

ONLINE LUM CLASSIFICATION WITH VARYING THRESHOLDS AND NON-IDENTICAL SAMPLING DISTRIBUTIONS

WANG Ze-xing

(*School of Mathematics and Statistics, Wuhan University, Wuhan 430072, China*)

Abstract: Large-margin Unified Machines (LUMs) have been widely studied in classification. LUMs is a family of large-margin classifiers and it offers a unique transition from soft to hard classification. In this paper, we are devoted to investigate the online binary classification algorithm with LUM loss function and non-identical sampling distributions, where each time a sample is drawn independently from different probability distributions. Especially, we also consider the LUM loss function with varying thresholds where parameter of the loss function decreases with the iteration process. The numerical convergence analysis of the algorithm associated with reproducing kernel Hilbert space (RKHS) is presented and the learning rate of this general framework is obtained.

Keywords: sampling with non-identical distributions; online classification; LUM loss with varying threshold; reproducing kernel Hilbert spaces

2010 MR Subject Classification: 62J99

Document code: A **Article ID:** 0255-7797(2023)03-0229-18

1 Introduction

In supervised learning, there is a large amount of literature on various classification methods. Among various classification methods, there are two main groups of methods: soft and hard classification which are defined in [1]. A new unified framework of large-margin classifiers which covers a broad range of methods from hard to soft classifiers is proposed in [2]. This family is called as Large-margin Unified Machines (LUMs). Correspondingly, the expression of its loss function is as follows:

Definition 1.1 The LUM loss function is defined as

$$V(u) = \begin{cases} 1 - u & \text{if } u \leq \frac{c}{1+c}, \\ \frac{1}{c+1} \left(\frac{a}{(1+c)u - c + a} \right)^a & \text{if } u > \frac{c}{1+c}. \end{cases}$$

where $0 \leq c \leq \infty$ and $a > 0$.

* **Received date:** 2022-01-12

Accepted date: 2022-04-06

Foundation item: Supported by National Natural Science Foundation of China(12071356).

Biography: Wang Zexing(1997–), male, born at Huainan, Anhui, postgraduate, major in machine learning.

By the definition, it's obvious that $V(\cdot)$ is a convex function. Besides, it is differentiable everywhere and has no zero for $0 \leq c < \infty$. When c is infinite, LUM loss function is standard hinge loss of SVM which is a kind of hard classification. However, when $c < \infty$ and a is fixed, we can estimate the class conditional probability by fisher consistency, see [2]. Obviously, the LUMs become soft classification now. Hence, this new framework offers a unique transition from soft to hard classification. Furthermore, LUM loss functions with $c = 0$ are the best case to estimate the class conditional probability. Hence, the study of the parameter c changing from a finite number to 0 in LUM loss function is very valuable. Inspired by [3], the methods of analysis about the varying parameter at each iteration have been established. In recent researches, the consistency of kernel based LUMs within the framework of learning theory has been proven, see [4]. But considering the varying parameter c , especially for finite c and fixed a , the quantitative convergence analysis of kernel is rarely studied. This is exactly the first novelty in this passage.

The second novelty is the analysis of sampling non-identical distributions with LUMs. In the literature on learning theory, samples are often drawn independently from an identical distribution. However, the data in real life are always not from an identical distribution. There are many researches about independent but non-identical samples for regression and classification, see [5–7]. Based on these studies and some consensus assumptions for sampling distribution, we consider the binary classification problem with LUM loss function in the setting that samples are drawn according to a non-identical sequence of probability distributions in this paper.

Machine learning can ultimately be boiled down to the optimization problems. The optimization algorithm is aim to get its solution step by step from the sample data. When the sample data is presented in a sequential manner, the stochastic gradient descent method is an option which is widely used in research. Online learning algorithm is a type of stochastic gradient descent methods which was first proposed in [8]. With linear complexity, online learning provides an important family of efficient and scalable machine learning algorithms for real applications. Thus, a variety of online learning paradigms have been introduced, see [3, 9–11]. In this paper, we aim to investigate the numerical convergence analysis of the online learning algorithms with LUM loss function with varying parameter c in a reproducing Kernel Hilbert space(RKHS) associated with non-identical distributions. Besides, we can show that the algorithm with varying LUM loss performs better than the situation with LUM loss where $c = 0$ from the simulation.

The rest of this paper is organized as follows. We begin with Section 2 by providing necessary background and notations which are required for later analysis of our algorithm. Then we present our main theorems to show the ability of our algorithm. Section 3 is devoted to present the key conclusions and their proofs which will be applied to the later proofs. Section 4 is the complete proof process of the main result. In Section 5, we give a simulation about online learning algorithm in the settings of this paper. Section 6 is a brief summary. The last part is Appendix which is the additional proof.

2 Backgrounds and Main Results

In the learning theory of binary classification, the usual setting is that (X, d) (input space) is a separable metric space and Y (output space) is $\{-1, 1\}$ representing the set of two classes. Let $Z = X \times Y$. The target is to search a binary classifier $\mathcal{C} : X \rightarrow Y$ which makes a prediction $y = \mathcal{C}(x) \in Y$ for every point $x \in X$. Let $f : X \rightarrow \mathbb{R}$ be a real valued function, then the classifier is $\mathcal{C}(x) = \text{sgn}(f(x))$ where $\text{sgn}(f(x)) = 1$ if $f(x) \geq 0$ and $\text{sgn}(f(x)) = -1$ if $f(x) < 0$. The loss function $V(yf(x))$ is to measure the prediction at every point for the real function f . In this paper, let $u = yf(x)$, we choose the LUM loss function with varying thresholds.

In this paper, we are devoted to investigate the LUM loss function with unchanged parameter a and the varying parameter c . What's more, we assume that c is gradually reduced to 0. Then the loss function is

$$V^c(u) = \begin{cases} 1 - u & \text{if } u \leq \frac{c}{1+c}, \\ \frac{1}{c+1} \left(\frac{a}{(1+c)u - c + a} \right)^a & \text{if } u > \frac{c}{1+c}. \end{cases} \tag{2.1}$$

Because of the parameter a is fixed, for the need of analysis, we assume $a = 1$ in this paper when we use the definition of $V^c(\cdot)$. When $c = c_0 = 0$ and $a = 1$, it is

$$V^{c_0}(u) = \begin{cases} 1 - u & \text{if } u \leq 0, \\ \frac{1}{1+u} & \text{if } u > 0. \end{cases} \tag{2.2}$$

2.1 Sampling with Non-identical Distributions

Differ from the i.i.d. samples, we assume that the sample $z_t = (x_t, y_t)$ is independently drawn from a distribution $\rho^{(t)}$ on Z at each step $t = 1, 2, \dots$. It is clear that the distributions of the sample $z = \{(x_t, y_t)\}_{t=1}^T$ are not identical. In the existing studies, we assume that the sequence $\{\rho_X^{(t)}\}_{t=1,2,\dots}$ of marginal distributions on X converges polynomially in the dual of the Hölder space $C^s(X)$ for some $0 < s \leq 1$. We define the Hölder space $C^s(X)$ is 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 2.2 We say that the sequence $\{\rho_X^{(t)}\}_{t=1,2,\dots}$ converges polynomially to a probability distribution ρ_X in $(C^s(X))^*(0 \leq s \leq 1)$ if there exist $C > 0$ and $b > 0$ such that

$$\|\rho_X^{(t)} - \rho_X\|_{(C^s(X))^*} \leq Ct^{-b}, \quad t \in \mathbb{N}. \tag{2.3}$$

With definition 5 and Proposition 6 in [5], the sequence of conditional distributions $\{\rho_x : x \in X\}$ is Lipschitz s in $(C^s(Y))^*$ if and only if $f_\rho \in C^s(X)$. The regression function f_ρ is defined by

$$f_\rho(x) = \int_Y y d\rho_x(y), \quad x \in X.$$

Definition 2.3 The set of distributions $\{\rho_x : x \in X\}$ is Lipschitz s in $(C^s(Y))^*$ if

$$\|\rho_x - \rho_u\|_{(C^s(Y))^*} \leq C_\rho (d(x, u))^s, \quad \forall x, u \in X. \quad (2.4)$$

where $C_\rho \geq 0$ is a constant.

2.2 The Reproducing Kernel Hilbert Space

Let $K : X \times X \rightarrow \mathbb{R}$ be a Mercer kernel if it is continuous, symmetric and positive semi-definite. We define the reproducing kernel Hilbert space (RKHS) \mathcal{H}_K with the kernel K is the completion of the linear span of the set of functions $\{K_x = K(x, \cdot) : x \in X\}$. The inner product $\langle \cdot, \cdot \rangle_K$ is given by $\langle K_x, K_y \rangle_K = K(x, y)$. We denote $\kappa = \sup_{x \in X} \sqrt{K(x, x)}$. The reproducing property of RKHS is

$$\langle K_x, f \rangle = f(x), \quad x \in X, f \in \mathcal{H}_K. \quad (2.5)$$

From (2.5), we have

$$\|f\|_{C(X)} \leq \kappa \|f\|_K, \quad \forall x \in X, f \in \mathcal{H}_K. \quad (2.6)$$

Definition 2.4 K satisfies the kernel condition of order $s (s > 0)$ if $K \in C^s(X \times X)$ and for some $\kappa_{2s} > 0$,

$$|K(x, x) - 2K(x, u) + K(u, u)| \leq \kappa_{2s}^2 (d(x, u))^{2s}, \quad \forall x, u \in X. \quad (2.7)$$

When $0 < s \leq \frac{1}{2}$ and $K \in C^{2s}(X \times X)$, (2.7) holds true.

Proposition 2.5 If (2.7) holds for K , then

$$\|g\|_{C^s(X)} \leq (\kappa + \kappa_{2s}) \|g\|_K, \quad \forall g \in \mathcal{H}_K. \quad (2.8)$$

The proof of this proposition can be found in [5].

2.3 Some Conventions

In analysis of classification, the classical framework is established on the basis of some conventions. Here we give some universal definitions. The error of the classifier \mathcal{C} is measured by the misclassification error $\mathcal{R}(\mathcal{C})$ and it is defined as follows:

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

The *Bayes rule* is the best classifier which minimizes the misclassification error $\mathcal{R}(\mathcal{C})$ and it is expressed as $f_{bayes} = \text{sgn}(f_\rho)$:

$$f_{bayes}(x) = \begin{cases} 1 & \text{if } \rho_x(1) \geq \rho_x(-1), \\ -1 & \text{if } \rho_x(1) < \rho_x(-1). \end{cases}$$

Definition 2.6 When (x, y) in Z is from ρ on Z and V^c has the formulation as (2.1), then we define the generalization error of f :

$$\varepsilon^c(f) = \int_Z V^c(f(x)y) d\rho.$$

Remark 2.7 Here we define that ρ on Z is a combination of the marginal distribution ρ_X and the conditional distribution ρ_x .

Besides, we define f_ρ^c is the minimizer of $\varepsilon^c(f)$.

$$f_\rho^c(x) = \arg \inf_{f \in \mathcal{H}_K} \int_Z V^c(f(x)y) d\rho = \arg \inf_{u \in \mathbb{R}} \int_Y V^c(uy) d\rho_x(y), \quad x \in X.$$

Definition 2.8 As defined in previous article [6], the regularizing function $f_\lambda^c \in \mathcal{H}_K$ is defined as

$$f_\lambda^c = \arg \inf_{f \in \mathcal{H}_K} \{ \varepsilon^c(f) + \frac{\lambda}{2} \|f\|_K^2 \}, \tag{2.9}$$

where $\lambda > 0$.

We now give the prior conditions on the distribution ρ and the space \mathcal{H}_k to measure the approximation error $\mathcal{D}^{c_0}(\lambda)$, let

$$\mathcal{D}^{c_0}(\lambda) = \inf_{f \in \mathcal{H}_K} \{ \varepsilon^{c_0}(f) - \varepsilon^{c_0}(f_\rho^{c_0}) + \frac{\lambda}{2} \|f\|_K^2 \}, \tag{2.10}$$

then we assume that

$$\mathcal{D}^{c_0}(\lambda) \leq \mathcal{D}_0 \lambda^\beta, \tag{2.11}$$

for some $0 \leq \beta \leq 1$ and $\mathcal{D}_0 \geq 0$

Definition 2.9 We say the convex and differentiable function $V(u) = V(yf)$ has incremental exponent $p \geq 0$ if there exists some $N_V > 0$ such that

$$\left| V'(u) \right| \leq N_V |u|^p, \quad V(u) \leq N_V |u|^{p+1} \quad \forall |u| \geq 1.$$

From the definition, we can know $V^c(\cdot)$ has the incremental exponent $p = 0$ and $\exists N_1 > 0, N_{V_c} \leq N_1$ holds true for any $c \geq 0$.

Now we introduce an elementary inequality used to the error analysis.

Proposition 2.10 Let $c > 0, q_2 \geq 0, t \geq 2 \in \mathbb{Z}$ and $0 < q_1 < 1$, then we have:

$$\sum_{i=1}^{t-1} i^{-q_2} \exp\{-c \sum_{j=i+1}^t j^{-q_1}\} \leq \left(\frac{2^{q_1+q_2}}{c}\right) + \left(\frac{1+q_2}{ec(1-2^{q_1-1})}\right)^{\frac{1+q_2}{1-q_1}} t^{q_1-q_2}. \tag{2.12}$$

The elementary inequality can be found in [12].

2.4 Online Learning Algorithm for Classification

Definition 2.11 The online learning algorithm is defined by $f_1 = 0$ and

$$f_{t+1} = f_t - \eta_t \{ \partial V^{c_t}(y_t f_t) K_{x_t} + \lambda_t f_t \}, \quad \text{for } t = 1, 2, \dots,$$

where $\lambda_t > 0$ is called the regularization parameter, $\eta_t > 0$ is called the step size, c_t is the parameter of the loss function which converge to zero as the step t increases, and $\partial V^{c_t}(y_t f_t) = V'_{c_t}(y_t f_t) y_t$ where $V'_{c_t}(\cdot)$ is the derivative of $V^{c_t}(\cdot)$ with default $a = 1$.

In other words, it is

$$f_{t+1} = \begin{cases} (1 - \eta_t \lambda_t) f_t + \left[\frac{a}{(1+c_t)y_t f_t - c_t + a} \right]^{a+1} \cdot \eta_t y_t K_{x_t} & \text{if } y_t f_t(x_t) > \frac{c_t}{1+c_t}, \\ (1 - \eta_t \lambda_t) f_t + \eta_t y_t K_{x_t} & \text{if } y_t f_t(x_t) \leq \frac{c_t}{1+c_t}, \end{cases} \quad (2.13)$$

where $a = 1$.

Proposition 2.12 If f_t is defined by (2.13) and λ_t decreases with iteration, we can conclude that

$$\|f_t\|_K \leq \frac{\kappa}{\lambda_t}. \quad (2.14)$$

Proof When $f_1 = 0$, it is easy to see that $\|f_1\| \leq \frac{\kappa}{\lambda_1}$. If we assume that $\|f_t\|_K \leq \frac{\kappa}{\lambda_t}$ holds, from the formulation of f_{t+1} , we can show :

$$\begin{aligned} \|f_{t+1}\|_K &\leq (1 - \eta_t \lambda_t) \|f_t\|_K + \eta_t \kappa \\ &\leq (1 - \eta_t \lambda_t) \frac{\kappa}{\lambda_t} + \eta_t \kappa = \frac{\kappa}{\lambda_t} \leq \frac{\kappa}{\lambda_{t+1}}. \end{aligned}$$

Then we can complete the proof because of the inductive method.

2.5 The Main Results

The online learning algorithm mentioned in this paper is based on LUM loss function with varying parameter c which is different from the normal unchanged loss function. Next we consider the convergence of this algorithm and the numerical result is given by the following theorems.

Theorem 2.13 Let $f_\rho \in C^s(X)$, $K \in C^{2s}(X \times X)$ for some $0 \leq s \leq \frac{1}{2}$. Suppose assumptions (2.3) and (2.11) hold, f_t is from (2.13) and $a=1$, if

$$\lambda_t = \lambda_1 t^{-\gamma}, \eta_t = \eta_1 t^{-\alpha}, c_t = c_1 t^{-\theta} \quad (2.15)$$

with $\lambda_1, \eta_1, c_1 > 0$ and

$$0 < \gamma < \frac{b}{2}, (5 - \beta)\gamma < 2, \max\{\gamma, \frac{\gamma(\beta - 3)}{2}\} < \alpha < 1 + \frac{\gamma(\beta - 3)}{2}, \theta > \max\{\beta\gamma, \alpha + 2\gamma - 1\}, \quad (2.16)$$

then

$$\mathbb{E}_{z_1, \dots, z_T} (\mathcal{R}(\text{sgn}(f_{T+1})) - \mathcal{R}(f_{\text{bayes}})) \leq C^* T^{-\omega^*},$$

where C^* is a constant independent of T and

$$\omega^* = \min\left\{\frac{\gamma\beta}{2}, \frac{\theta - \gamma}{2}, \frac{1}{2} - \frac{\gamma(3 - \beta)}{4} - \frac{\alpha}{2}, \frac{1 - 2\gamma + \theta - \alpha}{2}, \frac{\alpha - \gamma}{4}, \frac{b - 2\gamma}{4}\right\}.$$

The proof will be provided in Section 4.

From the theorem, the results tell us that the convergence rate is less than $O(T^{-\frac{1}{2}})$ which is the theoretically best limit. Besides, when $\gamma \geq \frac{4}{5}$ and $\beta > 1$, $\frac{1}{2} - \frac{\gamma(3-\beta)}{4} - \frac{\alpha}{2} < \min\{\frac{\gamma\beta}{2}, \frac{1-2\gamma+\theta-\alpha}{2}\}$ is provided by $\alpha > \frac{\gamma(\beta-3)}{2}$ and $\theta > \beta\gamma$. At the same time, $\gamma > \frac{4}{5}$ implies that $\frac{1-2\gamma+\theta-\alpha}{2} < \frac{\theta-\gamma}{2}$. And if we assume that $\alpha > \frac{2}{3} + \frac{(\beta-2)\gamma}{3}$, it is obvious that $\frac{1}{2} - \frac{\gamma(3-\beta)}{4} - \frac{\alpha}{2} < \frac{\alpha-\gamma}{4}$. In summary, the rate can be transformed into $\mathbb{E}_{z_1, \dots, z_T}(\mathcal{R}(\text{sgn}(f_{T+1})) - \mathcal{R}(f_{\text{bayes}})) \leq C^* T^{\{-\frac{1}{2} + \frac{\gamma(3-\beta)}{4} + \frac{\alpha}{2}, -\frac{b-2\gamma}{4}\}}$. Furthermore, when $b \leq 2 - 2\alpha + (\beta - 1)\gamma$, the limit of rate is $O(T^{-\frac{4}{5} + \frac{2}{5}})$ with $\gamma = \frac{4}{5}$. In this situation, we can conclude that the convergence rate depends heavily on parameter b where α, γ, θ are arbitrary values satisfying some conditions and β is big enough.

3 Key Analysis and Conclusions

In this section, we prove some theorems and lemmas that will be used in the proof of our main results.

First of all, we show comparison theorems of the LUM loss function which is the basis of the proof.

Theorem 3.14 For the loss function V^c ($c > 0$) and any measurable function $f : X \rightarrow \mathbb{R}$, the following holds :

$$\mathcal{R}(\text{sgn}(f)) - \mathcal{R}(f_{\text{bayes}}) \leq C_1 \{\varepsilon^c(f) - \varepsilon^c(f_\rho^c)\},$$

where $C_1 > 0$.

Theorem 3.15 For the loss function V^{c_0} ($c_0 = 0$) and any measurable function $f : X \rightarrow \mathbb{R}$, the following holds :

$$\mathcal{R}(\text{sgn}(f)) - \mathcal{R}(f_{\text{bayes}}) \leq C_2 \{\varepsilon^{c_0}(f) - \varepsilon^{c_0}(f_\rho^{c_0})\}^{\frac{1}{2}}, \tag{3.1}$$

where $C_2 > 0$.

The proof of the theorem can be found in [13].

Next, the key analysis for the varying loss functions is considered. Here we give three lemmas which are very important in the analysis and will be applied in the following proof.

Lemma 3.16 Let $\mathcal{D}^c(\lambda) = \inf_{f \in \mathcal{H}_K} \{\varepsilon^c(f) - \varepsilon^c(f_\rho^c) + \frac{\lambda}{2} \|f\|_K^2\}$, when the parameter a in loss function is 1, we have

$$\mathcal{D}^c(\lambda) \leq \mathcal{D}^{c_0}(\lambda) + 2c.$$

From the lemma, because of the definition of f_λ^c , (2.10) and (2.11), it's easy to show

$$\|f_\lambda^c\|_K \leq \sqrt{\frac{2\mathcal{D}_0\lambda^\beta + 4c}{\lambda}}. \tag{3.2}$$

Proof First of all we can show that $\|V^c - V^{c_0}\|_\infty \leq c$ for any $c \geq 0$. This proof can be found in Appendix.

Then, we have

$$|\varepsilon^{c_0}(f) - \varepsilon^c(f)| = \left| \int_Z V^{c_0}(f(x)y) - V^c(f(x)y) d\rho \right| \leq c.$$

It follows that

$$\begin{aligned} \varepsilon^c(f) - \varepsilon^c(f_\rho^c) &= \varepsilon^c(f) - \varepsilon^{c_0}(f) + \varepsilon^{c_0}(f) - \varepsilon^{c_0}(f_\rho^{c_0}) + \varepsilon^{c_0}(f_\rho^{c_0}) - \varepsilon^{c_0}(f_\rho^c) + \varepsilon^{c_0}(f_\rho^c) - \varepsilon^c(f_\rho^c) \\ &\leq \varepsilon^{c_0}(f) - \varepsilon^{c_0}(f_\rho^{c_0}) + 2c \end{aligned}$$

because of $\varepsilon^{c_0}(f_\rho^{c_0}) \leq \varepsilon^{c_0}(f_\rho^c)$. Then

$$\begin{aligned} \mathcal{D}^c(\lambda) &= \inf_{f \in \mathcal{H}_K} \{ \varepsilon^c(f) - \varepsilon^c(f_\rho^c) + \frac{\lambda}{2} \|f\|_K^2 \} \\ &\leq \inf_{f \in \mathcal{H}_K} \{ \varepsilon^{c_0}(f) - \varepsilon^{c_0}(f_\rho^{c_0}) + 2c + \frac{\lambda}{2} \|f\|_K^2 \} \\ &= \mathcal{D}^{c_0}(\lambda) + 2c. \end{aligned}$$

We have completed the proof.

Lemma 3.17 Let $\lambda > 0$ and V^c be the loss function with $a = 1$. Take $0 \leq \nu < \mu$ and put the expression of f_λ^c with $c = \mu, \nu$, then we have

$$\|f_\lambda^\mu - f_\lambda^\nu\|_K \leq \frac{6\kappa}{\lambda} |\mu - \nu|. \quad (3.3)$$

This lemma is crucial in our analysis and its proof is one of the main novelties.

Proof Taking the functional derivative of (2.9), we know that the regularization function $f_\lambda^c \in \mathcal{H}_K$ satisfies

$$\lambda(f_\lambda^c) + \int_Z \partial V^c(f_\lambda^c(x)y) K_x d\rho = 0.$$

Due to $\|f_\lambda^\mu - f_\lambda^\nu\|_K^2 = \langle f_\lambda^\mu - f_\lambda^\nu, f_\lambda^\mu - f_\lambda^\nu \rangle_K$, we know that

$$\begin{aligned} \|f_\lambda^\mu - f_\lambda^\nu\|_K^2 &= -\frac{1}{\lambda} \langle f_\lambda^\mu - f_\lambda^\nu, \int_Z \{ \partial V^\mu(f_\lambda^\mu(x)y) - \partial V^\nu(f_\lambda^\nu(x)y) \} K_x d\rho \rangle_K \\ &> 0. \end{aligned}$$

Applying the reproducing property (2.5) yields

$$\begin{aligned} \|f_\lambda^\mu - f_\lambda^\nu\|_K^2 &= -\frac{1}{\lambda} \int_Z \{ f_\lambda^\mu(x) - f_\lambda^\nu(x) \} \{ \partial V^\mu(f_\lambda^\mu(x)y) - \partial V^\nu(f_\lambda^\nu(x)y) \} d\rho \\ &:= I_1 + I_2, \end{aligned}$$

where

$$I_1 = -\frac{1}{\lambda} \int_Z \{f_\lambda^\mu(x) - f_\lambda^\nu(x)\} \{\partial V^\mu(f_\lambda^\mu(x)y) - \partial V^\mu(f_\lambda^\nu(x)y)\} d\rho,$$

$$I_2 = \frac{1}{\lambda} \int_Z \{f_\lambda^\mu(x) - f_\lambda^\nu(x)\} \{\partial V^\nu(f_\lambda^\nu(x)y) - \partial V^\mu(f_\lambda^\nu(x)y)\} d\rho.$$

By the convexity of the loss function $V^c(\cdot)$, we know that $I_1 \leq 0$.

To bound I_2 , we observe that

$$\partial V^\nu(f_\lambda^\nu(x)y) - \partial V^\mu(f_\lambda^\nu(x)y) = [V'_\nu(f_\lambda^\nu(x)y) - V'_\mu(f_\lambda^\nu(x)y)]y,$$

where

$$V'_c(u) = \begin{cases} -1 & \text{if } u \leq \frac{c}{1+c}, \\ -\left(\frac{1}{(1+c)u-c+1}\right)^2 & \text{if } u > \frac{c}{1+c}, \end{cases}$$

with $c = \mu, \nu$ and $u = f_\lambda^\nu(x)y$. Because of $\mathbb{E}_Z = \mathbb{E}_{X,Y} = \mathbb{E}_X \mathbb{E}_{Y|X}$, we have

$$I_2 = \frac{1}{\lambda} \int_X \{f_\lambda^\mu(x) - f_\lambda^\nu(x)\} \{\rho_x(1)h(1) + (\rho_x(-1) - 1)h(-1)\} d\rho_X(x),$$

where

$$h(1) = (1 - N)I_{\{\frac{\nu}{1+\nu} < f_\lambda^\nu(x) \leq \frac{\mu}{1+\mu}\}} + (M - N)I_{\{f_\lambda^\nu(x) > \frac{\mu}{1+\mu}\}},$$

$$h(-1) = (1 - N')I_{\{-\frac{\mu}{1+\mu} < f_\lambda^\nu(x) \leq -\frac{\nu}{1+\nu}\}} + (M' - N')I_{\{f_\lambda^\nu(x) < -\frac{\mu}{1+\mu}\}}.$$

and

$$M = \left(\frac{1}{(1 + \mu)f_\lambda^\nu(x) - \mu + 1}\right)^2, \quad N = \left(\frac{1}{(1 + \nu)f_\lambda^\nu(x) - \nu + 1}\right)^2,$$

$$M' = \left(\frac{1}{-(1 + \mu)f_\lambda^\nu(x) - \mu + 1}\right)^2, \quad N' = \left(\frac{1}{-(1 + \nu)f_\lambda^\nu(x) - \nu + 1}\right)^2.$$

Now we bound $|h(1)|$ and $|h(-1)|$.

(1) If $M \leq N$, we can get $f_\lambda^\nu(x) \geq 1$. It implies $(M - N)I_{\{f_\lambda^\nu(x) \geq 1\}} \leq 0$.

When $\frac{\mu}{1+\mu} < f_\lambda^\nu(x) < 1$, we denote $\varphi(x) = \left(\frac{1}{(1+x)f_\lambda^\nu(x)-x+1}\right)^2$, then

$$M - N = \varphi'(\eta)(\mu - \nu), \quad \exists \eta \in (\nu, \mu), \tag{3.4}$$

where

$$\varphi'(\eta) = 2 \left(\frac{1}{(1 + \eta)f_\lambda^\nu(x) - \eta + 1}\right)^3 (1 - f_\lambda^\nu(x)) \tag{3.5}$$

Due to $\frac{\mu}{1+\mu} < f_\lambda^\nu(x) < 1$, we have $\varphi'(\eta) \leq 2$ such that $M - N \leq 2(\mu - \nu)$. Hence, we can show

$$(M - N)I_{\{f_\lambda^\nu(x) > \frac{\mu}{1+\mu}\}} = (M - N)I_{\{\frac{\mu}{1+\mu} < f_\lambda^\nu(x) < 1\}} + (M - N)I_{\{f_\lambda^\nu(x) \geq 1\}}$$

$$\leq (M - N)I_{\{\frac{\mu}{1+\mu} < f_\lambda^\nu(x) < 1\}} < 2(\mu - \nu).$$

Similarly, we can get $(M' - N')I_{\{f_\lambda^\nu(x) < -\frac{\mu}{1+\mu}\}} \leq 2(\mu - \nu)$ in the same way.

(2) When $\frac{\nu}{1+\nu} < f_\lambda^\nu(x) \leq \frac{\mu}{1+\mu}$, we can show

$$N = \left(\frac{1}{(1 + \nu)f_\lambda^\nu(x) - \nu + 1}\right)^2 > \left(\frac{1}{\mu - \nu + 1}\right)^2$$

because of $(1 + \nu)f_\lambda^\nu(x) < (1 + \mu)f_\lambda^\nu(x) \leq (1 + \mu)\frac{\mu}{1+\mu} = \mu$.

Due to inequality $\frac{1}{1+x} \geq 1 - x$ for $x > -1$ and $(1 - x)^2 \geq 1 - 2x$, we know that

$$1 - N < 1 - \left(\frac{1}{\mu - \nu + 1}\right)^2 < 1 - (1 - (\mu - \nu))^2 \leq 1 - (1 - 2(\mu - \nu)) = 2(\mu - \nu).$$

It implies $(1 - N)I_{\{\frac{\nu}{1+\nu} < f_\lambda^\nu(x) \leq \frac{\mu}{1+\mu}\}} \leq 2(\mu - \nu)$. In the same way, $(1 - N')I_{\{-\frac{\mu}{1+\mu} < f_\lambda^\nu(x) \leq -\frac{\nu}{1+\nu}\}} \leq 2(\mu - \nu)$ can be proved. Hence, we have

$$h(1) \leq 2(\mu - \nu), \quad h(-1) \leq 2(\mu - \nu).$$

On the other hand, due to $1 - N > 0$, we know that

$$\begin{aligned} h(1) &= (1 - N)I_{\{\frac{\nu}{1+\nu} < f_\lambda^\nu(x) \leq \frac{\mu}{1+\mu}\}} + (M - N)I_{\{f_\lambda^\nu(x) > \frac{\mu}{1+\mu}\}} \\ &\geq (M - N)I_{\{f_\lambda^\nu(x) > \frac{\mu}{1+\mu}\}} \\ &= (M - N)I_{\{\frac{\mu}{1+\mu} < f_\lambda^\nu(x) \leq 1\}} + (M - N)I_{\{f_\lambda^\nu(x) > 1\}} \\ &\geq (M - N)I_{\{f_\lambda^\nu(x) > 1\}}. \end{aligned}$$

With (3.4) and (3.5), due to $\varphi'(\eta) \geq -2$ when $f_\lambda^\nu(x) > 1$, we have $(M - N)I_{\{f_\lambda^\nu(x) > 1\}} \geq -2(\mu - \nu)$. It implies $h(1) \geq -2(\mu - \nu)$. For the same reason, we know that $h(-1) \geq -2(\mu - \nu)$.

In summary, we have $|h(1)| \leq 2(\mu - \nu)$ and $|h(-1)| \leq 2(\mu - \nu)$.

Now we can show

$$\begin{aligned} |\rho_x(1)h(1) + (\rho_x(1) - 1)h(-1)| &= |\rho_x(1)h(1) + \rho_x(1)h(-1) - h(-1)| \\ &\leq |h(1)| + 2|h(-1)| \\ &\leq 6(\mu - \nu) \end{aligned}$$

With (2.6), it causes that

$$I_2 \leq \frac{\kappa}{\lambda} \|f_\lambda^\mu - f_\lambda^\nu\|_K \cdot 6(\mu - \nu).$$

It follows that

$$\|f_\lambda^\mu - f_\lambda^\nu\|_K \leq \frac{6\kappa}{\lambda} |\mu - \nu|.$$

We complete the proof now.

Lemma 3.18 Let $h, g \in C^s(X)$. If (2.4) holds, then

$$\begin{aligned} &\left| \int_Z V^c(yh(x)) - V^c(yg(x)) d[\rho^{(t)} - \rho] \right| \\ &\leq \{(\|h\|_{C^s(X)} + \|g\|_{C^s(X)}) + 2C_\rho \tilde{B}_{h,g}\} \left\| \rho_X^{(t)} - \rho_X \right\|_{(C^s(X))^*}, \end{aligned}$$

where

$$\tilde{B}_{h,g} = \sup \{ \|V^c(\cdot, f)\|_{C^s(Y)} : |f| \leq \max\{\|h\|_{C(X)}, \|g\|_{C(X)}\} \},$$

The proof of this lemma can be found in [5] because of $|\partial V^c(yf)| \leq 1$.

Now we consider the drift error. In (2.13), the regularization λ_t changes with the iteration and it affects the drift error which is estimated by the approximation error and has been well studied for classification (see [10]).

Definition 3.19 In this paper, the drift error is defined by

$$d_t = \left\| f_{\lambda_t}^{c_{t-1}} - f_{\lambda_{t-1}}^{c_{t-1}} \right\|_K.$$

Theorem 3.20 V^c is the LUM loss function and f_λ^c is defined by (2.9). If $\vartheta > \lambda > 0$, we have

$$\|f_\lambda^c - f_\vartheta^c\|_K \leq \frac{\vartheta}{2} \left(\frac{1}{\lambda} - \frac{1}{\vartheta} \right) (\|f_\lambda^c\|_K + \|f_\vartheta^c\|_K).$$

We can find the proof of this theorem from [5]. When $\lambda_t = \lambda_1 t^{-\gamma}$ and $\lambda = \lambda_t$, $\vartheta = \lambda_{t-1}$, $c = c_{t-1}$ for $t \geq 2$, because of $\|f_\lambda^c\|_K \leq \sqrt{\frac{2D_0\lambda^\beta + 4c}{\lambda}}$ in (3.2), we have

$$d_t \leq 2(t-1)^{\frac{\gamma}{2}-1} \sqrt{\frac{D_0\lambda_1^\beta(t-1)^{-\beta\gamma} + 2c_{t-1}}{\lambda_1}}.$$

4 The Proof of the Main Result

Now we give the proof of Theorem 2.13.

First of all, we assume the conditions of Theorem 2.13 and (2.15), (2.16) hold true.

With (3.1), we have $\mathcal{R}(\text{sgn}(f_{T+1})) - \mathcal{R}(f_{\text{bayes}}) \leq C_2 \{ \varepsilon^{c_0}(f_{T+1}) - \varepsilon^{c_0}(f_\rho^{c_0}) \}^{\frac{1}{2}}$. Besides, we can prove $\mathbb{E}(\sqrt{\xi}) \leq \sqrt{\mathbb{E}(\xi)}$ with Hölder inequality where ξ is a random variable. Then we have

$$\mathbb{E}_{z_1, \dots, z_T}(\mathcal{R}(\text{sgn}(f_{T+1})) - \mathcal{R}(f_{\text{bayes}})) \leq C_2 \{ \mathbb{E}_{z_1, \dots, z_T}(\varepsilon^{c_0}(f_{T+1}) - \varepsilon^{c_0}(f_\rho^{c_0})) \}^{\frac{1}{2}}. \tag{4.1}$$

Now we estimate $\varepsilon^{c_0}(f_{T+1}) - \varepsilon^{c_0}(f_\rho^{c_0})$. It can be displayed as

$$\varepsilon^{c_0}(f_{T+1}) - \varepsilon^{c_0}(f_\rho^{c_0}) = \varepsilon^{c_0}(f_{T+1}) - \varepsilon^{c_0}(f_{\lambda_T}^{c_T}) + \varepsilon^{c_0}(f_{\lambda_T}^{c_T}) - \varepsilon^{c_0}(f_{\lambda_T}^{c_0}) + \varepsilon^{c_0}(f_{\lambda_T}^{c_0}) - \varepsilon^{c_0}(f_\rho^{c_0}).$$

First, due to $|\partial V^{c_0}(\cdot)| \leq 1$ and $\|f\|_{C(X)} \leq \kappa \|f\|_K$, we have

$$\left| \varepsilon^{c_0}(f_{T+1}) - \varepsilon^{c_0}(f_{\lambda_T}^{c_T}) \right| = \left| \int_Z V^{c_0}(yf_{T+1}(x)) - V^{c_0}(yf_{\lambda_T}^{c_T}(x)) d\rho \right| \leq \kappa \|f_{T+1} - f_{\lambda_T}^{c_T}\|_K. \tag{4.2}$$

Next, with (3.3), the second term is

$$\left| \varepsilon^{c_0}(f_{\lambda_T}^{c_T}) - \varepsilon^{c_0}(f_{\lambda_T}^{c_0}) \right| \leq \kappa \|f_{\lambda_T}^{c_T} - f_{\lambda_T}^{c_0}\|_K \leq \kappa \cdot \frac{6\kappa}{\lambda_T} \cdot c_T = \frac{6\kappa^2 c_1}{\lambda_1} T^{\gamma-\theta}. \tag{4.3}$$

For the third term, from the definition (2.10), we have

$$|\varepsilon^{c_0}(f_{\lambda_T}^{c_0}) - \varepsilon^{c_0}(f_{\rho}^{c_0})| \leq \mathcal{D}^{c_0}(\lambda_T) \leq \mathcal{D}_0 \lambda_1^\beta T^{-\beta\gamma}. \quad (4.4)$$

Now our target is to estimate $\|f_{T+1} - f_{\lambda_T}^{c_T}\|_K$, we tackle it by estimating one-step iteration. First of all, we give the following lemma.

Lemma 4.21 Define $\{f_t\}$ by (2.13), we get

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

where Δ_t is defined by

$$\Delta_t = \int_Z \{V^{c_t}(y f_{\lambda_t}^{c_t}(x)) - V^{c_t}(y f_t(x))\} d[\rho^{(t)} - \rho].$$

The proof of the lemma is shown in [6].

At the same time,

$$\begin{aligned} \|f_t - f_{\lambda_t}^{c_t}\|_K &= \|f_t - f_{\lambda_{t-1}}^{c_{t-1}} + f_{\lambda_{t-1}}^{c_{t-1}} - f_{\lambda_t}^{c_{t-1}} + f_{\lambda_t}^{c_{t-1}} - f_{\lambda_t}^{c_t}\|_K \\ &\leq \|f_t - f_{\lambda_{t-1}}^{c_{t-1}}\|_K + g_t + h_t, \end{aligned}$$

where

$$g_t = \|f_{\lambda_t}^{c_{t-1}} - f_{\lambda_{t-1}}^{c_{t-1}}\|_K, \quad h_t = \|f_{\lambda_t}^{c_t} - f_{\lambda_t}^{c_{t-1}}\|_K.$$

Apply the elementary inequality $2xy \leq Ax^2 y^\tau + y^{2-\tau}/A$ with $0 < \tau_1 < 2$, $0 < \tau_2 < 2$, $A_1 > 0$ and $A_2 > 0$ to $x = \|f_t - f_{\lambda_{t-1}}^{c_{t-1}}\|_K$, $y = g_t$ and $A = A_1$ or $y = h_t$ and $A = A_2$. Then,

$$\begin{aligned} \|f_t - f_{\lambda_t}^{c_t}\|_K^2 &\leq (1 + A_1 g_t^{\tau_1} + A_2 h_t^{\tau_2}) \|f_t - f_{\lambda_{t-1}}^{c_{t-1}}\|_K^2 \\ &\quad + g_t^{2-\tau_1}/A_1 + h_t^{2-\tau_2}/A_2 + 2g_t^2 + 2h_t^2. \end{aligned}$$

Due to $(1 + A_1 g_t^{\tau_1} + A_2 h_t^{\tau_2})(1 - \eta_t \lambda_t) \leq (1 + A_1 g_t^{\tau_1} + A_2 h_t^{\tau_2} - \eta_t \lambda_t)$, we have

$$\begin{aligned} \mathbb{E}_{z_t}(\|f_{t+1} - f_{\lambda_t}^{c_t}\|_K^2) &\leq (1 + A_1 g_t^{\tau_1} + A_2 h_t^{\tau_2} - \eta_t \lambda_t) \|f_t - f_{\lambda_{t-1}}^{c_{t-1}}\|_K^2 + g_t^{2-\tau_1}/A_1 + h_t^{2-\tau_2}/A_2 \\ &\quad + 2g_t^2 + 2h_t^2 + 2\eta_t \Delta_t + \eta_t^2 \mathbb{E}_{z_t} \|\partial V^{c_t}(y_t f_t(x_t)) K_{x_t} + \lambda_t f_t\|_K^2. \end{aligned}$$

Now we estimate these items separately:

With Lemma 3.18, we have

$$\Delta_t \leq \{(\|f_t\|_{C^s(X)} + \|f_{\lambda_t}^{c_t}\|_{C^s(X)}) + 2C_\rho \tilde{B}_{f_t, f_{\lambda_t}^{c_t}}\} \left\| \rho_X^{(t)} - \rho_X \right\|_{(C^s(X))^*}.$$

Due to (2.6), (2.8), (2.14), (3.2) and $V^c(\cdot)$ with incremental exponent $p = 0$, we can show $\tilde{B}_{f_t, f_{\lambda_t}^{c_t}} \leq \frac{N_2}{\lambda_t}$ holds true for a constant $N_2 > N_1$ such that we have

$$\Delta_t \leq A_3 t^{\gamma-b},$$

where $A_3 = \{(\kappa + \kappa_{2s})(\frac{\kappa}{\lambda_1} + \sqrt{\frac{2D_0\lambda_1^\beta + 4c_1}{\lambda_1}}) + \frac{2C_\rho N_2}{\lambda_1}\}C$.

From drift error $d_t \leq 2(t-1)^{\frac{\gamma}{2}-1} \sqrt{\frac{D_0\lambda_1^\beta(t-1)^{-\beta\gamma} + 2c_{t-1}}{\lambda_1}}$, we have

$$g_t \leq A_4 t^{-\zeta},$$

where $A_4 = 2\sqrt{\frac{D_0\lambda_1^\beta + 2c_1}{\lambda_1}}$ and $\zeta = \min\{1 - \frac{\gamma(1-\beta)}{2}, 1 - \frac{\gamma-\theta}{2}\} = 1 - \frac{\gamma(1-\beta)}{2}$ because of $\theta > \beta\gamma$ in the condition of Theorem 2.13.

Apply Lemma 3.17 to $\nu = c_t = c_1 t^{-\theta}$ and $\mu = c_{t-1} = c_1(t-1)^{-\theta}$, due to $(t-1)^{-\theta} - t^{-\theta} \leq \theta(t-1)^{-\theta-1} \leq \theta 2^{\theta+1} t^{-\theta-1}$, we have

$$h_t \leq A_5 t^{-(1+\theta-\gamma)},$$

where $A_5 = \frac{12\kappa c_1 \theta 2^\theta}{\lambda_1}$.

Then we assume $\tau_1 = \frac{\alpha+\gamma}{1-\frac{\gamma(1-\beta)}{2}}$, $\tau_2 = \frac{\alpha+\gamma}{1+\theta-\gamma}$, $A_1 = \frac{\lambda_1 \eta_1}{4A_4^{\tau_1}}$ and $A_2 = \frac{\lambda_1 \eta_1}{4A_5^{\tau_2}}$, the condition (2.16) can assure $0 < \tau_1 < 1$ and $0 < \tau_2 < 1$. Therefore, we have

$$1 + A_1 g_t^{\tau_1} + A_2 h_t^{\tau_2} - \eta_t \lambda_t \leq 1 - \frac{\lambda_1 \eta_1}{2} t^{-(\alpha+\gamma)}.$$

The last term $\eta_t^2 \mathbb{E}_{z_t} \|\partial V^{c_t}(y_t f_t(x_t))K_{x_t} + \lambda_t f_t\|_K^2$ can be bound by

$$\begin{aligned} \|\partial V^{c_t}(y_t f_t(x_t))K_{x_t} + \lambda_t f_t\|_K^2 &\leq (\|\partial V^{c_t}(y_t f_t(x_t))K_{x_t}\|_K + \lambda_t \|f_t\|_K)^2 \\ &\leq (\kappa + \lambda_t \frac{\kappa}{\lambda_t})^2 = 4\kappa^2 \end{aligned}$$

Hence,

$$\eta_t^2 \mathbb{E}_{z_t} \|\partial V^{c_t}(y_t f_t(x_t))K_{x_t} + \lambda_t f_t\|_K^2 \leq 4\kappa^2 \eta_1^2 t^{-2\alpha}.$$

In summary, due to the independency of z_1, \dots, z_t , we get the one-step iteration as follows:

$$\mathbb{E}_{z_1, \dots, z_t} \|f_{t+1} - f_{\lambda_t}^{c_t}\|_K^2 \leq (1 - \frac{\lambda_1 \eta_1}{2} t^{-(\alpha+\gamma)}) \mathbb{E}_{z_1, \dots, z_{t-1}} (\|f_t - f_{\lambda_{t-1}}^{c_{t-1}}\|_K^2) + A_6 t^{-\varpi}, \quad (4.5)$$

Where

$$A_6 = \frac{A_4^{2-\tau_1}}{A_1} + \frac{A_5^{2-\tau_2}}{A_2} + 2A_4^2 + 2A_5^2 + 2\eta_1 A_3 + 4\kappa^2 \eta_1^2,$$

and

$$\varpi = \min\{2 - \gamma(2 - \beta) - \alpha, 2 - 3\gamma + 2\theta - \alpha, 2\alpha, 2(1 + \theta - \gamma), \alpha + b - \gamma\}.$$

Applying (4.5) iteratively for $t = 2, \dots, T$ implies

$$\begin{aligned} \mathbb{E}_{z_1, \dots, z_T} (\|f_{T+1} - f_{\lambda_T}^{c_T}\|_K^2) &\leq A_6 \sum_{t=2}^T \prod_{j=t+1}^T (1 - \frac{\lambda_1 \eta_1}{2} j^{-\alpha-\gamma}) t^{-\varpi} \\ &\quad + \{\prod_{t=2}^T (1 - \frac{\lambda_1 \eta_1}{2} t^{-\alpha-\gamma})\} \|f_2 - f_{\lambda_1}^{c_1}\|_K^2. \end{aligned}$$

For the first term, we apply the inequality (2.12) with $c = \frac{\eta_1}{2}$, $q_1 = \alpha + \gamma$, $q_2 = \varpi$ and $1 - u \leq e^{-u}$, we have

$$\sum_{t=2}^T \prod_{j=t+1}^T (1 - \frac{\lambda_1 \eta_1}{2} j^{-\alpha-\gamma}) t^{-\varpi} \leq \sum_{t=2}^T \exp\{-\frac{\lambda_1 \eta_1}{2} \prod_{j=t+1}^T j^{-\alpha-\gamma}\} t^{-\varpi} \leq A_7 T^{\alpha+\gamma-\varpi},$$

where $A_7 = \frac{2^{\alpha+\gamma+\varpi+1}}{\lambda_1 \eta_1} + 1 + (\frac{2+2\varpi}{e\lambda_1 \eta_1 (1-2^{\alpha+\gamma-1})})^{\frac{1+\varpi}{1-\alpha-\gamma}}$. As for the second term, due to

$$\begin{aligned} \prod_{t=2}^T (1 - \frac{\lambda_1 \eta_1}{2} t^{-\alpha-\gamma}) &\leq \exp\{-\frac{\lambda_1 \eta_1}{2} \prod_{t=2}^T t^{-\alpha-\gamma}\} \\ &\leq \exp\{-\frac{\lambda_1 \eta_1}{2} \int_2^{T+1} x^{-\alpha-\gamma} dx\} \\ &\leq \exp\{\frac{\lambda_1 \eta_1}{2(1-\alpha-\gamma)} 2^{1-\alpha-\gamma}\} \exp\{\frac{-\lambda_1 \eta_1}{2(1-\alpha-\gamma)} (T+1)^{1-\alpha-\gamma}\}. \end{aligned}$$

Apply the elementary inequality $\exp\{-cx\} \leq (\frac{v}{ec})^v x^{-v}$ with $c = \frac{\lambda_1 \eta_1}{2(1-\alpha-\gamma)}$, $v = \frac{2}{1-\alpha-\gamma}$ and $x = (T+1)^{1-\alpha-\gamma}$, we see that

$$\prod_{t=2}^T (1 - \frac{\lambda_1 \eta_1}{2} t^{-\alpha-\gamma}) \leq \exp\{\frac{\lambda_1 \eta_1}{2(1-\alpha-\gamma)} 2^{1-\alpha-\gamma}\} (\frac{4}{e\lambda_1 \eta_1})^{\frac{2}{1-\alpha-\gamma}} T^{-2}.$$

Combine the estimate of two term, it shows that

$$\mathbb{E}_{z_1, \dots, z_T} (\|f_{T+1} - f_{\lambda_T}^{c_T}\|_K^2) \leq A^* T^{-\omega}, \tag{4.6}$$

where $A^* = A_6 A_7 + \exp\{\frac{\lambda_1 \eta_1}{2(1-\alpha-\gamma)} 2^{1-\alpha-\gamma}\} (\frac{4}{e\lambda_1 \eta_1})^{\frac{2}{1-\alpha-\gamma}} \|f_2 - f_{\lambda_1}^{c_1}\|_K^2$ and

$$\begin{aligned} \omega &= \varpi - \alpha - \gamma \\ &= \min\{2 - \gamma(3 - \beta) - 2\alpha, 2 - 4\gamma + 2\theta - 2\alpha, \alpha - \gamma, 2 + 2\theta - \alpha - 3\gamma, b - 2\gamma\}. \end{aligned}$$

According to the above, we can know the bounds of $\varepsilon^{c_0}(f_{T+1}) - \varepsilon^{c_0}(f_{\rho}^{c_0})$ with (4.2), (4.3), (4.4) and (4.6). Combine it with (4.1), the main result can be expressed as

$$\mathbb{E}_{z_1, \dots, z_T} (\mathcal{R}(\text{sgn}(f_{T+1})) - \mathcal{R}(f_{\text{bayes}})) \leq C^* T^{-\omega^*},$$

where ω^* has the form in Theorem 2.13 and

$$C^* = C_2 \sqrt{\kappa \sqrt{A^*} + \frac{6\kappa^2 c_1}{\lambda_1} + \mathcal{D}_0 \lambda_1^\beta}.$$

The proof of the main result has been completed.

5 Simulation

We further demonstrate our theory by an illustrative example. Let ρ_X be the Lebesgue measure on $[-5, 5]$, then the marginal distribution sequence $\{\rho_X^{(t)}\}$ satisfies $d\rho_X^{(t)} = d\rho_X +$

$Ct^{-b}d\rho_X$ where $C = 1$ and $b = 2$. We assume that ρ_X is uniform distribution on $[-5, 5]$ and for each t , a sample x_t is drawn independently from the different distributions $\left\{ \rho_X^{(t)} \right\}$. According to x_t , label $y_t \in \{-1, 1\}$ is produced by the expectation of conditional distribution ρ_x :

$$f_\rho(x) = \sum_{i=1}^3 k_i \exp\left(-\frac{(x - p_i)^2}{2v_i^2}\right),$$

where the parameters are described as: $k_1, k_2, k_3 = 2.1, 3.3, -4.4$, $p_1, p_2, p_3 = 0, 0.1, 0.01$ and $v_1, v_2, v_3 = 0.6, 0.61, 0.62$. Hence, we get the sample (x_t, y_t) which are non-i.i.d.. In this simulation, we randomly draw 4000 samples which includes 1000 train data and 3000 test data. Especially, we take the Gaussian kernel $K(x, u) = \exp\left(-\frac{(x - u)^2}{2\sigma^2}\right)$ with variance $\sigma^2 = 0.6^2$.

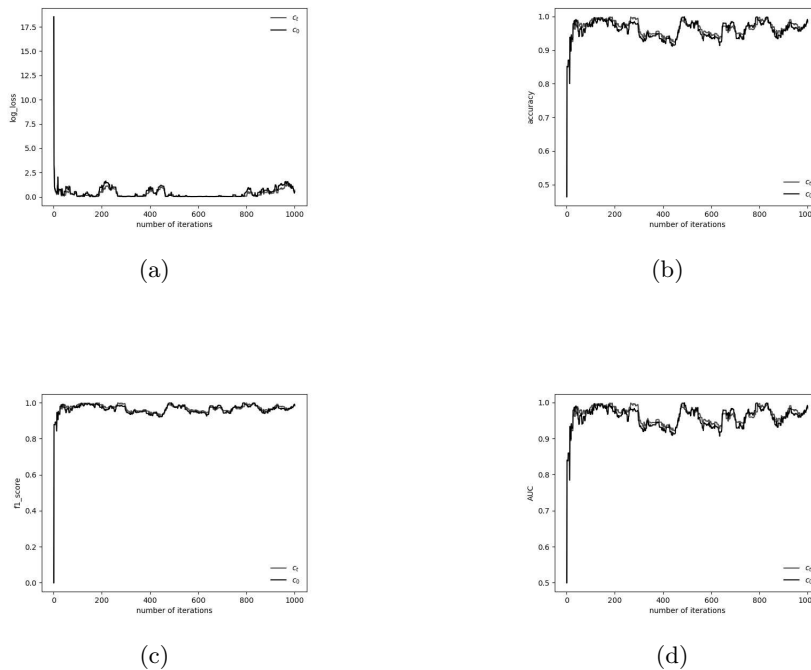


Figure 1 (a) Log-loss function on test samples as number of iterations varies. (b) Accuracy on test samples as number of iterations varies. (c) F1-score on test samples as number of iterations varies. (d) Auc-score on test samples as number of iterations varies.

By online learning algorithm (2.13) where $\lambda_1 = 0.01, \gamma = 0.04, \eta_1 = 0.4, \alpha = 0.1, c_1 = 5, \theta = 0.4$, we feed 1000 train samples to train the model and evaluate f_T on the test samples. The results of log-loss , accuracy, f1-score and auc-score are used to evaluate and the trends

of the four indicators with iteration are showed on Figure 1. From the figures, we apply two models to evaluate four indicators on the test data. The one of them is produced by online learning algorithm with LUM loss function parameter $c_0 = 0$ and another one is the c_t that changes with iteration mentioned in this paper. It is easy to show that the red line performs a little better than the black line in total four figures. Hence, we can see the online learning algorithm with LUMs loss parameter c_t is valid and similar to the convergence rate of c_0 with iteration. Furthermore, we can conclude that the algorithm introduced in this paper which uses a parameter of loss function that gradually approximates to c_0 behaves better than the normal loss function with c_0 . Thus, this strategy is valuable.

6 Conclusions

LUMs combine soft classification and hard classification because of the varying parameter c and LUM loss functions with $c = 0$ are the best case to estimate the class conditional probability. Hence, we aim to investigate parameter c changing from a finite number to 0 at each iteration in the online learning algorithm. The setting of the online learning algorithm is that sample data are drawn independently from a non-identical distribution with iteration.

In this paper, the general convergence analysis of online LUM classification in this setting has been shown and we give the numerical learning rate for the framework in main result. The limit of convergence rate can be provided by $O(T^{-\frac{1}{4} + \frac{2}{5}})$ with some conditions. In addition, we reveal that this algorithm converges on the actual non-i.i.d. data set and performs better than normal unchanged LUM function with parameter $c = 0$ from the simulation. This strategy of adjusting the loss function as the algorithm iterates can be introduced into other related research.

Appendix

Now we prove $\|V^c - V^{c_0}\|_\infty \leq c$.

Proof Due to the definition of $V^c(\cdot)$, when $a = 1$, it implies

$$V^c(u) = \begin{cases} 1 - u & \text{if } u \leq \frac{c}{1+c}, \\ \frac{1}{c+1} \left(\frac{1}{(1+c)u - c + 1} \right) & \text{if } u > \frac{c}{1+c}. \end{cases}$$

and

$$V^{c_0}(u) = \begin{cases} 1 - u & \text{if } u \leq 0, \\ \frac{1}{1+u} & \text{if } u > 0. \end{cases}$$

We denote $V^c - V^{c_0}$ as Z .

(1) When $u \leq 0$, we have $|Z| = 0 \leq c$.

(2) When $0 < u \leq \frac{c}{1+c}$, apply the elementary inequality $x + \frac{1}{x} \geq 2$, we get $u + 1 + u + \frac{1}{1+u} \geq 2$.

It follows that

$$Z = 1 - u - \frac{1}{1+u} < u.$$

Besides, $Z > -u$ holds. Then

$$|Z| \leq u \leq \frac{c}{1+c} \leq c.$$

(3) When $\frac{c}{1+c} < u \leq 1$, $|Z| = \left| \frac{1}{1+u} - \frac{1}{c+1} \left(\frac{1}{(1+c)u-c+1} \right) \right|$.

On the one hand, we have

$$\begin{aligned} \frac{1}{1+u} - \frac{1}{c+1} \left(\frac{1}{(1+c)u-c+1} \right) - c &= \frac{\{[1-c(1+u)][(1+c)u-c+1] - (1+u)\}}{(1+u)(1+c)[(1+c)u-c+1]} \\ &< \frac{(1+c)u-c+1 - (1+u)}{(1+u)(1+c)[(1+c)u-c+1]} \\ &= \frac{(1+u)c - 2c}{(1+u)(1+c)[(1+c)u-c+1]} \leq 0. \end{aligned}$$

On the other hand,

$$\begin{aligned} \frac{1}{1+u} - \frac{1}{c+1} \left(\frac{1}{(1+c)u-c+1} \right) + c &\geq \frac{1}{1+u} - \frac{1}{(1+c)u-c+1} - c \\ &= \frac{uc + c\{(1+u)[(1+c)u-c+1] - 1\}}{(1+u)[(1+c)u-c+1]} \geq 0. \end{aligned}$$

Hence, $|Z| \leq c$ holds true.

(4) When $u > 1$, we denote $l(x) = \frac{1}{1+x} \left(\frac{1}{(1+x)u-x+1} \right)$. It follows that

$$|Z| = |l(0) - l(c)| = |l'(\xi)| \cdot |c|, \quad \exists \xi \in (0, c).$$

We can see $|l'(\xi)| \leq 1$ with $u > 1$ and it implies that $|Z| \leq c$.

The above four cases show $\|V^c - V^{c_0}\|_\infty \leq c$.

References

- [1] Wahba G. Soft and hard classification by reproducing kernel Hilbert space methods[J]. Proceedings of the National Academy of Sciences, 2002, 99(26): 16524–16530.
- [2] Liu Y, Zhang H H, Wu Y. Hard or soft classification? large-margin unified machines[J]. Journal of the American Statistical Association, 2011, 106(493): 166–177.
- [3] Hu T, Xiang D H, Zhou D X. Online learning for quantile regression and support vector regression[J]. Journal of Statistical Planning and Inference, 2012, 142(12): 3107–3122.
- [4] Fan J, Xiang D H. Quantitative convergence analysis of kernel based large-margin unified machines[J]. Communications on Pure & Applied Analysis, 2020, 19(8): 4069.
- [5] Hu T, Zhou D X. Online learning with samples drawn from non-identical distributions[J]. Journal of Machine Learning Research, 2009, 10(12): 2873–2898.
- [6] Hu T, Yao Y. Online regression with varying Gaussians and non-identical distributions[J]. Analysis and Applications, 2011, 9(04): 395–408.
- [7] Guo Q, Ye P. Error analysis of the moving least-squares method with non-identical sampling[J]. International Journal of Computer Mathematics, 2019, 96(4): 767–781.

- [8] Kiefer J, Wolfowitz J. Stochastic estimation of the maximum of a regression function[J]. The Annals of Mathematical Statistics, 1952, 23(3): 462–466.
- [9] Smale S, Yao Y. Online learning algorithms[J]. Foundations of computational mathematics, 2006, 6(2): 145–170.
- [10] Ye G B, Zhou D X. Fully online classification by regularization[J]. Applied and Computational Harmonic Analysis, 2007, 23(2): 198–214.
- [11] Kivinen J, Smola A J, Williamson R C. Online learning with kernels[J]. IEEE transactions on signal processing, 2004, 52(8): 2165–2176.
- [12] Smale S, Zhou D X. Online learning with Markov sampling[J]. Analysis and Applications, 2009, 7(01): 87–113.
- [13] Benabid A, Fan J, Xiang D H. Comparison theorems on large-margin learning[J]. International Journal of Wavelets, Multiresolution and Information Processing, 2021, 19(05): 2150015.

基于非同分布样本的变参LUM在线分类

王泽兴

(武汉大学数学与统计学院, 湖北 武汉 430072)

摘要: LUMs(Large-margin Unified Machines)在分类学习中受到广泛关注, LUMs是一类最大化间隔分类器, 它提供了一种独特的软分类到硬分类转化的方式. 本文研究的是基于独立不同分布样本和LUM损失函数的二分类在线学习算法. 同时, 在线算法的每一步迭代, 涉及的LUM损失函数的参数是随着迭代在逐渐减小的. 在这种假设下, 我们基于再生核希尔伯特空间(RKHS), 给出了在线算法的收敛阶.

关键词: 非同分布样本; 在线分类算法; 变参数LUM损失函数; 再生核希尔伯特空间

MR(2010)主题分类号: 62J99 中图分类号: O29