Available online at www.sciencedirect.com
ScienceDirect
Appl. Comput. Harmon. Anal. 23 (2007) 198–214
Applied and Computational Harmonic Analysis
www.elsevier.com/locate/acha

Fully online classification by regularization ☆

Gui-Bo Ye [a], Ding-Xuan Zhou [b,*]

[a] *School of Mathematical Sciences, Fudan University, Shanghai 200433, PR China*
[b] *Department of Mathematics, City University of Hong Kong, Kowloon, Hong Kong, China*

Received 22 March 2006; revised 15 December 2006; accepted 21 December 2006

Available online 10 January 2007

Communicated by Charles K. Chui

Abstract

In this paper we consider fully online learning algorithms for classification generated from Tikhonov regularization schemes associated with general convex loss functions and reproducing kernel Hilbert spaces. For such a fully online algorithm, the regularization parameter in each learning step changes. This is the essential difference from the partially online algorithm which uses a fixed regularization parameter. We first present a novel approach to the drift error incurred by the change of the regularization parameter. Then we estimate the error of the learning process for the strong approximation in the reproducing kernel Hilbert space. Finally, learning rates are derived from decays of the regularization error. The convexity of the loss function plays an important role in our analysis. Concrete learning rates are given for the hinge loss and the support vector machine q-norm loss.

Keywords: Classification algorithm; Online learning; Reproducing kernel Hilbert spaces; Regularization; Error analysis

1. Introduction

This paper aims at fully online binary classification algorithms generated from Tikhonov regularization schemes associated with general convex loss functions and reproducing kernel Hilbert spaces.

A binary classification algorithm produces a binary *classifier* $\mathcal{C} : X \to Y$ which divides the input space X (a metric space such as a subset of $\mathbb{R}^n$) into two classes represented by $Y = \{1, -1\}$. The classifier $\mathcal{C}$ makes a prediction $y \in Y$ for each point $x \in X$ (a vector $x \in \mathbb{R}^n$ with n components corresponding to n practical measurements). A real valued function $f : X \to \mathbb{R}$ can be used to generate a classifier $\mathcal{C}(x) = \operatorname{sgn}(f(x))$ where

$$\operatorname{sgn}\big(f(x)\big) = \begin{cases} 1, & \text{if } f(x) \geqslant 0, \\ -1, & \text{if } f(x) < 0. \end{cases}$$

☆ Supported partially by the Research Grants Council of Hong Kong (Project No. CityU 103405), City University of Hong Kong (Project No. 7001816), National Science Fund for Distinguished Young Scholars of China (Project No. 10529101), and National Basic Research Program of China (Project No. 973-2006CB303102).

* Corresponding author.
E-mail addresses: yeguibo@hotmail.com (G.-B. Ye), mazhou@cityu.edu.hk (D.-X. Zhou).

1063-5203/$ – see front matter
doi:10.1016/j.acha.2006.12.001

A *loss function* $\phi:\mathbb{R}\to\mathbb{R}_+$ is often used for the real valued function f, to measure the local error $\phi(yf(x))$ suffered from the use of $\operatorname{sgn}(f)$ as a model for the process producing y at $x\in X$.

A *Mercer kernel* $K:X\times X\to\mathbb{R}$ is a continuous and symmetric function which is positive semidefinite, i.e., for any finite set of points $\{x_1,\dots,x_\ell\}\subset X$, the matrix $(K(x_i,x_j))_{i,j=1}^{\ell}$ is positive semidefinite. The *reproducing kernel Hilbert space* (RKHS) $\mathcal{H}_K$ associated with the kernel K is defined [1] to be the completion of the linear span of the set of functions $\{K_x=K(x,\cdot)\colon x\in X\}$ with the inner product $\langle\cdot,\cdot\rangle_K$ given by $\langle K_x,K_y\rangle_K=K(x,y)$. Its reproducing property plays a special role in learning theory:

$$\langle K_x,f\rangle_K=f(x),\quad x\in X,\ f\in\mathcal{H}_K. \tag{1.1}$$

We consider classification algorithms induced by regularization schemes learned from samples. Assume that ρ is a probability distribution on $Z=X\times Y$ and $\mathbf{z}=\{z_t=(x_t,y_t)\}_{t=1}^{T}\in Z^T$ is a set of random samples independently drawn according to ρ. The batch learning algorithm for classification produces a classifier $\operatorname{sgn}(f_{\mathbf{z},\lambda})$ by implementing an off-line regularization scheme in $\mathcal{H}_K$ involving the sample $\mathbf{z}$, $\lambda>0$ and the loss function ϕ as

$$f_{\mathbf{z},\lambda}=\arg\min_{f\in\mathcal{H}_K}\left\{\frac{1}{T}\sum_{t=1}^{T}\phi\big(y_t f(x_t)\big)+\frac{\lambda}{2}\|f\|_K^2\right\}. \tag{1.2}$$

This off-line classification algorithm has been extensively studied in the literature. In particular, the error analysis is well done; see, e.g., [8,16,21,22,25,27]. The main idea is to show that $f_{\mathbf{z},\lambda}$ has behaviors similar to the *regularizing function* $f_\lambda^\phi\in\mathcal{H}_K$ of scheme (1.2) defined by

$$f_\lambda^\phi=\arg\inf_{f\in\mathcal{H}_K}\left\{\mathcal{E}(f)+\frac{\lambda}{2}\|f\|_K^2\right\}. \tag{1.3}$$

Here $\mathcal{E}(f)$ is the *generalization error* defined as

$$\mathcal{E}(f)=\int_Z\phi\big(yf(x)\big)\,\mathrm{d}\rho.$$

Though the off-line algorithm (1.2) performs well in theory and in many applications, it might be practically challenging when the sample size T or data is very large. For example, if $\phi(x)=(1-x)_+=\max\{1-x,0\}$ or $(1-x)_+^2$ corresponding to the support vector machines (SVM), the scheme (1.2) is a quadratic optimization problem. Its standard complexity is about $O(T^3)$.

When the sample size is large, online learning algorithms with linear complexity $O(T)$ can be applied and provide efficient classifiers. These algorithms are generalizations of the perceptron which has a long history, see, e.g., [4,15].

Here we study a family of online learning algorithms associated with a general convex loss function. We assume throughout the paper that the loss function has the following form.

Definition 1. We say that $\phi:\mathbb{R}\to\mathbb{R}_+$ is an *admissible loss function* if it is convex and differentiable at 0 with $\phi'(0)<0$.

The convexity of ϕ tells us that the left derivative $\phi'_-(x)=\lim_{\delta\to0-}(\phi(x+\delta)-\phi(x))/\delta$ exists. In this paper we study the following (stochastic gradient descent) online algorithm for classification given in [4,10,17,26].

Definition 2. The *fully online algorithm* for classification is defined by $f_1=0$ and

$$f_{t+1}=f_t-\eta_t\big\{\phi'_-\big(y_t f_t(x_t)\big)y_t K_{x_t}+\lambda_t f_t\big\}\quad\text{for } t=1,\dots,T, \tag{1.4}$$

where $\lambda_t>0$ is called the *regularization parameter* and $\eta_t>0$ the *step size*. The classifier is given by the sign function $\operatorname{sgn}(f_{T+1})$.

In this fully online algorithm, the regularization parameter λ_t changes with the learning step t. Throughout the paper we assume that $\lambda_{t+1}\leqslant\lambda_t$ for each $t\in\mathbb{N}$. When the regularization parameter $\lambda_t\equiv\lambda_1$ is independent of the step t, we call the scheme (1.4) *partially online*.

The main purpose of this paper is to study the role of the regularization parameter in the fully online algorithm. A usual form is $\lambda_t = \lambda_1 t^{-\gamma}$ for some $\gamma > 0$.

The prediction power of classification algorithms are often measured by the *misclassification error* which is defined for a classifier $\mathcal{C}: X \to Y$ to be the probability of the event $\{\mathcal{C}(x) \neq y\}$:

$$\mathcal{R}(\mathcal{C}) = \operatorname{Prob}\big\{\mathcal{C}(x) \neq y\big\} = \int_X P\big(y \neq \mathcal{C}(x) \mid x\big)\,\mathrm{d}\rho_X. \tag{1.5}$$

Here ρ_X denotes the marginal distribution of ρ on X, and $P(\cdot \mid x)$ the conditional probability measure. The best classifier minimizing the misclassification error is called the *Bayes rule* (e.g., [7]) and can be expressed as $f_c = \operatorname{sgn}(f_\rho)$, where f_ρ is the *regression function*

$$f_\rho(x) = \int_Y y\,\mathrm{d}\rho(y \mid x) = P(y = 1 \mid x) - P(y = -1 \mid x), \quad x \in X. \tag{1.6}$$

Recall that for the online learning algorithm (1.4), we are interested in the classifier $\operatorname{sgn}(f_{T+1})$ produced by the real valued function f_{T+1} from $\mathbf{z} = \{z_t\}_{t=1}^T$. So the error analysis for the online classification algorithm (1.4) is aimed at the *excess misclassification error*

$$\mathcal{R}\big(\operatorname{sgn}(f_{T+1})\big) - \mathcal{R}(f_c). \tag{1.7}$$

To illustrate the special role played by the varying regularization parameter $\{\lambda_t\}$ in the fully online algorithm (1.4), we state a result, proved in Section 6, for the *hinge loss* $\phi(x) = (1-x)_+$. For this loss, the online algorithm (1.4) can be expressed as $f_1 = 0$ and

$$f_{t+1} = \begin{cases} (1-\eta_t\lambda_t) f_t, & \text{if } y_t f_t(x_t) > 1, \\ (1-\eta_t\lambda_t) f_t + \eta_t y_t K_{x_t}, & \text{if } y_t f_t(x_t) \leqslant 1. \end{cases} \tag{1.8}$$

Example 1. Let $\kappa := \sup_{x\in X}\sqrt{K(x,x)}$, $\phi(x) = (1-x)_+$ and for some $\lambda_1 > 0$, $0 < \eta_1 \leqslant \frac{1}{2\kappa^2+\lambda_1}$, $0 < \epsilon < \frac{1}{4}$, the parameters $\{\lambda_t, \eta_t\}$ take the form

$$\lambda_t = \lambda_1 t^{-\frac{1}{4}}, \quad \eta_t = \eta_1 t^{\epsilon-\frac{1}{2}} \quad \forall t \in \mathbb{N}. \tag{1.9}$$

If for some $0 < \beta \leqslant 1$ and $\mathcal{D}_0 > 0$, the pair (ρ, K) satisfies

$$\inf_{f\in\mathcal{H}_K}\left\{\|f - f_c\|_{L^1_{\rho_X}} + \frac{\lambda}{2}\|f\|_K^2\right\} \leqslant \mathcal{D}_0\lambda^\beta \quad \forall \lambda > 0, \tag{1.10}$$

then

$$\mathbb{E}_{z_1,\dots,z_T}\big(\mathcal{R}\big(\operatorname{sgn}(f_{T+1})\big) - \mathcal{R}(f_c)\big) \leqslant C_\epsilon T^{-\min\{\frac{\beta}{4}, \frac{1}{8}-\frac{\epsilon}{2}\}}, \tag{1.11}$$

where $C_\epsilon = C_{\epsilon,\eta_1,\lambda_1,\kappa,\mathcal{D}_0,\beta}$ is a constant depending on ϵ, η_1, λ_1, κ, $\mathcal{D}_0$ and β.

The condition (1.10) concerns the approximation of the function f_c in the L^1 space $L^1_{\rho_X}$ by functions from the RKHS $\mathcal{H}_K$. It can be characterized by requiring f_c to lie in an interpolation space of the pair $(L^1_{\rho_X}, \mathcal{H}_K)$, an intermediate space between the metric space $L^1_{\rho_X}$ and the much smaller approximating space $\mathcal{H}_K$. For details, see the discussion in [3].

Assumptions like (1.10) are necessary to determine the regularization parameter for achieving the learning rate (1.11). This can be seen from the literature [16,25,27] of the off-line algorithm (1.2): learning rates are obtained by suitable choices of the regularization parameter $\lambda = \lambda(T)$, according to the behavior of the approximation error estimated from a priori conditions on the distribution ρ and the space $\mathcal{H}_K$.

2. Main results

In this paper we investigate fully online algorithm (1.4) in the sense that the regularization parameter λ_t depends on the step t. This makes the regularizing function $f_\lambda^\phi = f_{\lambda_t}^\phi$ change with the step t.

2.1. Bounds for the drift error

Our first main result bounds the difference of the regularizing function f_λ^ϕ for different regularization parameters. Denote $f_{\lambda_0}^\phi = 0$.

Definition 3. The *drift error* associated with the pair (K,ϕ) and the regularization parameter sequence $\{\lambda_t\}_{t\in\mathbb{N}}$ is defined by means of the function f_λ^ϕ in (1.3) as

$$d_t := \big\| f_{\lambda_t}^\phi - f_{\lambda_{t-1}}^\phi \big\|_K, \quad t \in \mathbb{N}.$$

The drift error is an approximation-type error and does not depend on the sample drawn. To state our bound, we need the regularization error [3] or approximation error [18,19] defined as follows.

Definition 4. The *regularization error* $\mathcal{D}(\lambda)$ associated with the triple (K,ϕ,ρ) is

$$\mathcal{D}(\lambda) = \inf_{f\in\mathcal{H}_K}\left\{\mathcal{E}(f) - \mathcal{E}\big(f_\rho^\phi\big) + \frac{\lambda}{2}\|f\|_K^2\right\} = \mathcal{E}\big(f_\lambda^\phi\big) - \mathcal{E}\big(f_\rho^\phi\big) + \frac{\lambda}{2}\big\|f_\lambda^\phi\big\|_K^2, \quad \lambda > 0, \tag{2.1}$$

where f_ρ^ϕ is a minimizer of the generalization error $\mathcal{E}(f)$.

The regularization error measures the approximation ability of the space $\mathcal{H}_K$ with respect to the classification process involving ϕ and ρ. It is independent of the sample. If $\mathcal{H}_K$ is dense in $C(X)$, we know that $\lim_{\lambda\to 0}\mathcal{D}(\lambda) = 0$. So a natural assumption would be

$$\mathcal{D}(\lambda) \leqslant \mathcal{D}_0\lambda^\beta \quad \text{for some } 0 \leqslant \beta \leqslant 1 \text{ and } \mathcal{D}_0 > 0. \tag{2.2}$$

This is a fundamental assumption about the hypothesis space itself. Since $\mathcal{D}(\lambda) \leqslant \mathcal{E}(0) + 0 = \phi(0)$ for any $\lambda > 0$, we see that (2.2) always holds with $\beta = 0$ and $\mathcal{D}_0 = \phi(0)$. Moreover, β cannot be greater than 1, as proved in [3].

Theorem 1. *Let ϕ be an admissible loss function, f_λ^ϕ by* (1.3)*, and $\mu > \lambda > 0$. Then*

$$\big\|f_\lambda^\phi - f_\mu^\phi\big\|_K \leqslant \frac{\mu}{2}\left(\frac{1}{\lambda}-\frac{1}{\mu}\right)\big(\big\|f_\lambda^\phi\big\|_K + \big\|f_\mu^\phi\big\|_K\big) \leqslant \frac{\mu}{2}\left(\frac{1}{\lambda}-\frac{1}{\mu}\right)\left(\sqrt{\frac{2\mathcal{D}(\lambda)}{\lambda}} + \sqrt{\frac{2\mathcal{D}(\mu)}{\mu}}\right).$$

In particular, if with some $0 < \gamma \leqslant 1$ we take $\lambda_t = \lambda_1 t^{-\gamma}$ for $t \geqslant 1$, then

$$d_{t+1} \leqslant 2t^{\frac{\gamma}{2}-1}\sqrt{\mathcal{D}\big(\lambda_1 t^{-\gamma}\big)/\lambda_1} \leqslant 2t^{\frac{\gamma}{2}-1}\sqrt{\phi(0)/\lambda_1}.$$

Theorem 1 will be proved in Section 3. For the least-square regression, the drift error can be estimated by linear functional analysis and has been done in [23].

2.2. Strong convergence to the regularizing function

Our second main result provides some estimates for the strong approximation of the learning scheme (1.4) measured by $\|f_{T+1} - f_{\lambda_T}^\phi\|_K$ in the $\mathcal{H}_K$ norm. It is an estimation-type error depending on the sample.

Let us demonstrate our general result by considering the special case of hinge loss.

Proposition 1. *Let $\phi(x) = (1-x)_+$ and with some $\lambda_1, \eta_1 > 0$, $0 < \gamma, \alpha < 1$, we take*

$$\lambda_t = \lambda_1 t^{-\gamma}, \quad \eta_t = \eta_1 t^{-\alpha} \quad \forall t \in \mathbb{N}. \tag{2.3}$$

If $\eta_1 \leqslant \frac{1}{2\kappa^2+\lambda_1}$ and $\gamma < \frac{2}{5}$, then

$$\mathbb{E}_{z_1,\dots,z_T}\big(\big\|f_{T+1} - f_{\lambda_T}^\phi\big\|_K^2\big) \leqslant C_{\eta_1,\lambda_1,\kappa}T^{-\theta},$$

where $C_{\eta_1,\lambda_1,\kappa}$ is a constant depending on $\eta_1, \lambda_1, \kappa$ and

$$\theta = \begin{cases} \alpha - \gamma, & \text{if } \gamma < \alpha < \frac{2}{3}(1-\gamma), \\ 2 - 3\gamma - 2\alpha - \epsilon, & \text{if } \frac{2}{3}(1-\gamma) \leqslant \alpha < 1 - \frac{3}{2}\gamma \text{ and } 0 < \epsilon < 2 - 3\gamma - 2\alpha. \end{cases}$$

Proposition 1 follows from Theorem 2 below. Here the convergence rates can be of order $O(T^{\epsilon-2/3})$ for an arbitrary small $\epsilon > 0$ by taking γ to be small enough. The reason for choosing θ in two different cases will be seen in the next subsection.

To state the rates of strong convergence involving a general loss, we need the following constants measuring the increment of the (left) derivative of the loss ϕ.

Definition 5. Denote

$$N(\lambda) = \sup\left\{ \left|\phi'_-(x)\right| \colon |x| \leqslant \frac{\kappa^2 |\phi'(0)|}{\lambda} \right\}, \quad \lambda > 0. \tag{2.4}$$

We say that ϕ has *incremental exponent* $p \geqslant 0$ if for some $N_1 > 0$ and $\lambda_1 > 0$ we have

$$N(\lambda) \leqslant N_1 \left(\frac{1}{\lambda}\right)^p \quad \forall 0 < \lambda \leqslant \lambda_1. \tag{2.5}$$

We say that ϕ'_- is locally Lipschitz at the origin if

$$M_0 := \sup_{|x| \leqslant 1} \left\{ \frac{|\phi'_-(x) - \phi'(0)|}{|x|} \right\} < \infty. \tag{2.6}$$

For the least-square loss $\phi(x) = (1-x)^2$, $N(\lambda) = 2 + 4\kappa^2/\lambda$ and $M_0 = 2$.

The following explicit rates for the strong approximation will be verified in Section 5 where the constant $C_{\eta_1,\lambda_1,\kappa,p,\mathcal{D}_0,\beta}$ can be found explicitly. A brief description for deriving the convergence rate (2.8) below in two different cases (2.9) will be given in the next subsection.

Theorem 2. *Let $\{\lambda_t, \eta_t\}$ be given by* (2.3)*. Assume* (2.5) *and* (2.6) *for ϕ. If $p\gamma \leqslant \alpha$ and*

$$\eta_1 \leqslant \frac{1}{\kappa^2 M_0 + 2\kappa^2 N_1 \lambda_1^{-p} + \lambda_1}, \tag{2.7}$$

then $\|f_t\|_K \leqslant \frac{\kappa|\phi'(0)|}{\lambda_t}$ for each $t \in \mathbb{N}$. If moreover (2.2) *holds for the triple (K, ϕ, ρ) and $\gamma < \frac{2}{5+4p-\beta}$, then we have*

$$\mathbb{E}_{z_1,\ldots,z_T}\left(\left\| f_{T+1} - f_{\lambda_T}^{\phi} \right\|_K^2 \right) \leqslant C_{\eta_1,\lambda_1,\kappa,p,\mathcal{D}_0,\beta} T^{-\theta}, \tag{2.8}$$

where $C_{\eta_1,\lambda_1,\kappa,p,\mathcal{D}_0,\beta}$ is a constant depending on $\eta_1, \lambda_1, \kappa, p, \mathcal{D}_0, \beta$ and

$$\theta = \begin{cases} \alpha - (2p+1)\gamma, & \text{if } (2p+1)\gamma < \alpha < \frac{2+(2p-2+\beta)\gamma}{3}, \\ 2 - \gamma(1-\beta) - 2(\alpha+\gamma) - \epsilon, & \text{if } \frac{2+(2p-2+\beta)\gamma}{3} \leqslant \alpha < 1 - \frac{(3-\beta)\gamma}{2}. \end{cases} \tag{2.9}$$

Here in the second case ϵ is an arbitrary number satisfying $0 < \epsilon < 2 - (3-\beta)\gamma - 2\alpha$.

Consider the special case of hinge loss $\phi(x) = \max\{1-x, 0\}$. We have $N(\lambda) \equiv 1$ and $M_0 = 0$, hence $p = 0$. Moreover, (2.5) holds with $N_1 = 1$ and $\mathcal{D}(\lambda) \leqslant \phi(0) = 1$ for any $\lambda > 0$. Thus Theorem 2 with $\mathcal{D}_0 = 1, \beta = 0$ and $p = 0$ verifies Proposition 1.

The rate stated in Theorem 2 can be of order $T^{\epsilon-2/3}$ for an arbitrary small $\epsilon > 0$ when $\gamma \to 0$ and $\alpha \to \frac{2+(2p-2+\beta)\gamma}{3}$. It says that for the error of $f_{T+1} - f_{\lambda_T}^{\phi}$ to be small, we need small γ for the regularization parameter. More explicitly, we have

Corollary 1. *Under the conditions of Theorem* 2*, if $\alpha = \frac{2+(2p-2+\beta)\gamma}{3}$ and $0 < \epsilon < \frac{2-(5+4p-\beta)\gamma}{3}$, we have*

$$\mathbb{E}_{z_1,\ldots,z_T}\left(\left\| f_{T+1} - f_{\lambda_T}^{\phi} \right\|_K^2 \right) \leqslant C_{\eta_1,\lambda_1,\kappa,p,\mathcal{D}_0,\beta} T^{\epsilon + \frac{(5+4p-\beta)\gamma - 2}{3}}.$$

The above bound for the strong approximation error decays fast when the regularization parameter λ_t decays slowly with a small γ.

2.3. Outline of the key analysis

Our key analysis for deriving the bound in Theorem 2 for $\|f_{T+1} - f_{\lambda_T}^{\phi}\|_K$ consists of three steps (Sections 4 and 5). The first step is to bound $\|f_{t+1} - f_{\lambda_t}^{\phi}\|_K$ in terms of $\|f_t - f_{\lambda_t}^{\phi}\|_K$:

$$\mathbb{E}_{z_1,z_2,\ldots,z_t}\big(\big\|f_{t+1} - f_{\lambda_t}^{\phi}\big\|_K^2\big) \leqslant (1-\eta_t\lambda_t)\mathbb{E}_{z_1,z_2,\ldots,z_{t-1}}\big(\big\|f_t - f_{\lambda_t}^{\phi}\big\|_K^2\big) + \big(2\kappa N(\lambda_t)\eta_t\big)^2. \tag{2.10}$$

This inequality will be proved in Lemma 5. The convexity of the loss function ϕ plays an important role in this step. From (2.10) we can see the choice of the index α for $\eta_t = \eta_1 t^{-\alpha}$: for the last term of (2.10) to be small, we need large α; while for the middle term to be small, we should choose small α. This gives some clue why in Proposition 1 we should choose θ in two different cases in order to maximize a power index for the convergence rate.

The second step is to bound $\|f_t - f_{\lambda_t}^{\phi}\|_K^2$ in terms of $\|f_t - f_{\lambda_{t-1}}^{\phi}\|_K$ and $d_t = \|f_{\lambda_t}^{\phi} - f_{\lambda_{t-1}}^{\phi}\|_K$. Since d_t is expected to be much smaller than $\|f_t - f_{\lambda_{t-1}}^{\phi}\|_K$, we shall apply an uneven inequality $2ab \leqslant a^2b^s + b^{2-s}$ with a small $s \in (0,2)$ to $2\|f_t - f_{\lambda_{t-1}}^{\phi}\|_K d_t$ and obtain

$$\mathbb{E}_{z_1,z_2,\ldots,z_{t-1}}\big(\big\|f_t - f_{\lambda_t}^{\phi}\big\|_K^2\big) \leqslant \big(1+d_t^s\big)\mathbb{E}_{z_1,z_2,\ldots,z_{t-1}}\big(\big\|f_t - f_{\lambda_{t-1}}^{\phi}\big\|_K^2\big) + d_t^{2-s} + d_t^2.$$

Together with (2.10), this implies $\mathbb{E}_{z_1,\ldots,z_t}(\|f_{t+1} - f_{\lambda_t}^{\phi}\|_K^2)$ is bounded by

$$\big(1+d_t^s - \eta_t\lambda_t\big)\mathbb{E}_{z_1,\ldots,z_{t-1}}\big(\big\|f_t - f_{\lambda_{t-1}}^{\phi}\big\|_K^2\big) + d_t^{2-s} + d_t^2 + \big(2\kappa N(\lambda_t)\eta_t\big)^2. \tag{2.11}$$

To see how to choose s, we need the decay of d_t.

Lemma 1. *Let $\lambda_t = \lambda_1 t^{-\gamma}$ for some $0 < \gamma < 1$ and $\lambda_1 > 0$. If (2.2) holds, then*

$$d_t \leqslant 4\sqrt{\mathcal{D}_0\lambda_1^{\beta-1}}\, t^{\frac{\gamma(1-\beta)}{2}-1} \quad \forall t \in \mathbb{N}. \tag{2.12}$$

Proof. The assumption (2.2) in connection with Theorem 1 tells us that for any $t \geqslant 1$,

$$d_{t+1} \leqslant \frac{\lambda_1 t^{-\gamma}}{2}\left(\frac{1}{\lambda_1(t+1)^{-\gamma}} - \frac{1}{\lambda_1 t^{-\gamma}}\right)2\sqrt{2\mathcal{D}_0\lambda_1^{\beta-1}(t+1)^{\gamma(1-\beta)}}.$$

Since $(t+1)^{\gamma} - t^{\gamma} = \gamma\xi^{\gamma-1}$ for some $\xi \in (t, t+1)$, we have $(t+1)^{\gamma} - t^{\gamma} \leqslant t^{\gamma-1}$ and hence

$$d_{t+1} \leqslant \frac{\sqrt{2}}{t}\sqrt{\mathcal{D}_0\lambda_1^{\beta-1}}(t+1)^{\frac{\gamma(1-\beta)}{2}} \leqslant 4\sqrt{\mathcal{D}_0\lambda_1^{\beta-1}}(t+1)^{\frac{\gamma(1-\beta)}{2}-1}.$$

This inequality also holds for $t = 0$ since $d_1 = \|f_{\lambda_1}^{\phi}\|_K \leqslant \sqrt{2\mathcal{D}(\lambda_1)/\lambda_1} \leqslant 4\sqrt{\mathcal{D}_0\lambda_1^{\beta-1}}$. So the desired bound (2.12) holds true. □

Once the decay of d_t is obtained, we can compare the rates of $d_t^s = O(t^{s(\frac{\gamma(1-\beta)}{2}-1)})$ and $\eta_t\lambda_t = O(t^{-(\alpha+\gamma)})$ for the first term of (2.11). We shall require

$$s\left(1 - \frac{\gamma(1-\beta)}{2}\right) > \alpha + \gamma \tag{2.13}$$

so that the coefficient $1 + d_t^s - \eta_t\lambda_t$ in the first term of (2.11) behaves as $1 - ct^{-\alpha-\gamma}$.

On the other hand, we shall see that $N(\lambda_t)\eta_t$ behaves as $O(t^{p\gamma-\alpha})$. So to dominate the last term $d_t^{2-s} + d_t^2 + (2\kappa N(\lambda_t)\eta_t)^2$ of (2.11) by the quantity d_t^{2-s}, we require $(2-s)(1 - \frac{\gamma(1-\beta)}{2}) \leqslant 2(\alpha - p\gamma)$. That is,

$$s\left(1 - \frac{\gamma(1-\beta)}{2}\right) \geqslant 2\left(1 - \frac{\gamma(1-\beta)}{2}\right) - (2\alpha - 2p\gamma). \tag{2.14}$$

The choice of s in the second step of our key analysis for the bound (2.11) is seen from the restrictions (2.13) and (2.14). For details, see the proof of Theorem 2.

The third step is to applying the above recursive relation iteratively to get bounds for $\mathbb{E}_{z_1,\dots,z_T}(\|f_{T+1}-f_{\lambda_T}^{\phi}\|_K^2)$. Again for the last term of (2.11) to be small, we need large α, but for the product $\prod_{t=1}^{T}(1+d_t^s-\eta_t\lambda_t)$ (appearing after iterations) to be small, we require α to be large. This leads to the choice (2.9) for the learning rate (2.8) presented in Theorem 2.

2.4. Learning rates

The convergence rate stated in Theorem 2, together with a bound for the regularization error (requiring γ to be large, a trade-off) yields learning rate of the misclassification error of the fully online algorithm (1.4), taking suitable choices of the regularization parameter λ_t and the step size η_t. This is our last main result.

We first present the example of hinge loss again to illustrate the general result.

Corollary 2. *Let $\phi(x)=(1-x)_+$ and for some $\lambda_1,\eta_1>0$, $0<\gamma,\alpha<1$, we take*

$$\lambda_t=\lambda_1 t^{-\gamma},\quad \eta_t=\eta_1 t^{-\alpha}\quad \forall t\in\mathbb{N}. \tag{2.15}$$

Assume $\eta_1\leqslant\frac{1}{2\kappa^2+\lambda_1}$ and (1.10) holds for the pair (ρ,K). If $0<\gamma<\frac{2}{5-\beta}$ and $\gamma<\alpha<\frac{2-2\gamma+\beta\gamma}{3}$, then for any $T\in\mathbb{N}$ we have

$$\mathbb{E}_{z_1,\dots,z_T}\big(\mathcal{R}\big(\operatorname{sgn}(f_{T+1})\big)-\mathcal{R}(f_c)\big)\leqslant C_{\eta_1,\lambda_1,\kappa,\mathcal{D}_0,\beta}T^{-\min\{\beta\gamma,\frac{\alpha-\gamma}{2}\}}, \tag{2.16}$$

where $C_{\eta_1,\lambda_1,\kappa,\mathcal{D}_0,\beta}$ is a constant depending on $\eta_1,\lambda_1,\kappa,\mathcal{D}_0$ and β.

The general learning rate proved in Section 6 can be stated as follows.

Theorem 3. *Let ϕ be an admissible loss function such that $\phi''(0)$ exists and is positive. Under the assumptions of Theorem 2 and θ given by (2.9), if $\gamma<\frac{2}{5+10p-\beta}$, we have*

$$\mathbb{E}_{z_1,\dots,z_T}\big(\mathcal{R}\big(\operatorname{sgn}(f_{T+1})\big)-\mathcal{R}(f_c)\big)\leqslant\tilde{C}_{\phi}T^{-\min\{\frac{\beta\gamma}{2},\frac{\theta}{4}-\frac{p\gamma}{2}\}}, \tag{2.17}$$

where $\tilde{C}_{\phi}$ is a constant depending on $\eta_1,\lambda_1,\kappa,p,\mathcal{D}_0,\beta,N_1$ and ϕ.

There has been a vast literature on the partially online algorithm (that is, when $\lambda_t\equiv\lambda_1$ is independent of the step t). Let us mention some works relating to this paper. In [17], a stochastic gradient method in the Hilbert space $\mathcal{H}_K$ is considered. Let $SL(\mathcal{H}_K)$ be the space of positive definite linear operators on $\mathcal{H}_K$, and $A:Z\to SL(\mathcal{H}_K)$ and $B:Z\to\mathcal{H}_K$ be two maps. To learn a stationary point f^* satisfying

$$\mathbb{E}_{z\in Z}\big(A(z)f^*+B(z)\big)=0,$$

they proposed the learning sequence

$$f_{t+1}=f_t-\eta_t\big\{A(z_t)(f_t)+B(z_t)\big\}. \tag{2.18}$$

But the partially online scheme (1.4) involving the general loss function ϕ is in general nonlinear and is hard to write in the setting (2.18) except for the least-square loss.

The cumulative loss $\frac{1}{T}\sum_{t=1}^{T}\phi(y_t f_t(x_t))$ for partially online algorithms more general than (1.4) has been well studied in the literature; see, for example, [4,5,9] and references therein. In particular, cumulative loss bounds are derived for online linear regression with least-square loss in [4]. In Section 6 of [9], for a learning algorithm different from (1.4), the relative expected instantaneous loss, measuring the prediction ability of f_{T+1} in linear regression problem, is analyzed in detail.

A general regularized partially online learning scheme is introduced and analyzed in [10]. Assume the loss function ϕ is convex, uniformly Lipschitz continuous, the step size has the form $\eta_t=O(t^{-1/2})$, and $\lambda>0$ is fixed. It was proved there that the average instantaneous risk $\frac{1}{T}\sum_{t=1}^{T}(\phi(y_t f_t(x_t))+\frac{\lambda}{2}\|f_t\|_K^2)$ converges toward the regularized

generalization error $\mathcal{E}(f_\lambda^\phi) + \frac{\lambda}{2}\|f_\lambda^\phi\|_K^2$ with error bound $O(T^{-1/2})$. This result is about the average instantaneous risk and $\lambda_t \equiv \lambda_1$.

Different from estimating the cumulative loss bounds as done in many previous results (e.g., [5,10]), the strong approximation for the partially online algorithm was considered in [26], as done for the least-square regression in [20]. In particular, it provides estimates for the error $\|f_{T+1} - f_\lambda^\phi\|_K$ in the $\mathcal{H}_K$ norm with fixed $\lambda_t \equiv \lambda_1 > 0$, and then applies them to the analysis of the misclassification error. The learning rates are given in terms of suitable choices of the regularization parameter $\lambda_1 = \lambda_1(T)$ depending on T. But the results are not for the fully online algorithm.

Recently, fully online scheme (1.4) has been studied for least-square regression in [23] where $\phi(x) = (1-x)^2$. For this loss function, learning rates for the approximation of the regression function f_ρ by f_{T+1} in spaces $L^2_{\rho_X}$ and $\mathcal{H}_K$, similar to those [13,14,20,21] for off-line schemes, are derived. Our error bounds stated in Theorems 2 and 3 are for classification with a general convex loss function including the least-square loss as a special example. So our setting is more general.

3. Estimating the drift error

In this section, we prove Theorem 1 which estimates $\|f_\lambda^\phi - f_\mu^\phi\|_K$ (with $\lambda, \mu > 0$) for the drift error and plays an important role in deriving satisfactory learning rates.

We first prove the theorem for differentiable loss functions. Under this differentiability assumption, it was observed in [26] by taking a variational derivative of the functional (regularized generalization error) given in (1.3) that the minimizer f_λ^ϕ of the functional satisfies

$$\int_Z \phi'\big(yf_\lambda^\phi(x)\big)yK_x \,\mathrm{d}\rho + \lambda f_\lambda^\phi = 0. \tag{3.1}$$

Lemma 2. *Let $\mu > \lambda > 0$ and ϕ be a differentiable convex loss function. Then*

$$\big\|f_\lambda^\phi - f_\mu^\phi\big\|_K \leqslant \frac{\mu}{2}\left(\frac{1}{\lambda} - \frac{1}{\mu}\right)\big\{\big\|f_\lambda^\phi\big\|_K + \big\|f_\mu^\phi\big\|_K\big\}.$$

Proof. From (3.1), we know that

$$f_\lambda^\phi - f_\mu^\phi = \frac{1}{\mu}\int_Z \phi'\big(yf_\mu^\phi(x)\big)yK_x \,\mathrm{d}\rho - \frac{1}{\lambda}\int_Z \phi'\big(yf_\lambda^\phi(x)\big)yK_x \,\mathrm{d}\rho.$$

Combining with the reproducing property (1.1), we know $\|f_\lambda^\phi - f_\mu^\phi\|_K^2 = \langle f_\lambda^\phi - f_\mu^\phi, f_\lambda^\phi - f_\mu^\phi\rangle_K$ can be expressed as

$$\big\|f_\lambda^\phi - f_\mu^\phi\big\|_K^2 = \frac{1}{\mu}\int_Z \phi'\big(yf_\mu^\phi(x)\big)y\big(f_\lambda^\phi - f_\mu^\phi\big)(x) \,\mathrm{d}\rho - \frac{1}{\lambda}\int_Z \phi'\big(yf_\lambda^\phi(x)\big)y\big(f_\lambda^\phi - f_\mu^\phi\big)(x) \,\mathrm{d}\rho. \tag{3.2}$$

Since ϕ is a convex function on $\mathbb{R}$, we know that

$$\phi'(a)(b-a) \leqslant \phi(b) - \phi(a), \quad \forall a, b \in \mathbb{R}. \tag{3.3}$$

Thus

$$\phi'\big(yf_\mu^\phi(x)\big)\big(yf_\lambda^\phi(x) - yf_\mu^\phi(x)\big) \leqslant \phi\big(yf_\lambda^\phi(x)\big) - \phi\big(yf_\mu^\phi(x)\big)$$

and

$$\phi'\big(yf_\lambda^\phi(x)\big)\big(yf_\mu^\phi(x) - yf_\lambda^\phi(x)\big) \leqslant \phi\big(yf_\mu^\phi(x)\big) - \phi\big(yf_\lambda^\phi(x)\big).$$

Putting these two inequalities into (3.2), we get

$$\big\|f_\lambda^\phi - f_\mu^\phi\big\|_K^2 \leqslant \left(\frac{1}{\lambda} - \frac{1}{\mu}\right)\big(\mathcal{E}\big(f_\mu^\phi\big) - \mathcal{E}\big(f_\lambda^\phi\big)\big). \tag{3.4}$$

For $\mu > \lambda > 0$, we know that $\frac{1}{\lambda} - \frac{1}{\mu} > 0$ and hence $\mathcal{E}(f_\mu^\phi) - \mathcal{E}(f_\lambda^\phi) \geqslant 0$.

From the definition of f_μ^ϕ, we see that $\mathcal{E}(f_\mu^\phi) + \frac{\mu}{2}\|f_\mu^\phi\|_K^2 - (\mathcal{E}(f_\lambda^\phi) + \frac{\mu}{2}\|f_\lambda^\phi\|_K^2) \leqslant 0$. It follows that

$$\mathcal{E}\big(f_\mu^\phi\big) - \mathcal{E}\big(f_\lambda^\phi\big) \leqslant \frac{\mu}{2}\big(\big\|f_\lambda^\phi\big\|_K^2 - \big\|f_\mu^\phi\big\|_K^2\big) = \frac{\mu}{2}\big(\big\|f_\lambda^\phi\big\|_K - \big\|f_\mu^\phi\big\|_K\big)\big(\big\|f_\lambda^\phi\big\|_K + \big\|f_\mu^\phi\big\|_K\big). \tag{3.5}$$

Note that $\|f_\lambda^\phi\|_K - \|f_\mu^\phi\|_K \leqslant \|f_\lambda^\phi - f_\mu^\phi\|_K$. Then the desired inequality follows from (3.4). □

Now we can prove Theorem 1.

Proof of Theorem 1. For the admissible loss function ϕ, we define for each $0 < \varepsilon \leqslant 1$ a convex, differentiable function on $\mathbb{R}$ by

$$\phi_\varepsilon(x) = \int_0^1 \phi(x - \varepsilon\theta)\,\mathrm{d}\theta = \frac{1}{\varepsilon}\int_{x-\varepsilon}^{x} \phi(u)\,\mathrm{d}u.$$

It is a differentiable admissible loss function and the conclusion of Lemma 2 holds true.

We define, for $0 < \varepsilon \leqslant 1$, $\mathcal{E}^{(\varepsilon)}(f) = \int_Z \phi_\varepsilon(yf(x))\,\mathrm{d}\rho$ and $f_\lambda^{(\varepsilon)} = \arg\min_{f\in\mathcal{H}_K} \mathcal{E}^{(\varepsilon)}(f) + \frac{\lambda}{2}\|f\|_K^2$.

An intermediate step in the proof of Lemma 2 of [26] shows that there exists a sequence $\{\varepsilon_j > 0\}_{j=1}^\infty$ such that $\lim_{j\to\infty}\varepsilon_j = 0$ and for each $\zeta \in \{\lambda, \mu\}$, the sequence $\{f_\zeta^{(\varepsilon_j)}\}$ is uniformly bounded in $\mathcal{H}_K$ (with respect to j) and converges to f_ζ^ϕ weakly. Moreover, $\mathcal{E}(f_\zeta^\phi) = \lim_{j\to\infty}\mathcal{E}^{(\varepsilon_j)}(f_\zeta^{(\varepsilon_j)})$ and

$$\mathcal{E}\big(f_\zeta^\phi\big) + \frac{\zeta}{2}\big\|f_\zeta^\phi\big\|_K^2 = \underline{\lim}_{j\to\infty}\left\{\mathcal{E}^{(\varepsilon_j)}\big(f_\zeta^{(\varepsilon_j)}\big) + \frac{\zeta}{2}\big\|f_\zeta^{(\varepsilon_j)}\big\|_K^2\right\}.$$

It follows that $\underline{\lim}_{j\to\infty}\|f_\zeta^{(\varepsilon_j)}\|_K \leqslant \|f_\zeta^\phi\|_K$. The weak convergence implies that

$$\big\|f_\lambda^\phi - f_\mu^\phi\big\|_K^2 = \lim_{j\to\infty}\big\langle f_\lambda^{(\varepsilon_j)} - f_\mu^{(\varepsilon_j)}, f_\lambda^\phi - f_\mu^\phi\big\rangle_K \leqslant \underline{\lim}_{j\to\infty}\big\|f_\lambda^{(\varepsilon_j)} - f_\mu^{(\varepsilon_j)}\big\|_K\big\|f_\lambda^\phi - f_\mu^\phi\big\|_K.$$

Applying Lemma 2 to the modified loss function ϕ_{ε_j} yields

$$\big\|f_\lambda^{(\varepsilon_j)} - f_\mu^{(\varepsilon_j)}\big\|_K \leqslant \frac{\mu}{2}\left(\frac{1}{\lambda} - \frac{1}{\mu}\right)\big\{\big\|f_\lambda^{(\varepsilon_j)}\big\|_K + \big\|f_\mu^{(\varepsilon_j)}\big\|_K\big\}.$$

Thus, by letting $j \to \infty$, we see that

$$\big\|f_\lambda^\phi - f_\mu^\phi\big\|_K \leqslant \underline{\lim}_{j\to\infty}\big\|f_\lambda^{(\varepsilon_j)} - f_\mu^{(\varepsilon_j)}\big\|_K \leqslant \frac{\mu}{2}\left(\frac{1}{\lambda} - \frac{1}{\mu}\right)\big\{\big\|f_\lambda^\phi\big\|_K + \big\|f_\mu^\phi\big\|_K\big\}.$$

Using (2.1), we know that

$$\big\|f_\lambda^\phi\big\|_K \leqslant \sqrt{\frac{2\mathcal{D}(\lambda)}{\lambda}}, \qquad \big\|f_\mu^\phi\big\|_K \leqslant \sqrt{\frac{2\mathcal{D}(\mu)}{\mu}}.$$

So the first statement follows.

The second statement follows by observing $(t+1)^\gamma - t^\gamma \leqslant \gamma t^{\gamma-1} \leqslant t^{\gamma-1}$ and that $\mathcal{D}$ is a nondecreasing function. This verifies Theorem 1. □

4. Analyzing the remainder for online schemes

The first two steps in getting the convergence and rates of convergence for the fully online scheme are to analyze the remainder $f_{t+1} - f_{\lambda_t}^\phi$. To this end, we need to bound the learning sequence $\{f_t\}$ first. The following result can be proved following the same line of reasoning as [26]. A detailed proof is given in Appendix A.

Proposition 2. *Assume that ϕ'_- is locally Lipschitz at the origin. Define $\{f_t\}$ by* (1.4) *and denote*

$$M(\lambda) = \sup\left\{\frac{|\phi'_-(x) - \phi'(0)|}{|x|}\colon\ |x| \leqslant \frac{\kappa^2|\phi'(0)|}{\lambda}\right\}, \quad \lambda > 0. \tag{4.1}$$

If

$$\eta_t\big(\kappa^2 M(\lambda_t)+\lambda_t\big)\leqslant 1 \quad \forall t\geqslant 1, \tag{4.2}$$

then

$$\|f_t\|_K \leqslant \frac{\kappa|\phi'(0)|}{\lambda_t}. \tag{4.3}$$

Remark 1. The bound (4.3) seems artificial since the definition of f_t depends only on $\lambda_1,\dots,\lambda_{t-1}$, not on λ_t. Actually the proof of Proposition 2 ensures a tighter bound $\|f_t\|_K \leqslant \frac{\kappa|\phi'(0)|}{\lambda_{t-1}}$ for any $t\geqslant 2$ which implies (4.3) by our assumption $\lambda_t\leqslant\lambda_{t-1}$. But for our error analysis we only need the bound (4.3).

The following relation between the functions $M(\lambda)$ and $N(\lambda)$ will be used.

Lemma 3. *For any $\lambda>0$, we have*

$$M(\lambda)\leqslant \max\big\{M_0, N(\lambda)+\big|\phi'(0)\big|\big\}\leqslant \max\big\{M_0, 2N(\lambda)\big\}.$$

Proof. When $|x|\leqslant 1$, the definition (2.6) of M_0 yields $\frac{|\phi'_-(x)-\phi'(0)|}{|x|}\leqslant M_0$.

When $1<|x|\leqslant\frac{\kappa^2|\phi'(0)|}{\lambda}$, we have $\frac{|\phi'_-(x)-\phi'(0)|}{|x|}\leqslant|\phi'_-(x)-\phi'(0)|\leqslant|\phi'_-(x)|+|\phi'(0)|$. So the desired bound follows from the definition (2.4): $|\phi'_-(x)|\leqslant N(\lambda)$. □

The following are some commonly used examples of loss functions (e.g., [2,6,11,12,27]). The constants $N(\lambda)$ and $M(\lambda)$ are estimated.

Example 2. (1) For the least-square loss $\phi(x)=(1-x)^2$, we have $M(\lambda)\equiv 2$ and $N(\lambda)=2+\frac{4\kappa^2}{\lambda}$.

(2) For the q-norm SVM loss $\phi(x)=(1-x)_+^q$ with $1\leqslant q\leqslant 2$, we have $M(\lambda)\leqslant 4$ and $N(\lambda)=q(1+\kappa^2 q/\lambda)^{q-1}$.

(3) For the q-norm SVM loss with $q>2$, we have $M(\lambda)\leqslant q(q-1)(1+\kappa^2 q/\lambda)^{q-2}$ and $N(\lambda)=q(1+\kappa^2 q/\lambda)^{q-1}$.

(4) For the exponential loss $\phi(x)=e^{-x}$, we have $N(\lambda)=e^{\kappa^2/\lambda}$ and $M(\lambda)\leqslant e^{\kappa^2/\lambda}$.

Proof. (1) Note the least-square loss ϕ is twice continuously differentiable, $\phi'(x)=2(x-1)$ and $\phi''\equiv 2$. Then we have the desired expressions.

(2) When $1\leqslant q\leqslant 2$, the q-norm SVM loss $\phi(x)=(1-x)_+^q$ has $\phi'_-(x)=-q(1-x)_+^{q-1}$. The expression for $N(\lambda)$ follows, which also works for $q>2$. We have $M(\lambda)\leqslant 4$ for all $\lambda>0$ since

$$\frac{|\phi'_-(x)-\phi'(0)|}{|x|}\leqslant\begin{cases}\|\phi''\|_{L^\infty[0,1/2]}\leqslant 4, & \text{if } x\in[0,1/2],\\ |\phi'(0)|/|x|\leqslant 4, & \text{if } x>1/2,\\ \|\phi''\|_{L^\infty[-\infty,0]}\leqslant 2, & \text{if } x<0.\end{cases}$$

(3) When $q>2$, we have $\phi'(0)=-q$ and $\phi''_-(x)=q(q-1)(1-x)_+^{q-2}$. So we find that $M(\lambda)\leqslant q(q-1)(1+\kappa^2 q/\lambda)^{q-2}$.

(4) For the exponential loss, we have $\phi'(x)=-e^{-x}$ and $\phi''(x)=e^{-x}$. Then we have $N(\lambda)=e^{\kappa^2/\lambda}$ and $M(\lambda)\leqslant\|\phi''\|_{L^\infty[-\kappa^2|\phi'(0)|/\lambda,\kappa^2|\phi'(0)|/\lambda]}=e^{\kappa^2/\lambda}$. □

For the error analysis, we also need the following relation derived in [26].

Lemma 4. *Let ϕ be an admissible loss function and $\lambda>0$. Then*

$$\frac{\lambda}{2}\big\|f-f_\lambda^\phi\big\|_K^2\leqslant\left\{\mathcal{E}(f)+\frac{\lambda}{2}\|f\|_K^2\right\}-\left\{\mathcal{E}\big(f_\lambda^\phi\big)+\frac{\lambda}{2}\big\|f_\lambda^\phi\big\|_K^2\right\}\quad\forall f\in\mathcal{H}_K. \tag{4.4}$$

We are in a position to provide the first step of the key analysis: estimating $\|f_{t+1}-f_{\lambda_t}^\phi\|_K$ in terms of $\|f_t-f_{\lambda_t}^\phi\|_K$.

Lemma 5. *If the bound* (4.3) *is valid for some* $t \in \mathbb{N}$, *then* (2.10) *holds true.*

Proof. Denote $G_t := \phi'_-(y_t f_t(x_t)) y_t K_{x_t} + \lambda_t f_t$. The online scheme (1.4) can be written as $f_{t+1} = f_t - \eta_t G_t$. Then

$$\|f_{t+1} - f_{\lambda_t}^{\phi}\|_K^2 = \|f_t - f_{\lambda_t}^{\phi}\|_K^2 + \eta_t^2 \|G_t\|_K^2 + 2\eta_t \langle G_t, f_{\lambda_t}^{\phi} - f_t\rangle_K. \tag{4.5}$$

Applying (1.1) and (3.3) to part of the last term of (4.5), we see that the inner product $\langle \phi'_-(y_t f_t(x_t)) y_t K_{x_t}, f_{\lambda_t}^{\phi} - f_t\rangle_K$ equals to

$$\phi'_-\big(y_t f_t(x_t)\big)\big(y_t f_{\lambda_t}^{\phi}(x_t) - y_t f_t(x_t)\big) \leqslant \phi\big(y_t f_{\lambda_t}^{\phi}(x_t)\big) - \phi\big(y_t f_t(x_t)\big).$$

For the other part, we have

$$\langle f_t, f_{\lambda_t}^{\phi} - f_t\rangle_K \leqslant \frac{1}{2}\|f_{\lambda_t}^{\phi}\|_K^2 + \frac{1}{2}\|f_t\|_K^2 - \|f_t\|_K^2 = \frac{1}{2}\|f_{\lambda_t}^{\phi}\|_K^2 - \frac{1}{2}\|f_t\|_K^2.$$

Thus the last term of (4.5) can be bounded as

$$\langle G_t, f_{\lambda_t}^{\phi} - f_t\rangle_K \leqslant \left[\phi\big(y_t f_{\lambda_t}^{\phi}(x_t)\big) + \frac{\lambda_t}{2}\|f_{\lambda_t}^{\phi}\|_K^2\right] - \left[\phi\big(y_t f_t(x_t)\big) + \frac{\lambda_t}{2}\|f_t\|_K^2\right].$$

Since f_t depends on $\{z_1, z_2, \ldots, z_{t-1}\}$ but not on z_t, it follows that

$$\mathbb{E}_{z_t}\big(\langle G_t, f_{\lambda_t}^{\phi} - f_t\rangle_K\big) \leqslant \left[\mathcal{E}\big(f_{\lambda_t}^{\phi}\big) + \frac{\lambda_t}{2}\|f_{\lambda_t}^{\phi}\|_K^2\right] - \left[\mathcal{E}(f_t) + \frac{\lambda_t}{2}\|f_t\|_K^2\right].$$

This in connection with Lemma 4 implies

$$\mathbb{E}_{z_1,z_2,\ldots,z_t}\big(\langle G_t, f_{\lambda_t}^{\phi} - f_t\rangle_K\big) \leqslant -\frac{\lambda_t}{2}\mathbb{E}_{z_1,z_2,\ldots,z_{t-1}}\big(\|f_t - f_{\lambda_t}^{\phi}\|_K^2\big).$$

By (4.3) and the definition of $N(\lambda)$, we have $|\phi'_-(y_t f_t(x_t))| \leqslant N(\lambda_t)$. Hence

$$\|G_t\|_K \leqslant \kappa N(\lambda_t) + \lambda_t \frac{\kappa|\phi'(0)|}{\lambda_t} \leqslant 2\kappa N(\lambda_t).$$

The last holds because $|\phi'(0)| \leqslant N(\lambda_t)$. Therefore, we get (2.10) from (4.5). □

Now we can analyze the remainder $f_{t+1} - f_{\lambda_t}^{\phi}$. Recall Definition 3 for the drift error d_t. For simplicity, denote $\prod_{j=T+1}^{T}(1 + d_j^s - \eta_j\lambda_j) = 1$ and $\sum_{j=T+1}^{T}(d_j^s - \eta_j\lambda_j) = 0$.

Lemma 6. *Let* $0 < s < 2$. *Assume that for some* $t_0 \in \mathbb{N}$, *the bound* (4.3) *and* $\eta_t\lambda_t < 1$ *hold for* $t \geqslant t_0$. *Then for* $T \geqslant t_0$, *we have*

$$\begin{aligned}\mathbb{E}_{z_1,\ldots,z_T}\big(\|f_{T+1} - f_{\lambda_T}^{\phi}\|_K^2\big) \leqslant{}& \prod_{t=t_0}^{T}\big(1 + d_t^s - \eta_t\lambda_t\big)\mathbb{E}_{z_1,\ldots,z_{t_0-1}}\big(\|f_{t_0} - f_{\lambda_{t_0-1}}^{\phi}\|_K^2\big)\\ &+ \sum_{t=t_0}^{T}\big(d_t^{2-s} + d_t^2 + \big(2\kappa N(\lambda_t)\eta_t\big)^2\big)\prod_{j=t+1}^{T}\big(1 + d_j^s - \eta_j\lambda_j\big).\end{aligned} \tag{4.6}$$

Proof. Let $t \geqslant t_0$. Recall that $\|f_{\lambda_{t-1}}^{\phi} - f_{\lambda_t}^{\phi}\|_K = d_t$. Then $\|f_t - f_{\lambda_t}^{\phi}\|_K^2 \leqslant \|f_t - f_{\lambda_{t-1}}^{\phi}\|_K^2 + 2\|f_t - f_{\lambda_{t-1}}^{\phi}\|_K d_t + d_t^2$. Applying the elementary inequality $2ab = 2ab^{s/2}b^{1-s/2} \leqslant a^2b^s + b^{2-s}$ $(a, b \geqslant 0)$ to $a = \|f_t - f_{\lambda_{t-1}}^{\phi}\|_K$ and $b = d_t$, we know that

$$\mathbb{E}_{z_1,z_2,\ldots,z_{t-1}}\big(\|f_t - f_{\lambda_t}^{\phi}\|_K^2\big) \leqslant \big(1 + d_t^s\big)\mathbb{E}_{z_1,z_2,\ldots,z_{t-1}}\big(\|f_t - f_{\lambda_{t-1}}^{\phi}\|_K^2\big) + d_t^{2-s} + d_t^2.$$

Using this bound to (2.10) and noticing the inequalities $(1 - \eta_t\lambda_t)(1 + d_t^s) \leqslant 1 + d_t^s - \eta_t\lambda_t$, $1 - \eta_t\lambda_t \leqslant 1$, we obtain

$$\mathbb{E}_{z_1,\ldots,z_t}\big(\|f_{t+1} - f_{\lambda_t}^{\phi}\|_K^2\big) \leqslant \big(1 + d_t^s - \eta_t\lambda_t\big)\mathbb{E}_{z_1,\ldots,z_{t-1}}\big(\|f_t - f_{\lambda_{t-1}}^{\phi}\|_K^2\big) + d_t^{2-s} + d_t^2 + \big(2\kappa N(\lambda_t)\eta_t\big)^2.$$

Applying this recursive relation iteratively for $t = T, T-1, \ldots, t_0$, we see that $\mathbb{E}_{z_1,\ldots,z_T}(\|f_{T+1} - f_{\lambda_T}^{\phi}\|_K^2)$ is bounded by (4.6). This proves the statement. □

5. Convergence rates for the strong approximation

Now we turn to the last step of getting convergence rates for the strong approximation. We need the following lemma which modifies an inequality given in [26] and [17].

Lemma 7. *Let $\nu > 0$, $0 < p_1 < p_2 < 1$, $T > t_0 \in \mathbb{N}$. We have*

$$\sum_{j=t_0}^{T} j^{-p_2} \exp\left\{-\nu \sum_{i=j+1}^{T} i^{-p_1}\right\} \leqslant C_{\nu,p_1,p_2} T^{p_1-p_2}, \tag{5.1}$$

where C_{ν,p_1,p_2} is the constant $C_{\nu,p_1,p_2} := 1 + \frac{6}{\nu} + \frac{2^{p_2+1}(1-p_1)}{e\nu(1-2^{p_1-1})(1-p_2)}$.

Proof. Denote $I = \sum_{j=t_0}^{T-1} j^{-p_2} \exp\{-\nu \sum_{i=j+1}^{T} i^{-p_1}\}$. Observe that for $u \in [i, i+1]$ we have $i^{-p_1} \geqslant u^{-p_1}$. Then $\sum_{i=j+1}^{T} i^{-p_1} \geqslant \int_{j+1}^{T+1} u^{-p_1}\,\mathrm{d}u = \frac{(T+1)^{1-p_1}}{1-p_1} - \frac{(j+1)^{1-p_1}}{1-p_1}$. It follows that

$$I \leqslant \exp\left\{-\frac{\nu}{1-p_1}(T+1)^{1-p_1}\right\} \sum_{j=t_0}^{T-1} j^{-p_2} \exp\left\{\frac{\nu(j+1)^{1-p_1}}{1-p_1}\right\}.$$

For $x \in [j+1, j+2]$, we have $(j+1)^{1-p_1} \leqslant x^{1-p_1}$ and $x \leqslant 3j$ which implies $(x/j)^{p_2} \leqslant 3^{p_2}$ and hence $j^{-p_2} \leqslant 3x^{-p_2}$. So we can bound $j^{-p_2} \exp\{\frac{\nu(j+1)^{1-p_1}}{1-p_1}\}$ by $\int_{j+1}^{j+2} 3x^{-p_2} \exp\{\frac{\nu x^{1-p_1}}{1-p_1}\}\,\mathrm{d}x$. Hence

$$I \leqslant \exp\left\{-\frac{\nu}{1-p_1}(T+1)^{1-p_1}\right\} \int_{t_0+1}^{T+1} \frac{3}{x^{p_2}} \exp\left\{\frac{\nu x^{1-p_1}}{1-p_1}\right\} \mathrm{d}x. \tag{5.2}$$

Decompose the above integral into two parts. For the part with large x on the interval $[(T+1)/2, T+1]$, we see from the bound $x^{p_1-p_2} \leqslant \left(\frac{T+1}{2}\right)^{p_1-p_2}$ that

$$\int_{(T+1)/2}^{T+1} \frac{3}{x^{p_2}} \exp\left\{\frac{\nu x^{1-p_1}}{1-p_1}\right\} \mathrm{d}x \leqslant 3\left(\frac{T+1}{2}\right)^{p_1-p_2} \int_{(T+1)/2}^{T+1} x^{-p_1} \exp\left\{\frac{\nu x^{1-p_1}}{1-p_1}\right\} \mathrm{d}x.$$

The last integral equals $\frac{1}{\nu}\exp\{\frac{\nu(T+1)^{1-p_1}}{1-p_1}\} - \frac{1}{\nu}\exp\{\frac{\nu((T+1)/2)^{1-p_1}}{1-p_1}\} \leqslant \frac{1}{\nu}\exp\{\frac{\nu(T+1)^{1-p_1}}{1-p_1}\}$. This in connection with the bound $2^{p_2-p_1} \leqslant 2$ yields

$$\int_{(T+1)/2}^{T+1} \frac{3}{x^{p_2}} \exp\left\{\frac{\nu x^{1-p_1}}{1-p_1}\right\} \mathrm{d}x \leqslant \frac{6}{\nu}(T+1)^{p_1-p_2} \exp\left\{\frac{\nu(T+1)^{1-p_1}}{1-p_1}\right\}. \tag{5.3}$$

For the part with small x on the interval $[t_0+1, (T+1)/2]$, we have

$$\int_{t_0+1}^{(T+1)/2} \frac{3}{x^{p_2}} \exp\left\{\frac{\nu x^{1-p_1}}{1-p_1}\right\} \mathrm{d}x \leqslant \exp\left\{\frac{\nu}{1-p_1}\left(\frac{T+1}{2}\right)^{1-p_1}\right\} \int_{t_0+1}^{(T+1)/2} \frac{3}{x^{p_2}}\,\mathrm{d}x$$

which is bounded (by computing the integral) by $\exp\{\frac{\nu(T+1)^{1-p_1}}{2^{1-p_1}(1-p_1)}\} \frac{3}{1-p_2}\left(\frac{T+1}{2}\right)^{1-p_2}$. This in connection with (5.2) and (5.3) implies

$$I \leqslant \frac{6}{\nu}(T+1)^{p_1-p_2} + \frac{3}{1-p_2}\left(\frac{T+1}{2}\right)^{1-p_2} \exp\left\{-\frac{\nu(1-2^{p_1-1})}{1-p_1}(T+1)^{1-p_1}\right\}.$$

Now for bounding the last term, we need an elementary inequality involving an arbitrary $c > 0$:

$$\exp\{-cx\} \leqslant \frac{1}{ec} x^{-1} \quad \forall x > 0. \tag{5.4}$$

This inequality is verified by considering the function $f(x) = x\exp\{-cx\}$ which is maximized at $x = \frac{1}{c}$. Choose $c = \frac{\nu(1-2^{p_1-1})}{1-p_1}$ and $x = (T+1)^{1-p_1}$ in (5.4). Then

$$\exp\left\{-\frac{\nu(1-2^{p_1-1})}{1-p_1}(T+1)^{1-p_1}\right\} \leqslant \frac{1-p_1}{e\nu(1-2^{p_1-1})}(T+1)^{p_1-1}.$$

Thus $I \leqslant \big(\frac{6}{\nu} + \frac{2^{(p_2+1)}(1-p_1)}{e\nu(1-2^{p_1-1})(1-p_2)}\big)(T+1)^{p_1-p_2}$ and the desired inequality follows. □

We are in a position to prove Theorem 2 about convergence rates of the error $\|f_{T+1} - f_{\lambda_T}^{\phi}\|_K$ stated in Section 2. Note that

$$1 - u \leqslant e^{-u} \quad \forall u \in \mathbb{R}. \tag{5.5}$$

Proof of Theorem 2. We first claim that (4.2) holds. This follows from Lemma 3 together with the restrictions (2.7) on η_1 and $p\gamma \leqslant \alpha$:

$$\begin{aligned}\eta_t\big(\kappa^2 M(\lambda_t) + \lambda_t\big) &\leqslant \eta_1 t^{-\alpha}\big(\kappa^2\big(M_0 + 2N_1\lambda_t^{-p}\big) + \lambda_t\big)\\ &\leqslant \eta_1 t^{-\alpha}\big(\kappa^2 M_0 + 2\kappa^2 N_1\lambda_1^{-p}t^{p\gamma} + \lambda_1 t^{-\gamma}\big) \leqslant \eta_1\big(\kappa^2 M_0 + 2\kappa^2 N_1\lambda_1^{-p} + \lambda_1\big) \leqslant 1.\end{aligned}$$

Then by Proposition 2, we see that $\|f_t\|_K \leqslant \frac{\kappa|\phi'(0)|}{\lambda_t}$ for each $t \geqslant 1$.

The restrictions (2.7) on η_1 gives $\eta_1\lambda_1 < 1$, and hence $\eta_t\lambda_t < 1$ for $t \geqslant 1$. In Lemma 6 we take $t_0 = 1$ and according to a requirement (2.13) and (2.14) we choose

$$s = \begin{cases} 2 - \frac{2\alpha - 2p\gamma}{1-\gamma(1-\beta)/2}, & \text{if } 2(1-\frac{\gamma(1-\beta)}{2}) - (2\alpha - 2p\gamma) > \alpha + \gamma,\\ \frac{\alpha+\gamma+\epsilon}{1-\gamma(1-\beta)/2}, & \text{if } 2(1-\frac{\gamma(1-\beta)}{2}) - (2\alpha - 2p\gamma) \leqslant \alpha + \gamma,\end{cases} \tag{5.6}$$

where in the second case, ϵ is an arbitrary number satisfying $0 < \epsilon < 2(1-\frac{\gamma(1-\beta)}{2}) - 2(\alpha+\gamma)$. It can be easily seen that in either case, (2.13) holds and $s < 2$. Then the error bound (4.6) holds for $T > 1$. But $f_{t_0} = f_{\lambda_{t_0-1}}^{\phi} = 0$. So (4.6) becomes

$$\mathbb{E}_{z_1,\ldots,z_T}\big(\big\|f_{T+1} - f_{\lambda_T}^{\phi}\big\|_K^2\big) \leqslant \sum_{t=1}^{T}\big(d_t^{2-s} + d_t^2 + \big(2\kappa N(\lambda_t)\eta_t\big)^2\big)\prod_{j=t+1}^{T}\big(1 + d_j^s - \eta_j\lambda_j\big). \tag{5.7}$$

Apply Lemma 1, we see that

$$d_j^s \leqslant \big(16\mathcal{D}_0\lambda_1^{\beta-1}\big)^{s/2} \leqslant 1 + 16\mathcal{D}_0\lambda_1^{\beta-1} \quad \forall j \in \mathbb{N}. \tag{5.8}$$

Use this uniform bound for small j and then apply (2.13) and Lemma 1. We see that

$$d_j^s \leqslant 4^s\big(\mathcal{D}_0\lambda_1^{\beta-1}\big)^{s/2} j^{s(\frac{\gamma(1-\beta)}{2}-1)} \leqslant \frac{1}{2}\eta_j\lambda_j = \frac{1}{2}\eta_1\lambda_1 j^{-(\alpha+\gamma)}$$

for every $j \geqslant J_s$, where J_s is a positive integer satisfying

$$J_s^{s(1-\frac{\gamma(1-\beta)}{2})-(\alpha+\gamma)} \geqslant 2^{2s+1}\big(\mathcal{D}_0\lambda_1^{\beta-1}\big)^{s/2}/(\eta_1\lambda_1).$$

One can take J_s to be the smallest integer greater than $((32 + 32\mathcal{D}_0\lambda_1^{\beta-1})/(\eta_1\lambda_1))^{\frac{1}{\tau}}$ with

$$\tau = \begin{cases} 2 - 3\alpha - (2-2p-\beta)\gamma, & \text{if } 2(1-\frac{\gamma(1-\beta)}{2}) - (2\alpha - 2p\gamma) > \alpha + \gamma,\\ \epsilon, & \text{if } 2(1-\frac{\gamma(1-\beta)}{2}) - (2\alpha - 2p\gamma) \leqslant \alpha + \gamma.\end{cases} \tag{5.9}$$

Since $1 - \frac{1}{2}\eta_j\lambda_j \geqslant \frac{1}{2}$, we know that for any $1 \leqslant t < J_s$,

$$\prod_{j=t+1}^{T}\big(1 + d_j^s - \eta_j\lambda_j\big) \leqslant \prod_{j=t+1}^{J_s}\big(1 + d_j^s\big)\prod_{j=J_s+1}^{T}\left(1 - \frac{1}{2}\eta_j\lambda_j\right) \leqslant 2^{J_s}\prod_{j=t+1}^{J_s}\big(1 + d_j^s\big)\prod_{j=t+1}^{T}\left(1 - \frac{1}{2}\eta_j\lambda_j\right).$$

Applying (5.8) for $t+1 \leqslant j \leqslant J_s$, the first term above can be bounded as $\prod_{j=t+1}^{J_s}(1+d_j^s) \leqslant (2+16\mathcal{D}_0\lambda_1^{\beta-1})^{J_s}$. This in connection with (5.5) implies that for any $t \geqslant 1$,

$$\prod_{j=t+1}^{T}\left(1+d_j^s-\eta_j\lambda_j\right) \leqslant \left(4+32\mathcal{D}_0\lambda_1^{\beta-1}\right)^{J_s}\exp\left\{-\frac{1}{2}\eta_1\lambda_1\sum_{j=t+1}^{T} j^{-(\alpha+\gamma)}\right\}. \tag{5.10}$$

It provides an estimate for the last part of (5.7).

Next we estimate the other part of (5.7). For $t \geqslant 1$, we have

$$N(\lambda_t)\eta_t \leqslant N_1\left(\lambda_1 t^{-\gamma}\right)^{-p}\eta_1 t^{-\alpha} \leqslant N_1\lambda_1^{-p}\eta_1 t^{p\gamma-\alpha}.$$

Thus, $d_t^{2-s}+d_t^2+(2\kappa N(\lambda_{t-1})\eta_t)^2$ is bounded by

$$\left(4\sqrt{\mathcal{D}_0\lambda_1^{\beta-1}}\right)^{2-s}t^{(2-s)(\frac{\gamma(1-\beta)}{2}-1)}+16\mathcal{D}_0\lambda_1^{\beta-1}t^{\gamma(1-\beta)-2}+\left(2\kappa N_1\lambda_1^{-p}\eta_1\right)^2 t^{2p\gamma-2\alpha}.$$

According to (5.6), we see that the first term above dominates since

$$\begin{cases}(2-s)(1-\frac{\gamma(1-\beta)}{2})=2\alpha-2p\gamma, & \text{if } 2(1-\frac{\gamma(1-\beta)}{2})-(2\alpha-2p\gamma)>\alpha+\gamma,\\ (2-s)(1-\frac{\gamma(1-\beta)}{2})<2\alpha-2p\gamma, & \text{if } 2(1-\frac{\gamma(1-\beta)}{2})-(2\alpha-2p\gamma)\leqslant\alpha+\gamma.\end{cases}$$

Therefore, if we denote $p_2=(2-s)(1-\frac{\gamma(1-\beta)}{2})$ and the constant

$$\tilde{C}_s' := 16+32\mathcal{D}_0\lambda_1^{\beta-1}+\left(2\kappa N_1\lambda_1^{-p}\eta_1\right)^2,$$

then $d_t^{2-s}+d_t^2+(2\kappa N(\lambda_{t-1})\eta_t)^2 \leqslant \tilde{C}_s' t^{-p_2}$. This in connection with (5.7) and (5.10) yields

$$\mathbb{E}_{z_1,\dots,z_T}\left(\|f_{T+1}-f_{\lambda_T}^{\phi}\|_K^2\right) \leqslant \left(4+32\mathcal{D}_0\lambda_1^{\beta-1}\right)^{J_s}\tilde{C}_s'\sum_{t=1}^{T}t^{-p_2}\exp\left\{-\frac{1}{2}\eta_1\lambda_1\sum_{j=t+1}^{T}j^{-(\alpha+\gamma)}\right\}. \tag{5.11}$$

Observe from the choice (5.6) of s that

$$p_2-(\alpha+\gamma)=\begin{cases}\alpha-(2p+1)\gamma, & \text{if } 2(1-\frac{\gamma(1-\beta)}{2})-(2\alpha-2p\gamma)>\alpha+\gamma,\\ 2-\gamma(1-\beta)-2(\alpha+\gamma)-\epsilon, & \text{if } 2(1-\frac{\gamma(1-\beta)}{2})-(2\alpha-2p\gamma)\leqslant\alpha+\gamma.\end{cases}$$

This power exponent is positive and such an α exists because the restriction $(5+4p-\beta)\gamma<2$ ensures $(2p+1)\gamma<\alpha<\frac{2+(2p-2+\beta)\gamma}{3}$ in the first case, and $\frac{2+(2p-2+\beta)\gamma}{3}\leqslant\alpha<1-\frac{(3-\beta)\gamma}{2}$ with $0<\epsilon<2-\gamma(1-\beta)-2(\alpha+\gamma)$ in the second case. Then by applying Lemma 7, we see that

$$\mathbb{E}_{z_1,\dots,z_T}\left(\|f_{T+1}-f_{\lambda_T}^{\phi}\|_K^2\right) \leqslant \left(4+32\mathcal{D}_0\lambda_1^{\beta-1}\right)^{J_s}\tilde{C}_s' C_{\frac{1}{2}\eta_1\lambda_1,\alpha+\gamma,p_2}T^{\alpha+\gamma-p_2}. \tag{5.12}$$

This completes the proof of Theorem 2 since $\alpha+\gamma-p_2=-\theta$. □

6. Total error bounds and learning rates

To demonstrate our method, let us first prove Example 1.

Proof of Example 1. First we derive the decay (2.2). Observe that the hinge loss ϕ is uniformly Lipschitz satisfying $|\phi(u)-\phi(v)| \leqslant |u-v|$ for any $u, v \in \mathbb{R}$. Then for any functions f, g on X,

$$\left|\mathcal{E}(f)-\mathcal{E}(g)\right| = \left|\int_Z \phi\left(yf(x)\right)-\phi\left(yg(x)\right)\mathrm{d}\rho\right| \leqslant \|f-g\|_{L^1_{\rho_X}}. \tag{6.1}$$

This in connection with the assumption (1.10) and the fact that $f_\rho^\phi = f_c$ verified in [24] implies that (2.2) holds true.

Then we apply Theorem 2 to estimate $\|f_{T+1}-f_{\lambda_T}^{\phi}\|_K^2$. We only need to determine the indices. By (2.15) we have $\gamma=\frac{1}{4}$ and $\alpha=\frac{1}{2}-\epsilon$. The left derivative ϕ_-' of the hinge loss is given by $\phi_-'(x)=-1$ for $x \leqslant 1$ and 0 for

$x > 1$. So $N(\lambda) \equiv 1$ and (2.5) holds with $N_1 = 1$ and $p = 0$. Moreover, $M_0 = 0$. So the restriction (2.7) is the same as $0 < \eta_1 \leqslant \frac{1}{2\kappa^2+\lambda_1}$. Also, we have $\gamma = \frac{1}{4} < \frac{2}{5+4p-\beta}$. The index $\alpha = \frac{1}{2} - \epsilon$ corresponds to the first case of (2.9). Therefore, by the bound (2.8) in Theorem 2, we have

$$\mathbb{E}_{z_1,\dots,z_T}\big(\big\|f_{T+1} - f_{\lambda_T}^{\phi}\big\|_K^2\big) \leqslant C_{\eta_1,\lambda_1,\kappa,p,\mathcal{D}_0,\beta} T^{\epsilon-\frac{1}{4}}. \tag{6.2}$$

According to (1.1) and (6.1), we have $\mathcal{E}(f_{T+1}) - \mathcal{E}(f_{\lambda_T}^{\phi}) \leqslant \|f_{T+1} - f_{\lambda_T}^{\phi}\|_\infty \leqslant \kappa\|f_{T+1} - f_{\lambda_T}^{\phi}\|_K$. It follows from the Schwarz inequality and (6.2) that

$$\mathbb{E}_{z_1,\dots,z_T}\big(\mathcal{E}(f_{T+1}) - \mathcal{E}\big(f_{\lambda_T}^{\phi}\big)\big) \leqslant \kappa\sqrt{\mathbb{E}_{z_1,\dots,z_T}\big(\big\|f_{T+1} - f_{\lambda_T}^{\phi}\big\|_K^2\big)} \leqslant \kappa\sqrt{C_{\eta_1,\lambda_1,\kappa,p,\mathcal{D}_0,\beta}}\, T^{\frac{\epsilon}{2}-\frac{1}{8}}.$$

Also, we have $\mathcal{E}(f_{\lambda_T}^{\phi}) - \mathcal{E}(f_c) \leqslant \mathcal{D}(\lambda_T)$ since f_c is a minimizer of $\mathcal{E}(f)$. Thus we get a bound for the excess generalization error

$$\mathbb{E}_{z_1,\dots,z_T}\big(\mathcal{E}(f_{T+1}) - \mathcal{E}(f_c)\big) \leqslant \kappa\sqrt{C_{\eta_1,\lambda_1,\kappa,p,\mathcal{D}_0,\beta}}\, T^{\frac{\epsilon}{2}-\frac{1}{8}} + \mathcal{D}_0\lambda_1^{\beta} T^{-\frac{\beta}{4}} \leqslant C_\epsilon T^{-\min\{\frac{\beta}{4},\frac{1}{8}-\frac{\epsilon}{2}\}},$$

where $C_\epsilon = \kappa\sqrt{C_{\eta_1,\lambda_1,\kappa,p,\mathcal{D}_0,\beta}} + \mathcal{D}_0\lambda_1^{\beta}$.

An important relation concerning the hinge loss is the one [27] between the excess misclassification error and the excess generalization error given for any measurable function $f : X \to \mathbb{R}$ as

$$\mathcal{R}\big(\operatorname{sgn}(f)\big) - \mathcal{R}(f_c) \leqslant \mathcal{E}(f) - \mathcal{E}(f_c). \tag{6.3}$$

Then our conclusion of Example 1 follows. □

Turn to the general loss ϕ. To derive rates for the excess misclassification error, we show how the strong convergence of f_{T+1} to $f_{\lambda_T}^{\phi}$ and the regularization error yield the excess generalization error.

Lemma 8. *Under the assumptions of Theorem* 2 *and* θ *given by* (2.9)*, if* $\gamma < \frac{2}{5+10p-\beta}$*, then*

$$\mathbb{E}_{z_1,\dots,z_T}\big(\mathcal{E}(f_{T+1}) - \mathcal{E}\big(f_\rho^{\phi}\big)\big) \leqslant \tilde{C} T^{-\min\{\beta\gamma,\ \frac{\theta}{2}-p\gamma\}},$$

where $\tilde{C}$ *is a constant depending on* $\eta_1, \lambda_1, \kappa, p, \mathcal{D}_0, \beta$ *and* N_1.

Proof. Decompose $\mathcal{E}(f_{T+1}) - \mathcal{E}(f_\rho^{\phi})$ as $\mathcal{E}(f_{T+1}) - \mathcal{E}(f_{\lambda_T}^{\phi}) + \mathcal{E}(f_{\lambda_T}^{\phi}) - \mathcal{E}(f_\rho^{\phi})$. Observe that $\mathcal{E}(f_{\lambda_T}^{\phi}) - \mathcal{E}(f_\rho^{\phi}) \leqslant \mathcal{D}(\lambda_T) \leqslant \mathcal{D}_0\lambda_T^{\beta}$.

By Theorem 2, $\|f_{T+1}\|_\infty \leqslant \kappa\|f_{T+1}\|_K \leqslant \frac{\kappa^2|\phi'(0)|}{\lambda_{T+1}}$. From the definition of $\mathcal{D}(\lambda_T)$, we see that

$$\big\|f_{\lambda_T}^{\phi}\big\|_\infty \leqslant \kappa\big\|f_{\lambda_T}^{\phi}\big\|_K \leqslant \kappa\sqrt{2\mathcal{D}(\lambda_T)/\lambda_T} \leqslant \kappa\sqrt{2\mathcal{D}_0\lambda_T^{\beta}/\lambda_T} = \frac{\sqrt{2\mathcal{D}_0}\lambda_1^{\frac{\beta+1}{2}} T^{-\frac{\gamma(\beta+1)}{2}}}{\kappa|\phi'(0)|}\,\frac{\kappa^2|\phi'(0)|}{\lambda_T}.$$

It follows that $\|f_{\lambda_T}^{\phi}\|_\infty \leqslant \frac{\kappa^2|\phi'(0)|}{\lambda_T} \leqslant \frac{\kappa^2|\phi'(0)|}{\lambda_{T+1}}$ when $T \geqslant \lambda_1^{\frac{1}{\gamma}}\big(\frac{\sqrt{2\mathcal{D}_0}}{\kappa|\phi'(0)|}\big)^{\frac{2}{\gamma(\beta+1)}}$. Under this restriction of T, we know that for any $x \in X$ and $y \in Y$, the numbers $yf_{T+1}(x)$ and $yf_{\lambda_T}^{\phi}(x)$ are both bounded by $\frac{\kappa^2|\phi'(0)|}{\lambda_{T+1}}$. Thus by the definition of $N(\lambda_{T+1})$, we have

$$\big|\phi\big(yf_{T+1}(x)\big) - \phi\big(yf_{\lambda_T}^{\phi}(x)\big)\big| \leqslant N(\lambda_{T+1})\big|yf_{T+1}(x) - yf_{\lambda_T}^{\phi}(x)\big| \leqslant \kappa N(\lambda_{T+1})\big\|f_{T+1} - f_{\lambda_T}^{\phi}\big\|_K.$$

This in connection with the assumption $N(\lambda_{T+1}) \leqslant N_1\lambda_1^{-p}(T+1)^{p\gamma} \leqslant 2^{p\gamma}N_1\lambda_1^{-p}T^{p\gamma}$ implies

$$\mathcal{E}(f_{T+1}) - \mathcal{E}\big(f_{\lambda_T}^{\phi}\big) = \int_Z \phi\big(yf_{T+1}(x)\big) - \phi\big(yf_{\lambda_T}^{\phi}(x)\big)\,\mathrm{d}\rho \leqslant \kappa 2^{p\gamma}N_1\lambda_1^{-p}T^{p\gamma}\big\|f_{T+1} - f_{\lambda_T}^{\phi}\big\|_K.$$

Therefore, by Theorem 2, we have

$$\mathbb{E}_{z_1,\dots,z_T}\big(\mathcal{E}(f_{T+1}) - \mathcal{E}\big(f_\rho^{\phi}\big)\big) \leqslant \kappa 2^p N_1\lambda_1^{-p}\sqrt{C_{\eta_1,\lambda_1,\kappa,p,\mathcal{D}_0,\beta}}\, T^{p\gamma-\frac{\theta}{2}} + \mathcal{D}_0\lambda_1^{\beta}T^{-\beta\gamma}.$$

The requirement $\frac{\theta}{2} - p\gamma > 0$ restricts γ by $(5 + 10p - \beta)\gamma < 2$. This verifies the desired bound for the excess generalization error. □

We can now prove the learning rates for the online algorithm (1.8) with the hinge loss.

Proof of Corollary 2. The assumption (1.10) in connection with (6.1) implies $\mathcal{D}(\lambda) \leqslant \mathcal{D}_0 \lambda^\beta$. Note that (2.5) holds with $p = 0$. Then our conclusion follows from (6.3) and Lemma 8 by considering the first case of (2.9). □

Proof of Theorem 3. Since ϕ is an admissible loss function with $\phi''(0) > 0$, it was shown in [3] that there exists a constant depending c_ϕ only on ϕ such that for any measurable function $f : X \to \mathbb{R}$,

$$\mathcal{R}\big(\operatorname{sgn}(f)\big) - \mathcal{R}(f_c) \leqslant c_\phi \sqrt{\mathcal{E}(f) - \mathcal{E}\big(f_\rho^\phi\big)}. \tag{6.4}$$

Then the stated error bound follows from Lemma 8. This proves Theorem 3. □

To demonstrate further, we apply our main results to the q-norm SVM loss $\phi(x) = (1-x)_+^q$ with $q > 1$. It satisfies $\phi''(0) > 0$. According to the expression of $N(\lambda)$, we see that $p = q - 1$. So Theorem 3 yields the following learning rates.

Corollary 3. *Let $\phi(x) = (1-x)_+^q$ with $q > 1$. Assume* (2.2) *for the pair (K, ρ). Let $\gamma < \frac{2}{10q-5-\beta}$ and $(2q-1)\gamma < \alpha < \frac{2+(2q-4+\beta)\gamma}{3}$. If $\eta_1 \leqslant \frac{1}{4\kappa^2 q^2 + 2\kappa^2 q (1+\kappa^2 q/\lambda_1)^{q-1} + \lambda_1}$, then*

$$\mathbb{E}_{z_1,\dots,z_T}\big(\mathcal{R}\big(\operatorname{sgn}(f_{T+1})\big) - \mathcal{R}(f_c)\big) = O\big(T^{-\min\{\frac{\beta\gamma}{2}, \frac{\alpha-(4q-3)\gamma}{4}\}}\big).$$

Finally, let us state a general result for the strong approximation. The proof follows from Lemma 6, as that for Theorem 2.

Proposition 3. *Let the pairs $\{(\eta_t, \lambda_t)\}_{t=1}^\infty$ satisfy*

$$\sum_{t=1}^\infty \eta_t \lambda_t = +\infty, \quad \lim_{t\to\infty} \eta_t = 0, \ \lim_{t\to\infty} \lambda_t = 0$$

and the following restrictions (with respect to $\{d_t, N(\lambda_t)\}$)

$$\lim_{t\to\infty} \frac{d_t}{\eta_t \lambda_t} = 0, \qquad \lim_{t\to\infty} \frac{\eta_t (N(\lambda_t))^2}{\lambda_t} = 0.$$

If ϕ'_- is locally Lipschitz at the origin and $\eta_t(\kappa^2 M(\lambda_t) + \lambda_t) \leqslant 1$ for all $t \geqslant 1$, then

$$\mathbb{E}_{z_1,\dots,z_T}\big(\big\| f_{T+1} - f_{\lambda_T}^\phi \big\|_K\big) \to 0 \quad \text{as } T \to +\infty.$$

Appendix A. Proof of Proposition 2

We prove Proposition 2 by induction on $t \in \mathbb{N}$.

The case $t = 1$ is trivial since $f_1 = 0$. Suppose that (4.3) holds for t. Consider f_{t+1}. It can be expressed as

$$f_{t+1} = (1 - \eta_t \lambda_t) f_t - \eta_t \big[\phi'_-\big(y_t f_t(x_t)\big) - \phi'(0)\big] y_t K_{x_t} - \eta_t \phi'(0) y_t K_{x_t}.$$

Write the middle term as

$$\big[\phi'_-\big(y_t f_t(x_t)\big) - \phi'(0)\big] y_t K_{x_t} = \frac{\phi'_-(y_t f_t(x_t)) - \phi'(0)}{y_t f_t(x_t)} L_t f_t,$$

where $L_t : \mathcal{H}_K \to \mathcal{H}_K$ is a self-adjoint, rank-one, positive linear operator given by $L_t g = \langle g, K_{x_t}\rangle_K K_{x_t}$ and we have used the reproducing property for $f_t(x_t) K_{x_t} = \langle f_t, K_{x_t}\rangle_K K_{x_t}$. The operator norm of L_t can be bounded as $\|L_t\|_{\mathcal{H}_K \to \mathcal{H}_K} \leqslant \kappa^2$ since

$$\langle L_t g, g\rangle_K = \big|\langle g, K_{x_t}\rangle_K\big|^2 \leqslant \kappa^2 \|g\|_K^2 \quad \forall g \in \mathcal{H}_K.$$

The local Lipschitz condition tells us that $\frac{\phi'_-(y_t f_t(x_t))-\phi'(0)}{y_t f_t(x_t)}$ is well defined (set as zero when $f_t(x_t)=0$). It is bounded by $M(\lambda_t)$, since $|y_t f_t(x_t)| \leqslant \kappa \|f_t\|_K \leqslant \kappa^2 \frac{|\phi'(0)|}{\lambda_t}$ by our induction hypothesis. The convexity of ϕ implies that ϕ'_- is nondecreasing. Hence

$$0 \leqslant \frac{\phi'_-(y_t f_t(x_t)) - \phi'(0)}{y_t f_t(x_t)} \leqslant M(\lambda_t).$$

Therefore, $\frac{\phi'_-(y_t f_t(x_t))-\phi'(0)}{y_t f_t(x_t)} L_t$ is a self-adjoint, positive linear operator on $\mathcal{H}_K$ and its norm is bounded by $\kappa^2 M(\lambda_t)$. Since $\eta_t \kappa^2 M(\lambda_t) \leqslant 1-\eta_t\lambda_t$, the linear operator $T_t := (1-\eta_t\lambda_t)I - \eta_t \frac{\phi'_-(y_t f_t(x_t))-\phi'(0)}{y_t f_t(x_t)} L_t$ is self-adjoint, positive and $T_t \leqslant (1-\eta_t\lambda_t)I$. It follows that

$$\big\|(1-\eta_t\lambda_t) f_t - \eta_t\big[\phi'_-\big(y_t f_t(x_t)\big) - \phi'(0)\big] y_t K_{x_t}\big\|_K = \|T_t f_t\| \leqslant (1-\eta_t\lambda_t)\|f_t\|_K.$$

Hence

$$\|f_{t+1}\|_K \leqslant (1-\eta_t\lambda_t)\|f_t\|_K + \kappa\eta_t\big|\phi'(0)\big|.$$

This in connection with the induction on $\|f_t\|_K$ implies that

$$\|f_{t+1}\|_K \leqslant (1-\eta_t\lambda_t)\frac{\kappa|\phi'(0)|}{\lambda_t} + \kappa\eta_t\big|\phi'(0)\big| \leqslant \frac{\kappa|\phi'(0)|}{\lambda_{t+1}}.$$

Then the conclusion of Proposition 2 follows. □

References

[1] N. Aronszajn, Theory of reproducing kernels, Trans. Amer. Math. Soc. 68 (1950) 337–404.
[2] P.L. Bartlett, M.I. Jordan, J.D. McAuliffe, Convexity, classification, and risk bounds, J. Amer. Statist. Assoc. 101 (2006) 138–156.
[3] D.R. Chen, Q. Wu, Y. Ying, D.X. Zhou, Support vector machine soft margin classifiers: Error analysis, J. Mach. Learn. Res. 5 (2004) 1143–1175.
[4] N. Cesa-Bianchi, P. Long, M.K. Warmuth, Worst-case quadratic loss bounds for prediction using linear functions and gradient descent, IEEE Trans. Neural Networks 7 (1996) 604–619.
[5] N. Cesa-Bianchi, A. Conconi, C. Gentile, On the generalization ability of on-line learning algorithms, IEEE Trans. Inform. Theory 50 (2004) 2050–2057.
[6] E. De Vito, A. Caponnetto, L. Rosasco, Model selection for regularized least-squares algorithm in learning theory, Found. Comput. Math. 5 (2005) 59–85.
[7] L. Devroye, L. Györfi, G. Lugosi, A Probabilistic Theory of Pattern Recognition, Springer-Verlag, New York, 1997.
[8] T. Evgeniou, M. Pontil, T. Poggio, Regularization networks and support vector machines, Adv. Comput. Math. 13 (2000) 1–50.
[9] J. Forster, M.K. Warmuth, Relative expected instantaneous loss bounds, J. Comput. Syst. Sci. 64 (2002) 76–102.
[10] J. Kivinen, A.J. Smola, R.C. Williamson, Online learning with kernels, IEEE Trans. Signal Process. 52 (2004) 2165–2176.
[11] Y. Lin, Support vector machines and the Bayes rule in classification, Data Min. Knowl. Discov. 6 (2002) 259–275.
[12] G. Lugosi, N. Vayatis, On the Bayes-risk consistency of regularized boosting methods, Ann. Statist. 32 (2004) 30–55.
[13] S. Mukherjee, R. Rifkin, T. Poggio, Regression and classification with regularization, in: D.D. Denison, M.H. Hansen, C. Holmes, B. Mallick, B. Yu (Eds.), Nonlinear Estimation and Classification, Springer-Verlag, New York, 2002, pp. 107–124.
[14] A. Rakhlin, S. Mukherjee, T. Poggio, Stability results in learning theory, Anal. Appl. 3 (2005) 397–417.
[15] F. Rosenblatt, Principles of Neurodynamics, Spartan Book, New York, 1962.
[16] C. Scovel, I. Steinwart, Fast rates for support vector machines, in: Proc. 18th Conf. on Learning Theory (COLT-2005), Bertinoro, Italy, 2005.
[17] S. Smale, Y. Yao, Online learning algorithms, Found. Comput. Math. 6 (2006) 145–170.
[18] S. Smale, D.X. Zhou, Estimating the approximation error in learning theory, Anal. Appl. 1 (2003) 17–41.
[19] S. Smale, D.X. Zhou, Shannon sampling and function reconstruction from point values, Bull. Amer. Math. Soc. 41 (2004) 279–305.
[20] S. Smale, D.X. Zhou, Shannon sampling II: Connections to learning theory, Appl. Comput. Harmon. Anal. 19 (2005) 285–302.
[21] S. Smale, D.X. Zhou, Learning theory estimates via integral operators and their applications, Constr. Approx., in press.
[22] I. Steinwart, Support vector machines are universally consistent, J. Complexity 18 (2002) 768–791.
[23] P. Tarrès, Y. Yao, Online learning as stochastic approximations of regularization paths, preprint, 2005.
[24] G. Wahba, Spline Models for Observational Data, SIAM, Philadelphia, 1990.
[25] Q. Wu, Y. Ying, D.X. Zhou, Multi-kernel regularized classifiers, J. Complexity, in press.
[26] Y. Ying, D.X. Zhou, Online regularized classification algorithms, IEEE Trans. Inform. Theory 52 (2006) 4775–4788.
[27] T. Zhang, Statistical behavior and consistency of classification methods based on convex risk minimization, Ann. Statist. 32 (2004) 56–85.