系数正则化在线分类算法收敛性分析

田明党[1]，盛宝怀[2]

（1. 宁波大学理学院，浙江 宁波 315211;

2. 绍兴文理学院数理信息学院，浙江 绍兴 312000）

摘要：文章研究基于凸损失函数的系数正则化在线分类学习算法问题，给出了基于欧式空间的一种不依赖于样本容量的方法并且详细给出了该算法的误差分析过程。该分类算法的目的是构造一个不依赖于样本容量和样本概率分布的分类器来学习未知概率分布的样本空间。文章给出了基于欧式空间的梯度下降算法的具体产生过程，该方法的优点是对于样本容量较大的样本空间同样可以有效的构造分类器进行分类。文章对由梯度下降算法产生的学习序列进行了界的限定，在此过程中要求损失函数在原点是满足李普希茨条件的，并且对步长的具体形式也做了要求，再次基础上明确的给出了其错分类误差。最后，以铰链损失函数为例给出了该算法的误差界的分析.

关键词： 分类算法；在线学习；系数正则；误差分析

中图分类号：O212.2

Convergence Analysis of Coefficient Regularized Classification Online Algorithms

Tian Mingdang[1], Sheng Baohuai[2]

(1. Department of Science, Ningbo University, Zhejiang Ningbo 315211;

2. Department of Mathmatics, Shaoxing University, Zhejiang Shaoxing 312000)

Abstract: The present paper considers online classification learning algorithms associated with convex loss functions, and gives the error analysis of coefficient regularized online classification algorithms. The goal of classification is to consruct, on the basis of independent and identically distribution samples, a classifier which can predict the unknown distribution with small missclassification error. A novel capacity independent approach based on a Euclean Space is presented. It designs a gradient descent online learning algorithm, which is suitable for large size samples. It shows how a local Lipschitz condition on loss function at the origin and some restrictions on step size to ensure the uniform boundedness of learning sequence, a crucial assumption for the convergence of the online scheme states in the present paper. Explicit learning rates with respect to the misclassification error are given in terms of the choice of step sizes and the regularization parameter. Error bounds associated with the hinge loss is presented to illustrate the method.

Key words: Classification algorithm; online learning; coefficient regularization; error analysis

0 Introduction

Support vector machines (SVMs)[1] classification was introduced by Boser. Now, SVMs and related regularized methods have formed an important part of learning theory. They have been applied successfully to various practical problem in science and engineering, especially for classification problem. There are also many results and theories which can provide us to learn more knowledge about this aspect. Binary classification is one of the central application for machine learning methods in general and for SVMs in particular.

Let X be a compact metric space and $Y=\{1,-1\}$. A function $C:X\to Y$ is called a classifier if it divides the input space X into two classes. A real valued function $f:X\to R$

Brief author introduction:田明党、1983、女、硕士研究生、主要研究方向为学习理论. E-mail: tianjing2213@126.com can be used to generate a classifier $C(x)=\operatorname{sgn}(f(x))$, where the sign function is defined as

$\operatorname{sgn}(f(x))=1$, if $f(x)\geq 0$ and $\operatorname{sgn}(f(x))=-1$ for $f(x)<0$. For such a real valued

function f, a loss function $\phi:R\to R_+$ is often used to measure the error: $\phi(yf(x))$, which is the local error at the point (x,y), while $\operatorname{sgn}(f(x))\in Y$ is assigned to the event $x\in X$.

A Reproducing Kernel Hilbert Space (RKHS) H_K is defined as the completion of the span of $\{K_x=K(\cdot,x):x\in X\}$ with the inner product $\langle K_x, K_y\rangle_K = K(x,y)$,

where K is a Mercer kernel on $X\times X$, which is continuous, symmetric, and positive definite. The reproducing property $f(x)=\langle f,K_x\rangle_K$ holds for every $f\in H_K$. For more properties of H_K see [2].

Assume that ρ is an unknown probability distribution on $Z=X\times Y$ and $z=\{z_t =(x_t,y_t)\}_{t=1}^T\in Z^T$ is a set of random samples independently drawn according to ρ. The batch learning algorithms for classification is implemented by an off-line regularization scheme[3] in a reproducing kernel space involving sample z and a loss function ϕ:

$$f_{z,\lambda}=\arg\min_{f\in H_K}\{\frac{1}{T}\sum_{t=1}^{T}\phi(y_t f(x_t))+\frac{\lambda}{2}\|f\|_K^2\}, \tag{0.1}$$

where $\lambda>0$ is the regularizer parameter.

By [4, 5] we know $f_{\alpha_{z,\lambda}}$ has changed into the form of

$$f_{\alpha_{z,\lambda}}(x)=\sum_{t=1}^{m}\alpha_t K(x,x_t),\ x\in z,$$

(0.1) then becomes an infinite dimensional optimization problem on R^m. In particular, now we considered the following coefficient regularized algorithm[6]

$$\alpha_{z,\lambda}=\arg\min_{\alpha\in R^m}\{\frac{1}{T}\sum_{t=1}^{T}\phi(y_t f_\alpha(x_t))+\frac{\lambda}{2}\|\alpha\|_2^2\}, \tag{0.2}$$

where $f_\alpha(x)=\sum_{t=1}^{m}\alpha_t K(x,x_t)$ and $\|\alpha\|_2^2=\sum_{t=1}^{m}\alpha_t^2$.

Since $f_\alpha(x)$ is determined by its coefficients and the penalty is imposed on these coefficients. We call this regularization technique the coefficient regularization[3]. It was first introduced by Vapnik to design linear programming support vector machines.

Online algorithm with linear complexity $O(T)$ can be applied and provide efficient classifier, when the sample size is large. It is known that there are many online algorithms to solve the classification problems (0.1). For example, there is a kind of fully online algorithm[7] which needs that the regularization parameters λ_t changes with the learning step t. These algorithms require that the loss function is derivable near the origin to maintain the existence of the gradient. The gradient descent algorithm for binary classification[8] improves the learning rates by the empirical covering number and Rademacher average. Yiming Ying and Dingxuan Zhou gave a online algorithms[9] based on regularized schemes in reproducing kernel Hilbert space. In the present paper, we shall design a gradient descent online learning algorithm in the space R^m to show the error analysis for framework (0.2).

By the sample error analysis[10] of learning theory, we know that $\alpha_{z,\lambda}$ has an asymptotic behavior to the regularization function $\alpha_\lambda \in R^m$ defined by

$$\alpha_\lambda := \arg\min_{\alpha\in R^m}\{\varepsilon(f_\alpha)+\frac{\lambda}{2}\|\alpha\|_2^2\}\,, \tag{0.3}$$

where $\varepsilon(f_\alpha):=\int_Z \phi(yf_\alpha(x))d\rho(x,y)$. Define

$$\alpha_\phi := \arg\min_{\alpha\in R^m}\{\int_Z \phi(yf_\alpha(x))d\rho(x,y)\}\,.$$

If we regard $\alpha_\lambda \in R^m$ as a good learner of α_ϕ, then we can use the classical gradient descent method to learn step by step. To explain this fact, we introduce the regularized loss function θ defined for $\alpha \in R^m$ and $z=(x,y)\in Z$ by

$$\theta(\alpha,z)=\theta_\lambda(\alpha,z):=\phi(yf_\alpha(x))+\frac{\lambda}{2}\|\alpha\|_2^2$$

and the regularized generalization error $F(\alpha)$ defined for $\alpha \in R^m$ as

$$F(\alpha)=F_\lambda(\alpha)=\int_Z \phi(yf_\alpha(x))d\rho+\frac{\lambda}{2}\|\alpha\|_2^2\,.$$

Define the gradient of $F(\alpha)$ at α as $D_\alpha F(\alpha)=(\frac{\partial F}{\partial\alpha_1},\frac{\partial F}{\partial\alpha_2},\cdots,\frac{\partial F}{\partial\alpha_m})$. Then for $\alpha \in R^m$ and $z=(x,y)\in Z$, we have

$$D_\alpha\theta(\alpha,z):=\phi_-^{'}(yf_\alpha(x))yK_{\overline{X}}(x)+\lambda\alpha\,,$$

where $K_{\overline{X}}(x)=(K(x,x_1),\cdots,\ K(x,x_m))$, $\phi_-^{'}$ is the left derivative[8] of ϕ at the point $(yf_\alpha(x))$ with respect to α. The Hilbert space valued random variable $D_\alpha\theta(\alpha,z)$ plays the role of the gradient of the functional θ defined above. So

$$\begin{aligned}D_\alpha F(\alpha)&=(\frac{\partial F}{\partial\alpha_1},\frac{\partial F}{\partial\alpha_2},\cdots,\frac{\partial F}{\partial\alpha_m})\\&=\lambda\alpha+\int{}_Z\phi_-^{'}(yf_\alpha(x))yK_{\overline{X}}(x)d\rho\\&=(\lambda\alpha_1+\int{}_Z\phi_-^{'}(yf_\alpha(x))yK(x,x_1)d\rho(x,y),\cdots,\\&\qquad(\lambda\alpha_m+\int{}_Z\phi_-^{'}(yf_\alpha(x))yK(x,x_m)d\rho(x,y)).\end{aligned}$$

The classical gradient descent[11] tells us that the following sequence $\{g_t : g_t \in R^m$, $t\in N_{t+1}\}$ provides an approximation to α_λ,

$$g_1=(0,\cdots,0)\,,\ \ \forall t\in N_T\,,$$

$$g_{t+1}=g_t-\eta_t(\lambda g_t+\int_Z\phi_-^{'}(yf_{g_t}(x))yK_{\overline{X}}(x)d\rho(x,y))\,.$$

Unfortunately, the use of this algorithm requires knowledge of the distribution. However, As we only have the random values $\phi_-^{'}(y_t f_{\alpha_t}(x_t))y_t K_{\overline{X}}(x_t)$. Then we can give the following Stochastic Gradient Descent (SGD) online algorithm :

$$\alpha_1=(0,\cdots,0), t\in N_T, \alpha_{t+1}=\alpha_t-\eta_t(\lambda\alpha_t+\phi_-^{'}(y_t f_{\alpha_t}(x_t))y_t K_{\overline{X}}(x_t)). \tag{0.4}$$

For each $t\in N_T$, the function α_t is in general dependent on the inputs $\{Z_t : t\in N_T\}$.

Part of the paper deal with the expectation

$$\|\alpha_{T+1}-\alpha_\lambda\|_2 \quad with \quad fixed \quad \lambda>0 \tag{0.5}$$

over the random samples of regularized sample error.

Recall that for the online learning algorithm (0.4), we deal with the classifer $\mathrm{sgn}(f_{\alpha_{T+1}})$ produced by the real valued function $f_{\alpha_{T+1}}$ from a sample $z=\{z_t\}_{t=1}^T$. So the error analysis for the classification algorithm (0.4) will aim at the excess misclassification error

$$R(\mathrm{sgn}(f_{\alpha_{T+1}}))-R(\mathrm{sgn}(f_\rho)) \tag{0.6}$$

which can often be bounded by the excess generalization error[12]

$$\varepsilon(f_{\alpha_{T+1}})-\varepsilon(f_\rho), \tag{0.7}$$

where f_ρ is a minimizer of the generalization error

$$f_\rho=\arg\inf_{f\in H_K}\{\varepsilon(f): f \ \ is\ measurable\ on\ \ X\}.$$

1 Results

The first result stated in the present is that the sequence $\{\alpha_t\}$ defined by (0.4) converges in expectation to α_λ in R^m as long as the sequence is uniformly bounded.

Definition 1. We say that $\phi: R\to R_+$ is an admissible loss function if it is convex and differentiable at 0 with $\phi^{'}<0$. The convexity of ϕ tells us that the left derivative $\phi_-^{'}(x)$ $=\lim_{\delta<0}(\phi(x+\delta)-\phi(x))/\delta$ exists and equals $\sup_{\delta<0}(\phi(x+\delta)-\phi(x))/\delta$. It is the same as $\phi^{'}(x)$ when it coincides with the right derivative

$$\phi_+^{'}(x)=\lim_{\delta\to 0_+}(\phi(x+\delta)-\phi(x))/\delta=\inf_{\delta>0}(\phi(x+\delta)-\phi(x))/\delta.$$

Theorem 1. Let $\lambda>0$ and the sequence of positive step sizes $\{\eta_t\}$ satisfy

$$\sum_{t=1}^{\infty}\eta_t=+\infty, \ \ \lim_{t\to\infty}\eta_t=0. \tag{1.1}$$

If $\phi(x)$ is an admissible loss and the learning sequence α_t defined by (0.4) is uniformly bounded on R^m, then

$$E_{z\in Z^T}(\|\alpha_{T+1}-\alpha_\lambda\|_2)\to 0 \quad \text{as} \quad T\to+\infty. \tag{1.2}$$

Theorem 1 will be proved in Section 5. The assumption of uniform boundedness is mild and will be studied in Section 3. In particular, we shall verify that $\|\alpha_t\|_2$ is uniformly bounded when $\sup_{|x|\le R}|\phi^{'}(x)-\phi^{'}(0)|/|x|<\infty$ for any R and $\eta_t\le C_\lambda$ for some constant C_λ.

Our second result is the following relation between the excess regularized generalization error and the R^m metric which plays an essential role in proving Theorem 1.

Theorem 2. Let ϕ be an admissible loss function and $\lambda > 0$. For any $\alpha \in R^m$, there holds

$$\frac{\lambda}{2}\|\alpha-\alpha_\lambda\|_2^2 \le \{\varepsilon（f_\alpha)+\frac{\lambda}{2}\|\alpha\|_2^2\}-\{\varepsilon(f_{\alpha\lambda})+\frac{\lambda}{2}\|\alpha_\lambda\|_2^2\}. \tag{1.3}$$

Theorem 2 will be proved in Section 4. It will be used to derive convergence rates of $E_{z\in Z^T}(\|\alpha_{T+1}-\alpha_\lambda\|_2)$ when the step size decays in the form $\eta_t = t^{-\theta}/\mu(\lambda)$ with a constant $\mu(\lambda)$.

Definition 2. Define the regularization error with respect to (0.2) associated with the triple (K,ϕ,ρ) as

$$D(\lambda)=\inf_{\alpha\in R^m}\{\varepsilon(f_\alpha)-\varepsilon(f_\rho)+\frac{\lambda}{2}\|\alpha\|_2^2\}=\varepsilon(f_{\alpha_\lambda})-\varepsilon(f_\rho)+\frac{\lambda}{2}\|\alpha_\lambda\|_2^2,\ \lambda>0. \tag{1.4}$$

then, we have

$$\begin{aligned}\varepsilon(f_{\alpha_{T+1}})-\varepsilon(f_\rho)&=\varepsilon(f_{\alpha_{T+1}})-\varepsilon(f_{\alpha_\lambda})+\varepsilon(f_{\alpha_\lambda})-\varepsilon(f_\rho)\\ &\le\varepsilon(f_{\alpha_{T+1}})-\varepsilon(f_{\alpha_\lambda})+D(\lambda).\end{aligned} \tag{1.5}$$

The regularization error term $D(\lambda)$ in the error decomposition (1.5) is independent of the sample $z=\{z_t\}_{t=1}^T$. It can be estimated by the rich knowledge of approximation theory.

The first term $\varepsilon(f_{\alpha_{T+1}})-\varepsilon(f_{\alpha_\lambda})$ in (1.5) is called the sample error which may be bounded by $\left\|f_{\alpha_{T+1}}-f_{\alpha_\lambda}\right\|_K$, which will be bounded by $\|\alpha_{T+1}-\alpha_\lambda\|_2$.

2 Bounding the Learning Sequence

Definition 3. We say that $\phi_-^{'}$ is locally Lipschitz at the origin if the local Lipschitz constant

$$M(\lambda)=\sup\{\frac{|\phi_-^{'}(x)-\phi_-^{'}(0)|}{|x|}:|x|\le\frac{k^2|\phi^{'}(0)|}{\lambda}\} \tag{2.1}$$

is finite for any $\lambda>0$, where $k:=\sup_{x,t\in X}\sqrt{K(x,t)}<\infty$.

The above local Lipschitz condition is equivalent to the existence of some $\varepsilon>0$ and $L>0$ such that $|\phi_-^{'}(x)-\phi^{'}(0)|\le L(x)$ for every $x\in[-\varepsilon,\varepsilon]$. In fact, the latter requirement implies

$$M(\lambda)\le\max\{L,|\phi_-^{'}(-k^2|\phi^{'}(0)|/\lambda)-\phi^{'}(0)|/\varepsilon,(|\phi_-^{'}(k^2|\phi^{'}(0)|/\lambda)|+|\phi^{'}(0)|)/\varepsilon\}.$$

Thus, when ϕ is twice continuously differentiable on R, $\phi_-^{'}$ is locally Lipschitz at the origin with $M(\lambda)=\|\phi^{''}\|_{L^\infty[-k^2|\phi^{'}(0)|/\lambda,k^2|\phi^{'}(0)|/\lambda]}$. Examples of loss functions will be discussed after the following theorem on the boundedness of the sequence $\{\alpha_t\}$.

Theorem 3. Assume that $\phi_-^{'}$ is locally Lipschitz at the origin. Define $\{\alpha_t\}$ by (0.4). If the step size η_t satisfies $\eta_t(M(\lambda)k^2+\lambda)\le 1$ for each t, then

$$\|\alpha_t\|_2 \le \frac{k|\phi'(0)|}{\lambda}, \quad \forall t \in N. \qquad (2.2)$$

Proof. We prove (2.2) by induction. It is trivial that $\alpha_1=(0,\cdots,0)$ satisfies the bound (2.2). Suppose that this bound holds true for $\alpha_t \in R^m$. Consider α_{t+1}. It can be written as

$$(1-\eta_t\lambda)\ \alpha_t-\eta_t\phi'_-(y_t f_{\alpha_t}(x_t))y_t K_{\overline{X}}(x_t),$$

that is,

$$(1-\eta_t\lambda)\alpha_t-\eta_t\left[\phi'_-(y_t f_{\alpha_t}(x_t))-\phi'(0)\right]y_t K_{\overline{X}}(x_t)-\eta_t y_t\phi'(0)K_{\overline{X}}(x_t).$$

Since $y_t^2=1$, we can write the middle term as

$$\frac{[\phi'_-(y_t f_{\alpha_t}(x_t))-\phi'(0)]y_t^2 K_{\overline{X}}(x_t) f_{\alpha t}(x_t)}{y_t f_{\alpha_t}(x_t)}$$

$$=\frac{[\phi'_-(y_t f_{\alpha_t}(x_t))-\phi'(0)]}{y_t f_{\alpha_t}(x_t)}\times K_{\overline{X}}(x_t) f_{\alpha_t}(x_t).$$

Since ϕ'_- is nondecreasing, $(\phi'_-(u)-\phi'(0))/u\ge 0$ for any $u\in R$, we can have

$$\frac{\phi'_-(y_t f_{\alpha_t}(x_t))-\phi'(0)}{y_t f_{\alpha_t}(x_t)}\ge 0.$$

Therefore, it follows that

$$\begin{aligned}\|\alpha_{t+1}\|_2 &= \left\|(1-\eta_t\lambda)\alpha_t-\eta_t\left(\frac{\phi'_-(y_t f_{\alpha_t}(x_t))-\phi'(0)}{y_t f_{\alpha_t}(x_t)}\times f_{\alpha_t}(x_t)K_{\overline{X}}(x_t)\right)-\eta_t\phi' K_{\overline{X}}(x_t)\right\|_2\\ &\le \left\|(1-\eta_t\lambda)\alpha_t-\eta_t\left(\frac{\phi'_-(y_t f_{\alpha_t}(x_t))-\phi'(0)}{y_t f_{\alpha_t}(x_t)}\times f_{\alpha_t}(x_t)K_{\overline{X}}(x_t)\right)-\eta_t\phi' K_{\overline{X}}(x_t)\right\|_2\\ &+\left\|\eta_t\phi'(0)K_{\overline{X}}(x_t)\right\|_2.\end{aligned}$$

Define an operator $L_t: R^m\to R^m$ as $L_t(g)=(g,K_{\overline{X}}(x_t))K_{\overline{X}}(x_t)$, then, the middle term can be written as

$$[\phi'_-(y_t f_{\alpha_t}(x_t))-\phi'(0)]y_t K_{\overline{X}}(x_t)=\frac{\phi'_-(y_t f_{\alpha_t}(x_t))-\phi'(0)}{y_t f_{\alpha_t}(x_t)}\times L_t(\alpha_t),$$

here by means of the property for $f_{\alpha_t}(x_t)K_{\overline{X}}(x_t)=(\alpha_t,K_{\overline{X}}(x_t))K_{\overline{X}}(x_t)$. L_t is a self-adjoint, rank-one, positive linear operator. The operator norm can be bounded as $\|L_t\|_{R^m\to H_K}\le k^2$.

Since $(L_t g,g)=|(g,K_{\overline{X}})|^2=(f_g(x_t))^2\le k^2\|g\|_2^2$ for any $g\in R^m$. The local Lipschitz condition tell us that $\frac{\phi'_-(y_t f_{\alpha_t}(x_t))-\phi'(0)}{y_t f_{\alpha_t}(x_t)}$ is well defined (set as zero when $f_{\alpha_t}(x_t)=0$). It is bounded by $M(k^2|\phi^{'}(0)|/\lambda)$, since $|y_t f_{\alpha_t}(x_t)| = |(\alpha_t, K_{\overline{X}}(x))| \le k\|\alpha_t\|_2 \le k^2|\phi^{'}(0)|/\lambda$ by our induction hypothesis.

Since $\phi_-^{'}$ is nondecreasing, $(\phi_-^{'}(u) - \phi^{'}(0))/u \ge 0$ for any $u \in R$. Thus,

$$\frac{\phi_-^{'}(y_t f_{\alpha_t}(x_t)) - \phi^{'}(0)}{y_t f_{\alpha_t}(x_t)} \ge 0 .$$

Then,

$$\begin{aligned}
&\left\|(1-\eta_t\lambda)\alpha_t - \eta_t[\phi_-^{'}(y_t f_{\alpha_t}(x_t)) - \phi^{'}(0)]y_t K_{\overline{X}}(x_t)\right\|_2 \\
&= \left\|(1-\eta_t\lambda)\alpha_t - \eta_t\left(\frac{\phi_-^{'}(y_t f_{\alpha_t}(x_t)) - \phi^{'}(0)}{y_t f_{\alpha_t}(x_t)} \times L_t(\alpha_t)\right\|_2 \\
&\le (1-\eta_t\lambda)\|\alpha_t\|_2 .
\end{aligned}$$

This in connection with the induction hypothesis on $\|\alpha_t\|_2$ implies that

$$\|\alpha_{t+1}\|_2 \le (1-\eta_t\lambda)\frac{k|\phi^{'}(0)|}{\lambda} + k\eta_t|\phi^{'}(0)| = \frac{k|\phi^{'}(0)|}{\lambda} \quad .$$

This completes the induction procedure and proves the theorem.

Corollary 1. Define a hypothesis space as

$$H_{K,\overline{X}} = \left\{ f_\alpha = \sum_{i=1}^m \alpha_i K_{x_i} : \alpha \in R^m, m \in N \right\}$$

with the inner product $\langle\cdot,\cdot\rangle_{H_{K,\overline{X}}} = \langle\cdot,\cdot\rangle_K$, satisfying $\langle K_x, K_y\rangle_K = K(x,y)$, then for any $\alpha \in R^m$, $\|f_\alpha\|_K \le \sqrt{mk}\|\alpha\|_2$.

Proof. Let α， $\beta \in R^m$ Then,

$$\langle f_\alpha, f_\beta\rangle_K = \left\langle \sum_{i=1}^m \alpha_i K_{x_i}, \sum_{j=1}^m \beta_j K_{y_j} \right\rangle_K = \sum_{i,j=1}^m \alpha_i K(x_i, x_j)\beta_j \quad .$$

Thus,

$$\begin{aligned}
\|f_\alpha\|_K^2 &= \langle f_\alpha, f_\alpha\rangle_K \\
&= \sum_{i,j=1}^m \alpha_i K(x_i, x_j)\alpha_j \\
&\le \left(\sum_{i=1}^m \alpha_i^2\right)^{\frac{1}{2}} \left(\sum_{i=1}^m \left(\sum_{j=1}^m K(x_i, x_j)\alpha_j\right)^2\right)^{\frac{1}{2}} \le mk\|\alpha\|_2^2 .
\end{aligned}$$

3 Excess Generalization

In this section, we prove Theorem 2, a relation between $\|\alpha - \alpha_\lambda\|_2$ and

$$\left\{\varepsilon(f_\alpha) + \frac{\lambda}{2}\|\alpha\|_2^2\right\} - \left\{\varepsilon(f_{\alpha_\lambda}) + \frac{\lambda}{2}\|\alpha_\lambda\|_2^2\right\}.$$

This relation is very important for the proof of the general convergence result, Theorem 1, as well as for the error analysis done in the next section.

Lemma 1. Assume ϕ is differentiable, then α_λ satisfies

$$\int_Z \phi^{'}(f_{\alpha_\lambda}(x))y(f_\alpha(x)-f_{\alpha_\lambda}(x))d\rho+\lambda(\alpha_\lambda,\alpha-\alpha_\lambda)=0,\quad \forall\alpha\in R^m. \qquad (3.1)$$

Proof. As α_λ is a minimizer of the regularized generation error defined by (0.3), taking $\alpha=(0,\cdots,0)$ yields $\varepsilon(f_{\alpha_\lambda})+\frac{\lambda}{2}\|\alpha_\lambda\|_2^2\le\varepsilon(f_0)=\phi(0)$, then for any $\alpha\in R^m$ and $\theta>0$, we know that $\frac{1}{\theta}\left\{(\varepsilon(f_{\alpha_\lambda+\theta\alpha})+\frac{\lambda}{2}\|\alpha_\lambda+\theta\alpha\|_2^2)-(\varepsilon(f_{\alpha_\lambda})+\frac{\lambda}{2}\|\alpha_\lambda\|_2^2)\right\}$ is nonnegative and equals

$$\int_Z\frac{1}{\theta yf_\alpha(x)}\left\{\phi(yf_{\alpha_\lambda+\theta\alpha}(x))-\phi(yf_{\alpha_\lambda}(x))\right\}yf_\alpha(x)d\rho+\lambda(\alpha_\lambda,\alpha)+\frac{\lambda}{2}\theta\|\alpha\|_2^2.$$

Let $\theta\to 0^+$, by the Lebesgue Dominant Theorem, we see that,

$$\int_Z\phi^{'}(yf_{\alpha_\lambda}(x))yf_\alpha(x)d\rho+\lambda(\alpha_\lambda,\alpha)\ge 0.$$

It follows that

$$\int_Z\phi^{'}(yf_{\alpha_\lambda}(x))y(K_{\overline{X}}(x),\alpha)d\rho+\lambda(\alpha_\lambda,\alpha)=(\int_Z\phi^{'}(yf_{\alpha_\lambda}(x))yK_{\overline{X}}(x)d\rho+\lambda\alpha_\lambda,\alpha)\ge 0.$$

This is true for every $\alpha\in R^m$, which implies

$$\int_Z\phi^{'}(yf_{\alpha_\lambda}(x))yK_{\overline{X}}(x)d\rho+\lambda\alpha_\lambda=(0,\cdots,0). \qquad (3.2)$$

Taking product with $\alpha-\alpha_\lambda$ in R^m proves the lemma.

Lemma 2. Let $\lambda>0$ and ϕ be a differentiable convex loss function. Then for any $\alpha\in R^m$ there holds

$$\frac{\lambda}{2}\|\alpha-\alpha_\lambda\|_2^2\le\{(\varepsilon(f_\alpha)+\frac{\lambda}{2}\|\alpha\|_2^2)\}-\{\varepsilon(f_{\alpha_\lambda})+\frac{\lambda}{2}\|\alpha_\lambda\|_2^2\}.$$

Proof. Let $\alpha\in R^m$, define a univariate function $G=G_\alpha$ by

$$G(\theta)=\varepsilon(f_{\alpha_\lambda+\theta(\alpha-\alpha_\lambda)})+\frac{\lambda}{2}\|\alpha_\lambda+\theta(\alpha-\alpha_\lambda)\|_2^2,\theta\in R.$$

Then

$$G(1)=\varepsilon(f_\alpha)+\frac{\lambda}{2}\|\alpha\|_2^2, G(0)=\varepsilon(f_{\alpha_\lambda})+\frac{\lambda}{2}\|\alpha_\lambda\|_2^2. \qquad (3.3)$$

Since ϕ is differentiable, as a function of θ, G is differentiable. In fact, if we denote $\alpha_\theta=\alpha_\lambda+\theta(\alpha-\alpha_\lambda)$, then

$$\begin{aligned}G^{'}(\theta)&=\lim_{\Delta\theta\to 0}\frac{1}{\Delta\theta}\left\{G(\theta+\Delta\theta)-G(\theta)\right\}\\&=\lambda(\alpha_\theta,\alpha-\alpha_\lambda)+\lim_{\Delta\theta\to 0}\int_Z\frac{y(f_\alpha(x)-f_{\alpha_\lambda}(x))}{\Delta\theta y(f_\alpha(x)-f_{\alpha_\lambda}(x))}\\&\quad\times\left\{\phi(yf_{\alpha_\theta}(x)+\Delta\theta y(f_\alpha(x)-f_{\alpha_\lambda}(x))-\phi(yf_{\alpha_\theta}(x)))d\rho\right\}.\end{aligned}$$

The Lebegue Dominant Theorem ensures that

$$G^{'}(\theta)=\lambda(\alpha_\theta,\alpha-\alpha_\lambda)+\int_Z\phi^{'}(yf_{\alpha_\theta}(x))y(f_\alpha(x)-f_{\alpha_\lambda}(x))d\rho.$$

The first term of $G^{'}(\theta)$ can be written as $\lambda(\alpha_\lambda,\alpha-\alpha_\lambda)+\lambda\theta\|\alpha-\alpha_\lambda\|_2^2$. This in connection with Lemma 1 tells us that

$\lambda(\alpha_\lambda,\alpha-\alpha_\lambda)=-\int_Z\phi^{'}(y_t f_{\alpha_\lambda}(x))y(f_\alpha(x)-f_{\alpha_\lambda}(x))d\rho$, and $G^{'}(\theta)$ equals

$$\begin{aligned}&\lambda\theta\|\alpha-\alpha_\lambda\|_2^2+\int_Z[\phi^{'}(yf_{\alpha_\lambda}(x)+\theta y(f_\alpha(x)-f_{\alpha_\lambda}(x)))\\&\qquad-\phi^{'}(yf_{\alpha_\lambda}(x))]y(f_\alpha(x)-f_{\alpha_\lambda}(x))d\rho.\end{aligned}\tag{3.4}$$

Since ϕ is convex in R, it satisfies

$$(\phi^{'}(x_1)-\phi^{'}(x_2))(x_1-x_2)\ge 0,\ \forall x_1,x_2\in R.$$

Using this for $x_1=yf_{\alpha_\lambda}(x)+\theta y(f_\alpha(x)-f_{\alpha_\lambda}(x))$ and $x_2=yf_{\alpha_\lambda}(x)$, we see from (3.4) that for $\theta\in(0,1)$, $G^{'}(\theta)\ge\lambda\theta\|\alpha-\alpha_\lambda\|_2^2$.

Therefore,

$$\begin{aligned}G(1)-G(0)&=\left\{\varepsilon(f_\alpha)+\frac{\lambda}{2}\|\alpha\|_2^2\right\}-\left\{\varepsilon(f_{\alpha_\lambda})+\frac{\lambda}{2}\|\alpha_\lambda\|_2^2\right\}\\&=\int_0^1G^{'}(\theta)d\theta\ge\frac{\lambda}{2}\|\alpha-\alpha_\lambda\|_2^2.\end{aligned}$$

This proves the desired result.

If ϕ is differentiable, we approximate it by ϕ_ε which is convex, differentiable and defined for $0<\varepsilon\le1$ as

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

The approximation $|\phi_\varepsilon(x)-\phi(x)|=\left|\int_0^1\phi(x-\varepsilon u)-\phi(x)du\right|\le\|\phi_-^{'}\|_{L^\infty[x-\varepsilon,x]}\varepsilon$ is valid, hence for any $R>0$, there holds

$$\|\phi_\varepsilon-\phi\|_{C[-R,R]}=O(\varepsilon)\quad,\qquad\varepsilon\to0_+.\tag{3.5}$$

Now we can prove Theorem 2 for a general loss function.

Proof of Theorem 2. We define, for any $0\le\varepsilon\le1,\varepsilon^{(\varepsilon)}(f_\alpha)=\int_Z\phi_\varepsilon(yf_\alpha(x))d\rho$ and

$$\alpha_\lambda^{(\varepsilon)}=\arg\inf_{\alpha\in R^m}\left\{\varepsilon^{(\varepsilon)}(f_\alpha)+\frac{\lambda}{2}\|\alpha\|_2^2\right\}.\tag{3.6}$$

For $\varepsilon=0$, we have used the conventional notation $\varepsilon^{(0)}(f_\alpha)=\varepsilon(f_\alpha)$ and $\alpha_\lambda^{(0)}=\alpha_\lambda$. Since $\alpha_\lambda^{(\varepsilon)}$ is the minimizer of (3.6), by taking $\alpha=(0,\cdots,0)$ we get

$$\varepsilon^{(\varepsilon)}(f_{\alpha_\lambda^{(\varepsilon)}})+\frac{\lambda}{2}\|\alpha_\lambda^{(\varepsilon)}\|_2^2\le\varepsilon^{(\varepsilon)}(0)=\phi_\varepsilon(0)\le\|\phi\|_{C[-1,0]}<\infty$$

which implies

$$\|\alpha_\lambda^{(\varepsilon)}\|_2\le\sqrt{2\|\phi\|_{C[-1,0]}/\lambda}\quad,\ \forall0\le\varepsilon\le1.\tag{3.7}$$

Since any closed ball $B_R=\{\alpha\in R^m,\|\alpha\|_2\le R\}$ of the Hilbert space R^m is weekly compact, the estimate (3.7) tells us that there exists a sequence $\{\varepsilon_j > 0\}_{j=1}^{\infty}$ such that $\lim\limits_{j\to\infty}\varepsilon_j = 0$ and $\alpha_\lambda^{(\varepsilon_j)}$ converges to some $\alpha_\lambda^* \in R^m$ weekly, that is,

$$\lim_{j\to\infty}(\alpha_\lambda^{(\varepsilon_j)},\alpha) = (\alpha_\lambda^*,\alpha), \quad \forall\alpha\in R^m . \tag{3.8}$$

In particular, $\left\|\alpha_\lambda^*\right\|_2^2 = (\alpha_\lambda^*,\alpha_\lambda^*) = \lim\limits_{j\to\infty}(\alpha_\lambda^{(\varepsilon_j)},\alpha_\lambda^*) \le \left\|\alpha_\lambda^*\right\|_2 \lim\limits_{j\to\infty}\left\|\alpha_\lambda^{(\varepsilon_j)}\right\|_2$ and

$$\left\|\alpha_\lambda^*\right\|_2 \le \lim_{j\to\infty}\left\|\alpha_\lambda^{(\varepsilon_j)}\right\|_2 \le \sqrt{2\|\phi\|_{C[-1,0]}/\lambda} . \tag{3.9}$$

Let $\alpha = K_{\overline{X}}(x)$ in (3.8). Then by $(\alpha, K_{\overline{X}}(x)) = f_\alpha(x)$ one has $\lim\limits_{j\to\infty} f_{\alpha_\lambda^{\varepsilon_j}}(x) = \lim\limits_{j\to\infty}(\alpha_\lambda^{\varepsilon_j}, K_{\overline{X}}(x))$, then $\varepsilon(f_{\alpha_\lambda^*}) = \lim\limits_{j\to\infty}\int_Z \phi(yf_{\alpha_\lambda^{\varepsilon_j}}(x))d\rho$. The uniform bound (3.7) of $\left\|\alpha_\lambda^{(\varepsilon_j)}\right\|_2$ in connection with uniform convergence (3.5) of ϕ_ε to ϕ ensures that

$$\begin{aligned}\underline{\lim}_{j\to\infty}\varepsilon^{(\varepsilon_j)}(f_{\alpha_\lambda^{(\varepsilon_j)}}) &= \lim_{j\to\infty}\int_Z \phi^{(\varepsilon_j)}(yf_{\alpha_\lambda^{(\varepsilon_j)}}(x))d\rho \\ &= \lim_{j\to\infty}\int_Z \phi(yf_{\alpha_\lambda^{(\varepsilon_j)}}(x))d\rho \\ &= \varepsilon(f_{\alpha_\lambda^*}) .\end{aligned} \tag{3.10}$$

Therefore, by (3.9), we have

$$\varepsilon(f_{\alpha_\lambda^*}) + \frac{\lambda}{2}\left\|\alpha_\lambda^*\right\|_2^2 \le \underline{\lim}_{j\to\infty}\left\{\varepsilon^{(\varepsilon_j)}(f_{\alpha_\lambda^{\varepsilon_j}}) + \frac{\lambda}{2}\left\|\alpha_\lambda^{(\varepsilon_j)}\right\|_2^2\right\}.$$

Taking $\alpha = \alpha_\lambda$ in (3.6), we know that

$$\begin{aligned}\underline{\lim}_{j\to\infty}\left\{\varepsilon^{(\varepsilon_j)}(f_{\alpha_\lambda^{\varepsilon_j}}) + \frac{\lambda}{2}\left\|\alpha_\lambda^{(\varepsilon_j)}\right\|_2^2\right\} &\le \underline{\lim}_{j\to\infty}\left\{\varepsilon^{(\varepsilon_j)}(f_{\alpha_\lambda}) + \frac{\lambda}{2}\left\|\alpha_\lambda\right\|_2^2\right\} \\ &= \varepsilon(f_{\alpha_\lambda}) + \frac{\lambda}{2}\left\|\alpha_\lambda\right\|_2^2 ,\end{aligned}$$

which means

$$\varepsilon(f_{\alpha_\lambda^*}) + \frac{\lambda}{2}\left\|\alpha_\lambda^*\right\|_2^2 \le \varepsilon(f_{\alpha_\lambda}) + \frac{\lambda}{2}\left\|\alpha_\lambda\right\|_2^2 .$$

It tells us that α_λ^* is also a minimizer of $\varepsilon(f_{\alpha_\lambda}) + \frac{\lambda}{2}\left\|\alpha_\lambda\right\|_2^2$. The strict convexity of the functional $\varepsilon(f_{\alpha_\lambda}) + \frac{\lambda}{2}\left\|\alpha_\lambda\right\|_2^2$ on R^m verifies the uniqueness of α_λ which leads to $\alpha_\lambda^* = \alpha_\lambda$. That is, (3.8) and (3.9) hold with α_λ^* replaced by α_λ. Apply Lemma 2 to the modified loss function ϕ_{ε_j}, we have

$$\frac{\lambda}{2}\left\|\alpha - \alpha_\lambda^{(\varepsilon_j)}\right\|_2^2 \le \left\{\varepsilon^{(\varepsilon_j)}(f_\alpha) + \frac{\lambda}{2}\left\|\alpha\right\|_2^2\right\} - \left\{\varepsilon^{(\varepsilon_j)}(f_{\alpha_\lambda^{\varepsilon_j}}) + \frac{\lambda}{2}\left\|\alpha_\lambda^{(\varepsilon_j)}\right\|_2^2\right\}. \tag{3.11}$$

Apply (3.8) to $(\alpha, \alpha - \alpha_\lambda)$, we know that

$$\|\alpha-\alpha_\lambda\|_2^2=(\alpha,\alpha-\alpha_\lambda)-\lim_{j\to\infty}(\alpha_\lambda^{(\varepsilon_j)},\alpha-\alpha_\lambda)=\lim_{j\to\infty}(\alpha-\alpha_\lambda^{(\varepsilon_j)},\alpha-\alpha_\lambda),$$

which can be bounded by $\lim_{j\to\infty}\left\|\alpha-\alpha_\lambda^{(\varepsilon_j)}\right\|_2\|\alpha-\alpha_\lambda\|_2$, hence,

$$\frac{\lambda}{2}\|\alpha-\alpha_\lambda\|_2\le\frac{\lambda}{2}\lim_{j\to\infty}\left\|\alpha-\alpha_\lambda^{(\varepsilon_j)}\right\|_2.$$

This in connection with (3.11), (3.9) and (3.10) implies that $\frac{\lambda}{2}\|\alpha-\alpha_\lambda\|_2^2$ is bounded by

$$\left\{\lim_{j\to\infty}\varepsilon^{(\varepsilon_j)}(f_\alpha)+\frac{\lambda}{2}\|\alpha\|_2^2\right\}-\left\{\varepsilon(f_{\alpha_\lambda})+\frac{\lambda}{2}\|\alpha_\lambda\|_2^2\right\}.$$

Since $\lim_{j\to\infty}\varepsilon^{(\varepsilon_j)}(f_\alpha)=\varepsilon(f_\alpha)$, the conclusion of Theorem 2 is proved.

4 General Convergence Results

In this section we prove Theorem 1. The essential estimate in the proof will also be used in the next section to give convergence rates. Note that the uniform boundedness of $\|\alpha_t\|_2$ by B implies

$$\left\|\phi_-^{'}(y_t f_{\alpha_t}(x_t))y_tK_{\overline{X}}(x_t)+\lambda\alpha_t\right\|_2\le k\left\|\phi_-^{'}\right\|_{L^\infty[-kB,kB]}+\lambda B.$$

For simplicity, we denote $\prod_{j=T+1}^{T}(1-\eta_j\lambda)=1$ and $\sum_{j=T+1}^{T}\eta_j\lambda=0$.

Lemma 3. Assume for some $t_0\in N$ and $\tilde{C}_\lambda>0$, that holds $\eta_t\lambda<1$,

$$E_{z_1,\cdots,z_T}\left(\left\|\phi_-^{'}(y_t f_{\alpha_t}(x_t))y_tK_{\overline{X}}(x_t)+\lambda\alpha_t\right\|_2^2\right)\le\tilde{C}_\lambda \tag{4.1}$$

for any $t\ge t_0$. Then for $T>t_0$,

$$\begin{aligned}E_{z_1,\cdots,z_T}\|\alpha_{T+1}-\alpha_\lambda\|_2^2\le&\prod_{t=t_0}^{T}(1-\eta_t\lambda)E_{z_1,\cdots,z_{t_0-1}}\left(\|\alpha_{t_0}-\alpha_\lambda\|_2^2\right)\\&+\tilde{C}_\lambda\sum_{t=t_0}^{T}\eta_t^2\prod_{j=t+1}^{T}(1-\eta_j\lambda),\end{aligned}\tag{4.2}$$

which can be controlled further by

$$\exp\left\{-\sum_{t=t_0}^{T}\eta_t\lambda\right\}E_{z_1,\cdots,z_{t_0-1}}\left(\|\alpha_{t_0}-\alpha_\alpha\|_2^2\right)+\tilde{C}_\lambda\sum_{t=t_0}^{T}\eta_t^2\exp\left\{-\sum_{j=t+1}^{T}\eta_j\lambda\right\}\tag{4.3}$$

Proof. Recall that $\alpha_{t+1}=\alpha_t-\eta_t\alpha_t^\lambda$ where $\alpha_t^\lambda=\phi_-^{'}(yf_{\alpha t}(x_t))y_tK_{\overline{X}}(x_t)+\lambda\alpha_t$. Then

$$\|\alpha_{t+1}-\alpha_\lambda\|_2^2=\|\alpha_t-\alpha_\lambda\|_2^2+\eta_t^2\left\|\alpha_t^\lambda\right\|_2^2+2\eta_t(\alpha_t^\lambda,\alpha_\lambda-\alpha_t).\tag{4.4}$$

By taking $(K_{\overline{X}}(x_t),\alpha)=f_\alpha(x_t)$, part of the last term of (4.4) equals

$$(\phi_-^{'}(y_t f_{\alpha_t}(x_t))y_tK_{\overline{X}}(x_t),\alpha_\lambda-\alpha_t)=\phi_-^{'}(y_t f_{\alpha_t}(x_t))y_t(f_{\alpha_\lambda}(x_t)-f_{\alpha_t}(x_t)).\tag{4.5}$$

Since ϕ is a convex function on R, we know that

$$\phi_-^{'}(a)(b-a)\le\phi(b)-\phi(a),\ \forall a,\,b\in R.$$

Applying this reaction to $a = y_t f_{\alpha_t}(x_t)$ and $b = y_t f_{\alpha_\lambda}(x_t)$ together with (4.5) yields

$$(\phi_-^{'}(y_t f_{\alpha_t}(x_t)) y_t K_{\overline{X}}(x_t), \alpha_\lambda - \alpha_t) \le \phi(y_t f_{\alpha_\lambda}(x_t)) - \phi(y_t f_{\alpha_t}(x_t)) . \qquad (4.6)$$

The schwarz inequality $(\alpha_t, \alpha_\lambda) \le \|\alpha_t\|_2 \|\alpha_\lambda\|_2 \le \frac{1}{2}\|\alpha_t\|_2^2 + \frac{1}{2}\|\alpha_\lambda\|_2^2$ implies

$$\lambda(\alpha_t, \alpha_\lambda - \alpha_t) \le \frac{\lambda}{2}\|\alpha_\lambda\|_2^2 - \frac{\lambda}{2}\|\alpha_t\|_2^2 .$$

Putting this and (4.6) into the last term of (4.4), we know that $(\alpha_t^\lambda, \alpha_\lambda - \alpha_t)$ can be bounded by

$$\left[\phi(y_t f_{\alpha_\lambda}(x_t)) + \frac{\lambda}{2}\|\alpha_\lambda\|_2^2\right] - \left[\phi(y_t f_{\alpha_t}(x_t)) + \frac{\lambda}{2}\|\alpha_t\|_2^2\right].$$

Since α_t depends on $\{z_1, \cdots, z_{t-1}\}$, but not on z_t, It follows that $E_{z_1,z_2,\cdots z_t}(\alpha_t^\lambda, \alpha_\lambda - \alpha_t)$ can be bounded by

$$E_{z_1,\cdots,z_{t-1}}(E_{z_t}([\phi(y_t f_{\alpha_\lambda}(x_t)) + \frac{\lambda}{2}\|\alpha_\lambda\|_2^2] - [\phi(y_t f_{\alpha_t}(x_t)) + \frac{\lambda}{2}\|\alpha_t\|_2^2]))$$

$$= E_{z_1,\cdots,z_{t-1}}([\varepsilon(f_{\alpha_\lambda}) + \frac{\lambda}{2}\|\alpha_\lambda\|_2^2] - [\varepsilon(f_{\alpha_t}) + \frac{\lambda}{2}\|\alpha_t\|_2^2]) . \qquad (4.7)$$

This in connection with (4.1) and (4.4) gives

$$E_{z_1,\cdots,z_t}(\|\alpha_{t+1} - \alpha_\lambda\|_2^2) \le E_{z_1,\cdots,z_{t-1}}(\|\alpha_t - \alpha_\lambda\|_2^2)$$

$$+ \tilde{C}_\lambda \eta_t^2 + 2\eta_t E_{z_1,\cdots,z_t}([\varepsilon(f_{\alpha_\lambda}) + \frac{\lambda}{2}\|\alpha_\lambda\|_2^2] - [\varepsilon(f_{\alpha_t}) + \frac{\lambda}{2}\|\alpha_t\|_2^2] . \qquad (4.8)$$

By Theorem 2, this implies that $E_{z_1,\cdots,z_t}(\|\alpha_{t+1} - \alpha_\lambda\|_2^2)$ is bounded by

$$(1 - \eta_t\lambda) E_{z_1,\cdots,z_{t-1}}(\|\alpha_t - \alpha_\lambda\|_2^2) + \tilde{C}_\lambda \eta_t^2 . \qquad (4.9)$$

Applying this relation iteratively for $t = T, T-1, \cdots, t_0$, we see that $E_{z_1,\cdots,z_T}($ $\|\alpha_{T+1} - \alpha_\lambda\|_2^2)$ is bounded by

$$(1 - \eta_T\lambda)(1 - \eta_{T-1}\lambda) E_{z_1,\cdots,z_{T-2}}(\|\alpha_{T-1} - \alpha_\lambda\|_2^2) + \tilde{C}_\lambda \eta_T^2 + (1 - \eta_t\lambda)\tilde{C}_\lambda \eta_{T-1}^2$$

$$\le \cdots \le \prod_{t=t_0}^{T}(1 - \eta_t\lambda) E_{z_1,\cdots,z_{t_0-1}}(\|\alpha_{t_0} - \alpha_\lambda\|_2^2) + \tilde{C}_\lambda \sum_{t=t_0}^{T} \eta_t^2 \prod_{j=t+1}^{T}(1 - \eta_j\lambda) .$$

This proves the first statement.

The second statement follows from the inequality $1 - \mu \le e^{-\mu}$ for any $\mu \ge 0$.

We are in a position to prove Theorem 1 stated in the introduction. For this purpose we use (4.2) while the bound (4.3) will be used to derive explicit learning rates in the next section.

Proof of Theorem 1. By (1.1), there exists an integer such that $\eta_t\lambda \le 1/2$ for all $t \ge t_0$. Since $\{\alpha_t\}$ is uniformly bounded in R^m, (4.1) is true for some constant C_λ. Applying Lemma 3, it is sufficient to estimate the right side of (4.2).

According to the assumption (1.1) on the step size, we have

$$\prod_{t=t_0}^{T}(1-\eta_t\lambda)\le \exp\{-\sum_{t=t_0}^{T}\eta_t\lambda\}\to 0,\ \ \text{as}\ \ \mathrm{T}\to\infty.$$

So for any $\varepsilon>0$ there exists some $T_1\in N$ such that the second term of (4.2) is bounded by ε whenever $T\ge T_1$.

To deal with the first term, we use the assumption $\lim_{t\to\infty}\eta_t=0$ and know that there exists some $t(\varepsilon)$ such that $\eta_t\le\lambda\varepsilon$ for every $t\ge t(\varepsilon)$. Write $\sum_{t=t_0}^{T}\eta_t^2\prod_{j=t+1}^{T}(1-\eta_j\lambda)$ as

$$\sum_{t=t_0}^{t(\varepsilon)}\eta_t^2\prod_{j=t+1}^{T}(1-\eta_j\lambda)+\sum_{t=t(\varepsilon)+1}^{T}\eta_t^2\prod_{j=t+1}^{T}(1-\eta_j\lambda). \tag{4.10}$$

Since $t(\varepsilon)$ is fixed, we can find some $T_2\in N$ such that for each $T\ge T_2$, there holds $\sum_{j=t(\varepsilon)+1}^{T}\eta_j\ge\sum_{j=t(\varepsilon)+1}^{T_2}\eta_j\ge\frac{1}{\lambda}\log\frac{t(\varepsilon)}{4\lambda^2\varepsilon}$. It follows that for each $t_0\le t\le t(\varepsilon)$, there holds $\prod_{j=t+1}^{T}(1-\eta_j\lambda)\le\exp\{-\sum_{j=t+1}^{T}\eta_j\lambda\}\le\exp\{-\sum_{j=t(\varepsilon)+1}^{T}\eta_j\lambda\}\le\frac{4\lambda^2\varepsilon}{t(\varepsilon)}$. This in connection with the bound $\eta_t\lambda\le 1/2$ for each $t\ge t_0$ tells us that the first term of (4.10) is bounded by

$$\sum_{t=t_0}^{t(\varepsilon)}\eta_t^2\prod_{j=t+1}^{T}(1-\eta_j\lambda)\le\frac{4\lambda^2\varepsilon}{t(\varepsilon)}\sum_{t=t_0}^{t(\varepsilon)}\eta_t^2\le\varepsilon.$$

The second term of (4.10) is dominated by $\lambda\varepsilon\sum_{t=t(\varepsilon)+1}^{T-1}\eta_t\prod_{j=t+1}^{T}(1-\eta_j\lambda)$. But $\eta_t\lambda=1-(1-\eta_t\lambda)$. Then

$$\lambda\sum_{t=t(\varepsilon)+1}^{T}\eta_t\prod_{j=t+1}^{T}(1-\eta_j\lambda)=\sum_{t=t(\varepsilon)+1}^{T}[\prod_{j=t+1}^{T}(1-\eta_j\lambda)-\prod_{j=t}^{T}(1-\eta_j\lambda)]$$

$$=[1-\prod_{j=t(\varepsilon)+1}^{T}(1-\eta_j\lambda)]\le 1.$$

Therefore, when $T\ge\max\{T_1,T_2\}$, by Lemma 3, we have $E(\|\alpha_{T+1}-\alpha_\lambda\|_2^2)\le(1+2\tilde{C}_\lambda)\varepsilon$. This proves Theorem 1.

5 Convergence Rates

Now we can derive convergence rates for the error $\|\alpha_{T+1}-\alpha_\lambda\|_2$. The step size here is often of the form $\eta_t=\frac{1}{\mu(\lambda)t^\theta}$ for some $\theta\in(0,1]$ and $\mu(\lambda)>0$. So to apply Lemma 3 for getting error bounds, we need to estimate the summations in (4.3) and lead to the following lemmas.

Lemma 4[9]**.** For any $t<T$ and $0<\theta$, there holds

$$\sum_{j=t+1}^{T}j^{-\theta}\ge\begin{cases}\frac{1}{1-\theta}[(T+1)^{1-\theta}-(t+1)^{1-\theta}] & if\ \ \theta<1,\\ \log(T+1)-\log(t+1) & if\ \ \theta=1.\end{cases}\tag{5.1}$$

The proof follows from the simple inequality $\sum_{j=t+1}^{T} j^{-\theta} \geq \int_{t+1}^{T+1} u^{-\theta} du.$

The next lemma in a modified form was given in [13] for $1/2 < \theta < 1$.

Lemma 5[9]. Let $0 \leq v \leq 1$ and $0 < \theta \leq 1$. Then $\sum_{t=1}^{T-1} \frac{1}{t^{2\theta}} \exp\{-v \sum_{j=t+1}^{T} j^{-\theta}\}$ is bounded by

$$\begin{cases} \frac{18}{vT^{\theta}} + \frac{9T^{1-\theta}}{(1-\theta)2^{1-\theta}} \exp\{-\frac{v(1-2^{\theta-1})}{1-\theta}(T+1)^{1-\theta}\}, & \theta < 1, \\ \frac{8}{1-v}(T+1)^{-V}, & \theta = 1. \end{cases}$$

We are in a position to state the convergence rates for the online algorithm (0.4). To this end, we need the following constant depending on λ:

$$\widetilde{C}_{\lambda} = 4k^2 \left\| \phi_{-}^{'} \right\|_{L^{\infty}[-k^2|\phi^{'}(0)|/\lambda, k^2|\phi^{'}(0)|/\lambda]}^{2}. \tag{5.2}$$

Theorem 4. Assume that $\phi_{-}^{'}$ is locally Lipschitz at the origin. Choose the step sizes as $\eta_t = \frac{1}{\mu(\lambda)t^{\theta}}$ for some $\theta \in (0,1]$ and $\mu(\lambda) \geq M(\lambda)k^2 + \lambda$. Define $\{\alpha_t\}$ by (0.4) and $\widetilde{C}_{\lambda}$ by (5.2). Then there holds

(1) For $0 < \theta < 1$,

$$\begin{aligned} E(\|\alpha_{T+1} - \alpha_{\lambda}\|_2^2) \leq (&\frac{2D(\lambda)}{\lambda} + \frac{9\widetilde{C}_{\lambda}T^{1-\theta}}{(1-\theta)2^{1-\theta}(\mu(\lambda))^2}) \\ &\times \exp\{-\frac{(1-2^{\theta-1})\lambda}{(1-\theta)\mu(\lambda)}T^{1-\theta}\} + \frac{19\widetilde{C}_{\lambda}}{\mu(\lambda)\lambda T^{\theta}}. \end{aligned} \tag{5.3}$$

(2) For $\theta = 1$, $E(\|\alpha_{T+1} - \alpha_{\lambda}\|_2^2)$ can be bounded by

$$(\frac{2D(\lambda)}{\lambda} + \frac{9\widetilde{C}_{\lambda}}{\mu(\lambda)(\mu(\lambda)-\lambda)})T^{-\frac{\lambda}{\mu(\lambda)}}. \tag{5.4}$$

Proof. The condition on the step size tells us that $\eta_t(M(\lambda)k^2 + \lambda) \leq 1$ for each t. By Theorem 3, this yields $\|\alpha_t\|_2 \leq \frac{k|\phi^{'}(0)|}{\lambda}$, and hence $\|y_t f_{\alpha_t}(x_t)\|_2 \leq \frac{k^2|\phi^{'}(0)|}{\lambda}$ for each t. Consequently, both $y_t f_{\alpha_t}(x_t)$ and 0 lie in the interval $I_{\lambda} = [-k^2|\phi^{'}(0)|/\lambda, k^2|\phi^{'}(0)|/\lambda]$. It follows that $\|\phi_{-}^{`}(y_t f_{\alpha_t}(x_t))y_t K_{\bar{X}}(x_t)\|_2 \leq k\|\phi_{-}^{'}\|_{L^{\infty}(I_{\lambda})}^2$ and $\|\lambda f_{\alpha_t}(x_t)\|_2 \leq k|\phi^{'}(0)| \leq k\|\phi_{-}^{'}\|_{L^{\infty}(I_{\lambda})}^2$.

Then (4.1) holds with $t_0 = 1$ and $\widetilde{C}_{\lambda}$. This in connection with (4.3) of Lemma 3 tells us that $E(\|\alpha_{T+1} - \alpha_{\lambda}\|_2^2) \leq I_1 + I_2$, where

$$I_1=\exp\{-\sum_{t=1}^{T}\frac{\lambda}{\mu(\lambda)t^{\theta}}\}\|\alpha_1-\alpha_\lambda\|_2^2,\ I_2=\frac{\tilde{C}_\lambda}{(\mu(\lambda))^2}\sum_{t=1}^{T}\frac{1}{t^{2\theta}}\exp\left\{-\frac{\lambda}{\mu(\lambda)}\sum_{t=1}^{T}j^{-\theta}\right\}.$$

Since $\alpha_1=(0,\cdots,0)$, (1.4) gives $\|\alpha_1-\alpha_\lambda\|_2^2\le 2D(\lambda)/\lambda$. By Lemma 4, we know that

$$I_1\le\begin{cases}\dfrac{2D(\lambda)}{\lambda}\exp\left\{-\dfrac{(1-2^{\theta-1}\lambda)}{(1-\theta)\mu(\lambda)}(T+1)^{1-\theta}\right\}, & if\quad 0<\theta<1,\\[2ex] \dfrac{2D(\lambda)}{\lambda}(T+1)^{-\frac{\lambda}{\mu(\lambda)}}, & if\quad \theta=1.\end{cases}$$

By Lemma 5, I_2 can be bounded by

$$\begin{cases}\dfrac{\tilde{C}_\lambda}{(\mu(\lambda))^2}\left\{\dfrac{9T^{1-\theta}}{(1-\theta)2^{1-\theta}}\exp\left\{-\dfrac{(1-2^{\theta-1})\lambda}{(1-\theta)\mu(\lambda)}(T+1)^{1-\theta}\right\}+\dfrac{18\mu(\lambda)/\lambda}{T^{\theta}}+\dfrac{1}{T^{2\theta}}\right\}, & if\quad 0<\theta<1,\\[2ex] \dfrac{\tilde{C}_\lambda}{(\mu(\lambda))^2}\left\{\dfrac{8}{1-\lambda/\mu(\lambda)}(T+1)^{-\frac{\lambda}{\mu(\lambda)}}+\dfrac{1}{T^2}\right\}, & if\quad \theta=1.\end{cases}$$

This proves Theorem 4.

To apply Theorem 4 for deriving rates for the misclassification error, we need the constant $\tilde{C}_\lambda$ and $\mu(\lambda)$. They depend on the loss function ϕ and play an essential role in getting learning rates. When ϕ_-' satisfies the following increment condition with some $p\ge 0, c_p>0$:

$$|\phi_-'(x)|\le c_p|x|^p,\forall|x|\ge 1, \tag{5.5}$$

we can find $\tilde{C}_\lambda$ and $\mu(\lambda)$ explicitly and then derive learning rates for the total error from Theorem 4. Denote $C_{\phi,k}$ as a constant depending only on ϕ and k satisfying

$$\begin{aligned}C_{\phi,k}\ge\max\{&2k\|\phi_-'\|_{L^\infty[-1,1]}^{1/2},2k^{1+p}\sqrt{c_p}|\phi'(0)|^{p/2},\\ &1+k^2\|[\phi_-'(x)-\phi(0)']/x\|_{L^\infty[-1,1]},\\ &c_pk^{2p}|\phi'(0)|^{p-1}+k^2|\phi'(0)|+1\}.\end{aligned}\tag{5.6}$$

Corollary 2. Assume that ϕ satisfies (5.5) and ϕ_-' is locally Lipschitz at the origin. Let $0<\lambda\le 1$. Choose the step sizes as $\eta_t=\dfrac{\lambda^{\max\{p-1,0\}}}{C_{\phi,k}t^{\theta}}$ with some $0<\theta<1$ and $C_{\phi,k}$ given by (5.6). Define α_t by (0.4). Then

$$\begin{aligned}E(\|\alpha_{T+1}-\alpha_\lambda\|_2^2)\le(&\frac{2D(\lambda)}{\lambda}+\frac{9\lambda^{\max\{p-2,-p\}}T^{1-\theta}}{(1-\theta)2^{1-\theta}})\\ &\times\exp\{-\frac{(1-2^{\theta-1})\lambda^{\max\{p,1\}}}{(1-\theta)C_{\phi,k}}T^{1-\theta}\}+\frac{19C_{\phi,K}}{\lambda^{\min\{p+1,2\}}T^{\theta}}.\end{aligned}\tag{5.7}$$

Proof. The increment condition (5.5) for ϕ_-' tells us that

$$\tilde{C}_\lambda \le 4k^2 \max\{\|\phi'_-\|_{L^\infty[-1,1],} c_p (k^2 |\phi'(0)|/\lambda)^p\},$$

which implies $\tilde{C}_\lambda \le C^2_{\phi,k}\lambda^{-p}, \forall\, \lambda \le 1$. Also, the local Lipschitz constant can be bounded as

$$M(\lambda) \le \max\{\left\|\frac{\phi'_-(x)-\phi'(0)}{x}\right\|_{L^\infty[-1,1]},\ c_p(k^2|\phi'(0)|/\lambda)^{p-1}+|\phi'(0)|\}.$$

Choose $\mu(\lambda) = C_{\phi,k}\lambda^{-\max\{p-1,0\}}$. Then, $M(\lambda)k^2 + \lambda \le \mu(\lambda),\ \forall\, \lambda \le 1$, and our conclusion follows from Theorem 4.

6 Error Bounds and Learning Rates

Applying the above mentioned techniques, we can derive the learning rates of the excess misclassification error for the online algorithm (0.4) from our analysis on $\|\alpha_{T+1} - \alpha_\lambda\|_2$ together with the regularization error $D(\lambda)$.

The prediction power of classification algorithms are often measured by the misclassification error which is defined for a classifier C: $X \to Y$ to be the probability of the event

$$\{C(x) \ne Y\} : R(C) = \Pr ob\{C(x) \ne y\} = \int_X P(y \ne C(x) \mid x) d\rho_X, \qquad (6.1)$$

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*[13] and can be expressed as $f_c = \mathrm{sgn}(f_\rho)$, where f_ρ is the regression function

$$f_\rho(x) = \int_Y y d\rho(y \mid x) = P(y=1 \mid x) - P(y=-1 \mid x). \qquad (6.2)$$

In particular for the SVM 1-norm soft margin classifier with the hinge loss $\phi(x) = (1-x)_+$, we have an important relation[12] was given as

$$R(\mathrm{sgn}(f_\alpha)) - R(\mathrm{sgn}(f_\rho)) \le \varepsilon(f_\alpha) - \varepsilon(f_\rho). \qquad (6.3)$$

Such a relation is called a comparison theorem. For the general loss function, a simple comparison theorem was established in [10, 14].

Proposition A. Let ϕ be an admissible loss function such that $\phi''(0)$ exists and is positive. Then there is a constant c_ϕ such that for any measurable function f, there holds

$$R(\mathrm{sgn}(f_\alpha)) - R(\mathrm{sgn}(f_c)) \le c_\phi\sqrt{\varepsilon(f_\alpha) - \varepsilon(f_\rho)}. \qquad (6.4)$$

If moreover, for some $\tau \in [0.1]$ and $c > 0$, ρ satisfies a Tsybakov noise condition: for any measurable function f,

$$\rho_X(\mathrm{sgn}(f_\alpha) \ne f_c) \le c\{R(\mathrm{sgn}(f_\alpha)) - R(f_c)\}^\tau, \qquad (6.5)$$

then (6.4) can be improved as

$$R(\mathrm{sgn}(f_\alpha)) - R(f_c) \le \{2c_\phi c(\varepsilon(f_\alpha) - \varepsilon(f_\rho))\}^{1/(1-\tau)}. \qquad (6.6)$$

The Tsybakov noise condition (6.5) was introduced in [15] where the reader can find more details and explanation. The greater τ is, the smaller the noise of ρ is. In particular, any distribution ρ satisfies (6.5) with $\tau = 0$ and $c = 1$.

With a comparision theorem, it is sufficient for us to estimate the excess generalization error

(0.7). In order to do so, we need the regularization error between f_{α_λ} and f_ρ.

Let us present an example to illustrate the learning rates of (0.6) from suitable choices of the regularization parameter $\lambda = \lambda(T)$ and the step size η_t.

Now, we see an example corresponds to the classical support vector machine (SVM) with ϕ being the hinge loss $\phi(x) = (1-x)_+$. For this loss, the online algorithm (0.4) can be expressed as $\alpha_1 = (0,\cdots,0)$ and

$$\alpha_{t+1} = \begin{cases} (1-\eta_t\lambda)\alpha_t, & if \quad y_t f_{\alpha_t}(x_t) > 1, \\ (1-\eta_t\lambda)\alpha_t + \eta_t y_t K_{\overline{X}}(x_t), & if \quad y_t f_{\alpha_t}(x_t) \le 1. \end{cases} \tag{6.7}$$

Corollary 3. Let $\phi(x) = (1+x)_+$. Assume for some $0 < \beta \le 1$, the pair (ρ, K) satisfies

$$\inf_{\alpha \in R^m} \{\|f_\alpha - f_c\|_{L^1_{\rho_X}} + \lambda\|\alpha\|_2^2\} = O(\lambda^\beta). \tag{6.8}$$

For any $0 < \varepsilon < \frac{\beta}{2(\beta+1)}$, choose $\lambda = (T) = T^{\frac{\varepsilon}{\beta} - \frac{1}{2(\beta+1)}}$ and $\eta_t = \frac{1}{2+k^2} t^{\frac{(2\beta+1)\varepsilon}{\beta} - \frac{2\beta+1}{2(\beta+1)}}$,

then

$$E(R(\mathrm{sgn}(f_{\alpha_{T+1}})) - R(f_c)) = O(T^{\varepsilon - \frac{\beta}{2(\beta+1)}}). \tag{6.9}$$

In (6.9), the expectation E is taken with respect to the random sample $\mathrm{z} \in Z^T$. We shall use this notion throughout the paper, if the random variable for E is not specified. The condition (6.8) concerns the approximation of the function f_c in the L^1 space $L^1_{\rho_X}$ by functions from the RKHS H_K. It can be characterized by requiring f_c to lie in an interpolation space of the pair $(L^1_{\rho_X}, H_K)$ an intermediate space between the metric space $L^1_{\rho_X}$ and the much smaller approximation space H_K. For details, see the discussion in [10]. The assumption (6.8) is satisfied when we use the Gaussian kernels with flexible variances and the distribution satisfies some geometric noise condition.

Assumptions like (6.8) are necessary to determine the regularization parameter for achieving the learning rates (6.9). This can be seen from the literature [12] about the off-line algorithm (0.2), 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 condition on the distribution ρ and the space H_K.

Proof of corollary 3. Consider the hinge loss $\phi(x) = (1-x)_+$. Recall the relation (6.3) between the excess misclassification error and the excess generalization error. Then

$$R(\mathrm{sgn}(f_\alpha)) - R(f_c) \le \varepsilon(f_\alpha) - \varepsilon(f_{\alpha_\rho^\phi}) \le \varepsilon(f_\alpha) - \varepsilon(f_{\alpha_\lambda}) + D(\lambda). \tag{6.10}$$

Using the uniform Lipschitz continuity of hinge loss, we know that

$$\varepsilon(f_{\alpha_{T+1}}) - \varepsilon(f_{\alpha_\lambda}) \le \|f_{\alpha_{T+1}} - f_{\alpha_\lambda}\|_{L^1_{\rho_X}}.$$

Combined with the assumption on the regularization error $D(\lambda) \le c_\beta \lambda^\beta$ for some $c_\beta > 0$ and Corollary 1, it follows from (6.9) that

$$\begin{aligned} R(\mathrm{sgn}(f_{\alpha_{T+1}}))-R(f_c) &\le \left\|f_{\alpha_{T+1}}-f_{\alpha_\lambda}\right\|_{L^1_{\rho_X}}+c_\beta\lambda^\beta \\ &\le k\left\|f_{\alpha_{T+1}}-f_{\alpha_\lambda}\right\|_K+c_\beta\lambda^\beta \\ &\le k\sqrt{mk}\left\|\alpha_{T+1}-\alpha_\lambda\right\|_2+c_\beta\lambda^\beta. \end{aligned} \qquad (6.11)$$

Now we apply Corollary 2. It is easy to see that ϕ satisfies (5.5) with $p=0, c_p=1$ and $C_{\phi,k}=2+k^2$. For any $0<\varepsilon<\frac{\beta}{2(\beta+1)}$, choose $\theta=\frac{2\beta+1}{2(\beta+1)}-\frac{(2\beta+1)\varepsilon}{\beta}$ in Corollary 2. We know that $E(\left\|\alpha_{T+1}-\alpha_\lambda\right\|_2^2)$ is bounded by

$$[\frac{2D(\lambda)}{\lambda}+\frac{9T^{1-\theta}}{(1-\theta)2^{1-\theta}}]\times\exp\{-\frac{(1-2^{\theta-1})\lambda}{(1-\theta)(2+k^2)}T^{1-\theta}\}+\frac{19(2+k^2)}{\lambda T^\theta}.$$

Select $\lambda=T^{-\gamma}$ with $0<\gamma<\min(1-\theta,\theta)$. Since the asymptotic behavior $\exp\{-cT^\varepsilon\}=O(T^{-s})$ holds for any $\varepsilon>0, s>0$ and $c>0$. We know that there exists a constant $C_{\theta,\gamma,k}$ depending only on θ, k and γ such that $E(\left\|\alpha_{T+1}-\alpha_\lambda\right\|_2^2)\le C_{\theta,\ \gamma,\ k}T^{-\theta+\gamma}$.

Putting this back into (6.11), we have

$$E(R(\mathrm{sgn}(f_{\alpha_{T+1}}))-R(f_c))\le k\sqrt{mkC_{\theta,\gamma,k}}\,T^{-\frac{(\theta-\gamma)}{2}}+c_\beta T^{-\beta\gamma}.$$

Now take $\gamma=\frac{1}{2(\beta+1)}-\frac{\varepsilon}{\beta}$. We know that the following holds

$$E(R(\mathrm{sgn}(f_{\alpha_{T+1}}))-R(f_c))=O(T^{\varepsilon-\frac{\beta}{2(\beta+1)}}).$$

This proves our conclusion.

Corollary 4. Let ϕ be an admissible loss function and $f_{\alpha_\lambda}\in H_K$. There holds

$$\begin{aligned} \varepsilon(f_\alpha)-\varepsilon(f_{\alpha_\lambda}) &\le \kappa\left\|f_\alpha-f_\lambda\right\|_K\max\{\left\|\phi'_+\right\|_{L^\infty(I_\lambda)},\left\|\phi'_-\right\|_{L^\infty(I_\lambda)}\} \\ &\le k\sqrt{mk}\left\|\alpha-\alpha_\lambda\right\|_2\max\{\left\|\phi'_+\right\|_{L^\infty(I_\lambda)},\left\|\phi'_-\right\|_{L^\infty(I_\lambda)}\}, \end{aligned}$$

where I_λ is the interval $I_\lambda=\left[-C_{\lambda,f_\alpha},C_{\lambda,f_\alpha}\right]$ with

$$\begin{aligned} C_{\lambda,f_\alpha} &:= \max\{\kappa\left\|f_\alpha\right\|_K,\kappa\sqrt{m\kappa 2D(\lambda)/\lambda}\} \\ &= \max\{\kappa\sqrt{m\kappa}\left\|\alpha\right\|_2,\kappa\sqrt{2mD(\lambda)/\lambda}\}. \end{aligned}$$

Proof. Since $\varepsilon(f_{\alpha_\lambda})-\varepsilon(f_\rho)\ge 0$, the definition for $D(\lambda)$ tells us that

$$\left\|f_{\alpha_\lambda}\right\|_K\le\sqrt{2m\kappa D(\lambda)/\lambda}. \qquad (6.12)$$

Note the elementary inequality

$$\left|\phi(u)-\phi(s)\right|\le\max\{\left\|\phi'_+\right\|_{L^\infty(I)},\left\|\phi'_-\right\|_{L^\infty(I)}\}\left|u-s\right|,$$

where I is an interval containing u and s. Applying this inequality with $u=yf_\alpha(x)$ and $s=yf_{\alpha_\lambda}(x)$, we know that for any $y\in Y$, $x\in X$, $\left|\phi(yf_\alpha(x))-\phi(yf_{\alpha_\lambda}(x))\right|$ can be bounded by

$$\left|f_\alpha(x)-f_{\alpha_\lambda}(x)\right|\max\{\|\phi'_+\|_{L^\infty}(I),\|\phi'_-\|_{L^\infty}(I)\},$$

where I is an interval containing $yf_\alpha(x)$ and $yf_\lambda(x)$. But $\|f_\alpha\|_\infty\le k\|f_\alpha\|_K$ implies $|yf_\alpha(x)|\le\|f_\alpha\|_\infty\le k\|f_\alpha\|_K$ and $|yf_{\alpha_\lambda}(x)|\le k\|f_\alpha\|_K\le k\sqrt{2mkD(\lambda)/\lambda}$, where $\|f_\alpha\|_K\le\sqrt{mk}\|\alpha\|_2$. In connection with the fact $\varepsilon(f_\alpha)-\varepsilon(f_{\alpha_\lambda})\le\int_Z\left|\phi(yf_\alpha(x))-\phi(yf_{\alpha_\lambda}(x))\right|d\rho$ and the inequality $\|f_\alpha-f_{\alpha_\lambda}\|_\infty\le k\|f_\alpha-f_{\alpha_\lambda}\|_K$, connectting with Corollary 1, this proves our conclusion.

7 Conclusion

In this paper, we provided an online generalized gradient classification algorithms for the coefficient regularized classification algorithm. Different from the off-line algorithm, it's also suitable when the sample size or data becomes very large. It is based on regularization schemes in the space R^m associated with general convex loss functions. For each $t\in N_T$, the function α_t is in general dependent on the inputs $\{Z_t:t\in N_T\}$ and we investigated explicit capacity independent learning rates of the excess misclassification error for the hinge loss. Throughout the paper a novel relation between α_{T+1} and α_λ played an important role in the excess regularized generalization error. That is to say, the expectation $\|\alpha_{T+1}-\alpha_\lambda\|_2$ over the random samples of regularized sample error is aim at the excess misclassification error $R(\mathrm{sgn}(f_{\alpha_{T+1}}))-R(\mathrm{sgn}(f_\rho))$, which is bounded by the excess generalization error $\varepsilon(f_{\alpha_{T+1}})-\varepsilon(f_\rho)$.

References

[1] BOSER B E, GUYON I M, VAPNIK V N. A training algorithm for optimal margin classifiers [A]. Pittsburgh. Annual ACM Workshop on COLT[C]. New York: PA. ACM Press, 1992. 144-152.

[2] ARONSZAJN. Theory of reproducing kernels[J]. Thans Amer Math Soc, 1950, 68: 337-404.

[3] VAPNIK V. Statistical Learning Theory [M]. New York : Wiley, 1998.

[4] ARGYRIOU A, MICCHELLI C A, PONTIL M, When is there a representer theorem? Vector versus Matrix Regularizers[J], J Mach Leaning Res, 2009, 10: 2507-2529.

[5] DINUZZO F, NEVE M, NICOLAO G D, GIANAZZA U P. On the representer theorem and equivalent degrees of freedom of SVR[J]. J Mach Learning Res, 2007, 8: 2467-2495.

[6] WU Q, ZHOU D X. Learning with sample dependent hypothesis spaces[J]. Computers and Mathematics with Applications, 2008, 56: 2896-2907.

[7] YE G B, ZHOU D X. Fully online classification by regularization[J]. Appl Comput Harmon Anal, 2007, 23: 198-214.

[8] YING Y M. Convergence analysis of online algorithms[J]. Advances in Computational Mathematics, 2007, 27: 273-291.

[9] YING Y M, ZHOU D X. Online Regularized Classification Algorithms[J]. IEEE Trans Inform Theory, 2006, 52: 4775-4788.

[10] CHEN D R, WU Q, YING Y M, et al. Support vector machine soft margin classifiers: error analysis[J]. J Machine Learning Res, 2004, 5: 1143-1175.

[11] BOYD S, VANDEBBERGHE L. Convex optimization[M]. Cambridge: Cambridge University Press, 2004.

[12] ZHANG T. Statistical behavior and consistency of classification methods based on convex risk minimization[J]. Ann Stat, 2004, 32: 56-85.

[13] DEVROYE L, GYORFI L, LUGOSI G. A Probabilistic Theory of Pattern Recognition[M]. New York: Springer Verlag, 1997.

[14] BARTLETT P L, JORDAN M I, MCAULIFFE J D. Convexity classification and risk bounds[J]. J Amer Statist Assoc, 2006, 101: 138-156.

[15] TSYBAKOV A B. Optimal aggregation of classifiers in statistical learning[J]. Ann Stat, 2004, 32: 135-166.
SMALE S, ZHOU D X. Learning theory estimates via integral operators and their applications[J]. Constr Approx, 2007, 26: 153-172.