$, we get

$$\mathcal{E}_{\mathbf{z}}(f_{t+1}) \leq \mathcal{E}_{\mathbf{z}}(f_t) + \langle f_{t+1} - f_t, g_t\rangle_K + \frac{L\kappa^2}{2}\|f_{t+1} - f_t\|_K^2.$$

Introducing with (1) and noting that $\eta_t \leq (L\kappa^2)^{-1}$, we get

$$\mathcal{E}_{\mathbf{z}}(f_{t+1}) \leq \mathcal{E}_{\mathbf{z}}(f_t) - \frac{\eta_t}{2}\|g_t\|_K^2. \tag{45}$$

By the convexity of $V(y,\cdot)$, it is easy to prove that

$$\mathcal{E}_{\mathbf{z}}(f_t) \leq \mathcal{E}_{\mathbf{z}}(f_*) + \langle f_t - f_*, g_t\rangle_K.$$

Introducing this inequality into (45), we get

$$\begin{aligned}
\mathcal{E}_{\mathbf{z}}(f_{t+1}) &\leq \mathcal{E}_{\mathbf{z}}(f_*) + \frac{1}{2\eta_t}\left(2\eta_t\langle f_t - f_*, g_t\rangle_K - \eta_t^2\|g_t\|_K^2\right) \\
&= \mathcal{E}_{\mathbf{z}}(f_*) + \frac{1}{2\eta_t}\left(\|f_t - f_*\|_K^2 - \|f_t - f_* - \eta_t g_t\|_K^2\right) \\
&= \mathcal{E}_{\mathbf{z}}(f_*) + \frac{1}{2\eta_t}\left(\|f_t - f_*\|_K^2 - \|f_{t+1} - f_*\|_K^2\right),
\end{aligned}$$

so that,

$$2\eta_t(\mathcal{E}_{\mathbf{z}}(f_{t+1}) - \mathcal{E}_{\mathbf{z}}(f_*)) \leq \|f_t - f_*\|_K^2 - \|f_{t+1} - f_*\|_K^2. \tag{46}$$

Summing up over $t = 1 \cdots, T$, with $f_1 = 0$, we have

$$\sum_{t=1}^T 2\eta_t(\mathcal{E}_{\mathbf{z}}(f_{t+1}) - \mathcal{E}_{\mathbf{z}}(f_*)) \leq \|f_1 - f_*\|_K^2 - \|f_{T+1} - f_*\|_K^2 \leq \|f_*\|_K^2.$$

By (45), we have $\mathcal{E}_{\mathbf{z}}(f_{T+1}) \leq \mathcal{E}_{\mathbf{z}}(f_{t+1})$ for all $t \leq T$. It thus follows that

$$\sum_{t=1}^T 2\eta_t(\mathcal{E}_{\mathbf{z}}(f_{T+1}) - \mathcal{E}_{\mathbf{z}}(f_*)) \leq \sum_{t=1}^T 2\eta_t(\mathcal{E}_{\mathbf{z}}(f_{t+1}) - \mathcal{E}_{\mathbf{z}}(f_*)) \leq \|f_*\|_K^2,$$

which leads to the first argument of the lemma. The rest of the proof can be finished by noting that

$$\sum_{t=1}^T \eta_t \geq \eta_1 \int_1^{T+1} u^{-\theta} du \geq \eta_1 \frac{T^{1-\theta}}{\mathrm{e}}.$$

■

Appendix B. Proofs for the Sample Error

Proof of Lemma 22 We apply Bernstein inequality which asserts that, for a random variable ξ bounded by $\widetilde{M} > 0$ and for any $\epsilon > 0$,

$$\text{Prob}\left\{\frac{1}{m}\sum_{i=1}^{m}\xi(z_i) - \mathbb{E}(\xi) > \epsilon\right\} \le \exp\left\{-\frac{m\epsilon^2}{2\big(\sigma^2(\xi) + \frac{1}{3}\widetilde{M}\epsilon\big)}\right\}. \tag{47}$$

Here the random variable ξ on Z is given by $\xi(x,y) = V(y, f_*(x)) - V(y, f_\rho^V(x))$. The increment condition (4) implies that ξ is bounded by $M := C_1' R^{q+1}$, where C_1' is the constant given by

$$C_1' := c_q \left(\kappa + \kappa^{q+1} + \|f_\rho^V\|_\infty + \|f_\rho^V\|_\infty^{q+1}\right). \tag{48}$$

By condition (5), its variance $\sigma^2(\xi)$ is bounded by

$$c_\tau R^{2+q-\tau}\left\{\mathcal{E}(f_*) - \mathcal{E}(f_\rho^V)\right\}^\tau \le c_\tau R^{2+q-\tau}(\mathcal{A}(f_*))^\tau.$$

Solving the quadratic equation from Bernstein inequality (47), we see that with confidence at least $1 - \frac{\delta}{2}$, there holds

$$\begin{aligned}\mathcal{F}_{\mathbf{z}}(f_*) &\le \frac{2M\log\frac{2}{\delta}}{3m} + \sqrt{\frac{2\log\frac{2}{\delta}}{m}\sigma^2(\xi)} \\ &\le (C_1' + 2\sqrt{c_\tau})\log\frac{2}{\delta}\max\left\{\frac{R^{q+1}}{m},\ \frac{R^{1+\frac{q-\tau}{2}}(\mathcal{A}(f_*))^{\frac{\tau}{2}}}{\sqrt{m}}\right\}.\end{aligned}$$

Applying the elementary inequality

$$x^\tau y^{1-\tau} \le \tau x + (1-\tau)y, \qquad \forall\ \tau \in [0,1],\ x, y \ge 0, \tag{49}$$

yields

$$\begin{aligned}\frac{R^{1+\frac{q-\tau}{2}}(\mathcal{A}(f_*))^{\frac{\tau}{2}}}{\sqrt{m}} &= \left[\left(\frac{R^{2+q-\tau}}{m}\right)^{\frac{1}{2-\tau}}\right]^{1-\frac{\tau}{2}}(\mathcal{A}(f_*))^{\frac{\tau}{2}} \\ &\le \left(1-\frac{\tau}{2}\right)\left(\frac{R^{2+q-\tau}}{m}\right)^{\frac{1}{2-\tau}} + \frac{\tau}{2}\mathcal{A}(f_*).\end{aligned}$$

Then the desired result follows. ■

To bound the empirical process over the ball $B_{\widetilde{R}}$ for some $\widetilde{R} > 0$, we need the following concentration inequality. Its proof is similar to that of Proposition 6 in Wu et al. (2007), as well as applying Theorem 3.5 from Steinwart (2009) and Exercise 6.8 in Steinwart and Christmann (2008). We omit the proof.

Lemma 25 *Let $\mathcal{G}$ be a set of measurable functions on $\mathcal{Z}$, and $B, c > 0, \tau \in [0,1]$ be constants such that each function $f \in \mathcal{G}$ satisfies $\|f\|_\infty \le B$ and $\mathbb{E}(f^2) \le c(\mathbb{E}f)^\tau$. If for some $a \ge B^\zeta$ and $\zeta \in (0,2)$,*

$$\mathbb{E}_{\mathbf{z}}[\log\mathcal{N}(\mathcal{G}, \epsilon, d_{2,\mathbf{z}})] \le a\epsilon^{-\zeta}, \qquad \forall \epsilon > 0,$$

then there exists a positive constant c'_ζ depending only on ζ such that for any $b > 0$, with probability at least $1 - \mathrm{e}^{-b}$, there holds

$$\mathbb{E}f - \frac{1}{m}\sum_{i=1}^m f(z_i) \le \frac{1}{2}\eta^{1-\tau}(\mathbb{E}f)^\tau + c'_\zeta \eta + 2\Big(\frac{cb}{m}\Big)^{1/(2-\tau)} + \frac{18Bb}{m}, \qquad \forall f \in \mathcal{G},$$

where

$$\eta := \max\left\{ c^{\frac{2-\zeta}{4-2\tau+\zeta\tau}} \left(\frac{a}{m}\right)^{\frac{2}{4-2\tau+\zeta\tau}},\ B^{\frac{2-\zeta}{2+\zeta}} \left(\frac{a}{m}\right)^{\frac{2}{2+\zeta}} \right\}.$$

Now, we are ready to prove Lemma 23.

Proof of Lemma 23 We first apply Lemma 25 to the function set

$$\mathcal{G} = \left\{ f(x,y) = V(y, g(x)) - V(y, f_\rho^V(x)) :\ g \in B_{\widetilde{R}} \right\}.$$

Condition (5) tells us that with $c = c_\tau \widetilde{R}^{2+q-\tau}$, each function $f \in \mathcal{G}$ satisfies $\mathbb{E}(f^2) \le c(\mathbb{E}f)^\tau$. Also, condition (4) implies that $\|f\|_\infty$ is upper bounded by $\widetilde{M} := C'_1 \widetilde{R}^{q+1}$, with C'_1 given by (48). Noticing from (4) that for $f, f' \in \mathcal{G}$,

$$|f(x,y) - f'(x,y)| = |V(y, g(x)) - V(y, g'(x))| \le c_q(1+\kappa^q)\widetilde{R}^q |g(x) - g'(x)|,$$

there holds

$$\mathcal{N}(\mathcal{G}, \epsilon, d_{2,\mathbf{z}}) \le \mathcal{N}\left(B_{\widetilde{R}}, \frac{\epsilon}{c_q(1+\kappa^q)\widetilde{R}^q}, d_{2,\mathbf{z}} \right) \le \mathcal{N}\left(B_1, \frac{\epsilon}{c_q(1+\kappa^q)\widetilde{R}^{q+1}}, d_{2,\mathbf{z}} \right).$$

Hence, condition (9) yields the covering number condition in Lemma 25 with $a = C''\widetilde{R}^{(q+1)\zeta}$, where

$$C'' = \max\{c_\zeta c_q^\zeta (1+\kappa^q)^\zeta, (C'_1)^\zeta\}$$

So we apply Lemma 25 and find that with confidence at least $1 - \frac{\delta}{2}$, there holds for every $f \in \mathcal{G}$,

$$\mathbb{E}(f) - \frac{1}{m}\sum_{i=1}^m f(z_i) \le \frac{1}{2}\eta^{1-\tau}(\mathbb{E}f)^\tau + c'_\zeta \eta + 2\Big(\frac{c_\tau \widetilde{R}^{2+q-\tau}\log\frac{2}{\delta}}{m}\Big)^{\frac{1}{2-\tau}} + \frac{18\widetilde{M}\log\frac{2}{\delta}}{m},$$

where

$$\begin{aligned}
\eta &= \max\left\{ \left(c_\tau \widetilde{R}^{2+q-\tau}\right)^{\frac{2-\zeta}{4-2\tau+\zeta\tau}} \left(\frac{C''\widetilde{R}^{(q+1)\zeta}}{m}\right)^{\frac{2}{4-2\tau+\zeta\tau}},\right. \\
&\qquad \left. \widetilde{M}^{\frac{2-\zeta}{2+\zeta}} \left(\frac{C''\widetilde{R}^{(q+1)\zeta}}{m}\right)^{\frac{2}{2+\zeta}} \right\} \\
&\le C'_2 \max\left\{ \left(\frac{\widetilde{R}^{\frac{q(2+\zeta)+(4-2\tau+\zeta\tau)}{2}}}{m}\right)^{\frac{2}{4-2\tau+\zeta\tau}},\ \frac{\widetilde{R}^{q+1}}{m^{\frac{2}{2+\zeta}}} \right\},
\end{aligned}$$

where C_2' is the constant given by

$$\left(c_\tau^{\frac{2-\zeta}{2}} C''\right)^{\frac{2}{4-2\tau+\zeta\tau}} + C_1'^{\frac{2-\zeta}{2+\zeta}} \left(C''\right)^{\frac{2}{2+\zeta}}.$$

Apply the elementary inequality (49) which yields $\eta^{1-\tau}(\mathbb{E}f)^\tau \leq \eta + \mathbb{E}f$, and notice that $\mathbb{E}(f) = \mathcal{E}(g) - \mathcal{E}(f_\rho^V)$ while $\frac{1}{m}\sum_{i=1}^m f(z_i) = \mathcal{E}_{\mathbf{z}}(g) - \mathcal{E}_{\mathbf{z}}(f_\rho^V)$. We get that with confidence at least $1 - \frac{\delta}{2}$, there holds for every $g \in B_{\widetilde{R}}$,

$$\begin{aligned}
&\left(\mathcal{E}(g) - \mathcal{E}(f_\rho^V)\right) - \left(\mathcal{E}_{\mathbf{z}}(g) - \mathcal{E}_{\mathbf{z}}(f_\rho^V)\right) \\
&\leq \left(\frac{1}{2} + c_\zeta'\right)\eta + \frac{1}{2}\left(\mathcal{E}(g) - \mathcal{E}(f_\rho^V)\right) + 2\Big(\frac{c_\tau \widetilde{R}^{2+q-\tau}}{m}\Big)^{\frac{1}{2-\tau}} \log\frac{2}{\delta} + \frac{18\widetilde{M}\log\frac{2}{\delta}}{m},
\end{aligned}$$

which leads to (36) with

$$C_3' = \left(\frac{1}{2} + c_\zeta'\right) C_2' + 2c_\tau^{\frac{1}{2-\tau}} + 18\widetilde{M}.$$

Now, introducing (36) into the equality

$$\mathcal{E}_{\mathbf{z}}(f_\rho^V) - \mathcal{E}_{\mathbf{z}}(g) = \left\{\left(\mathcal{E}(g) - \mathcal{E}(f_\rho^V)\right) - \left(\mathcal{E}_{\mathbf{z}}(g) - \mathcal{E}_{\mathbf{z}}(f_\rho^V)\right)\right\} - \left(\mathcal{E}(g) - \mathcal{E}(f_\rho^V)\right),$$

with $\mathcal{E}(g) - \mathcal{E}(f_\rho^V) \geq 0$ and by recalling that $\mathcal{M}_{\mathbf{z}}(\widetilde{R})$ is given by (31), we can derive (37). The proof is completed.

■