Journal of Machine Learning Research 23 (2022) 1-36 Submitted 8/21; Revised 7/22; Published 8/22

Early Stopping for Iterative Regularization with General Loss Functions

Ting Hu TINGHU@XJTU.EDU.CN
Center for Intelligent Decision-Making and Machine Learning
School of Management
Xi'an Jiaotong University
Xi'an, China

Yunwen Lei YUNWEN.LEI@HOTMAIL.COM
Department of Mathematics
Hong Kong Baptist University
Kowloon, Hong Kong, China

Editor: Lorenzo Rosasco

Abstract

In this paper, we investigate the early stopping strategy for the iterative regularization technique, which is based on gradient descent of convex loss functions in reproducing kernel Hilbert spaces without an explicit regularization term. This work shows that projecting the last iterate of the stopping time produces an estimator that can improve the generalization ability. Using the upper bound of the generalization errors, we establish a close link between the iterative regularization and Tikhonov regularization scheme and explain theoretically why the two schemes have similar regularization paths in the existing numerical simulations. We introduce a data-dependent way based on cross-validation to select the stopping time. We prove that the a-posteriori selection way can retain the comparable generalization errors to those obtained by our stopping rules with a-prior parameters.

Keywords: iterative regularization, early stopping, reproducing kernel Hilbert spaces, stopping rule, cross-validation

1. Introduction

Early stopping is a well known regularization method to overcome the phenomenon of overfitting in the fields of science and engineering. In inverse problems, the original idea of early stopping can date back to the 1970's in the context of the Landweber iteration (Strand, 1974). Later on it received considerable study based on spectral filtering for solving linear or nonlinear operator equations and a series of subsequent papers can be found (Vito et al., 2005; Guo et al., 2017; Lu and Pereverzev, 2013). Various forms of early stopping also have been developed in the machine learning community as well as in the statistics community, such as boosting algorithms (Bühlmann and Yu, 2003; Wei et al., 2019), multi-pass stochastic gradient descent (Lin and Rosasco, 2017), conjugate gradient descent (Blanchard and Mathé, 2010; Blanchard and Krämer, 2016) and gradient descent (Yao et al., 2007; Lin et al., 2016). Apart from solving convex problems, early stopping can be applied to non-convex optimization problems, including robust learning (Guo et al., 2018), minimum error entropy criterion (Fan et al., 2016) and training neural networks

(LeCun et al., 2012). In these algorithms, they sought to minimize empirically a loss function in an iterative fashion and the number of iterations serves as a regularization parameter. For ensuring best generalization abilities, the iterative regularization against overfitting is realized by a suitable stopping strategy that provides the best possible convergence rate in the iteration process. Statistical results on generalization properties and the regularization effect of early stopping have been investigated in a variety of learning algorithms, especially when an iterative update is generated by the least squares loss (Yao et al., 2007; Raskutti et al., 2014; Bühlmann and Yu, 2003).

In this paper, we shall analyze stopping strategies applied to subgradient or gradient descent of general convex loss functions associated with a reproducing kernel Hilbert space (RKHS) in the supervised setting, where no constraint or extra penalty term is taken into consideration. Similar framework has been investigated by the paper (Lin et al., 2016), in which consistency and non-asymptotic bounds quantifying the generalization properties were achieved under some specific stopping rules. However, these rates are not satisfactory and worse than the error bounds achieved by other regularization schemes, such as support vector machines, kernel ridge regression. In the recent papers (Wei et al., 2019; Stankewitz et al., 2021), the early stopping strategies and convergence properties were studied when iterative regularization schemes are generated by a class of (locally) smooth losses. Thus, these works still left an open question whether the kind of iterative algorithms associated with general convex losses can produce the estimators that have comparative generalization errors when a suitable stopping strategy is taken. The first contribution of this paper is to answer this question in the affirmative. This work improves the error analysis for the iterative algorithm by a sharp estimate for the bound of the iterates before the stopping time and shows that with a refined stopping rule the error bound of the projected last iterate can match the best ones available in previous papers (to be discussed in Section 2). Secondly, we investigate the link between our early stopping procedure and Tikhonov regularization. Friedman and Popescu (2004) empirically reported that the regularization paths for early stopping of gradient descent and L_2-penalized least squares regression are similar, but did not provide any theoretical explanation. Our strict error analysis will illustrate this phenomenon and give a comparison between early stopping and Tikhonov regularization over RKHS. More precisely, we prove that if the penalty parameter of Tikhonov regularization is selected appropriately according to our stopping rule, then the generalization error satisfies the same type of bounds. Finally, it is shown in the next section, our stopping rules depend on some a-prior parameters that may be unknown in advance. Hence, we introduce a stopping rule for determining the number of iterations based on cross-validation and show that the same order of generalization errors can be achieved.

The rest of the paper is organized as follows. In Section 2, we introduce the necessary backgrounds and provide three main results mentioned above. Section 3 is devoted to deriving the explicit stopping time and the corresponding generalization error bounds depend on some a-prior parameters. Section 4 contains the proofs of error bounds based on cross-validation with the a-posteriori parameter selection. We close with some discussions and conclusions in Section 5.

2. Learning Algorithms and Main Results

We begin with some basic notations and assumptions required for a precise statement of our results. Let $\mathcal{X}$ be a separable metric space, $\mathcal{Y} \subset \mathbb{R}$ and ρ be a Borel probability measure on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Given a predictor $f : \mathcal{X} \to \mathcal{Y}$, a convex loss function $\phi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ is used to measure its local error by the value $\phi(y, f(x))$ for $(x, y) \in \mathcal{Z}$. The learning task is to minimize the *generalization error* $\mathcal{E}(f)$ associated with the pair (ϕ, ρ), given as $\mathcal{E}(f) = \int_{\mathcal{Z}} \phi(y, f(x)) d\rho$. Denote by $f_\rho^\phi : \mathcal{X} \to \mathcal{Y}$ the minimizer of $\mathcal{E}(f)$ over all measurable functions. The main goal of learning is to estimate f_ρ^ϕ according to the sample set $\mathbf{D} = \{(x_i, y_i)\}_{i=1}^m \subset \mathcal{Z}$ drawn from the unknown ρ.

Kernel methods provide efficient non-parametric learning algorithms for dealing with nonlinear features and RKHSs are used in this work as hypothesis spaces in the design of iterative algorithms. Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a Mercer kernel, that is, a continuous, symmetric and positive semi-definite function. The RKHS $\mathcal{H}_K$ associated with K is defined to be the completion of the linear span of the set of functions $\{K_x := K(x, \cdot), x \in \mathcal{X}\}$ equipped with the inner product $\langle \cdot, \cdot \rangle_K$ satisfying *the reproducing property*

$$\langle g, K_x \rangle_K = g(x), \ \textit{for any } x \in \mathcal{X}, g \in \mathcal{H}_K. \tag{2.1}$$

Denote $\kappa := \sup_{x \in \mathcal{X}} \sqrt{k(x, x)}$. This property implies that $\|g\|_\infty \leq \kappa \|g\|_K, \ \forall g \in \mathcal{H}_K$.

Throughout the paper, we assume that the loss ϕ is convex with respect to the second variable. That is, for any fixed $y \in \mathcal{Y}$, the univariate function $\phi(y, \cdot)$ on $\mathbb{R}$ is convex. Hence its left derivative $\phi'_-(y, f)$ exists at every $f \in \mathbb{R}$ and is non-decreasing. In what follows, we consider the iterative algorithm based on the subgradient method, or the gradient descent if the loss is smooth.

Definition 1 *Given an i.i.d. sample set $\mathbf{D} = \{(x_i, y_i)\}_{i=1}^m \subset \mathcal{Z}$ and a stopping time $T > 2$, the iterative algorithm is given by $f_1 = 0$ and*

$$f_{t+1} = f_t - \frac{\eta_t}{m} \sum_{j=1}^m \phi'_- \left(y_j, f_t(x_j)\right) K_{x_j}, \quad t = 1, \cdots, T, \tag{2.2}$$

where $\{\eta_t > 0, t = 1, \cdots, T\}$ is a step size sequence.

The primal purpose of this paper is to investigate the generalization ability of algorithm (2.2), which is usually measured by the *excess generalization error* $\mathcal{E}(f) - \mathcal{E}(f_\rho^\phi)$. To this end, we need to introduce some necessary assumptions.

Assumption 1 *We assume that $c_\phi := \sup_{y \in \mathcal{Y}} \phi(y, 0) < \infty$ and an increment condition holds for the left derivative ϕ'_-, that is, for some $q > 0$ and $c_q > 0$,*

$$|\phi'_-(y, f)| \leq c_q (1 + |f|)^q, \quad \forall f \in \mathbb{R}, \ y \in \mathcal{Y}. \tag{2.3}$$

Condition (2.3) is satisfied by a broad class of loss functions. For smooth losses, it holds with $q = 1$. For Lipschitz continuous losses, it holds with $q = 0$. For α-activating losses, it holds with $q = \alpha, \ 0 < \alpha < 1$. Concrete examples include the hinge loss $\phi(y, f) = \max\{1 - yf, 0\}$, the logistic loss $\phi(y, f) = \log(1 + \exp(-yf))$ for classification, the least

squares loss $\phi(y,f)=(y-f)^2$, the ϵ-insensitive loss $\phi(y,f)=\max\{|y-f|-\epsilon,0\}$ for regression. We also notice that Nemitski losses satisfy (2.3) when the output is bounded, which are introduced in Steinwart and Christmann (2008); Vito et al. (2004) and consist of most commonly used convex loss functions in various learning problems.

Assumption 2 *We assume that for any $B\geq 1$, there exists an exponent $\tau\in[0,1]$ and the positive constant $c_\tau=c_\tau(B)$ satisfying*

$$\mathbb{E}\left\{\phi(y,f(x))-\phi(y,f_\rho^\phi(x))\right\}^2\leq c_\tau\left\{\mathcal{E}(f)-\mathcal{E}(f_\rho^\phi)\right\}^\tau,\quad \forall f:\mathcal{X}\to[-B,B]. \tag{2.4}$$

The above inequality is usually referred to a *variance-expectation* bound and plays an important role to improve our error analysis in this work. Inequality (2.4) always holds with $\tau=0$ and c_τ depending on $\|f_\rho^\phi\|_\infty$ due to the continuity of convex losses. For the least squares loss, we can take $\tau=1$. See the works (Cucker and Zhou, 2007; Steinwart and Christmann, 2008).

Remark 1 *The value of the exponent τ has close relation with some noise conditions imposed on the distribution ρ or the convexity of the loss ϕ. The hinge loss is a prominent example for the noise condition. We can see that $\tau=0$ always holds for the hinge loss and an improved $\tau=\frac{s}{s+1}$ holds when the Tsybakov margin noise condition (to be stated in (2.20)) is valid with exponent s (Wu et al., 2007). Another example is the pinball loss, as shown in Section 9 of Steinwart and Christmann (2008), we can take $\tau=\frac{s'}{2s'+2}$ if the conditional distribution of ρ satisfies the quantile noise condition with exponent $s'>0$.*

For the convexity of ϕ, we assume the loss $\phi(y,f)=V(yf)$ with some convex function V. **The modulus of convexity** *of V is defined as (Bartlett et al., 2006)*

$$\hat{\delta}_{\mathcal{S}}(\varepsilon)=\inf\left\{\frac{V(a)+V(b)}{2}-V\left(\frac{a+b}{2}\right):a,b\in\mathcal{S},|a-b|\geq\varepsilon\right\}.$$

If $\hat{\delta}_{\mathcal{S}}(\varepsilon)>0$ for all $\varepsilon>0$, then V is strictly convex in $\mathcal{S}$. Assume that

$$\hat{\delta}_{\mathcal{S}}(\varepsilon)\geq c|\varepsilon|^\vartheta,\ \textit{for some } c>0 \textit{ and } \vartheta>0. \tag{2.5}$$

$\vartheta=2$ if V is strongly convex, e.g., the quadratic loss. Then by the work (Bartlett et al., 2006), we know that (2.4) holds with $\tau=\min\{1,\frac{2}{\vartheta}\}$ if (2.5) is valid.

Assumption 3 *Let $\lambda>0$ and the regularization function f_λ be the minimizer of the regularization error:*

$$f_\lambda:=\arg\min_{f\in\mathcal{H}_K}\left\{\mathcal{E}(f)+\lambda\|f\|_K^2\right\}. \tag{2.6}$$

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

$$\mathcal{D}(\lambda):=\min_{f\in\mathcal{H}_K}\left\{\mathcal{E}(f)-\mathcal{E}(f_\rho^\phi)+\lambda\|f\|_K^2\right\}=\mathcal{E}(f_\lambda)-\mathcal{E}(f_\rho^\phi)+\lambda\|f_\lambda\|_K^2. \tag{2.7}$$

We assume that for some $\beta\in(0,1]$ and $c_\beta>0$, there holds

$$\mathcal{D}(\lambda)\leq c_\beta\lambda^\beta,\quad \lambda>0. \tag{2.8}$$

The approximation error measures the approximation ability of the space $\mathcal{H}_K$ with respect to the learning process involving ϕ and ρ. Assumption (2.8) is standard in learning theory and always holds with $\beta = 0$. The denseness of $\mathcal{H}_K$ in $C(\mathcal{X})$ implies $\lim_{\lambda\to 0} \mathcal{D}(\lambda) = 0$. Thus, the decay of (2.8) can be estimated by $\mathcal{K}$-functionals from the knowledge of approximation theory. For more details, see discussions in Cucker and Zhou (2007); Wang and Hu (2021). Throughout the paper, denote the closed ball of radius $R > 0$ in $\mathcal{H}_K$ as $\mathcal{B}_R := \{f \in \mathcal{H}_K, \|f\|_K \le R\}$.

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

$$d_{2,\mathbf{D}}(f,g) := \left\{ \frac{1}{m} \sum_{i=1}^{m} (f(x_i) - g(x_i))^2 \right\}^{\frac{1}{2}}, \quad \forall f, g \in \mathcal{G}.$$

Recall that for a subset $\mathcal{G}$ of a metric space $(\mathcal{H}, d)$, the covering number $\mathcal{N}(\mathcal{G}, \epsilon, d)$ is defined by

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

We assume that for some $\zeta \in (0, 2)$, $c_\zeta > 0$, the covering number of the unit ball $\mathcal{B}_1$ in $\mathcal{H}_K$ with respect to $d_{2,\mathbf{D}}$ satisfies

$$\mathbb{E}_{\mathbf{D}} \left[\log \mathcal{N}(\mathcal{B}_1, \epsilon, d_{2,\mathbf{D}}) \right] \le c_\zeta \left(\frac{1}{\epsilon} \right)^{\zeta}, \quad \forall \epsilon > 0. \tag{2.9}$$

Empirical covering number $\mathcal{N}(\mathcal{G}, \epsilon, d_{2,\mathbf{D}})$ is a widely adopted tool to characterize the capacity of $\mathcal{H}_K$ in learning theory. Note that for any $\mathcal{G} \in C(\mathcal{X})$, $\mathcal{N}(\mathcal{G}, \epsilon, d_{2,\mathbf{D}})$ is bounded by the uniform covering number $\mathcal{N}(\mathcal{G}, \epsilon, d)$ under the metric $d = \|\cdot\|_\infty$ since $d_{2,\mathbf{D}}(f, g) \le \|f - g\|_\infty$. For example, when $\mathcal{X}$ is a bounded subset of $\mathbb{R}^n$ and the RKHS $\mathcal{H}_K$ is a Sobolev space $H^\alpha(X)$ with index α, it can be checked by Zhou (2002); Cucker and Zhou (2007) that the condition (2.9) holds true with $\zeta = 2n/\alpha$. If the kernel K lies in the smooth space $C^\infty(\mathcal{X} \times \mathcal{X})$, then (2.9) is satisfied for an arbitrarily small $\zeta > 0$. Recall that capacity of the RKHS may be measured by other concepts: (dyadic) entropy numbers, effective dimensions, decay of the eigenvalues of the integral operator associated with K. For their connections and estimations in various function spaces, one can refer to the works (Lin et al., 2016; Steinwart and Christmann, 2008; Lv et al., 2018).

Throughout the paper, we assume that for some constant $B > 0$ the output space $\mathcal{Y} \subset [-B, B]$. Thus, we shall choose the estimator of algorithm (2.2) by restricting its final output onto $[-B, B]$. We make full use of the projection operator, given as

$$\hat{f}(x) = \begin{cases} -B, & \text{if} \quad f(x) < -B, \\ f(x), & \text{if} \quad |f(x)| \le B, \\ B, & \text{if} \quad f(x) > B. \end{cases}$$

The idea of using a projected estimator to improve the learning performance was first introduced by the work (Bartlett, 1998) in training neural networks, and then developed

to the context of SVM (Bousquet and Elisseeff, 2002). In the subsequent works (Wu et al., 2007; Steinwart and Christmann, 2008), they proved that the projection technique can lead to sharp generalization errors for Tikhonov regularization. Inspired by their works, we extend it to the analysis of iterative regularization (2.2) and demonstrate its superiority by the following theoretical verification. We also assume that the loss function ϕ is B-*admissible*:

$$\phi(y, \hat{f}(x)) \leq \phi(y, f(x)), \ \textit{for any } f : \mathcal{X} \to \mathbb{R} \textit{ and } y \in [-B, B]. \tag{2.10}$$

For binary classification problems, we can take $\phi(y, f(x)) = V(yf(x))$ with a convex function $V : \mathbb{R} \to \mathbb{R}_+$. When V belongs to the class of *classifying loss functions* (that is, $V'(0) < 0$ and the smallest zero of V is 1), it is easy to check that $V(y\hat{f}(x)) \leq V(yf(x))$ with $B = 1$. Examples of classifying losses include

- the *hinge loss* $V(yf) = \max\{1 - yf, 0\}$,
- *the least squares loss* $V(yf) = (1 - yf)^2$,
- *the p-norm soft margin SVM loss* $V(yf) = \max\{1 - yf, 0\}^p$ $(1 < p < \infty)$.

For regression problems, many popular loss functions are also B-admissible, including

- *the least squares regression loss* $\phi(y, f) = (y - f)^2$,
- *the* ϵ*-insensitive loss* $\phi(y, f) = \max\{|y - f| - \epsilon, 0\}$,
- *the pinball loss*

$$\phi(y, f) = \begin{cases} (1 - \tau)(y - f), & \textit{if } y \geq f, \\ \tau(f - y), & \textit{otherwise}, \end{cases}$$

- *the logistic loss for regression* $\phi(y, f) = -\log\left(\frac{4e^{y-f}}{(1+e^{y-f})^2}\right)$,
- *the Huber loss*

$$\phi(y, f) = \begin{cases} (y - f)^2, & \textit{if } |y - f| \leq 1, \\ 2|y - f| - 1, & \textit{otherwise}. \end{cases}$$

2.1 Main results for generalization error

Our first main theorem establishes a stopping rule for (2.2) in the situation of the smooth loss functions and provides the corresponding generalization error bounds for the projected last iterate. We say that the loss ϕ is *smooth* if the derivative $\phi'(y, \cdot)$ exists for any $y \in \mathcal{Y}$ and is L-Lipschitz continuous for some constant $L > 0$,

$$|\phi'(y, a) - \phi'(y, b)| \leq L|a - b|, \ \forall a, b \in \mathbb{R}.$$

With the above smoothness assumption, we obtain the following generalization result whose proof will be given in Subsection 3.6.

Denote $\lfloor a \rfloor$ as the largest integer not exceeding $a, a \in \mathbb{R}$. The notation $\lesssim$ means that the inequality holds up to a multiplicative constant that depends on various parameters appearing in the assumptions, but not on the sample size m or the confidence level δ.

Theorem 1 (Smooth losses) *Suppose Assumptions 1, 2, 3 and 4 hold. Let* $\eta_t = \eta t^{-\theta}$ *with* $0 \le \theta < 1$ *and some* $0 < \eta < \min\left\{\frac{1-\theta}{2c_\phi}, \frac{1}{L\kappa^2}\right\}$. *Define the parameters* α *and* γ *as*

$$\alpha = \min\left\{\frac{2}{(1-\theta)(\zeta(1-\beta)+\beta(4-2\tau+\zeta\tau))}, \frac{2}{(1-\theta)(2\beta+(1-\beta)(q+1))}\right\}, \tag{2.11}$$

and

$$\gamma = \min\left\{\frac{2}{\zeta(1-\beta)+\beta(4-2\tau+\zeta\tau)}, \frac{2}{2\beta+(1-\beta)(q+1)}\right\}. \tag{2.12}$$

If $T = \lfloor m^\alpha \rfloor$, *then with confidence at least* $1-\delta$,

$$\mathcal{E}(\hat{f}_T) - \mathcal{E}(f_\rho^\phi) \lesssim m^{-\beta\gamma} \log\frac{\log m}{\delta}. \tag{2.13}$$

Under the above stopping rule, the upper bound (2.13) greatly improves the rate $O(m^{-\beta\gamma'})$, $\gamma' = \left(\beta(2-\tau+\zeta\tau/2) + \left\{\frac{2-\tau+\zeta\tau/2}{2} + \frac{q(1+\zeta/2)}{2}\right\}\right)^{-1}$ in Theorem 8, (Lin et al., 2016). It is clear that algorithm (2.2) with different step sizes can yield the same generalization bounds. Thus, a constant step size is chosen for smooth losses, which can considerably reduce the iteration time.

Remark 2 *In the recent work by Stankewitz et al. (2021), stopping strategies and the corresponding generalization errors were established for (locally) Lipschitz and smooth loss functions without capacity conditions on hypothesis spaces. They showed that when the target function* f_ρ^ϕ *lies in a separable Hilbert space, the error for the averaged iterate is of order* $O\big(m^{-\frac{1}{2}}\big)$ *if the stopping time* $T = O\big(m^{\frac{1}{2}}\big)$. *By Theorem 1, we take a universal* $\zeta = 2$ *for no capacity conditions on RKHSs and let* $\beta = 1$ *for* $f_\rho^\phi \in \mathcal{H}_K$, *then our best rate is* $O\big(m^{-\frac{1}{2}}\log(\log m)\big)$ *if* $T = O\big(m^{\frac{1}{2}}\big)$, *which is only slightly inferior. However, our analysis does not require that the minimizer of* $\min_{f\in\mathcal{H}_K} \mathcal{E}(f)$ *exists, as shown in the proof, which is more general.*

Let us now derive some explicit consequences of our theorem by considering the least squares loss for special choices of kernels that are of interest in practice. Denote by $\rho(y|x)$ the conditional probability for all $(x,y) \in \mathcal{Z}$ and by $\rho_\mathcal{X}$ the marginal distribution on $\mathcal{X}$. The target function f_ρ^ϕ is the regression function f_ρ, defined as the conditional mean $\int_{y\in\mathcal{Y}} y d\rho(y|x)$ for given $x \in \mathcal{X}$ and $\|f_\rho\|_\infty \le B$ since $|y| \le B$. The generalization error is

$$\mathcal{E}(f) - \mathcal{E}(f_\rho) = \|f - f_\rho\|^2_{L^2_{\rho_\mathcal{X}}}$$

where

$$L^2_{\rho_\mathcal{X}} = \left\{f : \mathcal{X} \to \mathbb{R} : \int_\mathcal{X} |f(x)|^2 d\rho_\mathcal{X} < \infty\right\}.$$

In this case, Assumption 2 holds true with $\tau = 1$ and $c_\tau = (4B)^2$ since

$$\int_\mathcal{Z} \left[(y-f(x))^2 - (y-f_\rho(x))^2\right]^2 d\rho = \int_\mathcal{Z} (2y - f(x) - f_\rho(x))^2 (f(x) - f_\rho(x))^2 d\rho$$
$$\le (4B)^2 \|f - f_\rho\|^2_{L^2_{\rho_\mathcal{X}}} = (4B)^2 (\mathcal{E}(f) - \mathcal{E}(f_\rho))$$

for any $f : \mathcal{X} \to [-B, B]$.

Define the integral operator associated with K as

$$L_K(f) := \int_{\mathcal{X}} f(x) K_x d\rho_{\mathcal{X}}, \ \forall f \in L^2_{\rho_{\mathcal{X}}}.$$

When the operator L_K is compact, the r-power of L_K by L_K^r is well-defined for any $r > 0$. If the regression function f_ρ belongs to the range space $L_K^{\frac{\beta}{2}}\left(L^2_{\rho_{\mathcal{X}}}\right)$ with some $0 < \beta \leq 1$, it is usually referred to as a *regularity condition* imposed on f_ρ and (2.8) is valid for the least squares loss. We can refer to Cucker and Zhou (2007) for details.

Our next example applies to the class of RKHSs whose eigenvalues $\{\mu_i\}_i$ of L_K satisfy a polynomial decay condition, meaning that

$$\mu_i \leq C i^{-\frac{2}{\zeta}}, \ \textit{for some } 0 < \zeta < 2 \textit{ and constant } C > 0. \tag{2.14}$$

Kernels with the polynomial decaying eigenvalues include those that underlie for the Sobolev spaces with different orders of smoothness (Birman and Solomjak, 1967; Lian et al., 2019). As a concrete example, the first-order Sobolev kernel $K(x, x') = 1 + \min\{x, x'\}$ generates an RKHS of Lipschitz functions with smoothness $\zeta = 1$. Other higher-order Sobolev kernels also exhibit polynomial eigendecay with smaller values of the power ζ.

Example 1 *Let $\phi(y, f(x)) = (y - f(x))^2$ and $|y| \leq B$ for some $B > 0$. Assume the polynomial eigenvalue decay (2.14) holds and the regression function f_ρ lies in the space $L_K^{\frac{\beta}{2}}\left(L^2_{\rho_{\mathcal{X}}}\right)$ with $1 - \frac{\zeta}{2} \leq \beta \leq 1$. Let $\eta_t = \eta t^{-\theta}$ with $0 \leq \theta < 1$. If $T = \left\lfloor m^{\frac{2}{(1-\theta)(\zeta+2\beta)}} \right\rfloor$, then with confidence at least $1 - \delta$, there holds*

$$\|\hat{f}_T - f_\rho\|^2_{L^2_{\rho_{\mathcal{X}}}} = O\left(m^{-\frac{2\beta}{2\beta+\zeta}} \log(\log m)\right).$$

Furthermore, if f_ρ belongs to $\mathcal{H}_K$ ($\beta = 1$), with $T = \left\lfloor m^{\frac{2}{\zeta+2}} \right\rfloor$ and $\eta_t \equiv \eta$ (let $\theta = 0$),

$$\|\hat{f}_T - f_\rho\|^2_{L^2_{\rho_{\mathcal{X}}}} = O\left(m^{-\frac{2}{2+\zeta}} \log(\log m)\right). \tag{2.15}$$

For regression learning, the known minimax bounds on estimation error in Sobolev spaces are of order $O\left(m^{-\frac{2}{2+\zeta}}\right)$. See the papers (Stone, 1982; Tsybakov, 2009; Caponnetto and Vito, 2007). The upper bound (2.15) is nearly optimal only up to a logarithm term and obviously superior to the result of Corollary 10 in Lin et al. (2016).

Remark 3 *Iterative techniques with the least squares can often be viewed as a member of the family of spectral algorithms in solving inverse problems, with a special filter function (Lu and Pereverzev, 2013; Vito et al., 2005; Yao et al., 2007; Guo et al., 2018). In those learning paradigms, error analyses were achieved by utilizing the linear structure of algorithms and integral operators techniques of L_K, which can not be applied in our work. Their best obtained bounds are of order $O\left(m^{-\frac{2\beta}{2\beta+\zeta}}\right)$, which match minimax lower bounds in the regression setting (Caponnetto and Vito, 2007). Besides, the same error bounds also*

have been established in other iterative schemes, such as conjugate gradient (Blanchard and Krämer, 2016), incremental gradient (Rosasco and Villa, 2015), bias correction (Sun and Wu, 2021), multi-pass stochastic gradient method (Lin and Rosasco, 2017; Lei et al., 2021). The error bound in Example 1 has shown to be comparable just with the logarithm term.

Next we state our second main result in the situation of general losses, where the smoothness condition in Theorem 1 is removed.

Theorem 2 (General losses) *Suppose Assumptions 1, 2, 3 and 4 hold. Let* $\eta_t = \eta t^{-\theta}$ *with* $\max\left\{\frac{q}{q+1}, \frac{1}{2}\right\} < \theta < 1$ *and* η *satisfying*

$$0 < \eta < \min\left\{\frac{\sqrt{(q^*-1)(1-\theta)}}{2\sqrt{5}c_q(\kappa+1)^{q+1}\sqrt{q^*}}, \frac{1-\theta}{4c_\phi+1}\right\}. \tag{2.16}$$

Define α *and* γ *as in (2.11) and (2.12). If* $T = \lfloor m^\alpha \rfloor$ *, then with confidence at least* $1-\delta$*, we have for any* $\epsilon > 0$*,*

$$\mathcal{E}(\hat{f}_T) - \mathcal{E}(f_\rho^\phi) \lesssim \begin{cases} m^{-\beta\gamma}\log\frac{\log m}{\delta}, & \text{if} \quad \theta \geq \frac{q+1}{q+2}, \\ m^{-\beta\gamma+\gamma-\alpha(\theta(1+q)-q)}\,(\log m)\log\frac{\log m}{\delta}, & \text{if} \quad \theta < \frac{q+1}{q+2}, \end{cases} \tag{2.17}$$

where

$$q^* = 2\theta - (1-\theta)\cdot\max\{0, q-1\} > 0. \tag{2.18}$$

In the following, we will further illustrate the above result by considering the case $q = 0$, which includes all Lipschitz losses.

Corollary 1 *Suppose Assumption 1 holds true with* $q = 0$*. Let* $\eta_t = \eta t^{-\theta}$ *with* $\frac{1}{2} < \theta < 1$ *and Assumptions 2, 3 and 4 hold. If* $T = \left\lfloor m^{\min\left\{\frac{2}{(1-\theta)(\zeta(1-\beta)+\beta(4-2\tau+\zeta\tau))}, \frac{2}{(1-\theta)(\beta+1)}\right\}}\right\rfloor$*, then with confidence at least* $1-\delta$*, we have*

$$\mathcal{E}(\hat{f}_T) - \mathcal{E}(f_\rho^\phi) = O\left(m^{-\min\left\{\frac{2\beta}{\zeta(1-\beta)+\beta(4-2\tau+\zeta\tau)}, \frac{2\beta}{\beta+1}\right\}}\log(\log m)\right). \tag{2.19}$$

Remark 4 *For the non-smooth Lipschitz loss* ϕ*, Assumption 2 always holds with* $\tau = 0$*, then our result (2.19) reduces to* $O\Big(m^{-\min\left\{\frac{2\beta}{\zeta(1-\beta)+4\beta}, \frac{2\beta}{\beta+1}\right\}}\log(\log m)\Big)$*. Let us compare it with the best results obtained for Tikhonov regularization with Lipschitz loss functions. By assuming (2.8) and* $\tau = 0$*, Theorem 7.23 in Steinwart and Christmann (2008) provides the generalization error rate* $O\Big(m^{-\min\left\{\frac{2\beta}{\zeta(1-\beta)+4\beta}, \frac{2\beta}{\beta+1}\right\}}\Big)$*. Hence, the two obtained error rates for Lipschitz loss functions are nearly the same by ignoring the logarithm term.*

A direct application of the above corollary can be done for the SVM classification with the hinge loss. In this case, $\mathcal{Y} = \{-1, 1\}$ and the corresponding *misclassification error* is given as

$$\mathcal{R}(f) = \mathbf{Prob}\left\{(x, y) \in \mathcal{Z} : f(x) \neq y\right\}.$$

The minimizer of the misclassification error over measurable functions is the Bayes rule

$$f_c(x) = 2\text{sign}(2\rho(1|x) - 1),\ \forall x \in \mathcal{X}.$$

Comparison theorems enable us to bound the *excess misclassification error* $\mathcal{R}(f) - \mathcal{R}(f_c)$ by estimates for the excess generalization error $\mathcal{E}(f) - \mathcal{E}(f_\rho^\phi)$. For the hinge loss, it holds that

$$\mathcal{R}(f) - \mathcal{R}(f_c) \le \mathcal{E}(f) - \mathcal{E}(f_\rho^\phi).$$

If further information about the distribution ρ is available, one can expect sharper error bound for the above estimate. For example, suppose ρ satisfies a Tsybakov margin condition (Tsybakov, 2009)

$$\rho_{\mathcal{X}}\left\{x \in \mathcal{X} : 0 < \left|\rho(1|x) - \frac{1}{2}\right| \le C\Delta\right\} \le \Delta^s,\ \forall \Delta > 0, \tag{2.20}$$

with an exponent $s > 0$ and $C > 0$. In classification, $\rho(1|x) = \frac{1}{2}$ is a critical point to characterize the amount of noise. Consequently, exponent s quantifies the size of the set of points that have noise in the labeling process. All distributions satisfy (2.20) with $s = 0$ and $C > 0$, whereas $s = \infty$ implies that $\rho(1|x)$ is far away from $\frac{1}{2}$ and ρ has a low noise level.

With these preliminaries, misclassification error bound, for algorithm (2.2) with the hinge loss, can then be obtained by applying Corollary 1.

Example 2 *Let $\phi(y, f) = \max\{1 - yf, 0\}$ and $\mathcal{Y} = \{-1, 1\}$. Suppose (2.8) holds for $\beta \in (0, 1]$ and (2.9) is valid for $\zeta \in (0, 2)$. Let $\eta_t = \eta t^{-\theta}$ with $\frac{1}{2} < \theta < 1$. Assume that ρ satisfies a Tsybakov margin condition (2.20) with exponent $s > 0$. If $T = \left\lfloor m^{\min\left\{\frac{2(s+1)}{(1-\theta)(2\beta(s+2)+\zeta(s+1-\beta))}, \frac{2}{(1-\theta)(1+\beta)}\right\}}\right\rfloor$, then with confidence at least $1 - \delta$, we have*

$$\mathcal{R}(sign(f_T)) - \mathcal{R}(f_c) = O\left(m^{-\min\left\{\frac{2\beta(s+1)}{2\beta(s+2)+\zeta(s+1-\beta)}, \frac{2\beta}{\beta+1}\right\}} \log(\log m)\right). \tag{2.21}$$

We know from the paper (Wu and Zhou, 2005) that (2.4) is valid with the exponent $\tau = \frac{s}{s+1}$ and $c_\tau = 8\left(\frac{1}{2C}\right)^{\frac{s}{s+1}}$. This example shows how noise condition (2.20) on ρ improves the error bounds for (2.2) with the hinge loss from $\tau = 0$ ($s = 0$, general distributions) to $\tau = \frac{s}{s+1} > 0$ ($s > 0$).

2.2 Comparison with Tikhonov regularization scheme

In this part, we will establish a close connection between the iterative regularization algorithm (2.2) and Tikhonov regularization algorithm. Given a data set $\mathbf{D} = \{(x_i, y_i)\}_{i=1}^m \subset \mathcal{Z}$, the *Tikhonov regularization scheme* associated with the loss ϕ and kernel K is defined as

$$f_{\mathbf{D},\lambda} := \arg\min_{f \in \mathcal{H}_K} \left\{\mathcal{E}_{\mathbf{D}}(f) + \lambda\|f\|_K^2\right\}, \tag{2.22}$$

where $\lambda > 0$ is a regularization parameter and the *empirical error* is

$$\mathcal{E}_{\mathbf{D}}(f) := \frac{1}{|\mathbf{D}|}\sum_{(x,y)\in\mathbf{D}} \phi(y, f(x)) = \frac{1}{m}\sum_{i=1}^m \phi(y_i, f(x_i)).$$

The statistical aspects of Tikhonov regularization for learning have been extensively studied in literature. In particular, the error analysis is well done due to many results. See e.g. Wu and Zhou (2005); Cucker and Zhou (2007); Caponnetto and Vito (2007); Raskutti et al. (2014). As pointed out in Introduction, Tikhonov scheme has regularization behavior similar to that of iterative regularization (2.2) in numerical experiments. For the least squares loss, Raskutti et al. (2014) provided a theoretical basis to illustrate this phenomenon by connecting their early stopping strategy with the choice of regularization parameter in (2.22). They proved that for various kernel classes if the inverse regularization parameter λ of (2.22) is chosen according to the same stopping criterion for (2.2), then scheme (2.22) by the least squares can produce a minimax optimal estimator. Here we shall show in theory that the two regularization schemes associated with general losses also have similar regularization paths.

In what follows, we first provide the upper bounds for the generalization error of Tikhonov regularization associated with general convex losses.

Theorem 3 *Under Assumptions 1 to 4, let $\lambda = m^{-\gamma}$ with γ given as (2.12), then with confidence at least $1-\delta$, we have that for any $\epsilon > 0$,*

$$\mathcal{E}(\hat{f}_{\mathbf{D},\lambda}) - \mathcal{E}(f_\rho^\phi) \lesssim m^{-(\beta-\epsilon)\gamma} \log \frac{12 l_\epsilon}{\delta}, \tag{2.23}$$

where the integer l_ϵ is independent of m, δ.

It is worth mentioning that the upper bounds (2.23) of this order have been obtained for multi-kernel based classification problems (Wu et al., 2007). By tracing their proofs carefully, we can immediately derive the above results. Here we omit its proofs for simplicity. The focus of the theorem is to connect the regularization paths of Tikhonov scheme (2.22) with those of iterative algorithm (2.2).

Remark 5 *To our knowledge, the result (2.23) provides the best error bounds for Tikhonov regularization with general losses, which are available in existing literature (Steinwart and Christmann, 2008; Wu et al., 2007; Steinwart et al., 2009). It is not yet fully clear about the optimality and sharpness of the generalization error by general supervised learning problems, so we are not sure whether the bound (2.23) can be improved further under the conditions of Theorem 3. However, when we apply it to the case where ϕ is the least squares and the regression function f_ρ lies in a Sobolev space $\mathcal{H}_K$, it has been shown in Example 1 that the error rate (2.15) is nearly optimal in a minimax sense. In addition, when f_ρ is out of $\mathcal{H}_K$, belonging to the range space $L_K^{\beta/2}(L^2_{\rho_\mathcal{X}})$ with $1 - \frac{\zeta}{2} \le \beta < 1$, the resulting error rate of (2.23) is $O\left(m^{-\frac{2\beta}{2\beta+\zeta}+\epsilon}\right)$ ($\epsilon > 0$ is arbitrarily small) provided that the eigenvalue polynomial decay (2.14) holds for $\mathcal{H}_K$. We can check by Theorem 15 of Steinwart et al. (2009) that the i-th entropy number $e_i\left(id : L_K^{\beta/2}(L^2_{\rho_\mathcal{X}}) \to L^2_{\rho_\mathcal{X}}\right) \sim i^{-\frac{\beta}{\zeta}}$. Thus, it is known by Theorem 2.2 of Temlyakov (2006) that this rate is nearly optimal.*

We find that the choice of the regularization parameter is $\lambda \simeq T^{-(1-\theta)}$ where T is the stopping time used in Theorems 1, 2. Note that θ is determined by the decreasing rate of step sizes $\eta t^{-\theta}$ and does not affect the error rates. For general losses, Theorem 2 establishes

upper bounds (2.17) of the same order only with a logarithm term and a slightly different leading constant when $\theta \geq \frac{q+1}{q+2}$. For smooth losses, it is relaxed to $0 \leq \theta < 1$ in Theorem 1. In conjunction, Theorems 1, 2 and 3 provide a theoretical explanation for why, as shown in past works, the paths of the iterates (2.2) and the Tikhonov estimate (2.22) are so similar.

2.3 Data-dependent stopping time

It should be noted that our stopping strategies in Subsection 2.1 depend on a-prior knowledge of parameters, which is unknown in almost all situations. Thus, an adaptive choice of stopping time T is desirable to obtain good learning rates. Here we recommend a data-dependent way to select stopping time T, by taking a suitable a-posteriori choice based on cross-validation.

We are now in a position of describing this selection procedure. In the rest of the section, we assume that the data size $m = 2n$ that is even with some $n \in \mathbb{N}$ and the data set $\mathbf{D}$ is the disjoint union of two data subsets, $\mathbf{D_1}$ (the training set) and $\mathbf{D_2}$ (the validation set), of equal cardinality $|\mathbf{D_1}| = |\mathbf{D_2}| = n$.

Definition 2 *Let $\mathbf{T} = \{T_k\}_k$ be an integer sequence of finite subsets $T_k \in [n]$. Then, we perform algorithm (2.2) over $\mathbf{D_1}$ and get the T-th iterate $f_T^{(D_1)}, T \in \mathbf{T}$, i.e.,*

$$f_{t+1}^{(\mathbf{D_1})} = f_t^{(\mathbf{D_1})} - \frac{\eta_t}{n} \sum_{j=1}^{n} \phi'_- \left(y_j, f_t^{(\mathbf{D_1})}(x_j) \right) K_{x_j}, \quad t = 1, \cdots, T.$$

We use $\mathbf{D_2}$ to determine the stoping time T^* by

$$T^* := \arg\min_{T \in \mathbf{T}} \mathcal{E}_{\mathbf{D_2}}(\hat{f}_T^{(\mathbf{D_1})}), \tag{2.24}$$

and take $\hat{f}_{T^*}^{(\mathbf{D_1})}$ as the final estimate for (2.2). This method is called a training/validation iterative regularization with respect to $\mathbf{T}$.

Cross-validation methods have been developed in various learning algorithms when a-priori knowledge on the problem is usually not available. It has been proved in various learning paradigms (Caponnetto and Yao, 2010; Lin and Zhou, 2017; Steinwart et al., 2009) that the obtained error bounds for using a-priori choices of the parameters can be attained using a suitable a-posteriori choice based on cross-validation. The following theorem aims to reach a similar conclusion for the data-dependent choice (2.24). For simplicity, it only considers the case of the least squares.

To state our results precisely, denote by $\mathbf{\Lambda} = \{\lambda_s\}_s$ the finite ε-net of $(\frac{1}{m}, 1]$ of cardinality $|\mathbf{\Lambda}|$ with $0 < \varepsilon \leq \frac{1}{m}$. Then, $\mathbf{T} = \{T_k\}_k$ is given by

$$\mathbf{T} = \{\lfloor \lambda_s^{-1} \rfloor, \lambda_s \in \mathbf{\Lambda}\}, \tag{2.25}$$

where $\lfloor \lambda_s^{-1} \rfloor$ denotes the largest integer not exceeding λ_s^{-1}. Note that cardinality $|\mathbf{T}| \leq |\mathbf{\Lambda}|$ since for any $T_k \in \mathbf{T}$, there may exist more than one elementary $\lambda_s \in \mathbf{\Lambda}$ such that $T_k = \lfloor \lambda_s^{-1} \rfloor$.

Theorem 4 *Let ϕ be the least squares loss. Assume (2.8) holds for $0 < \beta < 1$, and (2.9) is valid for $0 < \zeta < 2$. Take $\eta_t \equiv \eta$ to be small enough. Define the set $\mathbf{T}$ by (2.25). If we choose T^* by (2.24), then for any $0 < \delta < 1$, with confidence at least $1 - \delta$,*

$$\mathcal{E}(\hat{f}_{T^*}^{(\mathbf{D_1})}) - \mathcal{E}(f_\rho) \lesssim m^{-\min\left\{1, \frac{2}{2\beta+\zeta}\right\}\beta} \log \frac{\log m}{\delta} \log\left(\frac{(|\mathbf{\Lambda}|+1)(3|\mathbf{\Lambda}|+1)}{\delta}\right).$$

The proof of the theorem will be provided in Section 4. Under the same conditions in Theorem 4, Theorem 1 shows that the error rate with a-prior choice of T should be $O\Big(m^{-\min\left\{\frac{2\beta}{2\beta+\zeta}, \beta\right\}} \log\left(\frac{\log m}{\delta}\right)\Big)$. The two bounds are slightly different in the logarithm term. Thus, Theorem 4 confirms that the cross-validation based on (2.24) can retain the best possible rate, which is achieved by the a-priori choices of the parameters.

3. Error Decomposition and Technical Estimates

In the section, we present the error decomposition for algorithm (2.2) and some useful estimates, which are critical to prove the main results in Section 2.

Error decomposition is a standard method to estimate learning errors in various algorithms. Here we will use a decomposition for the generalization error $\mathcal{E}(\hat{f}_T) - \mathcal{E}(f_\rho^\phi)$ as follows

$$\begin{aligned}\mathcal{E}(\hat{f}_T) - \mathcal{E}(f_\rho^\phi) = \mathcal{E}_{\mathbf{D}}(\hat{f}_T) - \mathcal{E}_{\mathbf{D}}(f_\lambda) + \Big\{\Big[\mathcal{E}(\hat{f}_T) - \mathcal{E}(f_\lambda)\Big] - \Big[\mathcal{E}_{\mathbf{D}}(\hat{f}_T) - \mathcal{E}_{\mathbf{D}}(f_\lambda)\Big]\Big\} \\ + \mathcal{E}(f_\lambda) - \mathcal{E}(f_\rho^\phi). \qquad (3.1)\end{aligned}$$

Here λ is a regularization parameter that should be optimized to achieve the sharpest possible error bounds and can be chosen by a trade-off of error terms in (3.1). The last term $\mathcal{E}(f_\lambda) - \mathcal{E}(f_\rho^\phi)$ is called the *approximation error*, which is independent of the sample set and can be estimated by the decay rate of $\mathcal{D}(\lambda)$. The middle term is the *sample error*, which depends on the sample set. We will upper bound it by employing the laws of large numbers in empirical processes. The two types of error terms have also been tackled well in other regularization schemes. See e.g., Cucker and Zhou (2007); Wang and Hu (2021); Yao et al. (2007); Guo et al. (2018). Comparing with these works, decomposition (3.1) introduces the *computational error*, the first term $\mathcal{E}_{\mathbf{D}}(\hat{f}_T) - \mathcal{E}_{\mathbf{D}}(f_\lambda)$ that characterizes the empirical error between the final estimator and the regularization function. Lin et al. (2016) remarked that it is uncertain whether the best rate is preserved when such an error related to optimization is considered. Our following theoretical analysis will address this point by incorporating it into the total error estimate, and the obtained results in Theorems 1, 2 confirm that the best generalization error can be achieved by algorithm (2.2) when the trade-off among computational error, approximation error and sample error are taken properly.

3.1 Estimates for the computational error with general losses

In this subsection, we will first estimate the computational error $\mathcal{E}_{\mathbf{D}}(\hat{f}_T) - \mathcal{E}_{\mathbf{D}}(f_\lambda)$ in case of general convex losses. To this end, we introduce the following lemma.

Lemma 1 *Suppose Assumption 1 holds with $q > 0$. If $\eta_t = \eta t^{-\theta}$ with $\frac{q}{q+1} < \theta < 1$ and $0 < \eta \leq \min\left\{\frac{\sqrt{1-\theta}}{2^{\frac{1}{2}} c_q (\kappa+1)^{q+1}}, \frac{1-\theta}{4c_\phi}\right\}$, then we have*

$$\begin{aligned}\mathcal{E}_{\mathbf{D}}(\hat{f}_T) - \mathcal{E}_{\mathbf{D}}(f_\lambda) &\leq C_1 \|f_\lambda\|_K^2 \Lambda_{T,\theta} \\ &+ \frac{T^\theta}{2\eta} \sum_{k=1}^{T-1} \frac{1}{k+1} \left[2\eta_{T-k} - \frac{1}{k} \sum_{t=T-k+1}^{T} 2\eta_t\right] \left[\mathcal{E}_{\mathbf{D}}(f_\lambda) - \mathcal{E}_{\mathbf{D}}(\hat{f}_{T-k})\right] \end{aligned} \tag{3.2}$$

where

$$\Lambda_{T,\theta} = \begin{cases} T^{-(1-\theta)}, & \text{if} \quad \theta > \frac{q+1}{q+2}, \\ T^{-(1-\theta)} \left(\log T\right), & \text{if} \quad \theta = \frac{q+1}{q+2}, \\ T^{-(\theta(1+q)-q)} \left(\log T\right), & \text{if} \quad \theta < \frac{q+1}{q+2}. \end{cases}$$

and C_1 is a universal constant.

Proof Note that by (2.10), for any measurable $f : \mathcal{X} \to \mathbb{R}$,

$$\mathcal{E}(\hat{f}) \leq \mathcal{E}(f) \quad \textit{and} \quad \mathcal{E}_{\mathbf{D}}(\hat{f}) \leq \mathcal{E}_{\mathbf{D}}(f). \tag{3.3}$$

This together with Lemma 17 of Lin et al. (2016) yields the conclusion (3.2). ■

In what follows, we shall see how this lemma can be used in bounding the total error of (3.1). For notational simplicity, with $f^* \in \mathcal{H}_K$ we denote

$$\begin{aligned} \mathcal{F}_{\mathbf{D}}(f^*) &:= \left[\mathcal{E}_{\mathbf{D}}(f^*) - \mathcal{E}_{\mathbf{D}}(f_\rho^\phi)\right] - \left[\mathcal{E}(f^*) - \mathcal{E}(f_\rho^\phi)\right], \\ \mathcal{M}_{\mathbf{D}}(f) &:= \left[\mathcal{E}(\hat{f}) - \mathcal{E}(f_\rho^\phi)\right] - \left[\mathcal{E}_{\mathbf{D}}(\hat{f}) - \mathcal{E}_{\mathbf{D}}(f_\rho^\phi)\right], \\ \mathcal{A}(f^*) &:= \mathcal{E}(f^*) - \mathcal{E}(f_\rho^\phi), \\ \mathcal{A}_{\mathbf{D},T} &:= \max_{t=1,\cdots,T} \mathcal{E}_{\mathbf{D}}(f_\rho^\phi) - \mathcal{E}_{\mathbf{D}}(\hat{f}_t) \end{aligned} \tag{3.4}$$

Proposition 1 *Under the same conditions of Lemma 1, then for each $t = 1, \cdots, T$,*

$$\mathcal{E}(\hat{f}_T) - \mathcal{E}(f_\rho^\phi) \leq \mathcal{M}_{\mathbf{D}}(f_T) + \frac{4-\theta}{1-\theta} \left(\mathcal{F}_{\mathbf{D}}(f_\lambda) + \mathcal{A}(f_\lambda) + \mathcal{A}_{\mathbf{D},T}\right) + C_1 \|f_\lambda\|_K^2 \Lambda_{T,\theta}, \tag{3.5}$$

where C_1 and $\Lambda_{T,\theta}$ are given in Lemma 1.

Proof To bound the generalization error $\mathcal{E}(\hat{f}_T) - \mathcal{E}(f_\rho^\phi)$, we shall tackle the errors in (3.1) term by term.

For each $k = 1, \cdots, T-1$, decompose $\mathcal{E}_{\mathbf{D}}(f_\lambda) - \mathcal{E}_{\mathbf{D}}(\hat{f}_{T-k})$ into

$$\begin{aligned} \mathcal{E}_{\mathbf{D}}(f_\lambda) - \mathcal{E}_{\mathbf{D}}(\hat{f}_{T-k}) &= \left[\mathcal{E}_{\mathbf{D}}(f_\lambda) - \mathcal{E}_{\mathbf{D}}(f_\rho^\phi)\right] - \left[\mathcal{E}(f_\lambda) - \mathcal{E}(f_\rho^\phi)\right] \\ &+ \left[\mathcal{E}(f_\lambda) - \mathcal{E}(f_\rho^\phi)\right] + \mathcal{E}_{\mathbf{D}}(f_\rho^\phi) - \mathcal{E}_{\mathbf{D}}(\hat{f}_{T-k}) \leq \mathcal{F}_{\mathbf{D}}(f_\lambda) + \mathcal{A}(f_\lambda) + \mathcal{A}_{\mathbf{D},T}. \end{aligned}$$

Plugging it into (3.2), the first term of (3.1) is bounded by Lemma 7 in Appendix as

$$\mathcal{E}_{\mathbf{D}}(\hat{f}_T) - \mathcal{E}_{\mathbf{D}}(f_\lambda) \leq \frac{3}{1-\theta}\left\{\mathcal{F}_{\mathbf{D}}(f_\lambda) + \mathcal{A}(f_\lambda) + \mathcal{A}_{\mathbf{D},T}\right\} + C_1\|f_\lambda\|_K^2\Lambda_{T,\theta}.$$

For the second term of (3.1), with the above notations we have

$$\begin{aligned}&\left[\mathcal{E}(\hat{f}_T) - \mathcal{E}(f_\lambda)\right] - \left[\mathcal{E}_{\mathbf{D}}(\hat{f}_T) - \mathcal{E}_{\mathbf{D}}(f_\lambda)\right] = \left\{\left[\mathcal{E}(\hat{f}_T) - \mathcal{E}(f_\rho^\phi)\right] - \left[\mathcal{E}_{\mathbf{D}}(\hat{f}_T) - \mathcal{E}_{\mathbf{D}}(f_\rho^\phi)\right]\right\}\\ &+ \left\{\left[\mathcal{E}_{\mathbf{D}}(f_\lambda) - \mathcal{E}_{\mathbf{D}}(f_\rho^\phi)\right] - \left[\mathcal{E}(f_\lambda) - \mathcal{E}(f_\rho^\phi)\right]\right\} \leq \mathcal{M}_{\mathbf{D}}(f_T) + \mathcal{F}_{\mathbf{D}}(f_\lambda).\end{aligned}$$

Rewrite (3.1) as

$$\mathcal{E}(\hat{f}_T) - \mathcal{E}(f_\rho^\phi) = \mathcal{E}_{\mathbf{D}}(\hat{f}_T) - \mathcal{E}_{\mathbf{D}}(f_\lambda) + \left\{\left[\mathcal{E}(\hat{f}_T) - \mathcal{E}(f_\lambda)\right] - \left[\mathcal{E}_{\mathbf{D}}(\hat{f}_T) - \mathcal{E}_{\mathbf{D}}(f_\lambda)\right]\right\} + \mathcal{A}(f_\lambda).$$

Putting the above estimates into it, we get the desired conclusion. ■

3.2 Bounding the total error

According to the above proposition, we proceed estimating the quantities $\mathcal{A}(f_\lambda), \mathcal{F}_{\mathbf{D}}(f_\lambda)$, $\mathcal{A}_{\mathbf{D},T}, \mathcal{M}_{\mathbf{D}}(f)$. Obviously, $\mathcal{A}(f_\lambda)$ can be bounded by $\mathcal{D}(\lambda)$. The other three quantities can be estimated by the following lemmas 2, 3.

Lemma 2 *Suppose Assumptions 1, 2 and 4 hold. Then with confidence at least $1-\delta$ the following inequality holds for all $f \in \mathcal{B}_R$*

$$\mathcal{M}_{\mathbf{D}}(f) \leq \frac{1}{2}\left[\mathcal{E}(\hat{f}) - \mathcal{E}(f_\rho^\phi)\right] + \Omega_R(f) \quad \textit{and} \qquad \mathcal{E}_{\mathbf{D}}(f_\rho^\phi) - \mathcal{E}_{\mathbf{D}}(\hat{f}) \leq \Omega_R(f), \tag{3.6}$$

where

$$\Omega_R(f) = C_2 \max\left\{\left(\frac{\max\{1, \|f\|_K^\zeta\}}{m}\right)^{\frac{2}{4-2\tau+\zeta\tau}}, \left(\frac{\max\{1, \|f\|_K^\zeta\}}{m}\right)^{\frac{2}{2+\zeta}}, \left(\frac{1}{m}\right)^{\frac{1}{2-\tau}} \log\frac{\log R}{\delta}\right\}$$

and C_2 is a universal constant (depending on q, ζ, τ) and will be given in the proof.

Lemma 3 *Suppose Assumptions 1 and 2 hold. For any $f^* \in \mathcal{H}_K$ and $\|f^*\|_K \leq \tilde{R}$ with some constant $\tilde{R} \geq 1$, then for any $0 < \delta < 1$, with confidence at least $1-\delta$, there holds*

$$\mathcal{F}_{\mathbf{D}}(f^*) \leq C_3 \max\left\{\frac{\tilde{R}^{q+1}}{m}, \left(\frac{1}{m}\right)^{\frac{1}{2-\tau}}, \mathcal{A}(f^*)\right\} \log\frac{2}{\delta}. \tag{3.7}$$

where C_3 is a universal constant (depending on q, τ) and will be given in the proof.

Specially, with $f^* = f_\lambda$ and Assumption 3, we have with confidence at least $1-\delta$,

$$\mathcal{F}_{\mathbf{D}}(f_\lambda) \leq C_4 \max\left\{\frac{\lambda^{\frac{(q+1)(\beta-1)}{2}}}{m}, \left(\frac{1}{m}\right)^{\frac{1}{2-\tau}}, \lambda^\beta\right\} \log\frac{2}{\delta}, \tag{3.8}$$

where C_4 is a universal constant (depending on q, τ, β) and will be given in the proof.

The proofs of the above two lemmas can be found in Appendix. With their help, we can derive the following estimate, which plays a key role in deriving our main theorems. Note we will show $\|f_t\|_K \leq T^{\frac{1-\theta}{2}}$ for all $t \leq T$ and therefore we can apply Proposition 2 with $R = T^{\frac{1-\theta}{2}}$.

Proposition 2 *Let $R > 0$. Assume $\|f_t\|_K \leq R$ for any $1 \leq t \leq T$. Under the assumptions of Lemmas 1, 2, then for any $0 < \delta < 1$, with confidence at least $1 - \delta$,*

$$\begin{aligned}
&\mathcal{E}(\hat{f}_T) - \mathcal{E}(f_\rho^\phi) \\
&\leq C_5 \max\left\{ \left(\frac{R_T^\zeta}{m}\right)^{\frac{2}{4-2\tau+\zeta\tau}}, \left(\frac{R_T^\zeta}{m}\right)^{\frac{2}{2+\zeta}}, \left(\frac{1}{m}\right)^{\frac{1}{2-\tau}} \log\frac{2\log R}{\delta}, \frac{\lambda^{\frac{(q+1)(\beta-1)}{2}}}{m}\log\frac{4}{\delta}, \lambda^\beta \log\frac{4}{\delta} \right\} \\
&+ C_5 \lambda^{\beta-1} \Lambda_{T,\theta} \qquad (3.9)
\end{aligned}$$

where $R_T = \max\{1, \|f_t\|_K, t = 1, \cdots, T\}$ and C_5 is a universal constant (will be given in the proof).

Proof We shall prove (3.9) by Proposition 1. According to Lemma 2, with confidence at least $1 - \delta$,

$$\mathcal{M}_{\mathbf{D}}(f_T) \leq \frac{1}{2}\left[\mathcal{E}(\hat{f}_T) - \mathcal{E}(f_\rho^\phi)\right] + \Omega_R(f_T)$$

and

$$\mathcal{A}_{\mathbf{D},T} \leq \max_{t=1,\cdots,T}\{\Omega_R(f_t)\}$$

This together with Lemma 3 and Proposition 1 yields that with confidence at least $1 - 2\delta$,

$$\begin{aligned}
&\mathcal{E}(\hat{f}_T) - \mathcal{E}(f_\rho^\phi) \leq \frac{1}{2}\left[\mathcal{E}(\hat{f}_T) - \mathcal{E}(f_\rho^\phi)\right] + \left(\frac{5-2\theta}{1-\theta}\right) \max_{t=1,\cdots,T}\{\Omega_R(f_t)\} + \\
&C_4\left(\frac{4-\theta}{1-\theta}\right) \max\left\{ \frac{\lambda^{\frac{(q+1)(\beta-1)}{2}}}{m}, \left(\frac{1}{m}\right)^{\frac{1}{2-\tau}}, \lambda^\beta \right\} \log\frac{2}{\delta} + \left(\frac{4-\theta}{1-\theta}\right)\mathcal{A}(f_\lambda) + C_1\|f_\lambda\|_K^2 \Lambda_{T,\theta}.
\end{aligned}$$

Note that

$$\max_{t=1,\cdots,T}\{\Omega_R(f_t)\} \leq C_2 \max\left\{ \left(\frac{R_T^\zeta}{m}\right)^{\frac{2}{4-2\tau+\zeta\tau}}, \left(\frac{R_T^\zeta}{m}\right)^{\frac{2}{2+\zeta}}, \left(\frac{1}{m}\right)^{\frac{1}{2-\tau}} \log\frac{\log R}{\delta} \right\}$$

and $\mathcal{A}(f_\lambda) \leq \mathcal{D}(\lambda) \leq c_\beta \lambda^\beta$, $\|f_\lambda\|_K^2 \leq c_\beta \lambda^{\beta-1}$. Then subtracting $\frac{1}{2}\left[\mathcal{E}(\hat{f}_T) - \mathcal{E}(f_\rho^\phi)\right]$ from both sides of the above inequality and scaling 2δ to δ, then we can get the conclusion (3.9) with

$$C_5 = 2\max\left\{ \frac{5-2\theta}{1-\theta} C_2 + \frac{4-\theta}{1-\theta}(C_4 + c_\beta), C_1 c_\beta \right\}.$$

■

Observe from the above lemma that a crude bound $\|f_t\|_K \leq R$ on *the whole space* appears only in the logarithm term, which does no dominate the error analysis of $\mathcal{E}(\hat{f}_T) - \mathcal{E}(f_\rho^\phi)$. Meanwhile, a tight bound of $R_T = \max\{1, \|f_t\|_K, t = 1, \cdots, T\}$ will lead to a sharp estimate of $\mathcal{E}(\hat{f}_T) - \mathcal{E}(f_\rho^\phi)$. Thus, in the next subsection we shall improve the bound of the iterate sequence in probability.

3.3 Improving the iterate bound for general losses

In the paper (Lin et al., 2016), a uniform bound of $\mathcal{H}_K$-norm of $\{f_t\}_{t=1}^T$ has been established in the following lemma.

Lemma 4 *Suppose Assumption 1 holds. Let $\eta_t = \eta t^{-\theta}$ with $\frac{q}{q+1} \leq \theta < 1$ and η satisfying*

$$0 < \eta \leq \min\left\{\frac{\sqrt{1-\theta}}{\sqrt{2}c_q(\kappa+1)^{q+1}}, \frac{1-\theta}{4c_\phi}\right\},$$

then for $t = 1, \cdots, T$,

$$\|f_{t+1}\|_K \leq t^{\frac{1-\theta}{2}}. \tag{3.10}$$

One main contribution of this paper is to refine the above bound for the $\mathcal{H}_K$-norms of the sequence $\{f_t\}$ by algorithm (2.2), which is stated in the theorem below.

Theorem 5 *Let $0 < \lambda \leq T^{-(1-\theta)}$ and $\lambda = m^{-\gamma}$, where γ is defined in (2.12). If the stepsize η_t takes the form as $\eta_t = \eta t^{-\theta}$ with $\max\left\{\frac{1}{2}, \frac{q}{q+1}\right\} < \theta < 1$ and η satisfying (2.16), then with confidence at least $1-\delta$, there holds*

$$\|f_{t+1}\|_K \lesssim \lambda^{\frac{\beta-1}{2}} \left(\log\frac{\log T}{\delta}\right)^{\frac{1}{2}}, \quad t = 1, \cdots, T. \tag{3.11}$$

The notation $\lesssim$ means that the inequality holds up to a multiplicative constant that depends on various parameters appearing in the assumptions, but not on λ, T or δ.

It will be shown in the next subsection that the refined bound (3.11) essentially improves the error estimate for (3.1).

Remark 6 *This theorem asserts that $\|f_T\|_K$ has the bound of order $O\left(T^{\frac{(1-\beta)(1-\theta)}{2}}\right)$ (ignoring the log term) if we take $\lambda = T^{-(1-\theta)}$, which is obviously superior to $O\left(T^{\frac{1-\theta}{2}}\right)$ in (3.10). We also notice that this upper bound is nearly a constant when $\beta \to 1$. It implies that if $f_\rho^\phi \in \mathcal{H}_K$, the actual iterative process happens in a bounded region and the sequence $\{f_t\}$ is uniformly bounded (independent of t) with high probability.*

3.4 Proof of Theorem 5

This subsection is devoted to proving Theorem 5. The following lemma is key and will be used several times for our proofs, which has been proved in Lin et al. (2016).

Lemma 5 *For any fixed $f^* \in \mathcal{H}_K$, and $t = 1, \cdots, T$, there holds*

$$\|f_{t+1} - f^*\|_K^2 \leq \|f_t - f^*\|_K^2 + \eta_t^2 G_t^2 - 2\eta_t \left[\mathcal{E}_{\mathbf{D}}(f_t) - \mathcal{E}_{\mathbf{D}}(f^*)\right] \tag{3.12}$$

where

$$G_t^2 = \left\| \frac{1}{m} \sum_{j=1}^{m} \phi'_- \left(y_j, f_t(x_j)\right) K_{x_j} \right\|_K^2 \leq (1+\kappa)^{2q} c_q^2 \left(1 + \|f_t\|_K^{2q}\right). \tag{3.13}$$

Using the above lemma, we can bound the iterate sequence in probability as follows. This proposition allows us to control R_T^2 by a sub-quadratic function of R_T, from which we can get a bound on the norm of iterates.

Proposition 3 *If the stepsize η_t is taken as the form in Theorem 5, then for any $f^* \in \mathcal{H}_K$, with confidence at least $1-\delta$,*

$$\|f_{t+1}\|_K^2 \leq 5\left\{ \|f^*\|_K^2 + 1 + \Big(\mathcal{F}_{\mathbf{D}}(f^*) + \mathcal{A}(f^*) + \Delta_T\Big) t^{1-\theta} \right\}, \quad t = 1, \cdots, T, \tag{3.14}$$

where

$$\Delta_T = C_2 \max\left\{ \left(\frac{R_T^\zeta}{m}\right)^{\frac{2}{4-2\tau+\zeta\tau}}, \left(\frac{R_T^\zeta}{m}\right)^{\frac{2}{2+\zeta}}, \left(\frac{1}{m}\right)^{\frac{1}{2-\tau}} \log \frac{\log T^{\frac{1-\theta}{2}}}{\delta} \right\}$$

with $R_T = \max\{1, \|f_t\|_K, t = 1, \cdots, T\}$ (C_2 is a universal constant, given in Lemma 2).

Proof It follows from (3.12) and (3.3) that for $t = 1, \cdots, T$,

$$\begin{aligned}
\|f_{t+1} - f^*\|_K^2 &\leq \|f_t - f^*\|_K^2 + \eta_t^2 G_t^2 - 2\eta_t \left[\mathcal{E}_{\mathbf{D}}(f_t) - \mathcal{E}_{\mathbf{D}}(f_\rho^\phi)\right] + 2\eta_t \left[\mathcal{E}_{\mathbf{D}}(f^*) - \mathcal{E}_{\mathbf{D}}(f_\rho^\phi)\right] \\
&\leq \|f_t - f^*\|_K^2 + \eta_t^2 G_t^2 - 2\eta_t \left[\mathcal{E}_{\mathbf{D}}(\hat{f}_t) - \mathcal{E}_{\mathbf{D}}(f_\rho^\phi)\right] + 2\eta_t \left[\mathcal{E}_{\mathbf{D}}(f^*) - \mathcal{E}_{\mathbf{D}}(f_\rho^\phi)\right] \\
&= \|f_t - f^*\|_K^2 + \eta_t^2 G_t^2 + 2\eta_t \left\{ \left[\mathcal{E}(\hat{f}_t) - \mathcal{E}(f_\rho^\phi)\right] - \left[\mathcal{E}_{\mathbf{D}}(\hat{f}_t) - \mathcal{E}_{\mathbf{D}}(f_\rho^\phi)\right] \right\} \\
&\quad + 2\eta_t \mathcal{F}_{\mathbf{D}}(f^*) + 2\eta_t \mathcal{A}(f^*) - 2\eta_t \left[\mathcal{E}(\hat{f}_t) - \mathcal{E}(f_\rho^\phi)\right] \\
&= \|f_t - f^*\|_K^2 + \eta_t^2 G_t^2 + 2\eta_t \mathcal{M}_{\mathbf{D}}(f_t) \\
&\quad + 2\eta_t \mathcal{F}_{\mathbf{D}}(f^*) + 2\eta_t \mathcal{A}(f^*) - 2\eta_t \left[\mathcal{E}(\hat{f}_t) - \mathcal{E}(f_\rho^\phi)\right].
\end{aligned}$$

Note that by (3.10), $\|f_t\|_K \leq t^{\frac{1-\theta}{2}}$ for any $t \in \mathbb{N}$. It implies that for any $t \leq T$, $f_t \in \mathcal{B}_R$ with $R = T^{\frac{1-\theta}{2}}$. Then by Lemma 2, we have that there exists a subset $\mathcal{Z}_\delta^m \subset \mathcal{Z}^m$ with measure at least $1-\delta$ such that for arbitrary $\mathbf{D} \in \mathcal{Z}_\delta^m$,

$$\|f_{t+1} - f^*\|_K^2 \leq \|f_t - f^*\|_K^2 + \eta_t^2 G_t^2 + 2\eta_t \mathcal{F}_{\mathbf{D}}(f^*) + 2\eta_t \mathcal{A}(f^*) + 2\eta_t \Delta_T - \eta_t \left[\mathcal{E}(\hat{f}_t) - \mathcal{E}(f_\rho^\phi)\right]$$

holds for each $t = 1, \cdots, T$.

It follows from $\mathcal{E}(\hat{f}_t) - \mathcal{E}(f_\rho^\phi) > 0$ that when $\mathbf{D} \in \mathcal{Z}_\delta^m$,

$$\|f_{t+1} - f^*\|_K^2 \leq \|f_t - f^*\|_K^2 + \eta_t^2 G_t^2 + 2\eta t^{-\theta} \left(\mathcal{F}_{\mathbf{D}}(f^*) + \mathcal{A}(f^*) + \Delta_T\right) \tag{3.15}$$

holds for $t = 1, \cdots, T$.

Now we are in a position of estimating $\eta_t^2 G_t^2$. With Assumption 1, we have that

$$\eta_t^2 G_t^2 \leq c_q^2 \eta^2 (1 + \|f_t\|_\infty^q)^2 t^{-2\theta} \leq (1 + \kappa^{2q}) c_q^2 \eta^2 (1 + \|f_t\|_K^{2q}) t^{-2\theta}.$$

If $q \leq 1$, $\|f_t\|_K^{2q} \leq 1 + \|f_t\|_K^2$. If $q > 1$, by (3.10), we have $\|f_t\|_K^{2q} \leq \|f_t\|_K^2 t^{(q-1)(1-\theta)}$.

Combining the two cases above yields

$$\eta_t^2 G_t^2 \leq (1 + \kappa^{2q}) c_q^2 \eta^2 t^{-2\theta} (1 + (1 + \|f_t\|_K^2) t^{2\theta - q^*}) \leq 2(1 + \kappa^{2q}) c_q^2 \eta^2 (1 + \|f_t\|_K^2) t^{-q^*}$$

where q^* is defined in (2.18).

Plugging it into (3.15), then

$$\begin{aligned} \|f_{t+1} - f^*\|_K^2 \leq \|f_t - f^*\|_K^2 &+ 2(1 + \kappa^{2q}) c_q^2 \eta^2 (1 + \|f_t\|_K^2) t^{-q^*} \\ &+ 2\eta \bigg(\mathcal{F}_{\mathbf{D}}(f^*) + \mathcal{A}(f^*) + \Delta_T \bigg) t^{-\theta}, \quad t = 1, \cdots, T. \end{aligned}$$

Applying this inequality iteratively with $f_1 = 0$, we derive for each $t = 1, \cdots, T$,

$$\begin{aligned} \|f_{t+1} - f^*\|_K^2 &\leq \|f^*\|_K^2 + 2(1 + \kappa^{2q}) c_q^2 \eta^2 \sum_{j=1}^t (1 + \|f_j\|_K^2) j^{-q^*} + \frac{2\eta}{1-\theta} \bigg(\mathcal{F}_{\mathbf{D}}(f^*) + \mathcal{A}(f^*) + \Delta_T \bigg) t^{1-\theta} \\ &\leq \|f^*\|_K^2 + 2(1 + \kappa^{2q}) c_q^2 \eta^2 \left(1 + \max_{j=1,\cdots,t} \|f_j\|_K^2 \right) \sum_{j=1}^t j^{-q^*} + \frac{2\eta}{1-\theta} \bigg(\mathcal{F}_{\mathbf{D}}(f^*) + \mathcal{A}(f^*) + \Delta_T \bigg) t^{1-\theta}. \end{aligned}$$

The restriction on θ implies that $q^* > 1$. Applying the elementary inequality $\sum_{j=1}^t j^{-q^*} \leq \frac{q^*}{q^*-1}$ for $q^* > 1$, then

$$\|f_{t+1} - f^*\|_K^2 \leq \|f^*\|_K^2 + \frac{2(1 + \kappa^{2q}) c_q^2 \eta^2 q^*}{q^* - 1} \left(1 + \max_{j=1,\cdots,t} \|f_j\|_K^2 \right) + \frac{2\eta}{1-\theta} \bigg(\mathcal{F}_{\mathbf{D}}(f^*) + \mathcal{A}(f^*) + \Delta_T \bigg) t^{1-\theta}.$$

Thus, by $\|f_{t+1}\|_K^2 \leq 2\|f_{t+1} - f^*\|_K^2 + 2\|f^*\|_K^2$,

$$\|f_{t+1}\|_K^2 \leq 4\|f^*\|_K^2 + \frac{4(1 + \kappa^{2q}) c_q^2 \eta^2 q^*}{q^* - 1} \left(1 + \max_{j=1,\cdots,t} \|f_j\|_K^2 \right) + \frac{4\eta}{1-\theta} \bigg(\mathcal{F}_{\mathbf{D}}(f^*) + \mathcal{A}(f^*) + \Delta_T \bigg) t^{1-\theta}.$$

Suppose that for each $1 \leq j \leq t$, there holds

$$\|f_j\|_K^2 \leq 5 \bigg\{ \|f^*\|_K^2 + 1 + \Big(\mathcal{F}_{\mathbf{D}}(f^*) + \mathcal{A}(f^*) + \Delta_T \Big) (j-1)^{1-\theta} \bigg\}.$$

Then,

$$\begin{aligned}\|f_{t+1}\|_K^2 \leq\ & 4\|f^*\|_K^2 + \frac{4(1+\kappa^{2q})c_q^2\eta^2 q^*}{q^*-1}\left(6+5\|f^*\|_K^2+5\Big(\mathcal{F}_{\mathbf{D}}(f^*)+\mathcal{A}(f^*)+\Delta_T\Big)(t-1)^{1-\theta}\right)\\ &+\frac{4\eta}{1-\theta}\left(\mathcal{F}_{\mathbf{D}}(f^*)+\mathcal{A}(f^*)+\Delta_T\right)t^{1-\theta}\\ =\ & \left(4+\frac{20(1+\kappa^{2q})c_q^2\eta^2 q^*}{q^*-1}\right)\|f^*\|_K^2+\frac{24(1+\kappa^{2q})c_q^2\eta^2 q^*}{q^*-1}\\ &+\left(\frac{20(1+\kappa^{2q})c_q^2\eta^2 q^*}{q^*-1}+\frac{4\eta}{1-\theta}\right)\left(\mathcal{F}_{\mathbf{D}}(f^*)+\mathcal{A}(f^*)+\Delta_T\right)t^{1-\theta}.\end{aligned}$$

Then by (2.16),

$$\|f_{t+1}\|_K^2 \leq 5\left\{\|f^*\|_K^2+1+\Big(\mathcal{F}_{\mathbf{D}}(f^*)+\mathcal{A}(f^*)+\Delta_R\Big)t^{1-\theta}\right\}.$$

So, we have when $\mathbf{D}\in\mathcal{Z}_\delta^m$, for each $t=1,\cdots,T$,

$$\|f_{t+1}\|_K^2 \leq 5\left\{\|f^*\|_K^2+1+\Big(\mathcal{F}_{\mathbf{D}}(f^*)+\mathcal{A}(f^*)+\Delta_T\Big)t^{1-\theta}\right\}.$$

The proof is finished. ■

This proposition establishes the upper bound of $\{f_t\}_{t=1}^T$ in terms of a reference function $f^*\in\mathcal{H}_K$. It tells us with a suitable choice of f^*, the estimate of $\|f_t\|_K$ can be improved greatly, which will be shown in the following proof of Theorem 5.

Proof of Theorem 5. Recall $\mathcal{A}(f_\lambda)\leq\mathcal{D}(f_\lambda)\leq c_\beta\lambda^\beta$ and $\|f_\lambda\|_K^2\leq c_\beta\lambda^{\beta-1}$. Applying Proposition 3 with $f^*=f_\lambda$ and $t^{1-\theta}\leq T^{1-\theta}\leq\lambda^{-1}$, we have that with confidence at least $1-\delta$,

$$\|f_{t+1}\|_K^2 \leq 5\left\{2c_\beta\lambda^{\beta-1}+1+\Delta_T\lambda^{-1}+\mathcal{F}_{\mathbf{D}}(f_\lambda)\lambda^{-1}\right\}$$

holds for each $t=1,\cdots,T$.

By Lemma 3 then with confidence at least $1-2\delta$,

$$\begin{aligned}&\|f_{t+1}\|_K^2 \leq 5(2c_\beta+1+C_2+C_4)\lambda^{-1}\times\\ &\max\left\{\left(\frac{R_T^\zeta}{m}\right)^{\frac{2}{4-2\tau+\zeta\tau}},\left(\frac{R_T^\zeta}{m}\right)^{\frac{2}{2+\zeta}},\left(\frac{1}{m}\right)^{\frac{1}{2-\tau}}\log\frac{\log T^{\frac{1-\theta}{2}}}{\delta},\frac{\lambda^{\frac{(q+1)(\beta-1)}{2}}}{m}\log\frac{2}{\delta},\lambda^\beta\log\frac{2}{\delta}\right\}\end{aligned}$$

holds for each $t=1,\cdots,T$.

By the choice (2.12) of λ, scaling 2δ to δ, it follows that with confidence at least $1-\delta$,

$$\|f_{t+1}\|_K^2 \leq 5(2c_\beta+1+C_2+C_4)\lambda^{\beta-1}\max\left\{\left(R_T^2\lambda^{1-\beta}\right)^{\frac{\zeta}{4-2\tau+\zeta\tau}},\left(R_T^2\lambda^{1-\beta}\right)^{\frac{\zeta}{2+\zeta}},\log\frac{2\log T^{\frac{1-\theta}{2}}}{\delta}\right\}$$

holds for each $t = 1, \cdots, T$.

Denote $\widetilde{R}_T = \max_{t=1,\cdots,T} \|f_t\|_K$. By the above inequality, then with probability at least $1-\delta$,

$$\widetilde{R}_T^2 \le 5(2c_\beta + 1 + C_2 + C_4)\lambda^{\beta-1} \max\left\{ \left(R_T^2 \lambda^{1-\beta}\right)^{\frac{\zeta}{4-2\tau+\zeta\tau}}, \left(R_T^2\lambda^{1-\beta}\right)^{\frac{\zeta}{2+\zeta}}, \log\frac{2\log T^{\frac{1-\theta}{2}}}{\delta} \right\}.$$

Recall that $R_T = \max\{1, \|f_t\|_K, t = 1, \cdots, T\}$. When $\widetilde{R}_T \le 1$, the statement of Theorem 5 is obviously true.

Now we consider $\widetilde{R}_T > 1$. Then $R_T = \widetilde{R}_T$ and the above inequality implies

$$\widetilde{R}_T^2 \le 5(2c_\beta + 1 + C_2 + C_4)\lambda^{\beta-1} \max\left\{ \left(\widetilde{R}_T^2 \lambda^{1-\beta}\right)^{\frac{\zeta}{4-2\tau+\zeta\tau}}, \left(\widetilde{R}_T^2\lambda^{1-\beta}\right)^{\frac{\zeta}{2+\zeta}}, \log\frac{2\log T^{\frac{1-\theta}{2}}}{\delta} \right\}.$$

If $\widetilde{R}_T^2 \le 5(2c_\beta + 1 + C_2 + C_4)\lambda^{\beta-1}\left(\widetilde{R}_T^2\lambda^{1-\beta}\right)^{\frac{\zeta}{4-2\tau+\zeta\tau}}$, then

$$\widetilde{R}_T^2 \le (5(2c_\beta + 1 + C_2 + C_4))^{\frac{4-2\tau+\tau\zeta}{4-2\tau+\zeta\tau-\zeta}} \lambda^{\beta-1}.$$

Else if $\widetilde{R}_T^2 \le 5(2c_\beta + 1 + C_2 + C_4)\lambda^{\beta-1}\left(\widetilde{R}_T^2\lambda^{1-\beta}\right)^{\frac{\zeta}{2+\zeta}}$, then

$$\widetilde{R}_T^2 \le (5(2c_\beta + 1 + C_2 + C_4))^{\frac{2+\zeta}{2}} \lambda^{\beta-1}.$$

Otherwise we have $\widetilde{R}_T^2 \le 5(2c_\beta + 1 + C_2 + C_4)\lambda^{\beta-1}\log\frac{2\log T^{\frac{1-\theta}{2}}}{\delta}$.

Collecting the above analysis, we get the statement of Theorem 5 when $\widetilde{R}_T > 1$. The proof is finished. ■

3.5 Proof of Theorem 2

Now we can prove Theorem 2 with the help of Theorem 5.

Proof of Theorem 2. We shall prove Theorem 2 by Proposition 2. Recall (3.10), then we take $R = T^{\frac{1-\theta}{2}}$. Then by (3.9) with confidence at least $1-\delta$,

$$\begin{aligned}
&\mathcal{E}(\hat{f}_T) - \mathcal{E}(f_\rho^\phi) \\
&\le C_5 \max\left\{ \left(\frac{R_T^\zeta}{m}\right)^{\frac{2}{4-2\tau+\zeta\tau}}, \left(\frac{R_T^\zeta}{m}\right)^{\frac{2}{2+\zeta}}, \left(\frac{1}{m}\right)^{\frac{1}{2-\tau}} \log\frac{2\log T^{\frac{1-\theta}{2}}}{\delta}, \frac{\lambda^{\frac{(q+1)(\beta-1)}{2}}}{m}\log\frac{4}{\delta}, \lambda^\beta \log\frac{4}{\delta} \right\} \\
&+ C_5\lambda^{\beta-1}\Lambda_{T,\theta} \\
&\le C_5\lambda^\beta \max\left\{ \left(R_T^2\lambda^{1-\beta}\right)^{\frac{\zeta}{4-2\tau+\zeta\tau}}, \left(R_T^2\lambda^{1-\beta}\right)^{\frac{\zeta}{2+\zeta}}, \log\frac{2\log T^{\frac{1-\theta}{2}}}{\delta} \right\} + C_5\lambda^{\beta-1}\Lambda_{T,\theta} \qquad (3.16)
\end{aligned}$$

where the last inequality is obtained by $\lambda = m^{-\gamma}$ with γ in (2.12).

By (3.11), with confidence at least $1-\delta$, $R_T \le c_1 \lambda^{\frac{\beta-1}{2}} \left(\log \frac{\log T}{\delta}\right)^{\frac{1}{2}}$ ($c_1>0$ is a universal constant). Putting it into (3.16) yields that with confidence at least $1-2\delta$,

$$\mathcal{E}(\hat{f}_T) - \mathcal{E}(f_\rho^\phi) \le C_5 c_1^{\frac{2\zeta}{2+\zeta}} \lambda^\beta \log \frac{\log T}{\delta} + C_5 \lambda^{\beta-1} \Lambda_{T,\theta}. \tag{3.17}$$

By the choice of λ, T, we have that

$$\Lambda_{T,\theta} = \begin{cases} m^{-\gamma}, & \text{if} \quad \theta > \frac{q+1}{q+2}, \\ \alpha m^{-\gamma} (\log m), & \text{if} \quad \theta = \frac{q+1}{q+2}, \\ \alpha m^{-\alpha(\theta(1+q)-q)} (\log m), & \text{if} \quad \theta < \frac{q+1}{q+2}. \end{cases}$$

Plugging it into (3.17) and scaling 2δ to δ, then with confidence at least $1-\delta$,

$$\mathcal{E}(\hat{f}_T) - \mathcal{E}(f_\rho^\phi) \le$$

$$\begin{cases} 2(C_5 c_1^{\frac{2\zeta}{2+\zeta}} + C_5) m^{-\beta\gamma} \log \frac{\log T}{\delta}, & \text{if} \quad \theta > \frac{q+1}{q+2}, \\ 2(1+\alpha)(C_5 c_1^{\frac{2\zeta}{2+\zeta}} + C_5) \max\left\{ m^{-\beta\gamma} \log \frac{\log T}{\delta}, m^{-\beta\gamma} (\log m) \right\}, & \text{if} \quad \theta = \frac{q+1}{q+2}, \\ 2(1+\alpha)(C_5 c_1^{\frac{2\zeta}{2+\zeta}} + C_5) \max\left\{ m^{-\beta\gamma} \log \frac{\log T}{\delta}, m^{-\beta\gamma+\gamma-\alpha(\theta(1+q)-q)} (\log m) \right\}, & \text{if} \quad \theta < \frac{q+1}{q+2}. \end{cases}$$

Then we can get the statement of Theorem 2 by noting $T = \lfloor m^\alpha \rfloor$. ■

3.6 Proof of Theorem 1

This subsection proves error bounds for smooth losses stated in Theorem 1. Since ϕ' is Lipschitz continuous, its proof is shorter and easier than Theorem 2. Following the similar proof idea, we shall present the iterate bound and computational error for smooth losses in the following lemma.

Lemma 6 *Assume that ϕ'_- is Lipschitz continuous with a constant $L>0$. Then we have for $t \in \mathbb{N}$,*

$$\|f_{t+1}\|_K \le \sqrt{2c_\phi \sum_{k=1}^t \eta_k}. \tag{3.18}$$

In particular, if $\eta_t = \eta t^{-\theta}$ with $0 \le \theta < 1$ and $0 < \eta \le \frac{1-\theta}{2c_\phi}$, then for $t \in \mathbb{N}$,

$$\|f_{t+1}\|_K \le t^{\frac{1-\theta}{2}}. \tag{3.19}$$

If $0 < \eta < \frac{1}{L\kappa^2}$, then for $t \in \mathbb{N}$

$$\begin{aligned} \|f_{t+1} - f_\lambda\|_K^2 &\le \|f_t - f_\lambda\|_K^2 - 2\eta_t \left[\mathcal{E}_{\mathbf{D}}(f_{t+1}) - \mathcal{E}_{\mathbf{D}}(f_\lambda)\right] \\ &\le \|f_t - f_\lambda\|_K^2 - 2\eta_t \left[\mathcal{E}_{\mathbf{D}}(\hat{f}_{t+1}) - \mathcal{E}_{\mathbf{D}}(f_\lambda)\right], \end{aligned} \tag{3.20}$$

and

$$\mathcal{E}_{\mathbf{D}}(\hat{f}_T) - \mathcal{E}_{\mathbf{D}}(f_\lambda) \le \mathcal{E}_{\mathbf{D}}(f_T) - \mathcal{E}_{\mathbf{D}}(f_\lambda) \le \frac{2\|f_\lambda\|_K^2}{\eta} T^{\theta-1}. \tag{3.21}$$

The statements of this lemma can be derived by (3.3) and Lemmas 20, 21 in Lin et al. (2016). Then, we follow the similar proof procedure of Theorem 5 and obtain the refined iterate bound for smooth losses as follows.

Theorem 6 *Let $0 < \lambda \leq T^{-(1-\theta)}$ and $\lambda = m^{-\gamma}$, where γ is defined in (2.12). If the step size η_t takes the form as $\eta_t = \eta t^{-\theta}$ with $0 \leq \theta < 1$ and some $0 < \eta < \min\left\{\frac{1-\theta}{2c_\phi}, \frac{1}{L\kappa^2}\right\}$ then for any $\epsilon > 0$, with confidence at least $1-\delta$,*

$$\|f_{t+1}\|_K \lesssim \lambda^{\frac{\beta-1}{2}} \left(\log \frac{\log T}{\delta}\right)^{\frac{1}{2}}, \quad t = 1, \cdots, T.$$

Here the notation $\lesssim$ means that the inequality holds up to a multiplicative constant that depends on various parameters appearing in the assumptions, but not on λ, T or δ.

Proof of Theorem 6. By (3.20), for $t \in \mathbb{N}$,

$$\|f_{t+1} - f_\lambda\|_K^2 \leq \|f_t - f_\lambda\|_K^2 + 2\eta_t \mathcal{M}_{\mathbf{D}}(f_{t+1}) + 2\eta_t \mathcal{F}_{\mathbf{D}}(f_\lambda) + 2\eta_t \mathcal{A}(f_\lambda) - 2\eta_t \left[\mathcal{E}(\hat{f}_{t+1}) - \mathcal{E}(f_\rho^\phi)\right].$$

By Lemmas 2, 3, 6, with confidence at least $1-2\delta$,

$$\|f_{t+1} - f_\lambda\|_K^2 \leq \|f_t - f_\lambda\|_K^2 + 2\eta_t \tilde{\Delta}_T + 2\eta_t \mathcal{A}(f_\lambda), t = 1, \cdots, T, \tag{3.22}$$

with

$$\tilde{\Delta}_T = (C_2 + C_4) \max\left\{\left(\frac{R_T^\zeta}{m}\right)^{\frac{2}{4-2\tau+\zeta\tau}}, \left(\frac{R_T^\zeta}{m}\right)^{\frac{2}{2+\zeta}}, \lambda^\beta \log \frac{\log T^{\frac{1-\theta}{2}}}{\delta}\right\}.$$

Here $\tilde{\Delta}_T$ is obtained by $\lambda = m^{-\gamma}$ with (2.12) and $R_T = \max\{1, \|f_t\|_K, t = 1, \cdots, T\}$.

Applying this inequality iteratively with $f_1 = 0$, then with confidence at least $1-2\delta$,

$$\|f_{t+1} - f_\lambda\|_K^2 \leq \|f_\lambda\|_K^2 + 2\sum_{j=1}^{t} \eta_t \left(\tilde{\Delta}_T + \mathcal{A}(f_\lambda)\right)$$

holds for each $t = 1, \cdots, T$.

By $\|f_{t+1}\|_K^2 \leq 2\|f_{t+1} - f_\lambda\|_K^2 + 2\|f_\lambda\|_K^2$ and $t^{1-\theta} \leq \lambda^{-1}$, with confidence at least $1-2\delta$,

$$\begin{aligned}\|f_{t+1}\|_K^2 &\leq 4\|f_\lambda\|_K^2 + 4\sum_{j=1}^{t} \eta_t \left(\tilde{\Delta}_T + \mathcal{A}(f_\lambda)\right) \leq 4\|f_\lambda\|_K^2 + 4\eta \left(\tilde{\Delta}_T + \mathcal{A}(f_\lambda)\right) t^{1-\theta} \\ &\leq 4\|f_\lambda\|_K^2 + 4\eta \left(\tilde{\Delta}_T + \mathcal{A}(f_\lambda)\right) \lambda^{-1} \leq \left[4c_\beta + 4\eta(1+c_\beta)\right] \tilde{\Delta}_T \lambda^{-1}\end{aligned}$$

holds for each $t = 1, \cdots, T$.

The rest of proof is very similar to Theorem 5. Here we omit it for simplicity. ■

With the help of Theorem 6, we can prove Theorem 1.

Proof of Theorem 1. By (3.21) and $0 < \lambda \leq T^{-(1-\theta)}$, we have

$$\begin{aligned}\mathcal{E}(\hat{f}_T) - \mathcal{E}(f_\rho^\phi) &= \left[\mathcal{E}(\hat{f}_T) - \mathcal{E}(f_\rho^\phi)\right] - \left[\mathcal{E}_{\mathbf{D}}(\hat{f}_T) - \mathcal{E}_{\mathbf{D}}(f_\rho^\phi)\right] + \mathcal{E}_{\mathbf{D}}(\hat{f}_T) - \mathcal{E}_{\mathbf{D}}(f_\lambda) + \mathcal{F}_{\mathbf{D}}(f_\lambda) + \mathcal{A}(f_\lambda) \\ &\leq \left[\mathcal{E}(\hat{f}_T) - \mathcal{E}(f_\rho^\phi)\right] - \left[\mathcal{E}_{\mathbf{D}}(\hat{f}_T) - \mathcal{E}_{\mathbf{D}}(f_\rho^\phi)\right] + \frac{2\|f_\lambda\|_K^2}{\eta} T^{\theta-1} + \mathcal{F}_{\mathbf{D}}(f_\lambda) + \mathcal{D}(\lambda) \\ &\leq \mathcal{M}_{\mathbf{D}}(f_T) + \mathcal{F}_{\mathbf{D}}(f_\lambda) + (2\eta^{-1} + 1) c_\beta \lambda^\beta. \end{aligned} \tag{3.23}$$

This together with Lemmas 2, 3 yields with confidence at least $1-2\delta$,

$$\begin{aligned}\mathcal{E}(\hat{f}_T)-\mathcal{E}(f_\rho^\phi)\leq&\frac{1}{2}\left[\mathcal{E}(\hat{f})-\mathcal{E}(f_\rho^\phi)\right]+\Omega_R(f_T)\\&+C_4\max\left\{\frac{\lambda^{\frac{(q+1)(\beta-1)}{2}}}{m},\left(\frac{1}{m}\right)^{\frac{1}{2-\tau}},\lambda^\beta\right\}\log\frac{2}{\delta}+(2\eta^{-1}+1)c_\beta\lambda^\beta.\end{aligned}$$

Note $\lambda=m^{-\gamma}$ with (2.12) and $\|f_T\|_K\leq T^{\frac{1-\theta}{2}}$. Subtracting $\frac{1}{2}\left[\mathcal{E}(\hat{f}_T)-\mathcal{E}(f_\rho^\phi)\right]$ from both sides of the above inequality, then with confidence at least $1-2\delta$

$$\begin{aligned}\mathcal{E}(\hat{f}_T)-\mathcal{E}(f_\rho^\phi)\leq&2(C_2+C_4+(2\eta^{-1}+1)c_\beta)\lambda^\beta\\&\max\left\{\left(R_T^2\lambda^{1-\beta}\right)^{\frac{\zeta}{4-2\tau+\zeta\tau}},\left(R_T^2\lambda^{1-\beta}\right)^{\frac{\zeta}{2+\zeta}},\log\frac{\log T^{\frac{1-\theta}{2}}}{\delta}\right\}.\end{aligned}$$

By Theorem 6, we know that with confidence at least $1-\delta$,

$$\|f_T\|_K\leq c_2\lambda^{\frac{\beta-1}{2}}\left(\log\frac{\log T}{\delta}\right)^{\frac{1}{2}}\text{ with some universal constant }c_2>0.$$

Putting it into the above inequality yields that with confidence at least $1-3\delta$,

$$\mathcal{E}(\hat{f}_T)-\mathcal{E}(f_\rho^\phi)\leq 2(C_2+C_4+(2\eta^{-1}+1)c_\beta)c_2^{\frac{2\xi}{2+\xi}}\lambda^\beta\log\frac{\log T}{\delta}.$$

Scaling 3δ to δ, by $T=\lfloor m^\alpha\rfloor$, then with confidence at least $1-\delta$

$$\mathcal{E}(\hat{f}_T)-\mathcal{E}(f_\rho^\phi)\leq 2(C_2+C_4+(2\eta^{-1}+1)c_\beta)c_2^{\frac{2\xi}{2+\xi}}\lambda^\beta\log\frac{3\alpha\log m}{\delta}.$$

This finishes the proof of Theorem 1. ■

4. Generalization Error for Cross-validation

This section will derive the explicit error rate for cross-validation, stated in Theorem 4.

Proof of Theorem 4. For the least squares loss, it is easy to prove that for each $y\in\mathcal{Y}$, $\phi'_-(y,\cdot)$ is differentiable and $\phi'_-(y,\cdot)$ is Lipschitz continuous with constant $L=2$.

By (3.20), we have

$$\begin{aligned}\|f_{t+1}^{(\mathbf{D_1})}-f_\lambda\|_K^2\leq&\|f_t^{(\mathbf{D_1})}-f_\lambda\|_K^2+2\eta_t\mathcal{M}_{\mathbf{D_1}}(f_{t+1}^{(\mathbf{D_1})})\\&+2\eta_t\mathcal{F}_{\mathbf{D_1}}(f_\lambda)+2\eta_t\mathcal{A}(f_\lambda)-2\eta\left[\mathcal{E}(\hat{f}_{t+1}^{(\mathbf{D_1})})-\mathcal{E}(f_\rho)\right].\end{aligned}$$

Note that by (3.18), $\|f_{t+1}\|_K\leq\sqrt{2c_\phi\eta}t^{\frac{1}{2}}$ for any $t\in\mathbb{N}$. Meanwhile, for the least squares, (2.3) holds with $q=1$ and (2.4) is valid with $\tau=1$. Then by Lemmas 2, 3, we get with confidence at least $1-2\delta$

$$\|f_{t+1}^{(\mathbf{D_1})}-f_\lambda\|_K^2\leq\|f_t^{(\mathbf{D_1})}-f_\lambda\|_K^2+C_6\eta\max\left\{\left(\frac{R_T^\zeta}{n}\right)^{\frac{2}{2+\zeta}},\lambda^\beta\log\frac{\log T^{\frac{1}{2}}}{\delta}\right\},t=1,\cdots,T,$$

for a fixed $\lambda \in \mathbf{\Lambda}$, $T = \lfloor \lambda^{-1} \rfloor$. Here C_6 denotes a universal constant and

$$R_T = \max\{1, \|f_{t+1}^{(\mathbf{D_1})}\|_K, t = 1, \cdots, T\}.$$

Note that $T = \lfloor \lambda^{-1} \rfloor$ satisfies $t \leq T \leq \lambda^{-1}$. Scaling 2δ to δ, and applying the above inequality iteratively with $f_1 = 0$, we get that with confidence at least $1 - \delta$,

$$\begin{aligned}
\|f_{t+1}^{(\mathbf{D_1})} - f_\lambda\|_K^2 &\leq \|f_\lambda\|_K^2 + C_6 \eta t \max\left\{ \left(\frac{R_T^\zeta}{n}\right)^{\frac{2}{2+\zeta}}, \lambda^\beta \log\frac{\log T}{\delta} \right\} \\
&\leq \|f_\lambda\|_K^2 + C_6 \eta \max\left\{ \left(\frac{R_T^\zeta}{n}\right)^{\frac{2}{2+\zeta}} \lambda^{-1}, \lambda^{\beta-1} \log\frac{\log T}{\delta} \right\} \\
&\leq \|f_\lambda\|_K^2 + 2C_6 \eta \max\left\{ \left(\frac{R_T^\zeta}{m}\right)^{\frac{2}{2+\zeta}} \lambda^{-1}, \lambda^{\beta-1} \log\frac{\log m}{\delta} \right\}, t = 1, \cdots, T,
\end{aligned}$$

for any $\lambda \in \mathbf{\Lambda}$, $T = \lfloor \lambda^{-1} \rfloor$. The last inequality is obtained by $m = 2n$ and $T \leq m$ for any $T \in \mathbf{T}$.

Thus, by $\|f_{t+1}^{(\mathbf{D_1})}\|_K^2 \leq 2\|f_{t+1}^{(\mathbf{D_1})} - f_\lambda\|_K^2 + 2\|f_\lambda\|_K^2$ and $\|f_\lambda\|_K^2 \leq \mathcal{D}(\lambda)/\lambda \leq c_\beta \lambda^{\beta-1}$, we have that with confidence at least $1 - \delta$, there holds

$$\|f_{t+1}^{(\mathbf{D_1})}\|_K^2 \leq 4(C_6\eta + c_\beta) \max\left\{ \lambda^{-1} \left(\frac{R_T^\zeta}{m}\right)^{\frac{2}{2+\zeta}}, \lambda^{\beta-1} \left(\log\frac{\log m}{\delta}\right) \right\}, \ t = 1, \cdots, T, \quad (4.1)$$

for a fixed $\lambda \in \mathbf{\Lambda}$, $T \in \mathbf{T}$.

Then following the similar proof in Theorem 6, we have with confidence at least $1 - \delta$,

$$\|f_{t+1}^{(\mathbf{D_1})}\|_K \leq 2\sqrt{C_6\eta + c_\beta} \max\left\{ \lambda^{-\frac{2+\zeta}{4}} m^{-\frac{1}{2}}, \lambda^{\frac{\beta-1}{2}} \left(\log\frac{\log m}{\delta}\right)^{\frac{1}{2}} \right\} := R_\lambda \quad (4.2)$$

holds for each $t = 1, \cdots, T$.

By (3.23), we know that

$$\mathcal{E}(\hat{f}_T^{(\mathbf{D_1})}) - \mathcal{E}(f_\rho) \leq \mathcal{M}_{\mathbf{D_1}}(\hat{f}_T^{(\mathbf{D_1})}) + \mathcal{F}_{\mathbf{D_1}}(f_\lambda) + (2\eta^{-1} + 1)c_\beta \lambda^\beta.$$

Note that $\hat{f}_T^{(\mathbf{D_1})} \leq \sqrt{2c_\phi \eta} T^{\frac{1}{2}} \leq \sqrt{2c_\phi \eta} m^{\frac{1}{2}}$. Then we can use Lemma 2 with $R = \sqrt{2c_\phi \eta} m^{\frac{1}{2}}$ and $R_T \leq R_\lambda$ to estimate $\mathcal{M}_{\mathbf{D_1}}(\hat{f}_T^{(\mathbf{D_1})})$. This together with Lemma 3 and (4.2) yields with confidence at least $1 - 3\delta$,

$$\mathcal{E}(\hat{f}_T^{(\mathbf{D_1})}) - \mathcal{E}(f_\rho) \leq C_7 \max\left\{ \left(\frac{R_\lambda^\zeta}{m}\right)^{\frac{2}{2+\zeta}}, \lambda^\beta \log\frac{\log m}{\delta} \right\} \leq C_7 \left[\left(\frac{R_\lambda^\zeta}{m}\right)^{\frac{2}{2+\zeta}} + \lambda^\beta \log\frac{\log m}{\delta} \right] \quad (4.3)$$

where C_7 denotes a universal constant.

It implies that

$$\inf_{T_k\in\mathbf{T}}\mathcal{E}(\hat{f}_{T_k}^{(\mathbf{D_1})})-\mathcal{E}(f_\rho)\le C_7\inf_{\lambda_s\in\mathbf{\Lambda}}\left[\left(\frac{R_{\lambda_s}^{\zeta}}{m}\right)^{\frac{2}{2+\zeta}}+\lambda_s^{\beta}\log\frac{\log m}{\delta}\right] \tag{4.4}$$

holds with confidence at least $1-3|\mathbf{\Lambda}|\delta$.

Recall the definition of T^* in Definition 2. Note that (2.4) holds for $\tau=1$ and

$$\left\|\phi(y,\hat{f}_{T_k}^{(\mathbf{D_1})})(x)-\phi(y,\hat{f}_{T_k}^{(\mathbf{D_1})})(x)\right\|_\infty\le 4B^2,\ T_k\in\mathbf{T}.$$

Then by Theorem 7.2 in Steinwart and Christmann (2008) we find that for a fixed $\mathbf{D_1}$,

$$\mathcal{E}(\hat{f}_{T^*}^{(\mathbf{D_1})})-\mathcal{E}(f_\rho)\le 6\inf_{T_k\in\mathbf{T}}\left(\mathcal{E}(\hat{f}_{T_k}^{(\mathbf{D_1})})-\mathcal{E}(f_\rho)\right)+64(4B^2+1)m^{-1}\left(\log\frac{1+|\mathbf{T}|}{\delta}\right).$$

holds with confidence at least $1-\delta$. This together with (4.4) and $|\mathbf{T}|\le|\mathbf{\Lambda}|$ implies

$$\begin{aligned}\mathcal{E}(\hat{f}_{T^*}^{(\mathbf{D_1})})-\mathcal{E}(f_\rho)\le{}&\max\left\{6C_7,64(4B^2+1)\right\}\\&\left\{\inf_{\lambda_s\in\mathbf{\Lambda}}\left[\left(\frac{R_{\lambda_s}^{\zeta}}{m}\right)^{\frac{2}{2+\zeta}}+\lambda_s^{\beta}\log\frac{\log m}{\delta}\right]+m^{-1}\left(\log\frac{1+|\mathbf{\Lambda}|}{\delta}\right)\right\}\end{aligned} \tag{4.5}$$

holds with confidence at least $1-(3|\mathbf{\Lambda}|+1)\delta$.

Notice that $\left(\frac{R_\lambda^\zeta}{m}\right)^{\frac{2}{2+\zeta}}$ is continuous and non-increasing with respect to $\lambda\in\left[\frac{1}{m},1\right]$. So, there exists a $\lambda^*\in\left[\frac{1}{m},1\right]$ such that

$$\left(\frac{R_{\lambda^*}^{\zeta}}{m}\right)^{\frac{2}{2+\zeta}}+(\lambda^*)^{\beta}\log\frac{\log m}{\delta}=\min_{\lambda\in[\frac{1}{m},1]}\left[\left(\frac{R_\lambda^{\zeta}}{m}\right)^{\frac{2}{2+\zeta}}+\lambda^{\beta}\log\frac{\log m}{\delta}\right].$$

Write $\lambda_0:=\frac{1}{m}$ and let $\mathbf{\Lambda}$ be of the form

$$\mathbf{\Lambda}=\{\lambda_1,\cdots,\lambda_{|\mathbf{\Lambda}|}\},\ \text{with}\ \lambda_{s-1}<\lambda_s, s=2,\cdots,|\mathbf{\Lambda}|.$$

Since $\mathbf{\Lambda}\bigcup\{\lambda_0\}$ is the finite ε-net of $[\frac{1}{m},1]$, we have that

$$\lambda_s-\lambda_{s-1}\le 2\varepsilon,\ \text{for}\ s=1,\cdots,|\mathbf{\Lambda}|.$$

Then we can find an index $s^*\in\{1,\cdots,|\mathbf{\Lambda}|\}$ such that $\lambda_{s^*-1}\le\lambda^*\le\lambda_{s^*}$ and conclude that

$$\lambda^*\le\lambda_{s^*}\le\lambda^*+2\varepsilon.$$

By monotonicity and $\varepsilon \leq \frac{1}{m}$, we have that

$$\inf_{\lambda_s \in \mathbf{\Lambda}} \left[\left(\frac{R_\lambda^\zeta}{m} \right)^{\frac{2}{2+\zeta}} + \lambda_s^\beta \log \frac{\log m}{\delta} \right] \leq \left(\frac{R_{\lambda_{s^*}}^\zeta}{m} \right)^{\frac{2}{2+\zeta}} + (\lambda_{s^*})^\beta \log \frac{\log m}{\delta}$$
$$\leq \left(\frac{R_{\lambda^*}^\zeta}{m} \right)^{\frac{2}{2+\zeta}} + (\lambda^* + 2\varepsilon)^\beta \log \frac{\log m}{\delta} \leq \left(\frac{R_{\lambda^*}^\zeta}{m} \right)^{\frac{2}{2+\zeta}} + (\lambda^*)^\beta \log \frac{\log m}{\delta} + 2^\beta m^{-\beta} \log \frac{\log m}{\delta}$$
$$= \min_{\lambda \in [\frac{1}{m}, 1]} \left[\left(\frac{R_\lambda^\zeta}{m} \right)^{\frac{2}{2+\zeta}} + \lambda^\beta \log \frac{\log m}{\delta} \right] + 2^\beta m^{-\beta} \log \frac{\log m}{\delta}. \tag{4.6}$$

If $\frac{2}{\zeta+2\beta} \geq 1$, then $m^{-\frac{2}{\zeta+2\beta}} \leq m^{-1}$ and

$$\min_{\lambda \in [\frac{1}{m}, 1]} \left[\left(\frac{R_\lambda^\zeta}{m} \right)^{\frac{2}{2+\zeta}} + \lambda^\beta \log \frac{\log m}{\delta} \right] \leq \left[\left(\frac{R_\lambda^\zeta}{m} \right)^{\frac{2}{2+\zeta}} + \lambda^\beta \log \frac{\log m}{\delta} \right]_{\lambda = m^{-1}}$$
$$\leq \left(1 + (2\sqrt{C_6 \eta + c_\beta})^{\frac{2\zeta}{2+\zeta}} \right) m^{-\beta} \log \frac{\log m}{\delta}.$$

If $\frac{2}{\zeta+2\beta} \leq 1$, then $m^{-\frac{2}{\zeta+2\beta}} \in [m^{-1}, 1]$ and

$$\min_{\lambda \in [\frac{1}{m}, 1]} \left[\left(\frac{R_\lambda^\zeta}{m} \right)^{\frac{2}{2+\zeta}} + \lambda^\beta \log \frac{\log m}{\delta} \right] \leq \left[\left(\frac{R_\lambda^\zeta}{m} \right)^{\frac{2}{2+\zeta}} + \lambda^\beta \log \frac{\log m}{\delta} \right]_{\lambda = m^{-\frac{2}{\zeta+2\beta}}}$$
$$\leq \left(1 + (2\sqrt{C_6 \eta + c_\beta})^{\frac{2\zeta}{2+\zeta}} \right) m^{-\frac{2\beta}{2\beta+\zeta}} \log \frac{\log m}{\delta}.$$

So, we conclude that

$$\min_{\lambda \in [\frac{1}{m}, 1]} \left[\left(\frac{R_\lambda^\zeta}{m} \right)^{\frac{2}{2+\zeta}} + \lambda^\beta \log \frac{\log m}{\delta} \right] \leq \left(1 + (2\sqrt{C_6 \eta + c_\beta})^{\frac{2\zeta}{2+\zeta}} \right) m^{-\min\{1, \frac{2}{2\beta+\zeta}\}\beta} \log \frac{\log m}{\delta}.$$

This in connection with (4.5) and (4.6) yields

$$\mathcal{E}(\hat{f}_{T^*}^{(\mathbf{D_1})}) - \mathcal{E}(f_\rho) \leq 2^\beta \max\left\{6C_7, 64(4B^2+1)\right\} \left(1 + (2\sqrt{C_6 \eta + c_\beta})^{\frac{2\zeta}{2+\zeta}} \right)$$
$$\left\{ m^{-\min\{1, \frac{2}{2\beta+\zeta}\}\beta} \log \frac{\log m}{\delta} + m^{-1} \left(\log \frac{1+|\mathbf{\Lambda}|}{\delta} \right) \right\}$$
$$\lesssim m^{-\min\{1, \frac{2}{2\beta+\zeta}\}\beta} \log \frac{\log m}{\delta} \left(\log \frac{1+|\mathbf{\Lambda}|}{\delta} \right)$$

holds with confidence at least $1 - (3|\mathbf{\Lambda}| + 1)\delta$.

Then, we get the statement of Theorem 4 by scaling $(3|\mathbf{\Lambda}| + 1)\delta$ to δ.

■

5. Conclusion and Discussion

This paper considers the iterative regularization in the RKHSs for both classification and regression learning problems associated with a broad class of loss functions. We established the early stopping rules and derived corresponding generalization errors by only taking the projected last iterate as the final estimator. We show that these error bounds are comparable to the best obtained rates of Tikhonov regularization counterparts. Besides improving the existing results in Lin et al. (2016), our work provides a theoretical basis for the link between regularization paths of early stopping rules and and Tikhonov regularization. The novelty of our analysis is that utilizing the projection approach (or clipping idea) (Steinwart and Christmann, 2008; Gorbunov et al., 2020) provides a tight bound of the learning sequence $\{f_t\}_t$ for the sharpest possible convergence rates.

Iterative regularization with some specified losses have been shown to be closely related to spectral algorithms in inverse problems (see Remark 3 for details) or boosting-type algorithms. For the squared loss, it can be referred to as L^2-boosting and their consistency and convergence properties have been studied extensively (Blanchard and Krämer, 2016; Bühlmann and Yu, 2003; Yao et al., 2007; Caponnetto and Yao, 2010). In the work (Raskutti et al., 2014), the authors proposed a computable-from-data stopping rule that also produced mini-max optimal estimators for kernel classes. In contrast, our optimal stopping time is determined by some unknown parameters in practice. Furthermore, the cross-validation strategy by Definition 2 ensures comparable error rates, which has been proved in Theorem 4. Moreover, we note that by Example 1, our optimal stopping time depends only on ζ that can be derived directly from the smoothness of the kernel if the regression function f_ρ belongs to the known RKHS.

For other losses, the study on the regularization effect of stopping strategies is relatively fewer. We can refer to the papers (Bickel et al., 2006; Jiang, 2004; Bühlmann and Hothorn, 2007; Bartlett and Traskin, 2007; Zhang and Yu, 2005). In the work by Wei et al. (2019), the optimal stopping rules were established for a relatively broad class of loss functions in the context of kernel boosting algorithms. In the random design case their main theorem assumes that loss functions satisfy that

$$c_3\|f-g\|_{L^2_{\rho_{\mathcal{X}}}}^2 \le \mathcal{E}(f)-\mathcal{E}(g)-\langle\nabla\mathcal{E}(g), f-g\rangle \le c_4\|f-g\|_{L^2_{\rho_{\mathcal{X}}}}^2,\ \forall f,g\in\mathcal{B}(f_\phi^*,\sigma)$$

where f_ϕ^* is the minimizer of $\mathcal{E}(f)$ over the RKHS, $\mathcal{B}(f_\phi^*,\sigma)$ denotes a RKHS ball centred at f_ϕ^* with some specified radius $\sigma>0$, $c_3, c_4>0$ are universal constants. This assumption considers the least squares loss, logistic loss and exponential loss, but is hardly satisfied by some commonly used loss functions. For example, the hinge SVM loss $\phi(y,f)=\max\{1-yf\}_+$ for classification and the pinball loss for quantile regression (Steinwart and Christmann, 2008) are excluded, which can be addressed in our paper.

However, we should point out that our current analysis requires condition (2.10) for ϕ, which excludes loss functions of the form $V(yf)$ where $V(u)>0$ for all $u\in\mathbb{R}$, such as the logistic loss $V(yf)=\log(1+e^{-yf})$ for classification (Cucker and Zhou, 2007; Wu et al., 2007). Since the resulting analysis is more involved, we leave this as future research.

Acknowledgments

The work described in this paper is partially supported by National Natural Science Foundation of China (Project 12071356). The corresponding author is Yunwen Lei.

Appendix A. Some technical lemmas and proofs

Lemma 7 (Page 24, Lin et al. 2016) *Take $\eta_t = \eta t^{-\theta}$ with $\eta > 0$ and $0 < \theta < 1$. Then for $T \geq 2$,*

$$0 < \sum_{k=1}^{T-1} \frac{1}{k+1} \left[2\eta_{T-k} - \frac{1}{k} \sum_{t=T-k+1}^{T} 2\eta_t \right] \leq \frac{6\eta}{1-\theta} T^{-\theta}.$$

Lemma 8 (Bernstein inequality) *Let $\{\xi(z_i)\}_{i=1}^m$ be a set of random variables defined on $\mathcal{Z}$ and $\widetilde{M} > 0$ be the constant such that $\|\xi\|_\infty \leq \widetilde{M}$ and the variance $\sigma^2(\xi) < \infty$, then for any $\epsilon > 0$,*

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

Lemma 9 *Let $\{\xi(z_i)\}_{i=1}^m$ be a set of random variables defined on $\mathcal{Z}$ and $\widetilde{M}, c > 0, \tau \in [0,1]$ be the constant such that $\|\xi\|_\infty \leq \widetilde{M}$ and $\mathbb{E}(\xi)^2 \leq c(\mathbb{E}\xi)^\tau$, then for any $0 < \delta < 1$, with confidence at least $1-\delta$, there holds*

$$\frac{1}{m} \sum_{i=1}^{m} \xi(z_i) - \mathbb{E}(\xi) \leq 2 \log \frac{1}{\delta} \max \left\{ \frac{\widetilde{M}}{m}, \left(\frac{c}{m}\right)^{\frac{1}{2-\tau}}, \mathbb{E}\xi \right\}.$$

Proof Set the right hand of (A.1) as $\delta := \exp \left\{ -\frac{m\epsilon^2}{2(\sigma^2(\xi) + \frac{1}{3}\widetilde{M}\epsilon)} \right\}$. Solving it, we get that with confidence at least $1-\delta$, there holds

$$\begin{aligned} &\frac{1}{m} \sum_{i=1}^{m} \xi(z_i) - \mathbb{E}(\xi) \leq \frac{2\widetilde{M} \log \frac{1}{\delta}}{3m} + \sqrt{\frac{2 \log \frac{1}{\delta}}{m} \sigma^2(\xi)} \leq 2 \log \frac{1}{\delta} \max \left\{ \frac{\widetilde{M}}{m}, \frac{\sqrt{c}(\mathbb{E}\xi)^{\tau/2}}{\sqrt{m}} \right\} \\ &\leq 2 \log \frac{1}{\delta} \max \left\{ \frac{\widetilde{M}}{m}, \left[\left(\frac{c}{m}\right)^{\frac{1}{2-\tau}} \right]^{1-\frac{\tau}{2}} (\mathbb{E}\xi)^{\tau/2} \right\}. \end{aligned}$$

Applying the elementary inequality

$$x^{\tau/2} y^{1-\tau/2} \leq \frac{\tau}{2} x + (1 - \frac{\tau}{2}) y, \quad x, y \in \mathbb{R}, \tag{A.2}$$

then

$$\left[\left(\frac{c}{m}\right)^{\frac{1}{2-\tau}} \right]^{1-\frac{\tau}{2}} (\mathbb{E}\xi)^{\tau/2} \leq \left(1 - \frac{\tau}{2}\right) \left(\frac{c}{m}\right)^{\frac{1}{2-\tau}} + \frac{\tau}{2} \mathbb{E}\xi.$$

Thus, the desired conclusion holds. ■

Lemma 10 (Wu et al. 2007) *Let $\mathcal{G}$ be a set of measurable functions on $\mathcal{Z}$, and $M, c > 0$, $\tau \in [0,1]$ be constants such that each function $g \in \mathcal{G}$ satisfies $\|g\|_\infty \leq M$ and $\mathbb{E}(g^2) \leq c\,(\mathbb{E}g)^\tau$. If for some $a \geq M^\zeta$ and $\zeta \in (0,2)$,*

$$\mathbb{E}_{\mathbf{D}}\left[\log \mathcal{N}(\mathcal{G}, \epsilon, d_{2,\mathbf{D}})\right] \leq a\left(\frac{1}{\epsilon}\right)^\zeta, \quad \forall \epsilon > 0. \tag{A.3}$$

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

$$\mathbb{E}g - \frac{1}{m}\sum_{i=1}^m g(z_i) \leq \frac{1}{2}\eta^{1-\tau}\,(\mathbb{E}g)^\tau + c'_\zeta \eta + 2\left(\frac{cb}{m}\right)^{1/(2-\tau)} + \frac{18Mb}{m}, \quad \forall g \in \mathcal{G},$$

where

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

Proof of Lemma 3. Decompose $\mathcal{F}_{\mathbf{D}}(f^*)$ into

$$\begin{aligned}\mathcal{F}_{\mathbf{D}}(f^*) = &\left\{\left[\mathcal{E}_{\mathbf{D}}(f^*) - \mathcal{E}_{\mathbf{D}}(\hat{f}^*)\right] - \left[\mathcal{E}(f^*) - \mathcal{E}(\hat{f}^*)\right]\right\} \\ &+ \left\{\left[\mathcal{E}_{\mathbf{D}}(\hat{f}^*) - \mathcal{E}_{\mathbf{D}}(f_\rho^\phi)\right] - \left[\mathcal{E}(\hat{f}^*) - \mathcal{E}(f_\rho^\phi)\right]\right\} := S_1 + S_2.\end{aligned}$$

Firstly, we consider S_1. Let $\xi_1(z) = \phi(y, f^*(x)) - \phi(y, \hat{f}^*(x))$. Notice that $\xi_1(z) \geq 0$, then

$$\|\xi_1\|_\infty \leq \sup_{(x,y)\in\mathcal{Z}} |\phi(y, f^*(x))| \leq c_q\left(1 + \kappa^q \|f^*\|_K^q\right)\kappa\|f^*\|_K \leq c_q\kappa(1+\kappa^q)\tilde{R}^{q+1},$$

and

$$\mathbb{E}\xi_1^2 \leq \|\xi_1\|_\infty \mathbb{E}\xi_1 \leq c_q\kappa\,(1+\kappa^q)\,\tilde{R}^{q+1}\mathbb{E}\xi_1.$$

Applying Lemma 9 with $\xi = \xi_1$, $c = c_q\kappa\,(1+\kappa^q)\,\tilde{R}^{q+1}$, $\widetilde{M} = c_q\kappa\,(1+\kappa^q)\,\tilde{R}^{q+1}$ and $\tau = 1$, we know that there exists a subset $\mathcal{Z}_{\delta,1}^m \subset \mathcal{Z}^m$ with the measure at least $1-\delta$ such that for arbitrary $\mathbf{D} \in \mathcal{Z}_{\delta,1}^m$ we have

$$S_1 \leq 2c_q\kappa\,(1+\kappa^q)\log\frac{1}{\delta}\max\left\{\frac{\widetilde{R}^{q+1}}{m}, \mathbb{E}\xi_1\right\}. \tag{A.4}$$

Next, we estimate S_2. Let $\xi_2(z) = \phi(y, \hat{f}^*(x)) - \phi(y, f_\rho^\phi(x))$. Assumptions 1 and 2 imply that

$$\|\xi_2\|_\infty \leq c_q\kappa(1+\kappa^q)(B^{q+1} + \|f_\rho^\phi\|_K^{q+1}) := M_q$$

and $\mathbb{E}\xi_2^2 \leq c_\tau(\mathbb{E}\xi_2)^\tau$. Applying Lemma 9 with $\xi = \xi_2$ again, we know that there exists a subset $\mathcal{Z}_{\delta,2}^m \subset \mathcal{Z}^m$ with measure at least $1-\delta$ such that arbitrary $\mathbf{D} \in \mathcal{Z}_{\delta,2}^m$, there holds

$$S_2 \leq 2(M_q + c_\tau)\log\frac{1}{\delta}\max\left\{\left(\frac{1}{m}\right)^{\frac{1}{2-\tau}}, \mathbb{E}\xi_2\right\}. \tag{A.5}$$

Noticing the fact $\mathbb{E}\xi_1 \leq \mathcal{A}(f^*)$ and $\mathbb{E}\xi_2 \leq \mathcal{A}(f^*)$, by (A.4) and (A.5), then

$$\mathcal{F}_{\mathbf{D}}(f^*) = S_1 + S_2 \leq (2M_q + 2c_\tau + 2c_q\kappa\,(1+\kappa^q))\log\frac{1}{\delta}\max\left\{\frac{\widetilde{R}^{q+1}}{m}, \mathcal{A}(f^*)\right\}$$

holds with measure at least $1-2\delta$. Scaling 2δ to δ, then proof of (3.7) is finished with

$$C_3 = 2M_q + 2c_\tau + 2c_q\kappa\,(1+\kappa^q)\,.$$

For the estimate of (3.8), let $f^* = f_\lambda$, then by Assumption 3,

$$\mathcal{A}(f_\lambda) \leq \mathcal{D}(f_\lambda) \leq c_\beta\lambda^\beta,\ \|f_\lambda\|_K \leq c_\beta^{\frac{1}{2}}\lambda^{\frac{\beta-1}{2}}.$$

Applying (3.7) with $\widetilde{R} = c_\beta^{\frac{1}{2}}\lambda^{\frac{\beta-1}{2}}$, then the proof of (3.8) is finished with

$$C_4 = (2M_q + 2c_\tau + 2c_q\kappa\,(1+\kappa^q))\max\{c_\beta^{\frac{q+1}{2}}, c_\beta\}.$$

■

To prove Lemma 2, we need the following Lemma.

Lemma 11 *Suppose Assumptions 1, 2 and 4 hold. For a fixed $f \in \mathcal{B}_R$ with $R \geq 1$, then with confidence at least $1-\delta$,*

$$\mathcal{M}_{\mathbf{D}}(f) \leq \frac{1}{2}\left[\mathcal{E}(\hat{f}) - \mathcal{E}(f_\rho^\phi)\right] + C\max\left\{\left(\frac{R^\zeta}{m}\right)^{\frac{2}{4-2\tau+\zeta\tau}}, \left(\frac{R^\zeta}{m}\right)^{\frac{2}{2+\zeta}}, \left(\frac{1}{m}\right)^{\frac{1}{2-\tau}}\log\frac{1}{\delta}\right\} \quad \text{(A.6)}$$

and

$$\mathcal{E}_{\mathbf{D}}(f_\rho^\phi) - \mathcal{E}_{\mathbf{D}}(\hat{f}) \leq C\max\left\{\left(\frac{R^\zeta}{m}\right)^{\frac{2}{4-2\tau+\zeta\tau}}, \left(\frac{R^\zeta}{m}\right)^{\frac{2}{2+\zeta}}, \left(\frac{1}{m}\right)^{\frac{1}{2-\tau}}\log\frac{1}{\delta}\right\} \quad \text{(A.7)}$$

where C is a universal constant (depending on q, τ, ζ) and will be given in the proof.

Proof We apply Lemma 10 to the function set

$$\mathcal{G} = \left\{g(x,y) := \phi(y, \hat{f}(x)) - \phi(y, f_\rho^\phi(x)), f \in \mathcal{B}_R\right\}.$$

Assumption 1 implies that

$$\|g\|_\infty \leq M := c_q\kappa(1+\kappa^q)(B^{q+1} + \|f_\rho^\phi\|_K^{q+1}).$$

Assumption 2 tells us that with $c = c_\tau$, each $g \in \mathcal{G}$ satisfies that $\mathbb{E}(g^2) \leq c\,(\mathbb{E}g)^\tau$ by the fact $|\hat{f}(x)| \leq B$. Since

$$|\phi(y, \hat{f}(x)) - \phi(y, \hat{h}(x))| \leq \max_{\delta\in[-B,B]}|\phi'_-(y,\delta)||\hat{f}(x)) - \hat{h}(x)| \leq c_q(1+B)|f(x) - h(x)|,$$

then

$$\mathcal{N}(\mathcal{G},\epsilon,d_{2,\mathbf{D}}) \le \mathcal{N}(\mathcal{B}_R,\frac{\epsilon}{c_q(1+B)},d_{2,\mathbf{D}}) \le \mathcal{N}(\mathcal{B}_1,\frac{\epsilon}{c_q(1+B)R},d_{2,\mathbf{D}}).$$

Hence, Assumption 4 yields the covering number condition with $a = c_\zeta(c_q(1+B)R)^\zeta$. Thus all the conditions in Lemma 10 hold. We have that with confidence at least $1-\delta$, for all $g \in \mathcal{G}$, there holds

$$\mathbb{E}g - \frac{1}{m}\sum_{i=1}^m g(z_i) \le \frac{1}{2}\eta^{1-\tau}\left(\mathbb{E}g\right)^\tau + c'_\zeta\eta + 2\left(\frac{c_\tau \log\frac{1}{\delta}}{m}\right)^{1/(2-\tau)} + \frac{18M}{m}\log\frac{1}{\delta}$$

where

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

and

$$C' = (c_\tau)^{\frac{2-\zeta}{4-2\tau+\zeta\tau}}\left(c_\zeta(c_q(1+B))^\zeta\right)^{\frac{2}{4-2\tau+\zeta\tau}} + M^{\frac{2-\zeta}{2+\zeta}}\left(c_\zeta(c_q(1+B))^\zeta\right)^{\frac{2}{2+\zeta}}.$$

Applying the elementary inequality (A.2) with $x = \mathbb{E}g$ and $y = \eta$, we get that with confidence at last $1-\delta$, there holds for $f \in B_R$,

$$\begin{aligned}&\left[\mathcal{E}(\hat{f}) - \mathcal{E}(f_\rho^\phi)\right] - \left[\mathcal{E}_{\mathbf{D}}(\hat{f}) - \mathcal{E}_{\mathbf{D}}(f_\rho^\phi)\right]\\ &\le \left(\frac{1}{2}+c\zeta'\right)\eta + \frac{1}{2}\left[\mathcal{E}(\hat{f}) - \mathcal{E}(f_\rho^\phi)\right] + 2\left(\frac{c_\tau\log\frac{1}{\delta}}{m}\right)^{1/(2-\tau)} + \frac{18M}{m}\log\frac{1}{\delta}.\end{aligned}$$

It yields the desired conclusion (A.6) with

$$C = \left(\frac{1}{2}+c'_\zeta\right)C' + 2c_\tau^{\frac{1}{2-\tau}} + 18M.$$

For (A.7),

$$\begin{aligned}\mathcal{E}_{\mathbf{D}}(f_\rho^\phi) - \mathcal{E}_{\mathbf{D}}(\hat{f}) &= \left[\mathcal{E}(\hat{f}) - \mathcal{E}(f_\rho^\phi)\right] - \left[\mathcal{E}_{\mathbf{D}}(\hat{f}) - \mathcal{E}_{\mathbf{D}}(f_\rho^\phi)\right] - \left[\mathcal{E}(\hat{f}) - \mathcal{E}(f_\rho^\phi)\right]\\ &= \mathcal{M}_{\mathbf{D}}(f) - \left[\mathcal{E}(\hat{f}) - \mathcal{E}(f_\rho^\phi)\right].\end{aligned}$$

By $\mathcal{E}(\hat{f}) - \mathcal{E}(f_\rho^\phi) > 0$ and (A.6), we get the conclusion (A.7). ■

Proof of Lemma 2. Let $r_j = R\cdot 2^{-j}$ for $j = 0,1,\ldots,J$, where $J = \lfloor \log_2 R\rfloor + 1$. It is clear that $r_J \le 1$. For any $j \in \{0,\ldots,J-1\}$, define

$$B_j = \left\{f \in \mathcal{H}_K : r_{j+1} < \|f\|_K \le r_j\right\}.$$

and $B_J = \{f \in \mathcal{H}_K : \|f\|_K \leq r_J\}$.

For any $j \in \{0, 1, \ldots, J\}$, according to Lemma 11, with probability at least $1-\delta/(J+1)$ the following inequality holds for all $f \in B_j$

$$\left[\mathcal{E}(\hat{f}) - \mathcal{E}(f_\rho^\phi)\right] - \left[\mathcal{E}_{\mathbf{D}}(\hat{f}) - \mathcal{E}_{\mathbf{D}}(f_\rho^\phi)\right] \leq \frac{1}{2}\left[\mathcal{E}(\hat{f}) - \mathcal{E}(f_\rho^\phi)\right] + C \max\left\{\left(\frac{r_j^\zeta}{m}\right)^{\frac{2}{4-2\tau+\zeta\tau}}, \left(\frac{r_j^\zeta}{m}\right)^{\frac{2}{2+\zeta}}, \left(\frac{1}{m}\right)^{\frac{1}{2-\tau}} \log\frac{1}{\delta_J}\right\},$$

$$\mathcal{E}_{\mathbf{D}}(f_\rho^\phi) - \mathcal{E}_{\mathbf{D}}(\hat{f}) \leq C \max\left\{\left(\frac{r_j^\zeta}{m}\right)^{\frac{2}{4-2\tau+\zeta\tau}}, \left(\frac{r_j^\zeta}{m}\right)^{\frac{2}{2+\zeta}}, \left(\frac{1}{m}\right)^{\frac{1}{2-\tau}} \log\frac{1}{\delta_J}\right\},$$

where we introduce $\delta_J = \delta/(J+1)$.

It is clear that if $f \in B_j$ for $j < J$, then $r_j \leq 2\|f\|_K$. If $f \in B_J$, then $\|f\|_K \leq 1$. Therefore we have if $f \in B_j$,

$$r_j \leq \max\left\{1, 2\|f\|_K\right\}, \quad \forall j = 1, 2, \ldots, J.$$

Note that $\mathcal{B}_R = \bigcup_{j=0}^J B_j$. By the union of probability, with confidence at least $1-\delta$ the following holds for all $f \in \mathcal{B}_R$

$$\left[\mathcal{E}(\hat{f}) - \mathcal{E}(f_\rho^\phi)\right] - \left[\mathcal{E}_{\mathbf{D}}(\hat{f}) - \mathcal{E}_{\mathbf{D}}(f_\rho^\phi)\right] \leq \frac{1}{2}\left[\mathcal{E}(\hat{f}) - \mathcal{E}(f_\rho^\phi)\right] + C \max\left\{\left(\frac{\max\left\{1, (2\|f\|_K)^\zeta\right\}}{m}\right)^{\frac{2}{4-2\tau+\zeta\tau}}, \left(\frac{\max\left\{1, (2\|f\|_K)^\zeta\right\}}{m}\right)^{\frac{2}{2+\zeta}}, \left(\frac{1}{m}\right)^{\frac{1}{2-\tau}} \log\frac{1}{\delta_J}\right\},$$

$$\mathcal{E}_{\mathbf{D}}(f_\rho^\phi) - \mathcal{E}_{\mathbf{D}}(\hat{f}) \leq C \max\left\{\left(\frac{\max\left\{1, (2\|f\|_K)^\zeta\right\}}{m}\right)^{\frac{2}{4-2\tau+\zeta\tau}}, \left(\frac{\max\left\{1, (2\|f\|_K)^\zeta\right\}}{m}\right)^{\frac{2}{2+\zeta}}, \left(\frac{1}{m}\right)^{\frac{1}{2-\tau}} \log\frac{1}{\delta_J}\right\}.$$

The stated bound then follows directly by taking $C_2 = 2^{\frac{2\zeta}{2+\zeta}} C$. The proof is completed. ■

References

P. L. Bartlett. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of network. *IEEE Transactions on Information Theory*, 1998.

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

P. L. Bartlett, M. I. Jordan, and J. D. Mcauliffe. Convexity, classification, and risk bounds. *Journal of the American statistical association*, 101(473):138–156, 2006.

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

M. Birman and M. Z. Solomjak. Piecewise-polynomial approximations of functions of the classes W_p^α. *Mathematics of the USSR-Sbornik*, 2(3):295–317, 1967.

G. Blanchard and N. Krämer. Convergence rates of kernel conjugate gradient for random design regression. *Analysis and Applications*, 14(06):763–794, 2016.

G. Blanchard and P. Mathé. Conjugate gradient regularization under general smoothness and noise assumptions. *Journal of Inverse and Ill-Posed Problems*, 18(6):701–726, 2010.

O. Bousquet and L. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 02:499–526, 2002.

P. Bühlmann and T. Hothorn. Boosting algorithms: Regularization, prediction and model fitting. *Statictical Science*, pages 477–505, 2007.

P. Bühlmann and B. Yu. Boosting with the L-2 loss. *Journal of the American Statistical Association*, 462:324–339, 2003.

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

A. Caponnetto and Y. Yao. Cross-validation based adaptation for regularization operators in learning theory. *Analysis and Applications*, 8(2):161–183, 2010.

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

J. Fan, T. Hu, Q. Wu, and D. X. Zhou. Consistency analysis of an empirical minimum error entropy algorithm. *Applied and computational Harmonic Analysis*, 41:164–189, 2016.

J. H. Friedman and B. Popescu. Gradient directed regularization. *Technical report, Stanford University,*, 2004.

E. Gorbunov, M. Danilova, and A. Gasnikov. Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. In *Advances in Neural Information Processing Systems 33*, 2020.

Z. C. Guo, S. B. Lin, and D. X. Zhou. Learning theory of distributed spectral algorithms. *Inverse Problems*, 33(7):074009, 2017.

Z. C. Guo, T. Hu, and L. Shi. Gradient descent for robust kernel-based regression. *Inverse Problems*, 34:29pp, 2018.

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

Y. A. LeCun, L. Bottou, G. B. Orr, and K. R. Müller. Efficient backprop. In *Neural networks: Tricks of the trade*, pages 9–48. Springer, 2012.

Y. W. Lei, T. Hu, and K. Tang. Generalization performance of multi-pass stochastic gradient descent with convex loss functions. *Journal of Machine Learning Research*, 25:1–41, 2021.

H. Lian, K. Zhao, and S. Lv. Projected spline estimation of the nonparametric function in high-dimensional partially linear models for massive data. *The Annals of Statistics*, 47 (5):2922–2949, 2019.

J. Lin and L. Rosasco. Optimal learning for multi-pass stochastic gradient methods. *Journal of Machine Learning Research*, 18:1–47, 2017.

J. Lin, L. Rosasco, and D. X. Zhou. Iterative regularization for learning with convex loss functions. *Journal of Machine Learning Research*, 17:1–38, 2016.

S. B. Lin and D. X. Zhou. Optimal learning rates for kernel partial least squares. *Journal of Fourier Analysis and Applications*, pages 1–26, 2017.

S. Lu and S. V. Pereverzev. *Regularization theory for ill-posed problems. Selected topics.* Walter de Gruyter GmbH, 2013.

S. Lv, H. Lin, H. Lian, and J. Huang. Oracle inequalities for sparse additive quantile regression in reproducing kernel Hilbert space. *The Annals of Statistics*, 46(2):781–813, 2018.

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. In *Neural Information Processing Systems*, pages 1621–1629, 2015.

B. Stankewitz, N. Mücke, and L. Rosasco. From inexact optimization to learning via gradient concentration. *arXiv:2106.05397v3*, 2021.

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

I. Steinwart, D. R. Hush, and C. Scovel. Optimal rates for regularized least squares regression. In *COLT 2009 - The 22nd Conference on Learning Theory, Montreal, Quebec, Canada, June 18-21, 2009*, 2009.

C. J. Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10(4):1040–1053, 1982.

O. N. Strand. Theory and methods related to the singular. *SIAM Journal on Numerical Analysis*, 11, 1974.

H. Sun and Q. Wu. Optimal rates of distributed regression with imperfect kernels. *Journal of Machine Learning Research*, 171:1–34, 2021.

V. Temlyakov. Optimal estimators in learning theory. *Banach Center Publications, Inst. Math. Polish Academy of Sciences*, 72:341–366, 2006.

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

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. De Giovannini, F. Odone, and P. Bartlett. Learning from examples as an inverse problem. *Journal of Machine Learning Research*, 6(6):883–904, 2005.

B. B. Wang and T. Hu. Unregularized online algorithms with varying Gaussians. *Constructive Approximation*, 53:403–440, 2021.

Y. T. Wei, F. Yang, and M. J. Wainwright. Early stopping for kernel boosting algorithms: A general analysis with localized complexities. *IEEE Transactions on Information Theory*, 65(10):6685–6703, 2019.

Q. Wu and D. X. Zhou. SVM soft margin classifiers: Linear programming versus quadratic programming. *Neural computation*, 17(5):1160–1187, 2005.

Q. Wu, Y. Ying, and D. X. Zhou. Multi-kernel regularized classifiers. *Journal of Complexity*, 23(1):108–134, 2007.

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

T. Zhang and B. Yu. Boosting with early stopping: Convergence and consistency. *Annal of Statistics*, 33(4):1538–1579, 2005.

D. X. Zhou. The covering number in learning theory. *Journal of Complexity*, 18(3):739–767, 2002.