Journal of Machine Learning Research 10 (2009) 2873-2898 Submitted 1/08; Revised 7/09; Published 12/09

Online Learning with Samples Drawn from Non-identical Distributions

Ting Hu TING_HU2005@163.COM
Ding-Xuan Zhou MAZHOU@CITYU.EDU.HK
Department of Mathematics
City University of Hong Kong
Tat Chee Avenue, Kowloon, Hong Kong, China

Editor: Peter Bartlett

Abstract

Learning algorithms are based on samples which are often drawn independently from an identical distribution (i.i.d.). In this paper we consider a different setting with samples drawn according to a non-identical sequence of probability distributions. Each time a sample is drawn from a different distribution. In this setting we investigate a fully online learning algorithm associated with a general convex loss function and a reproducing kernel Hilbert space (RKHS). Error analysis is conducted under the assumption that the sequence of marginal distributions converges polynomially in the dual of a Hölder space. For regression with least square or insensitive loss, learning rates are given in both the RKHS norm and the L^2 norm. For classification with hinge loss and support vector machine q-norm loss, rates are explicitly stated with respect to the excess misclassification error.

Keywords: sampling with non-identical distributions, online learning, classification with a general convex loss, regression with insensitive loss and least square loss, reproducing kernel Hilbert space

1. Introduction

In the literature of learning theory, samples for algorithms are often assumed to be drawn independently from an identical distribution. Here we consider a setting with samples drawn from non-identical distributions. Such a framework was introduced in Smale and Zhou (2009) and Steinwart et al. (2008) where online learning for least square regression and off-line support vector machines are investigated. We shall follow this framework and study a kernel based online learning algorithm associated with a general convex loss function. Our analysis can be applied for various purposes including regression and classification.

1.1 Sampling with Non-identical Distributions

Let (X,d) be a metric space called an input space for the learning problem. Let Y be a compact subset of $\mathbb{R}$ (output space) and $Z = X \times Y$.

In our online learning setting, at each step $t = 1,2,\ldots$, a pair $z_t = (x_t, y_t)$ is drawn from a probability distribution $\rho^{(t)}$ on Z. The sampling sequence of probability distributions $\{\rho^{(t)}\}_{t=1,2,\cdots}$ is not identical. For convergence analysis, we shall assume that the sequence $\{\rho_X^{(t)}\}_{t=1,2,\cdots}$ of marginal distributions on X converges polynomially in the dual of the Hölder space $C^s(X)$ for some $0 < s \leq 1$.

Here the Hölder space $C^s(X)$ is defined to be the space of all continuous functions on X with the norm $\|f\|_{C^s(X)} = \|f\|_{C(X)} + |f|_{C^s(X)}$ finite, where $|f|_{C^s(X)} := \sup_{x\neq y} \frac{|f(x)-f(y)|}{(d(x,y))^s}$.

Definition 1 *We say that the sequence $\{\rho_X^{(t)}\}_{t=1,2,\cdots}$ converges polynomially to a probability distribution ρ_X in $(C^s(X))^*$ ($0 \le s \le 1$) if there exist $C > 0$ and $b > 0$ such that*

$$\|\rho_X^{(t)} - \rho_X\|_{(C^s(X))^*} \le Ct^{-b}, \qquad t \in \mathbb{N}. \tag{1}$$

By the definition of the dual space $(C^s(X))^*$, decay condition (1) can be expressed as

$$\left| \int_X f(x) d\rho_X^{(t)} - \int_X f(x) d\rho_X \right| \le Ct^{-b} \|f\|_{C^s(X)}, \qquad \forall f \in C^s(X), t \in \mathbb{N}. \tag{2}$$

What measures quantitatively differences between our non-identical setting and the i.i.d. case is the power index b. Its impact on performance of online learning algorithms will be studied in this paper. The i.i.d. case corresponds to $b = \infty$.

We describe three situations in which decay condition (1) is satisfied. The first is when a distribution ρ_X is perturbed by some noise and the noise level decreases as t increases.

Example 1 *Let $\{h^{(t)}\}$ be a sequence of bounded functions on X such that $\sup_{x\in X} |h^{(t)}(x)| \le Ct^{-b}$. Then the sequence $\{\rho_X^{(t)}\}_{t=1,2,\cdots}$ defined by $d\rho_X^{(t)} = d\rho_X + h^{(t)}(x)d\rho_X$ satisfies (1) for any $0 \le s \le 1$.*

The proof follows from $\left|\int_X f(x)h^{(t)}(x)d\rho_X\right| \le \sup_{x\in X} |h^{(t)}(x)| \|f\|_{C(X)} \le Ct^{-b}\|f\|_{C^s(X)}$. In this example, $h^{(t)}$ is the density function of the noise distribution and we assume its bound (noise level) to decay polynomially as t increases.

The second situation when decay condition (1) is satisfied is generated by iterative actions of an integral operator associated with a stochastic density kernel. We demonstrate this situation by an example on $X = S^{n-1}$, the unit sphere of $\mathbb{R}^n$ with $n \ge 2$. Let dS be the normalized surface element of S^{n-1}. The corresponding space $L^2(S^{n-1})$ has an orthonormal basis $\{Y_{\ell,k} : \ell \in \mathbb{Z}_+, k = 1, \ldots, N(n,\ell)\}$ with $N(n,0) = 1$ and $N(n,\ell) = \frac{2\ell+n-2}{\ell} \frac{(\ell+n-3)!(n-2)!}{(\ell-1)!}$. Here $Y_{\ell,k}$ is a spherical harmonic of order ℓ which is the restriction onto S^{n-1} of a homogeneous polynomial in $\mathbb{R}^n$ of degree ℓ satisfying the Laplace equation $\Delta f = 0$. In particular, $Y_{0,0} \equiv 1$.

Example 2 *Let $X = S^{n-1}$, $0 < \alpha < 1$, and $\psi \in C(X \times X)$ be given by*

$$\psi(x,u) = 1 + \sum_{\ell=1}^{\infty} \sum_{k=1}^{N(n,\ell)} a_{\ell,k} Y_{\ell,k}(x) Y_{\ell,k}(u) \quad \textit{where} \quad 0 \le a_{\ell,k} \le \alpha, \sum_{\ell=1}^{\infty} \sum_{k=1}^{N(n,\ell)} a_{\ell,k} \|Y_{\ell,k}\|_{C(X)}^2 < 1.$$

If $h^{(1)}$ is a square integrable density function on X and a sequence of density functions $\{h^{(t)}\}$ is defined by

$$h^{(t+1)}(x) = \int_X \psi(x,u) h^{(t)}(u) dS(u), \qquad x \in X,\ t \in \mathbb{N},$$

then we know $h^{(t)} = Y_{0,0} + \sum_{\ell=1}^{\infty} \sum_{k=1}^{N(n,\ell)} a_{\ell,k}^{t-1} \langle h^{(1)}, Y_{\ell,k} \rangle_{L^2(S^{n-1})} Y_{\ell,k}$ and $\|h^{(t)} - Y_{0,0}\|_{L^2(S^{n-1})} \le \alpha^{t-1} \|h^{(1)}\|_{L^2(S^{n-1})}$. It follows that the sequence $\{\rho_X^{(t)} = h^{(t)}(x)dS\}_{t=1,2,\cdots}$ of probability distributions on X converges polynomially to the uniform distribution $d\rho_X = dS$ on X and satisfies (1) for any $0 \le s \le 1$.

In general, if ν is a strictly positive probability distribution on X, and if $\psi \in C(X \times X)$ is strictly positive satisfying $\int_X \psi(x,u)d\nu(u) = 1$ for each $x \in X$, then the sequence $\{\rho_X^{(t)}\}$ defined by

$$\rho_X^{(t)}(\Gamma) = \int_\Gamma \left\{ \int_X \psi(x,u) d\rho_X^{(t-1)}(x) \right\} d\nu(u) \qquad \text{on Borel sets} \quad \Gamma \subseteq X$$

satisfies $\|\rho_X^{(t)} - \rho_X\|_{(C(X))^*} \leq C\alpha^t$ for some (strictly positive) probability distribution ρ_X on X and constants $C > 0, 0 < \alpha < 1$. Hence decay condition (1) is valid for any $0 \leq s \leq 1$. For details, see Smale and Zhou (2009).

The third situation to realize decay condition (1) is to induce distributions by dynamical systems. Here we present a simple example.

Example 3 *Let $X = [-1/2, 1/2]$ and for each $t \in \mathbb{N}$, the probability distribution $\rho_X^{(t)}$ on X has support $[-2^{-t}, 2^{-t}]$ and uniform density 2^{t-1} on its support. Then with δ_0 being the delta distribution at the origin, for each $0 < s \leq 1$ we have*

$$\left| \int_X f(x) d\rho_X^{(t)} - \int_X f(x) d\delta_0 \right| \leq 2^{t-1} \int_{-2^{-t}}^{2^{-t}} |f(x) - f(0)| dx \leq \left(2^{-s}\right)^t \|f\|_{C^s(X)}.$$

Remark 2 *Since $\|f\|_{C(X)} \leq \|f\|_{C^s(X)}$, we see from (2) that decay condition (1) with any $0 < s \leq 1$ is satisfied when this polynomial convergence requirement is valid in the case $s = 0$. This happens in Examples 1 and 2. Note that when $s = 0$, the dual space $(C(X))^*$ is exactly the space of signed finite measures on X. Each signed finite measure μ on X lies in $(C(X))^* \subset (C^s(X))^*$ and satisfies $\|\mu\|_{(C^s(X))^*} \leq \|\mu\|_{(C(X))^*} \leq \int_X d|\mu|$.*

1.2 Fully Online Learning Algorithm

In this paper we study a family of online learning algorithms associated with reproducing kernel Hilbert spaces and a general convex loss function.

A *reproducing kernel Hilbert space* (RKHS) is induced by a *Mercer kernel* $K : X \times X \to \mathbb{R}$ which is a continuous and symmetric function such that the matrix $(K(x_i, x_j))_{i,j=1}^\ell$ is positive semidefinite for any finite set of points $\{x_1, \cdots, x_\ell\} \subset X$. The RKHS $\mathcal{H}_K$ is the completion (Aronszajn, 1950) of the span of the set of functions $\{K_x = K(x, \cdot) : x \in X\}$ with the inner product given by $\langle K_x, K_y \rangle_K = K(x,y)$.

Definition 3 *We say that $V : Y \times \mathbb{R} \to \mathbb{R}_+$ is a convex loss function if for each $y \in Y$, the univariate function $V(y, \cdot) : \mathbb{R} \to \mathbb{R}_+$ is convex.*

The convexity tells us (Rockafellar, 1970) that for each $f \in \mathbb{R}$ and $y \in Y$, the left derivative $\lim_{\delta \to 0-}(V(y, f+\delta) - V(y,f))/\delta$ exists and is no more than the right derivative $\lim_{\delta \to 0+}(V(y, f+\delta) - V(y,f))/\delta$. An arbitrary number between them (which is a gradient) will be taken and denoted as $\partial V(y,f)$ in our algorithm.

For the least square regression problem, we can take $V(y,f) = (y-f)^2$. For the binary classification problem, we can take $V(y,f) = \phi(yf)$ with $\phi : \mathbb{R} \to \mathbb{R}_+$ a convex function.

The online algorithm associated with the RKHS $\mathcal{H}_K$ and the convex loss V is a stochastic gradient descent method (Cesa-Bianchi et al., 1996; Kivinen et al., 2004; Smale and Yao, 2006; Ying and Zhou, 2006; Ying, 2007).

Definition 4 *The fully online learning algorithm is defined by* $f_1 = 0$ *and*

$$f_{t+1} = f_t - \eta_t \{\partial V(y_t, f_t(x_t))K_{x_t} + \lambda_t f_t\}, \text{ for } t = 1, 2, \ldots, \tag{3}$$

where $\lambda_t > 0$ *is called the regularization parameter and* $\eta_t > 0$ *the step size.*

In this fully online algorithm, the regularization parameter λ_t changes with the learning step t. Throughout the paper we assume that $\lambda_{t+1} \leq \lambda_t$ for each $t \in \mathbb{N}$. When the regularization parameter $\lambda_t \equiv \lambda_1$ does not change as the step t develops, we call scheme (3) *partially online*.

The goal of this paper is to investigate the fully online learning algorithm (3) when the sampling sequence is not identical. We will show that learning rates in the non-identical setting can be the same as those in the i.i.d. case when the power index b in polynomial decay condition (1) is large enough, that is, $\{\rho_X^{(t)}\}$ converges fast to ρ_X. When b is small, the non-identical effect becomes crucial and the learning rates will depend essentially on b.

2. Error Bounds for Regression and Classification

As in the work on least square regression (Smale and Zhou, 2009), we assume for the sampling sequence $\{\rho^{(t)}\}_{t=1,2,\cdots}$ that the conditional distribution $\rho^{(t)}(y|x)$ of each $\rho^{(t)}$ at $x \in X$ is independent of t, denoted as ρ_x.

Throughout the paper we assume independence of the sampling, that is, $\{z_t = (x_t, y_t)\}_t$ is a sequence of samples drawn from the product probability space $\Pi_{t=1,2,\cdots}(Z, \rho^{(t)})$.

Error analysis will be conducted for fully online learning algorithm (3) under polynomial decay condition (1) for the sequence of marginal distributions $\{\rho_X^{(t)}\}$. Let ρ be the probability distribution on Z given by the marginal distribution ρ_X and the conditional distributions ρ_x. Essential difficulty in our non-identical setting is caused by the deviation of $\{\rho^{(t)}\}$ from ρ.

The first novelty of our analysis is to deal with an error quantity Δ_t involving $\rho^{(t)} - \rho$ (defined by (15) below) which occurs only in the non-identical setting. This is handled for a general loss function V and output space Y by Lemma 18 in Section 3 under decay condition (1) for marginal distributions $\{\rho_X^{(t)}\}$ and Lipschitz s continuity of conditional distributions $\{\rho_x : x \in X\}$.

Definition 5 *We say that the set of distributions* $\{\rho_x : x \in X\}$ *is Lipschitz* s *in* $(C^s(Y))^*$ *if there exists a constant* $C_\rho \geq 0$ *such that*

$$\|\rho_x - \rho_u\|_{(C^s(Y))^*} \leq C_\rho (d(x,u))^s, \qquad \forall x, u \in X. \tag{4}$$

Notice that on the compact subset Y of $\mathbb{R}$, the Hölder space $C^s(Y)$ and its dual $(C^s(Y))^*$ are well defined. Each ρ_x belongs to $(C^s(Y))^*$.

The second novelty of our analysis is to show for the least square loss (described in Section 3) and binary classification that Lipschitz s continuity (4) of $\{\rho_x : x \in X\}$ is the same as requiring $f_\rho \in C^s(X)$ where f_ρ is the *regression function* defined by

$$f_\rho(x) = \int_Y y d\rho_x(y), \qquad x \in X. \tag{5}$$

Proposition 6 *Let* $0 < s \leq 1$. *Condition (4) implies* $f_\rho \in C^s(X)$ *with* $|f_\rho|_{C^s(X)} \leq C_\rho(1 + 2^{1-s}) \sup_{y \in Y} |y|$. *When* $Y = \{1, -1\}$, $f_\rho \in C^s(X)$ *also implies (4) and* $C_\rho \leq |f_\rho|_{C^s(X)}$.

The two-point nature of the output space Y for binary classification plays a crucial role in our observation. The second statement of Proposition 6 is not true for general output space Y. Here is one example.

Example 4 *Let $0 < s \leq 1$ and $Y = \{1, -1, 0\}$. Then condition (4) holds if and only if $f_\rho \in C^s(X)$ and $f_{\rho,-1} \in C^s(X)$ where $f_{\rho,-1}$ is the function on X given by $f_{\rho,-1}(x) = \rho_x(\{-1\})$.*

Proofs of Proposition 6 and Example 4 will be given in the appendix.

Our third novelty is to understand some essential differences between our non-identical setting and the classical i.i.d. setting by pointing out the key role played by the power index b of polynomial decay condition $\|\rho_X^{(t)} - \rho_X\|_{(C^s(X))^*} \leq Ct^{-b}$ in derived convergence rates in Theorems 7 and 10 for regression and Theorem 11 for classification. Even for least square regression our result improves the error analysis in Smale and Zhou (2009) where a stronger exponential decay condition $\|\rho_X^{(t)} - \rho_X\|_{(C^s(X))^*} \leq C\alpha^t$ is assumed.

Our error bounds for fully online algorithm (3) are comparable with those for a batch learning algorithm generated by the off-line regularization scheme in $\mathcal{H}_K$ defined with a sample $\mathbf{z} := \{z_t = (x_t, y_t)\}_{t=1}^T$ and a regularization parameter $\lambda > 0$ as

$$f_{\mathbf{z},\lambda} = \arg\min_{f \in \mathcal{H}_K} \left\{ \frac{1}{T} \sum_{t=1}^{T} V(y_t, f(x_t)) + \frac{\lambda}{2} \|f\|_K^2 \right\}. \tag{6}$$

Let us demonstrate our error analysis by learning rates for regression with least square loss and insensitive loss and for binary classification with hinge loss.

2.1 Learning Rates for Least Square Regression

Here we take $Y = [-M, M]$ for some $M > 0$ and the least square loss $V = V_{ls}$ as $V_{ls}(y, f) = (y - f)^2$. Then the algorithm takes the form

$$f_{t+1} = f_t - 2\eta_t \left\{ (f_t(x_t) - y_t) K_{x_t} + \frac{\lambda_t}{2} f_t \right\}, \text{ for } t = 1, 2, \ldots$$

The following learning rates are derived by the procedure in Smale and Zhou (2009) where an exponential decay condition is assumed. Here we only impose a much weaker polynomial decay condition (1). We also assume the regularity condition (of order $r > 0$)

$$f_\rho = L_K^r(g_\rho) \text{ for some } g_\rho \in L^2_{\rho_X}(X), \tag{7}$$

where L_K is the integral operator $L^2_{\rho_X}$ defined by

$$L_K f(x) = \int_X K(x, v) f(v) d\rho_X(v), \qquad x \in X$$

with L_K^r well-defined as a compact operator.

Theorem 7 *Let $0 < s \leq \frac{1}{2}$ and $\frac{1}{2} < r \leq \frac{3}{2}$. Assume $K \in C^{2s}(X \times X)$, regularity condition (7) for f_ρ and (1) with $b > \frac{2r+2}{2r+1}$ for $\{\rho_X^{(t)}\}$. Take*

$$\lambda_t = \lambda_1 t^{-\frac{1}{2r+1}}, \quad \eta_t = \eta_1 t^{-\frac{2r}{2r+1}}$$

with $\lambda_1\eta_1 > \frac{2r-1}{4r+2}$, *then*

$$\mathbb{E}_{z_1,\dots,z_T}\left(\|f_{T+1}-f_\rho\|_K\right) \le \widetilde{C}T^{-\frac{2r-1}{4r+2}},$$

where $\widetilde{C}$ *is a constant independent of* T.

Denote the constant $\kappa = \max_{x\in X}\sqrt{K(x,x)}$. From the reproducing property

$$\langle K_x, f\rangle_K = f(x), \qquad x\in X,\ f\in\mathcal{H}_K \tag{8}$$

of the RKHS $\mathcal{H}_K$, we see that

$$\|f\|_{C(X)} \le \kappa\|f\|_K, \qquad \forall x\in X, f\in\mathcal{H}_K.$$

Most error analysis in the literature of least square regression (Zhang, 2004; De Vito et al., 2005; Smale and Yao, 2006; Wu et al., 2007) is about the L^2-norm $\|f_{T+1}-f_\rho\|_{L^2_{\rho_X}}$ or risk in the i.i.d. case. From a predictive viewpoint, in the non-identical setting, the error $f_{T+1}-f_\rho$ should be measured with respect to the distribution $\rho_X^{(T)}$, not the limit ρ_X. This can be done by bounding $\|f_{T+1}-f_\rho\|_{C(X)}$ (since $\rho_X^{(T)}$ changes with T), which follows from estimates for $\|f_{T+1}-f_\rho\|_K$. So our bounds for the error in the $\mathcal{H}_K$-norm provides useful predictive information about learning ability of fully online algorithm (3) in the non-identical setting.

Remark 8 *When* $X\subset\mathbb{R}^n$ *and* $K\in C^{2m}(X\times X)$ *for some* $m\in\mathbb{N}$, *we know from Zhou (2003, 2008) and Theorem 7 that* $\mathbb{E}_{z_1,\dots,z_T}\left(\|f_{T+1}-f_\rho\|_{C^m(X)}\right) = O(T^{-\frac{2r-1}{4r+2}})$. *So the regression function is learned efficiently by the online algorithm not only in the usual* $L^2_{\rho_X}$ *space, but also strongly in the space* $C^m(X)$ *implying the learning of gradients (Mukherjee and Wu, 2006).*

In the special case of $r=\frac{3}{2}$, the learning rate in Theorem 7 is $\mathbb{E}_{z_1,\dots,z_T}\left(\|f_{T+1}-f_\rho\|_K\right) = O(T^{-\frac{1}{4}})$, the same as those in the literature (Smale and Zhou, 2009; Tarrès and Yao, 2005; Smale and Zhou, 2007). Here we assume polynomial convergence condition (1) with a large index $b > \frac{2r+2}{2r+1}$. So the influence of the non-identical distributions $\{\rho_X^{(t)}\}$ does not appear in the learning rates (it is involved in the constant $\widetilde{C}$). Instead of refining the analysis for smaller b in Theorem 7, we shall show the influence of the index b on learning rates by the settings of regression with insensitive loss and binary classification.

2.2 Learning Rates for Regression with Insensitive Loss

A large family of loss functions for regression take the form $V(y,f)=\psi(y-f)$ where $\psi:\mathbb{R}\to\mathbb{R}_+$ is an even, convex and continuous function satisfying $\psi(0)=0$. One example is the ε-insensitive loss (Vapnik, 1998) with $\varepsilon\ge 0$ where $\psi(u)=\max\{|u|-\varepsilon,0\}$. We consider the case when $\varepsilon=0$. In this case the loss is called least absolute deviation or least absolute error in the literature of statistics and finds applications in some important problems because of robustness.

Definition 9 *The insensitive loss* $V=V_{in}$ *is given by* $V_{in}(y,f)=|y-f|$.

Algorithm (3) now takes the form

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

The following learning rates are new and will be proved in Section 5.

Theorem 10 *Let $0 < s \le \frac{1}{2}$ and $K \in C^{2s}(X \times X)$. Assume regularity condition (7) for f_ρ with $r > \frac{1}{2}$, and polynomial convergence condition (1) for $\{\rho_X^{(t)}\}$. Suppose that for each $x \in X$, ρ_x is the uniform distribution on the interval $[f_\rho(x)-1, f_\rho(x)+1]$. If $\lambda_1 \le (\kappa\|g_\rho\|_{L^2_{\rho_X}})^{2/(1-2r)}/2$ and $\eta_1 > 0$, then with a constant $\widetilde{C}$ independent of T, when $1 < r \le \frac{3}{2}$ we have*

$$\mathbb{E}_{z_1,\dots,z_T}\left(\|f_{T+1}-f_\rho\|_K\right) \le \widetilde{C}T^{-\min\{\frac{2r-1}{6r+1},\frac{b}{2}-\frac{2}{6r+1}\}} \quad \text{by taking } \lambda_t = \lambda_1 t^{-\frac{2}{6r+1}},\ \eta_t = \eta_1 t^{-\frac{4r}{6r+1}}$$

and when $\frac{1}{2} < r \le 1$, we have

$$\mathbb{E}_{z_1,\dots,z_T}\left(\|f_{T+1}-f_\rho\|_{L^2_{\rho_X}}\right) \le \widetilde{C}T^{-\min\{\frac{r}{3r+2},\frac{b}{2}-\frac{1}{3r+2}\}} \quad \text{with } \lambda_t = \lambda_1 t^{-\frac{1}{3r+2}},\ \eta_t = \eta_1 t^{-\frac{2r+1}{3r+2}}.$$

Again, when $b > \frac{4r+2}{6r+1}$, $X \subset \mathbb{R}^n$ and $K \in C^{2m}(X \times X)$ for some $m \in \mathbb{N}$, Theorem 10 tells us that $\mathbb{E}_{z_1,\dots,z_T}\left(\|f_{T+1}-f_\rho\|_{C^m(X)}\right) = O(T^{-\frac{2r-1}{6r+1}})$. So the regression function is learned efficiently by the online algorithm not only with respect to the risk, but also strongly in the space $C^m(X)$ implying the learning of gradients.

Consider the case $r = \frac{3}{2}$. When $\{\rho_X^{(i)}\}$ converges slowly with $b < \frac{4}{5}$, we see that $\mathbb{E}_{z_1,\dots,z_T}\left(\|f_{T+1}-f_\rho\|_K\right) = O(T^{-(\frac{b}{2}-\frac{1}{5})})$ which heavily depends on the index b representing quantitatively the deviation of $\{\rho_X^{(t)}\}$ from ρ_X. When b is large enough with $b \ge \frac{4}{5}$, the learning rate $\mathbb{E}_{z_1,\dots,z_T}\left(\|f_{T+1}-f_\rho\|_K\right)$ is of order $T^{-\frac{1}{5}}$ which is independent of b.

It would be interesting to extend Theorem 10 to situations when $\{\rho_x : x \in X\}$ are more general bounded symmetric distributions.

2.3 Learning Rates for Binary Classification

The output space Y for the binary classification problem is $Y = \{1,-1\}$ representing the set of two classes. A binary *classifier* $C : X \to Y$ makes a prediction $y = C(x) \in Y$ for each point $x \in X$.

A real valued function $f : X \to \mathbb{R}$ can be used to generate a classifier $C(x) = \mathrm{sgn}(f(x))$ where $\mathrm{sgn}(f(x)) = 1$ if $f(x) \ge 0$ and $\mathrm{sgn}(f(x)) = -1$ if $f(x) < 0$. A classifying convex loss $\phi : \mathbb{R} \to \mathbb{R}_+$ is often used for the real valued function f, to measure the local error $\phi(yf(x))$ suffered from the use of $\mathrm{sgn}(f)$ as a model for the process producing y at $x \in X$. Take $V(y,f) = \phi(yf)$ in out setting. Off-line classification algorithm (6) has been extensively studied in the literature. In particular, the error analysis is well done when the sample $\mathbf{z}$ is assumed to be an identical and independent drawer from a probability measure ρ on Z. See, for example, Evgeniou et al. (2000), Steinwart (2002), Zhang (2004) and Wu et al. (2007). Some analysis for off-line support vector machines with dependent samples can be found in Steinwart et al. (2008).

When the sample size T is very large, algorithm (6) might be practically challenging. Then online learning algorithms can be applied, which provide more efficient methods for classification. These algorithms are generalizations of the perceptron which has a long history. The error analysis for online algorithm (3) has also been conducted for classification in the i.i.d. setting, see, for example, Cesa-Bianchi et al. (2004), Ying and Zhou (2006), Ying (2007) and Ye and Zhou (2007).

Here we are interested in the error analysis for fully online algorithm (3) in the non-identical setting. The error is measured by the *misclassification error* $\mathcal{R}(C)$ defined to be the probability of the event $\{C(x) \ne y\}$ for a classifier C

$$\mathcal{R}(C) = \int_X \rho_x(y \ne C(x))d\rho_X(x) = \int_X \rho_x(\{-C(x)\})d\rho_X(x).$$

The best classifier minimizing the misclassification error is called the *Bayes rule* (e.g., Devroye et al. 1997) and can be expressed as $f_c = \mathrm{sgn}(f_\rho)$.

We are interested in the classifier $\mathrm{sgn}(f_{T+1})$ produced by the real valued function f_{T+1} using fully online learning algorithm (3). So our error analysis aims at the *excess misclassification error* $\mathcal{R}(\mathrm{sgn}(f_{T+1})) - \mathcal{R}(f_c)$.

We demonstrate our error analysis by a result, proved in Section 5, for the *hinge loss* $\phi(x) = (1-x)_+ = \max\{1-x, 0\}$. For this loss, the online algorithm (3) can be expressed as $f_1 = 0$ and

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

Theorem 11 *Let $V(y,f) = (1-yf)_+$ and $K \in C^{2s}(X \times X)$ for some $0 < s \le \frac{1}{2}$. Assume $f_\rho \in C^s(X)$ and the triple (ρ_X, f_c, K) satisfies*

$$\inf_{f \in \mathcal{H}_K} \{\|f - f_c\|_{L^1_{\rho_X}} + \frac{\lambda}{2}\|f\|_K^2\} \le \mathcal{D}_0 \lambda^\beta \qquad \forall \lambda > 0 \tag{9}$$

for some $0 < \beta \le 1$ and $\mathcal{D}_0 > 0$. Take

$$\lambda_t = \lambda_1 t^{-\frac{1}{4}}, \eta_t = \eta_1 t^{-(\frac{1}{2}+\frac{\beta}{12})}$$

where $\lambda_1 > 0$ and $0 < \eta_1 \le \frac{1}{2\kappa^2+\lambda_1}$. If $\{\rho_X^{(t)}\}_{t=1,2,\cdots}$ satisfies (1) with $b > \frac{1}{2}$, then

$$\mathbb{E}_{z_1,\ldots,z_T}(\mathcal{R}(sgn(f_{T+1})) - \mathcal{R}(f_c)) \le C_{\beta,s,b} T^{-\min\{\frac{\beta}{4}, \frac{1}{8}+\frac{\beta}{24}, \frac{b}{2}-\frac{1}{4}\}}, \tag{10}$$

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

In the i.i.d. case with $b = \infty$, the learning rate for fully online algorithm (3) we state in Theorem 11 is of form $O(T^{-\min\{\frac{\beta}{4}, \frac{1}{8}+\frac{\beta}{24}\}})$ which is better than that in Ye and Zhou (2007) of order $O(T^{-\min\{\frac{\beta}{4}, \frac{1}{8}-\frac{\varepsilon}{2}\}})$ with an arbitrarily small $\varepsilon > 0$. This improvement is realized by technical novelty in our error analysis, as pointed out in Remark 27. So even in the i.i.d. case, our learning rate for fully online classification with the hinge loss under approximation error assumption (9) is the best in the literature.

Let us discuss the role of the power index b for the convergence of $\{\rho_X^{(t)}\}$ to ρ_X played in the learning rate in Theorem 11. Consider the case $\beta \le \frac{3}{5}$. When $b \ge \frac{1+\beta}{2}$ meaning fast convergence of $\{\rho_X^{(t)}\}$ to ρ_X, learning rate (10) takes the form $O(T^{-\frac{\beta}{4}})$ which depends only on the approximation ability of $\mathcal{H}_K$ with respect to the function f_c and ρ_X. When $\frac{1}{2} < b < \frac{1+\beta}{2}$ representing some slow convergence of $\{\rho_X^{(t)}\}$ to ρ_X, the learning rate takes the form $O(T^{-\frac{2b-1}{4}})$ which depends only on b.

When $\beta > \frac{3}{5}$, learning rate (10) is of order $T^{-(\frac{1}{8}+\frac{\beta}{24})}$ for $b > \frac{3}{4} + \frac{\beta}{12}$ (fast convergence of $\{\rho_X^{(t)}\}$) and of order $T^{-(\frac{b}{2}-\frac{1}{4})}$ for $\frac{1}{2} < b \le \frac{3}{4} + \frac{\beta}{12}$ (slow convergence of $\{\rho_X^{(t)}\}$). It would be interesting to know how online learning algorithms adapt when the time dependent distribution drifts sufficiently slowly (corresponding to very small b).

2.4 Approximation Error Involving a General Loss

Condition (9) concerns the approximation of the function f_c in the space $L^1_{\rho_X}$ by functions from the RKHS $\mathcal{H}_K$. In particular, when $\mathcal{H}_K$ is a dense subset of $C(X)$ (i.e., K is a universal kernel), the quantity on the left-hand side of (9) tends to 0 as $\lambda \to 0$. So (9) is a reasonable assumption, which can be characterized by an interpolation space condition (Smale and Zhou, 2003; Chen et al., 2004).

Assumptions like (9) are necessary to determine the regularization parameter for achieving learning rate (10). This can be seen from the literature (Wu et al., 2007; Zhang, 2004; Caponnetto et al., 2007) of off-line algorithm (6): learning rates are obtained by suitable choices of the regularization parameter $\lambda \equiv \lambda_1 = \lambda_1(T)$, according to the behavior of the approximation error estimated from a priori conditions on the distribution ρ and the space $\mathcal{H}_K$.

For a general loss function V, conditions like (9) can be stated as the approximation error or regularization error.

Definition 12 *The approximation error $\mathcal{D}(\lambda)$ associated with the triple (K,V,ρ) is*

$$\mathcal{D}(\lambda) = \inf_{f\in\mathcal{H}_K} \left\{ \mathcal{E}(f) - \mathcal{E}(f_\rho^V) + \frac{\lambda}{2}\|f\|_K^2 \right\}, \tag{11}$$

where we define the generalization error for $f : X \to \mathbb{R}$ as

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

and f_ρ^V is a minimizer of $\mathcal{E}(f)$.

The approximation error measures the approximation ability of the space $\mathcal{H}_K$ with respect to the learning process involving V and ρ. The denseness of $\mathcal{H}_K$ in $C(X)$ implies $\lim_{\lambda\to 0}\mathcal{D}(\lambda) = 0$. A natural assumption would be

$$\mathcal{D}(\lambda) \le \mathcal{D}_0\lambda^\beta \qquad \text{for some } 0 \le \beta \le 1 \text{ and } \mathcal{D}_0 > 0. \tag{12}$$

Throughout the paper we assume for the general loss function V that $\|V\| := \sup_{y\in Y} V(y,0) + \sup\{|V(y,f) - V(y,0)|/|f| : y \in Y, |f| \le 1\} < \infty$.

Remark 13 *Since $\mathcal{D}(\lambda) \le \mathcal{E}(0) + 0 \le \|V\|$ for any $\lambda > 0$, we see that (12) always holds with $\beta = 0$ and $\mathcal{D}_0 = \|V\|$.*

For the least square loss $V(y,f) = (y-f)^2$, the minimizer f_ρ^V of $\mathcal{E}(f)$ is exactly the regression function defined by (5) and approximation error (11) takes the from $\mathcal{D}(\lambda) = \inf_{f\in\mathcal{H}_K}\{\|f - f_\rho\|^2_{L^2_{\rho_X}} + \frac{\lambda}{2}\|f\|_K^2\}$ which measures the approximation of f_ρ in $L^2_{\rho_X}$ by functions from the RKHS $\mathcal{H}_K$.

For a general loss function V, the minimizer f_ρ^V of $\mathcal{E}(f)$ is in general different from f_ρ. Moreover, the approximation error $\mathcal{D}(\lambda)$ involves the approximation of f_ρ^V by $\mathcal{H}_K$ in some function spaces which need not be $L^2_{\rho_X}$.

Example 5 *For the hinge loss $V(y,f) = (1-yf)_+$, the minimizer f_ρ^V is the Bayes rule $f_\rho^V = f_c$ (Devroye et al., 1997). Moreover the uniform Lipschitz continuity of $V(\cdot,f)$ implies (Chen et al., 2004)*

$$\mathcal{E}(f) - \mathcal{E}(f_\rho^V) = \int_Z \phi(yf(x)) - \phi(yf_c(x))d\rho \le \int_X |f(x) - f_c(x)|d\rho_X = \|f - f_c\|_{L^1_{\rho_X}}.$$

So approximation error (11) can be estimated by approximation in the space $L^1_{\rho_X}$ and condition (9) implies (12).

Consider the insensitive loss $V = V_{in}$. We can easily see that for each $x \in X$, $f_\rho^{V_{in}}(x)$ equals the median of the probability distribution ρ_x on Y. That is, $f_\rho^{V_{in}}(x)$ is uniquely determined by

$$\rho_x(\{y \in Y : y \le f_\rho^{V_{in}}(x)\}) \ge \frac{1}{2} \quad \text{and} \quad \rho_x(\{y \in Y : y \ge f_\rho^{V_{in}}(x)\}) \ge \frac{1}{2}.$$

When ρ_x is symmetric about its mean $f_\rho(x)$, we have $f_\rho^{V_{in}}(x) = f_\rho(x)$.

Example 6 *Let $V = V_{in}$, $f_\rho \in C(X)$ and $p \ge 1$. If for each $x \in X$, the conditional distribution ρ_x is given by*

$$d\rho_x(y) = \begin{cases} \frac{p}{2}|y - f_\rho(x)|^{p-1}dy, & \text{if } |y - f_\rho(x)| \le 1, \\ 0, & \text{if } |y - f_\rho(x)| > 1, \end{cases}$$

then we have

$$\begin{aligned} \mathcal{E}(f) - \mathcal{E}(f_\rho^{V_{in}}) &= \frac{1}{p+1}\|f - f_\rho\|^{p+1}_{L^{p+1}_{\rho_X}} - \frac{1}{p+1}\int_{\{x \in X : |f(x) - f_\rho(x)| > 1\}} \\ &\quad |f(x) - f_\rho(x)|^{p+1} + p - (1+p)|f(x) - f_\rho(x)|d\rho_X. \end{aligned}$$

It follows that

$$\mathcal{D}(\lambda) \le \inf_{f \in \mathcal{H}_K} \left\{ \|f - f_\rho\|^{p+1}_{L^{p+1}_{\rho_X}} + \frac{\lambda}{2}\|f\|^2_K \right\}.$$

The conclusion of Example 6 will be proved in the appendix. The following general result follows from the same argument as in Wu et al. (2006).

Proposition 14 *If the loss function V satisfies $|V(y, f_1) - V(y, f_2)| \le |f_1 - f_2|^c$ for some $0 < c \le 1$ and any $y \in Y$, $f_1, f_2 \in \mathbb{R}$, then*

$$\mathcal{D}(\lambda) \le \inf_{f \in \mathcal{H}_K} \{\|f - f_\rho^V\|^c_{L^1_{\rho_X}} + \frac{\lambda}{2}\|f\|^2_K\} \le \inf_{f \in \mathcal{H}_K} \{\|f - f_\rho^V\|^c_{L^2_{\rho_X}} + \frac{\lambda}{2}\|f\|^2_K\}.$$

If the univariate function $V(y, \cdot)$ is C^1 and satisfies $|\partial V(y, f_1) - \partial V(y, f_2)| \le |f_1 - f_2|^c$ for some $0 < c \le 1$ and any $y \in Y$, $f_1, f_2 \in \mathbb{R}$, then

$$\mathcal{D}(\lambda) \le \inf_{f \in \mathcal{H}_K} \{\|f - f_\rho^V\|^{1+c}_{L^{1+c}_{\rho_X}} + \frac{\lambda}{2}\|f\|^2_K\} \le \inf_{f \in \mathcal{H}_K} \{\|f - f_\rho^V\|^{1+c}_{L^2_{\rho_X}} + \frac{\lambda}{2}\|f\|^2_K\}.$$

3. Key Analysis for the Fully Online Non-identical Setting

Learning rates for fully online learning algorithm (3) such as those stated in the last section for regression and classification are obtained through analysis for approximation error and sample error. The approximation error has been well understood (Smale and Zhou, 2003; Chen et al., 2004). The sample error for (3) will be estimated in the following two sections. It is expressed as $\|f_{T+1} - f^V_{\lambda_T}\|_K$ where $f^V_{\lambda_T}$ is a regularizing function.

Definition 15 *For $\lambda > 0$ the regularizing function $f_\lambda^V \in \mathcal{H}_K$ is defined by*

$$f_\lambda^V = \arg\inf_{f\in\mathcal{H}_K}\left\{\mathcal{E}(f) + \frac{\lambda}{2}\|f\|_K^2\right\}. \tag{13}$$

Observe that f_λ^V is a sample-free limit of $f_{\mathbf{z},\lambda}$ defined by (6). It is natural to expect that the function f_{T+1} produced by online algorithm (3) approximates the regularizing function $f_{\lambda_T}^V$ well. This is actually the case. Our main result on sample error analysis estimates the difference $f_{T+1} - f_{\lambda_T}^V$ in the $\mathcal{H}_K$-norm for quite general situations. The estimation follows from an iteration procedure, developed for online classification algorithms in Ying and Zhou (2006), Ying (2007) and Ye and Zhou (2007), together with some novelty provided in the proof of Theorem 26 in the next section. It is based on a one-step iteration bounding $\|f_{t+1} - f_{\lambda_t}^V\|_K$ in terms of $\|f_t - f_{\lambda_{t-1}}^V\|_K$. The goal of this section is to present our key analysis for one-step iteration in tackling two barriers arising from the fully online non-identical setting.

3.1 Bounding Error Term Caused by Non-identical Sequence of Distributions

The first part of our key analysis for one-step iteration is to tackle an extra error term Δ_t caused by the non-identical sequence of distributions when bounding $\|f_{t+1} - f_{\lambda_t}^V\|_K$ by means of $\|f_t - f_{\lambda_t}^V\|_K$ with the fixed regularization parameter λ_t.

Lemma 16 *Define $\{f_t\}$ by (3). Then we have*

$$\mathbb{E}_{z_t}(\|f_{t+1} - f_{\lambda_t}^V\|_K^2) \le (1-\eta_t\lambda_t)\|f_t - f_{\lambda_t}^V\|_K^2 + 2\eta_t\Delta_t + \eta_t^2\mathbb{E}_{z_t}\|\partial V(y_t, f_t(x_t))K_{x_t} + \lambda_t f_t\|_K^2, \tag{14}$$

where Δ_t is defined by

$$\Delta_t = \int_Z \left\{V(y, f_{\lambda_t}^V(x)) - V(y, f_t(x))\right\} d\left[\rho^{(t)} - \rho\right]. \tag{15}$$

Lemma 16 follows from the same procedure as in the proof of Lemma 3 in Ying and Zhou (2006) and Ye and Zhou (2007). A crucial estimate is

$$\langle \partial V(y_t, f_t(x_t))K_{x_t} + \lambda_t f_t, f_{\lambda_t}^V - f_t\rangle_K \le [V(y_t, f_{\lambda_t}^V(x_t)) + \frac{\lambda_t}{2}\|f_{\lambda_t}^V\|_K^2] - [V(y_t, f_t(x_t)) + \frac{\lambda_t}{2}\|f_t\|_K^2].$$

When we take the expectation with respect to $z_t = (x_t, y_t)$, we get $\int_Z V(y, f_{\lambda_t}^V(x))d\rho^{(t)}$ (and $\int_Z V(y, f_t(x))d\rho^{(t)}$) on the right-hand side, not $\mathcal{E}(f_{\lambda_t}^V) = \int_Z V(y, f_{\lambda_t}^V(x))d\rho$. So compared with results in the i.i.d. case (Smale and Yao, 2006; Ying and Zhou, 2006; Ying, 2007; Tarrès and Yao, 2005; Ye and Zhou, 2007), an extra term Δ_t involving the difference measure $\rho^{(t)} - \rho$ appears. This is the first barrier we need to tackle here.

When V is the least square loss, it can be easily handled by assuming $f_\rho \in C^s(X)$. In fact, from $V(y, f) = (y-f)^2$, we see that

$$\begin{aligned}\Delta_t &= \int_X \left\{\int_Y [f_{\lambda_t}^V(x)) - f_t(x)][f_{\lambda_t}^V(x) + f_t(x) - 2y]d\rho_x(y)\right\} d\left[\rho_X^{(t)} - \rho_X\right] \\ &= \int_X [f_{\lambda_t}^V(x)) - f_t(x)][f_{\lambda_t}^V(x) + f_t(x) - 2f_\rho(x)]d\left[\rho_X^{(t)} - \rho_X\right].\end{aligned}$$

This together with the relation $\|hg\|_{C^s(X)} \le \|h\|_{C^s(X)}\|g\|_{C^s(X)}$ yields the following bound.

Proposition 17 *Let $V(y,f)=(y-f)^2$ and $f_\rho \in C^s(X)$. Then we have*

$$\Delta_t \le \left\{ \| f_{\lambda_t}^V\|_{C^s(X)} + \| f_t\|_{C^s(X)} + 2\| f_\rho\|_{C^s(X)} \right\} \|\rho_X^{(t)} - \rho_X\|_{(C^s(X))^*} \| f_{\lambda_t}^V - f_t\|_{C^s(X)}.$$

For a general loss function and output space, Δ_t can be bounded under assumption (4). It is a special case of the following general result when $h = f_{\lambda_t}^V$ and $g = f_t$.

Lemma 18 *Let $h,g \in C^s(X)$. If (4) holds, then we have*

$$\begin{aligned} &\left| \int_Z V(y,h(x)) - V(y,g(x)) d\left[\rho^{(t)} - \rho\right] \right| \\ &\quad \le \left\{ B_{h,g} \left(\|h\|_{C^s(X)} + \|g\|_{C^s(X)} \right) + 2C_\rho \widetilde{B}_{h,g} \right\} \|\rho_X^{(t)} - \rho_X\|_{(C^s(X))^*}, \end{aligned}$$

where $B_{h,g}$ and $\widetilde{B}_{h,g}$ are constants given by

$$B_{h,g} = \sup\left\{ |\partial V(y,f)| : y \in Y, |f| \le \max\{\|h\|_{C(X)}, \|g\|_{C(X)}\} \right\}$$

and

$$\widetilde{B}_{h,g} = \sup\left\{ \|V(\cdot,f)\|_{C^s(Y)} : |f| \le \max\{\|h\|_{C(X)}, \|g\|_{C(X)}\} \right\}.$$

Proof By decomposing the probability distributions on Z into marginal and conditional distributions, we see

$$\int_Z V(y,h(x)) - V(y,g(x)) d\left[\rho^{(t)} - \rho\right] = \int_X \int_Y V(y,h(x)) - V(y,g(x)) d\rho_x(y) d\left[\rho_X^{(t)} - \rho_X\right].$$

By the definition of the norm in $(C^s(X))^*$, we obtain

$$\left| \int_Z V(y,h(x)) - V(y,g(x)) d\left[\rho^{(t)} - \rho\right] \right| \le \|\rho_X^{(t)} - \rho_X\|_{(C^s(X))^*} \|J\|_{C^s(X)},$$

where J is a function on X defined by

$$J(x) = \int_Y V(y,h(x)) - V(y,g(x)) d\rho_x(y), \qquad x \in X.$$

The notion of $B_{h,g}$ tells us that $|V(y,h(x)) - V(y,g(x))| \le B_{h,g}|h(x) - g(x)|$ for each $y \in Y$. Hence $\|J\|_{C(X)} \le B_{h,g}\|h-g\|_{C(X)}$.

To bound $|J|_{C^s(X)}$, let $x,u \in X$. We can decompose $J(x) - J(u)$ as

$$\begin{aligned} J(x) - J(u) &= \int_Y \{[V(y,h(x)) - V(y,g(x))] - [V(y,h(u)) - V(y,g(u))]\} d\rho_x(y) \\ &\quad + \int_Y [V(y,h(u)) - V(y,g(u))] d[\rho_x - \rho_u](y). \end{aligned}$$

By the notion $B_{h,g}$, part of the first term above can be bounded as

$$|V(y,h(x)) - V(y,h(u))| \le B_{h,g}|h(x) - h(u)| \le B_{h,g}|h|_{C^s(X)} (d(x,u))^s.$$

The same bound holds for the other part of the first term. So we get

$$\left|\int_Y \{[V(y,h(x))-V(y,g(x))]-[V(y,h(u))-V(y,g(u))]\}\,d\rho_x(y)\right| \le B_{h,g}\left\{|h|_{C^s(X)}+|g|_{C^s(X)}\right\}(d(x,u))^s.$$

For the other term of the expression for $J(x)-J(u)$, we apply condition (4) to

$$\left|\int_Y [V(y,h(u))-V(y,g(u))]d[\rho_x-\rho_u](y)\right| \le \|\rho_x-\rho_u\|_{(C^s(Y))^*}\|V(y,h(u))-V(y,g(u))\|_{C^s(Y)}$$

and find that

$$\left|\int_Y [V(y,h(u))-V(y,g(u))]d[\rho_x-\rho_u](y)\right| \le C_\rho (d(x,u))^s 2\widetilde{B}_{h,g}.$$

Combining the above two bounds, we see that

$$|J|_{C^s(X)} = \sup_{x\neq u\in X}\frac{|J(x)-J(u)|}{(d(x,u))^s} \le B_{h,g}\left\{|h|_{C^s(X)}+|g|_{C^s(X)}\right\}+2C_\rho\widetilde{B}_{h,g}.$$

Then the desired bound follows and the lemma is proved. ■

3.2 Bounding Error Term Caused by Varying Regularization Parameters

The second part of our key analysis is to estimate the error term called drift error $\|f^V_{\lambda_t}-f^V_{\lambda_{t-1}}\|_K$ caused by the change of the regularization parameter from λ_{t-1} to λ_t in our fully online algorithm. This is the second barrier we need to tackle here.

Definition 19 *The drift error is defined as*

$$d_t = \|f^V_{\lambda_t}-f^V_{\lambda_{t-1}}\|_K.$$

The drift error can be estimated by the approximation error, which has been studied for regression (Smale and Zhou, 2009) and for classification (Ye and Zhou, 2007).

Theorem 20 *Let V be a convex loss function, f^V_λ by (13) and $\mu>\lambda>0$. We have*

$$\|f^V_\lambda-f^V_\mu\|_K \le \frac{\mu}{2}\left(\frac{1}{\lambda}-\frac{1}{\mu}\right)\left(\|f^V_\lambda\|_K+\|f^V_\mu\|_K\right) \le \frac{\mu}{2}\left(\frac{1}{\lambda}-\frac{1}{\mu}\right)\left(\sqrt{\frac{2\mathcal{D}(\lambda)}{\lambda}}+\sqrt{\frac{2\mathcal{D}(\mu)}{\mu}}\right).$$

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

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

Proof Taking derivative of the functional $\mathcal{E}(f)+\frac{\lambda}{2}\|f\|_K^2$ at the minimizer f^ϕ_λ defined in (13), we see that

$$\int_Z \partial V(y, f^V_\lambda(x))K_x d\rho + \lambda f^V_\lambda = 0.$$

It follows that

$$f_\lambda^V - f_\mu^V = \frac{1}{\mu}\int_Z \partial V(y, f_\mu^V(x))K_x d\rho - \frac{1}{\lambda}\int_Z \partial V(y, f_\lambda^V(x))K_x d\rho.$$

Combining with the reproducing property (8), we know $\|f_\lambda^V - f_\mu^V\|_K^2 = \langle f_\lambda^V - f_\mu^V, f_\lambda^V - f_\mu^V\rangle_K$ can be expressed as

$$\|f_\lambda^V - f_\mu^V\|_K^2 = \frac{1}{\mu}\int_Z \partial V(y, f_\mu^V(x))\left(f_\lambda^V - f_\mu^V\right)(x)d\rho - \frac{1}{\lambda}\int_Z \partial V(y, f_\lambda^V(x))\left(f_\lambda^V - f_\mu^V\right)(x)d\rho.$$

The convexity of the function $V(y,\cdot)$ on $\mathbb{R}$ tells us that

$$\partial V(y, f_\mu^V(x))\left(f_\lambda^V - f_\mu^V\right)(x) \le V(y, f_\lambda^V(x)) - V(y, f_\mu^V(x))$$

and

$$\partial V(y, f_\lambda^V(x))\left(f_\mu^V - f_\lambda^V\right)(x) \le V(y, f_\mu^V(x)) - V(y, f_\lambda^V(x)).$$

Hence

$$\|f_\lambda^V - f_\mu^V\|_K^2 \le (\frac{1}{\lambda} - \frac{1}{\mu})(\mathcal{E}(f_\mu^V) - \mathcal{E}(f_\lambda^V)).$$

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

$$\mathcal{E}(f_\mu^V) - \mathcal{E}(f_\lambda^V) \le \frac{\mu}{2}(\|f_\lambda^V\|_K^2 - \|f_\mu^V\|_K^2) \le \frac{\mu}{2}\|f_\lambda^V - f_\mu^V\|_K(\|f_\lambda^V\|_K + \|f_\mu^V\|_K).$$

Then the desired inequality follows. ■

4. Bounds for Sample Error in $\mathcal{H}_K$-norm

We are in a position to present our main result on the sample error measured with the $\mathcal{H}_K$-norm of the difference $f_{T+1} - f_{\lambda_T}^V$. This will be done by applying iteratively the key analysis in Lemma 16, Lemma 18 and Theorem 20.

When applying Lemma 18, we need to bound $\|g\|_{C^s(X)}$ in terms of $\|g\|_K$.

Definition 21 *We say that the Mercer kernel K satisfies the kernel condition of order s if $K \in C^s(X\times X)$ and for some $\kappa_{2s} > 0$,*

$$|K(x,x) - 2K(x,u) + K(u,u)| \le \kappa_{2s}^2 (d(x,u))^{2s}, \qquad \forall x,u \in X. \tag{16}$$

When $0 < s \le \frac{1}{2}$ and $K \in C^{2s}(X\times X)$, (16) holds true. The following result follows directly from Zhou (2003), Zhou (2008) and Smale and Zhou (2009).

Lemma 22 *If K satisfies the kernel condition of order s with (16) valid, then we have*

$$\|g\|_{C^s(X)} \le (\kappa + \kappa_{2s})\|g\|_K, \qquad \forall g \in \mathcal{H}_K.$$

When using Lemma 16 and Lemma 18, we need incremental behaviors of the loss function V to bound $\|f_t\|_K$.

Definition 23 *Denote*

$$N(\lambda) = \sup\left\{|\partial V(y,f)| : y \in Y, |f| \le \max\left\{\kappa^2\|V\|/\lambda, \kappa\sqrt{2\|V\|/\lambda}\right\}\right\}$$

and

$$\widetilde{N}(\lambda) = \sup\left\{\|V(\cdot,f)\|_{C^s(Y)} : y \in Y, |f| \le \max\left\{\kappa^2\|V\|/\lambda, \kappa\sqrt{2\|V\|/\lambda}\right\}\right\}.$$

We say that V has incremental exponent $p \ge 0$ if for some $N_1 > 0$ and $\lambda_1 > 0$ we have

$$N(\lambda) \le N_1(\frac{1}{\lambda})^p \quad \text{and} \quad \widetilde{N}(\lambda) \le N_1(\frac{1}{\lambda})^{p+1} \qquad \forall 0 < \lambda \le \lambda_1. \tag{17}$$

We say that ∂V is locally Lipschitz at the origin if

$$M_0 := \sup\left\{\frac{|\partial V(y,f) - \partial V(y,0)|}{|f|} : y \in Y, |f| \le 1\right\} < \infty. \tag{18}$$

The following result can be proved by exactly the same procedure as those in Ying and Zhou (2006), Ying (2007) and Ye and Zhou (2007).

Lemma 24 *Assume that ∂V is locally Lipschitz at the origin. Define $\{f_t\}$ by (3). If*

$$\eta_t\left(\kappa^2(M_0 + 2N(\lambda_t)) + \lambda_t\right) \le 1 \tag{19}$$

for $t = 1, \ldots, T$, then we have

$$\|f_t\|_K \le \frac{\kappa\|V\|}{\lambda_t}, \qquad t = 1, \ldots, T+1. \tag{20}$$

For the insensitive loss, (18) is not satisfied, but $\partial V(y,f)$ is uniformly bounded by 1. For such loss functions we can apply the following bound.

Lemma 25 *Assume $\|\partial V(y,f)\|_\infty := \sup_{y\in Y, f\in\mathbb{R}} |\partial V(y,f)| < \infty$. Define $\{f_t\}$ by (3). If for some $\lambda_1, \eta_1 > 0$ and $\gamma, \alpha > 0$ with $\lambda + \alpha < 1$, we take $\lambda_t = \lambda_1 t^{-\gamma}, \eta_t = \eta_1 t^{-\alpha}$ with $t = 1, \ldots, T$, then we have*

$$\|f_t\|_K \le \frac{C_{V,\gamma,\alpha}}{\lambda_t}, \qquad t = 1, \ldots, T+1, \tag{21}$$

where $C_{V,\gamma,\alpha}$ is the constant given by

$$C_{V,\gamma,\alpha} = \kappa\|\partial V(y,f)\|_\infty\left\{\lambda_1\eta_1 + \lambda_1\left(2^{\gamma+2\alpha}/(\lambda_1\eta_1) + \left((1+\alpha)/[e\lambda_1\eta_1(1-2^{\gamma+\alpha-1})]\right)^{\frac{1+\alpha}{1-\gamma-\alpha}}\right)\right\}.$$

Proof By taking norms in (3) we see that

$$\|f_{t+1}\|_K \le (1 - \eta_t\lambda_t)\|f_t\|_K + \eta_t\kappa\|\partial V(y,f)\|_\infty.$$

By iterating and the choice $f_1 = 0$ we find

$$\|f_{t+1}\|_K \le \sum_{i=1}^{t} \Pi_{j=i+1}^{t}(1 - \eta_j\lambda_j)\eta_i\kappa\|\partial V(y,f)\|_\infty.$$

But $1-\eta_j\lambda_j \le \exp\{-\eta_j\lambda_j\}$. It follows that

$$\|f_{t+1}\|_K \le \sum_{i=1}^{t} \Pi_{j=i+1}^{t} \exp\left\{-\sum_{j=i+1}^{t}\eta_j\lambda_j\right\}\eta_i\kappa\|\partial V(y,f)\|_\infty.$$

Now we need the following elementary inequality with $c>0, q_2 \ge 0$ and $0<q_1<1$:

$$\sum_{i=1}^{t-1} i^{-q_2}\exp\left\{-c\sum_{j=i+1}^{t} j^{-q_1}\right\} \le \left(\frac{2^{q_1+q_2}}{c}+\left(\frac{1+q_2}{ec(1-2^{q_1-1})}\right)^{\frac{1+q_2}{1-q_1}}\right)t^{q_1-q_2}. \tag{22}$$

This elementary inequality can be found in, for example, Smale and Zhou (2009). Taking $q_2 = \alpha, q_1 = \gamma+\alpha$ and $c=\lambda_1\eta_1$ we know that $\|f_{t+1}\|_K$ is bounded by

$$\kappa\|\partial V(y,f)\|_\infty\left\{\eta_1 t^{-\alpha}+\left(2^{\gamma+2\alpha}/(\lambda_1\eta_1)+\left((1+\alpha)/[e\lambda_1\eta_1(1-2^{\gamma+\alpha-1})]\right)^{\frac{1+\alpha}{1-\gamma-\alpha}}\right)t^{\gamma}\right\}.$$

Then our desired bound holds true. ■

Now we can present our bound for the sample error $\|f_{T+1}-f_{\lambda_T}^V\|_K$.

Theorem 26 *Suppose the following assumptions hold:*

1. *the kernel K satisfies the kernel condition of order s ($0<s\le 1$) with (16) valid.*
2. *V has incremental exponent $p\ge 0$ with (17) valid and ∂V satisfies (18).*
3. *$\{\rho_X^{(t)}\}_{t=1,2,\cdots}$ converges polynomially to ρ_X in $(C^s(X))^*$ with (1) valid.*
4. *the distributions $\{\rho_x : x\in X\}$ is Lipschitz s in $(C^s(Y))^*$ with (4) valid.*
5. *the triple (K,V,ρ) has the approximation ability of power $0<\beta\le 1$ stated by (12).*

Take

$$\lambda_t = \lambda_1 t^{-\gamma}, \eta_t = \eta_1 t^{-\alpha} \tag{23}$$

with some $\lambda_1,\eta_1>0$ and γ,α satisfying

$$0<\gamma<\frac{2}{5+4p-\beta},\ (2p+1)\gamma<\alpha<1-\frac{\gamma(3-\beta)}{2},\ \eta_1\le\frac{1}{\kappa^2M_0+2\kappa^2N_1\lambda_1^{-p}+\lambda_1}. \tag{24}$$

Then we have

$$\mathbb{E}_{z_1,z_2,\cdots,z_T}(\|f_{T+1}-f_{\lambda_T}^V\|_K^2) \le C_{K,V,\rho,b,\beta,s}T^{-\theta} \tag{25}$$

where the power index θ is given by

$$\theta := \min\left\{2-\gamma(3-\beta)-2\alpha, \alpha-\gamma(2p+1), b-\gamma(2+p)\right\}, \tag{26}$$

and $C_{K,V,\rho,b,\beta,s}$ is a constant independent of T given explicitly in the proof.

Proof We divide the proof into four steps.

First we bound Δ_t. Since λ_t and η_t take the form (23), we see from the lower bound for α in (24) that $p\gamma \le (2p+1)\gamma \le \alpha$. Hence for $t \in \mathbb{N}$,

$$\eta_t(\kappa^2(M_0+2N(\lambda_t))+\lambda_t) \le \eta_1 t^{-\alpha}(\kappa^2 M_0 + 2\kappa^2 N_1 \lambda_1^{-p} t^{p\gamma} + \lambda_1 t^{-\gamma}) \le \eta_1(\kappa^2 M_0 + 2\kappa^2 N_1 \lambda_1^{-p} + \lambda_1).$$

So the last restriction of (24) implies (19), and by Lemma 24, we know that (20) holds true.

Taking $f = 0$ in (13) yields

$$\|f_{\lambda_t}^V\|_K^2 \le \frac{2\|V\|}{\lambda_t}, \qquad t = 1, \ldots, T.$$

Putting these bounds into the definition of constant $B_{h,g}, \widetilde{B}_{h,g}$, we see by the notion $N(\lambda)$ that

$$B_{f_{\lambda_t}^V, f_t} \le N(\lambda_t), \qquad \widetilde{B}_{f_{\lambda_t}^V, f_t} \le N(\lambda_t).$$

So by Lemma 18 and Lemma 22,

$$\Delta_t \le B_t^* := \left\{(\kappa+\kappa_{2s})(\sqrt{2\|V\|/\lambda_t}+\kappa\|V\|/\lambda_t)N(\lambda_t) + 2C_\rho \widetilde{N}(\lambda_t)\right\} \|\rho_X^{(t)} - \rho_X\|_{(C^s(X))^*}.$$

Putting this bound and (20) into (14) yields

$$\mathbb{E}_{z_1,z_2,\cdots,z_t}(\|f_{t+1} - f_{\lambda_t}^V\|_K^2) \le (1-\eta_t\lambda_t)\mathbb{E}_{z_1,z_2,\cdots,z_{t-1}}\left(\|f_t - f_{\lambda_t}^V\|_K^2\right) + \kappa^2\eta_t^2(N(\lambda_t)+\|V\|)^2 + 2\eta_t B_t^*.$$

Next we derive explicit bounds for the one-step iteration.

Recall that $d_t = \|f_{\lambda_t}^V - f_{\lambda_{t-1}}^V\|_K$. It gives $\|f_t - f_{\lambda_t}^V\|_K^2 \le \|f_t - f_{\lambda_{t-1}}^V\|_K^2 + 2\|f_t - f_{\lambda_{t-1}}^V\|_K d_t + d_t^2$.

Take $\tau = \frac{\gamma+\alpha}{1-\gamma(1-\beta)/2}$. By the upper bound for α in (24), we find $0 < \tau < 1$.

Take $A_1 = \frac{\eta_1 \lambda_1^{1+\tau(1-\beta)/2}}{2^{1+2\tau}\mathcal{D}_0^{\tau/2}} > 0$. Applying the elementary inequality

$$2ab = 2[\sqrt{A_1}ab^{\tau/2}][b^{1-\tau/2}/\sqrt{A_1}] \le A_1 a^2 b^\tau + b^{2-\tau}/A_1 \tag{27}$$

to $a = \|f_t - f_{\lambda_{t-1}}^V\|_K$ and $b = d_t$, we know that

$$\mathbb{E}_{z_1,z_2,\cdots,z_{t-1}}\left(\|f_t - f_{\lambda_t}^V\|_K^2\right) \le (1+A_1 d_t^\tau)\mathbb{E}_{z_1,z_2,\cdots,z_{t-1}}\left(\|f_t - f_{\lambda_{t-1}}^V\|_K^2\right) + d_t^{2-\tau}/A_1 + d_t^2.$$

Using this bound and noticing the inequality $(1-\eta_t\lambda_t)(1+A_1 d_t^\tau) \le 1 + A_1 d_t^\tau - \eta_t\lambda_t$, we obtain

$$\begin{aligned}\mathbb{E}_{z_1,z_2,\cdots,z_t}(\|f_{t+1} - f_{\lambda_t}^V\|_K^2) \le{}& (1+A_1 d_t^\tau - \eta_t\lambda_t)\mathbb{E}_{z_1,z_2,\cdots,z_{t-1}}\left(\|f_t - f_{\lambda_{t-1}}^V\|_K^2\right) \\ &+ d_t^{2-\tau}/A_1 + d_t^2 + \kappa^2\eta_t^2(N(\lambda_t)+\|V\|)^2 + 2\eta_t B_t^*.\end{aligned}$$

By Theorem 20, condition (12) for the approximation error yields

$$d_t \le 2\sqrt{\mathcal{D}_0}\lambda_1^{(\beta-1)/2}(t-1)^{\gamma(1-\beta)/2-1} \le A_2 t^{\gamma(1-\beta)/2-1} \quad \text{where } A_2 = 4\sqrt{\mathcal{D}_0}\lambda_1^{(\beta-1)/2}.$$

Inserting the parameter form $\lambda_t = \lambda_1 t^{-\gamma}$ into assumption (17) and applying condition (1), we can bound B_t^* as

$$B_t^* \le A_3 t^{\gamma(1+p)-b} \quad \text{where } A_3 = \left\{(\kappa+\kappa_{2s})(\sqrt{2\|V\|/\lambda_1}+\kappa\|V\|/\lambda_1) + 2C_\rho/\lambda_1\right\} N_1 \lambda_1^{-p} C.$$

Therefore, for the one-step iteration, by denoting $f^V_{\lambda_1}$ as $f^V_{\lambda_0}$ when $t=1$, we have for each $t=1,\dots,T$,

$$\mathbb{E}_{z_1,z_2,\cdots,z_t}(\|f_{t+1}-f^V_{\lambda_t}\|^2_K)\le(1+A_1d_t^{\tau}-\eta_t\lambda_t)\mathbb{E}_{z_1,z_2,\cdots,z_{t-1}}\left(\|f_t-f^V_{\lambda_{t-1}}\|^2_K\right)+A_4t^{-\widetilde{\theta}},\qquad(28)$$

where

$$\widetilde{\theta}=\min\left\{2-\gamma(2-\beta)-\alpha,\ 2(\alpha-p\gamma),\ \alpha+b-\gamma(1+p)\right\}$$

and

$$A_4=A_2^{2-\tau}/A_1+A_2^2+\kappa^2\eta_1^2(N_1\lambda_1^{-p}+\|V\|)^2+2\eta_1A_3.$$

Then we iterate the above one-step analysis. Inserting the parameter forms for λ_t and η_t, we see from the definition of the constant A_1 that

$$1+A_1d_t^{\tau}-\eta_t\lambda_t\le1+A_1A_2^{\tau}t^{\tau(\gamma(1-\beta)/2-1)}-\eta_1\lambda_1t^{-\gamma-\alpha}=1-\frac{\eta_1\lambda_1}{2}t^{-\gamma-\alpha}.\qquad(29)$$

So the one-step analysis (28) yields

$$\mathbb{E}_{z_1,z_2,\cdots,z_t}(\|f_{t+1}-f^V_{\lambda_t}\|^2_K)\le\left(1-\frac{\eta_1\lambda_1}{2}t^{-\gamma-\alpha}\right)\mathbb{E}_{z_1,z_2,\cdots,z_{t-1}}\left(\|f_t-f^V_{\lambda_{t-1}}\|^2_K\right)+A_4t^{-\widetilde{\theta}}.$$

Applying this bound iteratively for $t=1,\dots,T$ implies

$$\begin{aligned}\mathbb{E}_{z_1,z_2,\cdots,z_T}(\|f_{T+1}-f^V_{\lambda_T}\|^2_K)&\le A_4\sum_{t=1}^{T}\Pi_{j=t+1}^{T}(1-\frac{\eta_1\lambda_1}{2}j^{-\gamma-\alpha})t^{-\widetilde{\theta}}\\&+\left\{\Pi_{t=1}^{T}(1-\frac{\eta_1\lambda_1}{2}t^{-\gamma-\alpha})\right\}\|f_1-f^V_{\lambda_1}\|^2_K.\end{aligned}$$

Finally we bound the above expressions by two elementary inequalities. The first one is (22). Applying this inequality with $c=\frac{\eta_1\lambda_1}{2},q_1=\gamma+\alpha$ and $q_2=\widetilde{\theta}$, since $1-u\le e^{-u}$ for any $u\ge0$, the first expression above can be bounded as

$$\sum_{t=1}^{T}\Pi_{j=t+1}^{T}(1-\frac{\eta_1\lambda_1}{2}j^{-\gamma-\alpha})t^{-\widetilde{\theta}}\le\sum_{t=1}^{T}\exp\left\{-\frac{\eta_1\lambda_1}{2}\sum_{j=t+1}^{T}j^{-\gamma-\alpha}\right\}t^{-\widetilde{\theta}}\le A_5T^{\gamma+\alpha-\widetilde{\theta}},$$

where A_5 is the constant given by

$$A_5=\frac{2^{\gamma+\alpha+\widetilde{\theta}+1}}{\eta_1\lambda_1}+1+\left(\frac{2+2\widetilde{\theta}}{e\eta_1\lambda_1(1-2^{\gamma+\alpha-1})}\right)^{\frac{1+\widetilde{\theta}}{1-\gamma-\alpha}}.$$

For the second expression above, we have

$$\begin{aligned}&\Pi_{t=1}^{T}(1-\frac{\eta_1\lambda_1}{2}t^{-\gamma-\alpha})\le\exp\left\{-\frac{\eta_1\lambda_1}{2}\sum_{j=1}^{T}t^{-\gamma-\alpha}\right\}\le\exp\left\{-\frac{\eta_1\lambda_1}{2}\int_1^{T+1}x^{-\gamma-\alpha}dx\right\}\\&\le\exp\left\{\frac{\lambda_1\eta_1}{2(1-\gamma-\alpha)}\right\}\exp\left\{-\frac{\lambda_1\eta_1}{2(1-\gamma-\alpha)}(T+1)^{1-\gamma-\alpha}\right\}.\end{aligned}$$

Applying another elementary inequality

$$\exp\{-cx\} \le \left(\frac{a}{ec}\right)^a x^{-a}, \qquad \forall c,a,x>0$$

with $c=\frac{\lambda_1\eta_1}{2(1-\gamma-\alpha)}$, $a=\frac{2}{1-\gamma-\alpha}$ and $x=(T+1)^{1-\gamma-\alpha}$ yields

$$\Pi_{t=1}^T(1-\frac{\eta_1\lambda_1}{2}t^{-\gamma-\alpha}) \le \exp\left\{\frac{\lambda_1\eta_1}{2(1-\gamma-\alpha)}\right\}\left(\frac{4}{e\lambda_1\eta_1}\right)^{\frac{2}{1-\gamma-\alpha}} T^{-2}.$$

The above two estimates give the desired bound (25) with $\theta=\widetilde{\theta}-\gamma-\alpha$ and the constant $C_{K,V,\rho,b,\beta,s}$ given by

$$C_{K,V,\rho,b,\beta,s} = A_4A_5 + \exp\left\{\frac{\lambda_1\eta_1}{2(1-\gamma-\alpha)}\right\}\left(\frac{4}{e\lambda_1\eta_1}\right)^{\frac{2}{1-\gamma-\alpha}} \frac{2\|V\|}{\lambda_1}.$$

This proves the theorem. ■

Remark 27 *Some ideas in the above proof are from Ying and Zhou (2006), Ye and Zhou (2007) and Smale and Zhou (2009). Two novel points are presented for the first time here. One is the bound for Δ_t, dealing with $\rho_X^{(t)}-\rho_X$, given in the first step of our proof in order to tackle the technical difficulty arising from the non-identical sampling process. The same difficulty for the least square regression was overcome in Smale and Zhou (2009) by the special linear feature and explicit expressions offered by the least square loss. The second technical novelty is to introduce a parameter A_1 into elementary inequality (27). With this parameter, we can bound $1+A_1d_t^\tau-\eta_t\lambda_t$ by $1-\frac{1}{2}\eta_t\lambda_t$, shown in (29). This improves the error bound even in the i.i.d. case presented in Ye and Zhou (2007) for the fully online algorithm.*

Let us discuss the role of parameters in Theorem 26. When γ is small enough and $b>\frac{2}{3}$, fully online algorithm (3) becomes very close to the partially online scheme with $\lambda_t\equiv\lambda_1$. By taking $\alpha=\frac{2}{3}$, the rates in (25) are of order $O(T^{-(\frac{2}{3}-\varepsilon)})$ with ε arbitrarily small, which is a nice bound for the sample error $\|f_{T+1}-f_{\lambda_T}^V\|_K$. In this case, the difference between $f_{\lambda_T}^V$ and f_ρ^V, measured by the approximation error, increases since $\lambda_T=\lambda_1T^{-\gamma}$. To estimate the total error between f_{T+1} and f_ρ^V, we should take a balance for the index γ of the regularization parameter, as shown in Theorem 11.

For the insensitive loss, (18) is not satisfied. We can apply Lemma 25 and obtain bounds for $\|f_{T+1}-f_{\lambda_T}^V\|_K$ by the same proof as that for Theorem 26.

Proposition 28 *Assume $\|\partial V(y,f)\|_\infty<\infty$ and all the conditions of Theorem 26 except (18). Take $\{\lambda_t,\eta_t\}$ by (23) with the restriction (24) without the last inequality. Then the same convergence rate (25) holds true with the power index θ given by (26).*

5. Bounds for Binary Classification and Regression with Insensitive Loss

We demonstrate how to apply Theorem 26 by deriving learning rates of fully online algorithm (3) for binary classification and regression with insensitive loss.

Theorem 29 *Let $V(y,f)=\phi(yf)$ where $\phi:\mathbb{R}\to\mathbb{R}_+$ is a convex function satisfying*

$$|\phi'_-(u)|\le N_\phi|u|^p,\quad \phi(u)\le N_\phi u^{p+1}\qquad \forall |u|\ge 1 \tag{30}$$

for some $p\ge 0$ and $N_\phi>0$. Suppose $M_\phi:=\sup\{|\phi'_-(u)-\phi'_-(0)|/|u| : |u|\le 1\}<\infty$. Assume (16) for K, (1) for $\{\rho_X^{(t)}\}$, and (12) for (K,ϕ,ρ). If $f_\rho\in C^s(X)$ and we choose $\{\lambda_t,\eta_t\}$ as (23) with $0<\gamma<\frac{2}{5+10p-\beta},\alpha=\frac{2+\gamma(2p-2+\beta)}{3}$ and $b>\gamma(2+p)$, and $\eta_1<\eta_0$, $\lambda_1\le\kappa^2(\phi(0)+\|\phi'_-\|_{L^\infty[-1,1]})/2$, then we have

$$\mathbb{E}_{z_1,\dots,z_T}\left(\mathcal{E}(f_{T+1})-\mathcal{E}(f_\rho^V)\right)\le \widetilde{C}_{\phi,\beta,\gamma}T^{-\min\left\{\frac{2-\gamma(5+10p-\beta)}{6},\frac{b-\gamma(2+3p)}{2},\gamma\beta\right\}},$$

where $\eta_0:=\frac{1}{\kappa^2M_\phi+2\kappa^2N_1\lambda_1^{-p}+\lambda_1}$ with $N_1=(M_\phi+N_\phi+\phi(0)+\|\phi'_-\|_{L^\infty[-1,1]})\kappa^{2p}(\phi(0)+\|\phi'_-\|_{L^\infty[-1,1]})^p$ and $\widetilde{C}_{\phi,\beta,\gamma}$ is a constant depending on $\eta_1,\lambda_1,\kappa,\mathcal{D}_0,\beta,\phi,\beta$ and s.

Proof By the bounds for $\|f_{\lambda_T}^V\|_K$ and $\|f_{T+1}\|_K$, we know from (30) that

$$\begin{aligned}\left|\mathcal{E}(f_{T+1})-\mathcal{E}(f_{\lambda_T}^V)\right| &= \left|\int_Z \phi(yf_{T+1}(x))-\phi(yf_{\lambda_T}^V(x))d\rho\right| \\ &\le C_{K,\phi}\lambda_T^{-p}\|f_{T+1}-f_{\lambda_T}^V\|_\infty\le \kappa C_{K,\phi}\lambda_T^{-p}\|f_{T+1}-f_{\lambda_T}^V\|_K,\end{aligned}$$

where $C_{K,\phi}$ is a constant depending on K and ϕ.

It is easy to check that the loss $V(y,f)=\phi(yf)$ satisfies (17) with incremental exponent p.

By Theorem 26 with $0<\gamma<\frac{2}{5+10p-\beta},\alpha=\frac{2+\gamma(2p-2+\beta)}{3}$ and $b>\gamma(2+p)$, we have

$$\mathbb{E}_{z_1,z_2,\cdots,z_T}\left(\left\|f_{T+1}-f_{\lambda_T}^V\right\|_K\right)\le\sqrt{C_{K,V,\rho,b,\beta,s}}T^{-\min\{[2-\gamma(5+4p-\beta)]/6,[b-\gamma(2+p)]/2\}}.$$

Also, we have $\mathcal{E}(f_{\lambda_T}^\phi)-\mathcal{E}(f_\rho^V)\le\mathcal{D}(\lambda_T)\le\mathcal{D}_0\lambda_T^\beta$. Thus we get a bound for the excess generalization error

$$\mathbb{E}_{z_1,\dots,z_T}(\mathcal{E}(f_{T+1})-\mathcal{E}(f_\rho^V))\le\widetilde{C}_{\phi,\beta,\gamma}T^{-\min\{[2-\gamma(5+10p-\beta)]/6,\gamma\beta,[b-\gamma(2+3p)]/2\}},$$

where $\widetilde{C}_{\phi,\beta,\gamma}=\kappa C_{K,\phi}\lambda_1^{-p}\sqrt{C_{K,V,\rho,b,\beta,s}}+\mathcal{D}_0\lambda_1^\beta$. This verifies the desired bound. ■

Theorem 29 yields concrete learning rates with various loss functions. When ϕ is chosen to be the hinge loss $\phi(x)=(1-x)_+$, we can prove Theorem 11.

5.1 Proof of Theorem 11

When $0<s\le\frac{1}{2}$ and $K\in C^{2s}(X\times X)$, (16) holds true.

The loss function $\phi(x)=(1-x)_+$ satisfies $\phi'_-(x)=-1$ for $x\le 1$ and 0 otherwise. It follows that (17) holds true with $p=0$ and $M_\phi=0$. By Example 5, (9) implies (12).

Thus all conditions in Theorem 29 are satisfied and by taking $p=0$ and $\gamma=\frac{1}{4}$, we have

$$\mathbb{E}_{z_1,\dots,z_T}\left(\mathcal{E}(f_{T+1})-\mathcal{E}(f_\rho^V)\right)\le\widetilde{C}_{\phi,\beta,\gamma}T^{-\min\left\{\frac{1}{8}+\frac{\beta}{24},\frac{\beta}{4},\frac{b}{2}-\frac{1}{4}\right\}}.$$

An important relation concerning the hinge loss is the one (Zhang, 2004) between the excess misclassification error and the excess generalization error given for any measurable function $f : X \to \mathbb{R}$ as

$$\mathcal{R}(\mathrm{sgn}(f)) - \mathcal{R}(f_c) \le \mathcal{E}(f) - \mathcal{E}(f_c).$$

Combining this relation with the above bound for the excess generalization error proves the conclusion of Theorem 11. ■

Turn to the general loss ϕ. We give an additional assumption that $\phi''(0)$ exists and is positive. Under this assumption it was proved in Chen et al. (2004) and Bartlett et al. (2006) that there exists a constant depending c_ϕ only on ϕ such that for any measurable function $f : X \to \mathbb{R}$,

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

Then Theorem 29 gives the following learning rate.

Corollary 30 *Let ϕ be a loss function such that $\phi''(0)$ exists and is positive. Under the assumptions of Theorem 29, if $\gamma = \frac{2}{5+10p+5\beta}$, we have*

$$\mathbb{E}_{z_1,\dots,z_T}\left(\mathcal{R}(sgn(f_{T+1})) - \mathcal{R}(f_c)\right) \le \widetilde{C}_{\phi,\beta} T^{-\min\{\frac{\beta}{5+10p+5\beta}, \frac{b}{4} - \frac{2+3p}{10+20p+10\beta}\}},$$

where $\tilde{C}_{\phi,\beta}$ is a constant independent of T.

As an example, the q-norm SVM loss $\phi(x) = ((1-x)_+)^q$ with $q > 1$ satisfies $\phi''(0) > 0$ and (17) with $p = q - 1$. So Corollary 30 yields the following rates.

Example 7 *Let $\phi(x) = ((1-x)_+)^q$ with $q > 1$. Under the assumptions of Theorem 29, if $\gamma = \frac{2}{10q-5+5\beta}$, $\alpha = \frac{8q-6+4\beta}{10q-5+5\beta}$ and $b > \frac{6q-2}{10q-5+5\beta}$, then*

$$\mathbb{E}_{z_1,\dots,z_T}\left(\mathcal{R}(sgn(f_{T+1})) - \mathcal{R}(f_c)\right) = O\left(T^{-\min\{\frac{\beta}{10q-5+5\beta}, \frac{b}{4} - \frac{3q-1}{20q-10+10\beta}\}}\right).$$

Finally we verify the learning rates for regression with insensitive loss stated in Section 2.

5.2 Proof of Theorem 10

We need the regularizing function $f_\lambda^{V_{ls}}$ defined by (13) with the least square loss $V = V_{ls}$. It can be found, for example, in Smale and Zhou (2007) that regularity condition (7) implies

$$\|f_\lambda^{V_{ls}} - f_\rho\|_K \le \left(\frac{\lambda}{2}\right)^{r-\frac{1}{2}} \|g_\rho\|_{L^2_{\rho_X}}, \qquad \text{when } \frac{1}{2} < r \le \frac{3}{2}$$

and

$$\|f_\lambda^{V_{ls}} - f_\rho\|_{L^2_{\rho_X}} \le \left(\frac{\lambda}{2}\right)^{r} \|g_\rho\|_{L^2_{\rho_X}}, \qquad \text{when } 0 < r \le 1.$$

It follows that when $\lambda \le 2(\kappa\|g_\rho\|_{L^2_{\rho_X}})^{2/(1-2r)}$, we have $\|f_\lambda^{V_{ls}} - f_\rho\|_{C(X)} \le 1$. Thus by the special form of the conditional distribution ρ_x, we see from the conclusion of Example 6 with $p = 1$ that

$f_\lambda^{V_{in}} = f_{2\lambda}^{V_{ls}}$ and bounds for $\|f_\lambda^{V_{in}} - f_\rho\|_K$ and $\|f_\lambda^{V_{in}} - f_\rho\|_{L^2_{\rho_X}}$ follow. Moreover, condition (12) for $\mathcal{D}(\lambda)$ is valid with $\beta = 1$.

Now we check other conditions of Proposition 28.

Condition (16) is valid because $K \in C^{2s}(X \times X)$ with $0 < s \le \frac{1}{2}$.

By a simple computation, incremental condition (17) is verified with exponent $p = 0$.

Note that $\|f_\rho\|_K \le \kappa^{2r-1}\|g_\rho\|_{L^2_{\rho_X}}$ and $\|f_\rho\|_{C^s(X)} \le (\kappa + \kappa_{2s})\|f_\rho\|_K$. Then for any $x, u \in X$ and $g \in C^s(Y)$, we see from the uniform distribution ρ_x and ρ_u that

$$\begin{aligned}\left|\int_Y g(y)d(\rho_x - \rho_u)\right| &= \frac{1}{2}\left|\int_{f_\rho(x)-1}^{f_\rho(x)+1} g(y)dy - \int_{f_\rho(u)-1}^{f_\rho(u)+1} g(y)dy\right| \le \|g\|_{C(Y)}|f_\rho(x) - f_\rho(u)| \\ &\le \|g\|_{C(Y)}(\kappa + \kappa_{2s})\kappa^{2r-1}\|g_\rho\|_{L^2_{\rho_X}}(d(x,u))^s.\end{aligned}$$

This verifies Lipschitz s continuous condition (4) for $\{\rho_x\}$ with constant $C_\rho = (\kappa + \kappa_{2s})\kappa^{2r-1}\|g_\rho\|_{L^2_{\rho_X}}$.

Thus all conditions of Proposition 28 are satisfied and we obtain

$$\mathbb{E}_{z_1,z_2,\cdots,z_T}(\|f_{T+1} - f_{\lambda_T}^{V_{in}}\|_K^2) \le C_{K,V,\rho,b,\beta,s}T^{-\theta}$$

where

$$\theta := \min\left\{2 - 2\gamma - 2\alpha, \alpha - \gamma, b - 2\gamma\right\}.$$

Finally we get

$$\mathbb{E}_{z_1,z_2,\cdots,z_T}(\|f_{T+1} - f_\rho\|_K) \le \sqrt{C_{K,V,\rho,b,\beta,s}}T^{-\theta/2} + \left(\frac{\lambda_{T+1}}{2}\right)^{r-\frac{1}{2}}\|g_\rho\|_{L^2_{\rho_X}}, \quad \text{when } \frac{1}{2} < r \le \frac{3}{2}$$

and

$$\mathbb{E}_{z_1,z_2,\cdots,z_T}(\|f_{T+1} - f_\rho\|_{L^2_{\rho_X}}) \le \kappa\sqrt{C_{K,V,\rho,b,\beta,s}}T^{-\theta/2} + \left(\frac{\lambda_{T+1}}{2}\right)^{r}\|g_\rho\|_{L^2_{\rho_X}}, \quad \text{when } 0 < r \le 1.$$

Then our desired learning rates follow. ■

Acknowledgments

We would like to thank the referees for their constructive suggestions and comments. The work described in this paper was partially supported by a grant from the Research Grants Council of Hong Kong [Project No. CityU 104007] and National Science Fund for Distinguished Young Scholars of China [Project No. 10529101]. The corresponding author is Ding-Xuan Zhou.

Appendix A.

This appendix includes some detailed proofs.

A.1 Proof of Proposition 6

The first statement follows by taking $g(y) = y$ on Y because

$$|f_\rho(x) - f_\rho(u)| = \left| \int_Y g(y) d(\rho_x - \rho_u)(y) \right| \le \|\rho_x - \rho_u\|_{(C^s(Y))^*} \|g\|_{C^s(Y)}$$

and $\|g\|_{C^s(Y)} = \|g\|_{C(Y)} + |g|_{C^s(Y)} \le \sup_{y\in Y} |y| + 2^{1-s} \sup_{y\in Y} |y|$. Actually the above estimates tell us that f_ρ is continuous and belongs to $C^s(X)$ with $|f_\rho|_{C^s(X)} \le C_\rho(1+2^{1-s}) \sup_{y\in Y} |y|$.

For the second statement, since $Y = \{1,-1\}$, we have $f_\rho(x) = \rho_x(\{1\}) - \rho_x(\{-1\})$. It follows that for each $y \in Y$ and $x \in X$, there holds $\rho_x(\{y\}) = \frac{1+y f_\rho(x)}{2}$. So for any $g \in C^s(Y)$ and $x, u \in X$,

$$\int_Y g(y) d(\rho_x - \rho_u)(y) = \sum_{y\in Y} g(y) \frac{y[f_\rho(x) - f_\rho(u)]}{2} = \sum_{y\in Y} \frac{y g(y)}{2} [f_\rho(x) - f_\rho(u)].$$

Now the conclusion follows from

$$\left| \int_Y g(y) d(\rho_x - \rho_u)(y) \right| \le \|g\|_{C(Y)} |f_\rho(x) - f_\rho(u)| \le |f_\rho|_{C^s(X)} (d(x,u))^s \|g\|_{C^s(Y)}.$$

This proves Proposition 6. ■

A.2 Proof of Example 4

For $x \in X$, we have $\rho_x(\{1\}) = f_{\rho,-1}(x) + f_\rho(x)$ and $\rho_x(\{0\}) = 1 - 2f_{\rho,-1}(x) - f_\rho(x)$. Hence for any $g \in C^s(Y)$,

$$\int_Y g(y) d\rho_x = f_{\rho,-1}(x) \{g(1) - 2g(0) + g(-1)\} + f_\rho(x) \{g(1) - g(0)\} + g(0)$$

and for $u \in X$,

$$\int_Y g(y) d(\rho_x - \rho_u) = [f_{\rho,-1}(x) - f_{\rho,-1}(u)] \{g(1) - 2g(0) + g(-1)\} + [f_\rho(x) - f_\rho(u)] \{g(1) - g(0)\}.$$

Then our statement follows from the first part of Proposition 6. This proves the conclusion of Example 4. ■

A.3 Proof of Example 6

Let $x \in X$. When $f(x) \ge f_\rho^{V_{in}}(x)$, we see from the explicit form of the insensitive loss V_{in} that

$$\begin{aligned} & \int_Y V_{in}(y, f(x)) d\rho_x(y) - \int_Y V_{in}(y, f_\rho^{V_{in}}(x)) d\rho_x(y) \\ = & \int_{y \ge f(x)} f_\rho^{V_{in}}(x) - f(x) d\rho_x + \int_{y \le f_\rho^{V_{in}}(x)} f(x) - f_\rho^{V_{in}}(x) d\rho_x \\ & + \int_{f_\rho^{V_{in}}(x) < y < f(x)} f(x) + f_\rho^{V_{in}}(x) - 2y d\rho_x. \end{aligned}$$

It follows that

$$\mathcal{E}(f) - \mathcal{E}(f_\rho^{V_{in}}) = \int_X \left\{ |f(x) - f_\rho^{V_{in}}(x)| \Delta_x + 2 \int_{I_x} |f(x) - y| d\rho_x \right\} d\rho_X \tag{31}$$

where

$$\Delta_x := \left|\rho_x\left(\{y \in Y : y \le f_\rho^{V_{in}}(x)\}\right) - \rho_x\left(\{y \in Y : y > f_\rho^{V_{in}}(x)\}\right)\right|$$

and I_x in the open interval between $f_\rho^{V_{in}}(x)$ and $f(x)$. The same relation (31) also holds when $f(x) < f_\rho^{V_{in}}(x)$.

Now we use the special assumption on the conditional distributions and see that the median and mean of ρ_x are equal: $f_\rho^{V_{in}}(x) = f_\rho(x)$ for each $x \in X$. Moreover, $\Delta_x = 0$ and when $f_\rho(x) \le f(x) \le f_\rho(x) + 1$, we have

$$2\int_{I_x} |f(x) - y| d\rho_x = 2\int_0^{f(x)-f_\rho(x)} (f(x) - f_\rho(x) - u)\frac{p}{2}u^{p-1}du = \frac{|f(x) - f_\rho(x)|^{p+1}}{p+1}.$$

The same expression holds true when $f_\rho(x) - 1 \le f(x) < f_\rho(x)$. When $|f(x) - f_\rho(x)| > 1$, since ρ_x vanishes outside $[-f_\rho(x) - 1, f_\rho(x) + 1]$, we have $2\int_{I_x} |f(x) - y| d\rho_x = |f(x) - f_\rho(x)| - \frac{p}{p+1}$. Therefore, (31) is the same as

$$\begin{aligned}\mathcal{E}(f) - \mathcal{E}(f_\rho^{V_{in}}) &= \int_{\{x \in X : |f(x) - f_\rho(x)| \le 1\}} \frac{|f(x) - f_\rho(x)|^{p+1}}{p+1} d\rho_X \\ &\quad + \int_{\{x \in X : |f(x) - f_\rho(x)| > 1\}} |f(x) - f_\rho(x)| - \frac{p}{p+1} d\rho_X.\end{aligned}$$

This proves the desired expression for $\mathcal{E}(f) - \mathcal{E}(f_\rho^{V_{in}})$ and hence the bound for $\mathcal{D}(\lambda)$. ■

References

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

P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101:138–156, 2006.

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

D. R. Chen, Q. Wu, Y. Ying and D. X. Zhou. Support vector machine soft margin classifiers: error analysis. *Journal of Machine Learning Research*, 5:1143–1175, 2004.

N. Cesa-Bianchi, P. Long and M. K. Warmuth. Worst-case quadratic loss bounds for prediction using linear functions and gradient descent. *IEEE Transactions on Neural Networks*, 7:604–619, 1996.

N. Cesa-Bianchi, A. Conconi and C. Gentile. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50:2050-2057, 2004.

E. De Vito, A. Caponnetto, and L. Rosasco. Model selection for regularized least-squares algorithm in learning theory. *Foundations of Computational Mathematics*, 5:59–85, 2005.

E. De Vito, L. Rosasco, A. Caponnetto, M. Piana, and A. Verri. Some properties of regularized kernel methods. *Journal of Machine Learning Research*, 5:1363–1390, 2004.

L. Devroye, L. Györfi and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, New York, 1997.

T. Evgeniou, M. Pontil and T. Poggio. Regularization networks and support vector machines. *Advances in Computational Mathematics*, 13:1–50, 2000.

D. Hardin, I. Tsamardinos, and C. F. Aliferis. A theoretical characterization of linear SVM-based feature selection. *Proc. of the 21st Int. Conf. on Machine Learning*, Banff, Canada, 2004.

J. Kivinen, A. J. Smola, and R. C. Williamson. Online learning with kernels. *IEEE Trans. Signal Processing* 52:2165–2176, 2004.

S. Mukherjee and Q. Wu. Estimation of gradients and coordinate covariation in classification. *Journal of Machine Learning Research*, 7:2481–2514, 2006.

R. T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, 1970.

S. Smale and Y. Yao. Online learning algorithms. *Foundations of Computational Mathematics*, 6:145–170, 2006.

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

S. Smale and D.X. Zhou. Shannon sampling and function reconstruction from point values. *Bulletin of the American Mathematical Society*, 41:279–305, 2004.

S. Smale and D.X. Zhou. Learning theory estimates via integral operators and their applications. *Constructive Approximation*, 26:153–172, 2007.

S. Smale and D. X. Zhou. Online learning with Markov sampling. *Analysis and Applications*, 7:87–113, 2009.

I. Steinwart. Support vector machines are universally consistent. *Journal of Complexity*, 18:768–791, 2002.

I. Steinwart, D. Hush and C. Scovel. Learning from dependent observations. *Journal of Multivariate Analysis*, 100:175–194, 2009.

P. Tarrès and Y. Yao. Online learning as stochastic approximations of regularization paths. preprint, 2005.

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

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

Q. Wu, Y. Ying and D. X. Zhou. Learning theory: From regression to classification. *Topics in Multivariate Approximation and Interpolation* (K. Jetter et al. eds.), Elsivier, pp. 257–290, 2006.

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

Y. Ying. Convergence analysis of online algorithms. *Advances in Computational Mathematics*, 27:273–291, 2007.

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

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

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

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

D. X. Zhou. Derivative reproducing properties for kernel methods in learning theory. *Journal of Computational and Applied Mathematics*, 220:456–463, 2008.