International Journal of Machine Learning and Cybernetics (2021) 12:1867–1877
https://doi.org/10.1007/s13042-021-01281-0

ORIGINAL ARTICLE

Universal consistency of twin support vector machines

Weixia Xu[1] · Dingjiang Huang[2] · Shuigeng Zhou[3]

Received: 7 November 2020 / Accepted: 30 January 2021 / Published online: 14 March 2021
© The Author(s) 2021

Abstract
A classification problem aims at constructing a best classifier with the smallest risk. When the sample size approaches infinity, the learning algorithms for a classification problem are characterized by an asymptotical property, i.e., universal consistency. It plays a crucial role in measuring the construction of classification rules. A universal consistent algorithm ensures that the larger the sample size of the algorithm is, the more accurately the distribution of the samples could be reconstructed. Support vector machines (SVMs) are regarded as one of the most important models in binary classification problems. How to effectively extend SVMs to twin support vector machines (TWSVMs) so as to improve performance of classification has gained increasing interest in many research areas recently. Many variants for TWSVMs have been proposed and used in practice. Thus in this paper, we focus on the universal consistency of TWSVMs in a binary classification setting. We first give a general framework for TWSVM classifiers that unifies most of the variants of TWSVMs for binary classification problems. Based on it, we then investigate the universal consistency of TWSVMs. To do this, we give some useful definitions of risk, Bayes risk and universal consistency for TWSVMs. Theoretical results indicate that universal consistency is valid for various TWSVM classifiers under some certain conditions, including covering number, localized covering number and stability. For applications of our general framework, several variants of TWSVMs are considered.

Keywords Binary classification · Twin support vector machine (TWSVM) · Bayes risk · Universal consistency · Regularization

1 Introduction

As sample size increases gradually to infinity, there is an asymptotical property for learning algorithms, called *consistency*. It is an extremely important part in statistical learning theory. In fact, though the sample size is always finite for practical problems, the *consistency* of learning algorithms guarantees that by using more samples, a more accurate distribution could be reconstructed. Since the concept of *consistency* was first proposed by Vapnik and Chervonenkis [27–29], the *consistency* of various learning algorithms has been extensively explored in statistical learning and machine learning areas.

According to different settings for learning machines, the consistency could be summarized as the following types. The consistency of empirical risk minimization (ERM) method [29] is a classical type of consistency. The loss function minimizing empirical risk is used to approximate the loss function minimizing true risk. For example, Chen et al. [5] studied the consistency of ERM method based on convex losses of multi-class classification problems. Brownlees et al. [4] investigated the performance bound of heavy-tailed losses from the view of consistency of ERM method. Berner et al. [1] analyzed the generalization error of ERM method based on deep artificial neural network hypothesis. Xu et al. [30] proposed the general framework for statistical learning with group invariance, and paid attention to

✉ Shuigeng Zhou
sgzhou@fudan.edu.cn

Weixia Xu
20190092@lixin.edu.cn

Dingjiang Huang
djhuang@dase.ecnu.edu.cn

[1] School of Information Management, Shanghai Lixin University of Accounting and Finance, Shanghai 201209, China

[2] School of Data Science and Engineering, East China Normal University, Shanghai 200062, China

[3] School of Computer Science and Shanghai Key Lab of Intelligent Information Processing, Fudan University, Shanghai 200433, China

the consistency of the general framework. Fisher consistency [10] strengthens the property of unbiasedness for parameters of functions. It estimates the parameters of functions directly, and uses the estimated values of parameters to approximate their true values. For instance, Liu [17] established the Fisher consistency theory for different loss functions of multi-category support vector machine (SVM) algorithms. Fathony et al. [8] proposed an adversarial bipartite matching algorithm with the computational efficiency and Fisher consistency properties.

In addition to the two types, universal consistency is also a typical type of consistency, which measures the consistency of learning algorithms with the base of structural risk minimization (SRM) method. Indeed, the ERM method is only about empirical risk, and is not about regularization term. However in many practical situations, in order to control the generalization ability of learning machines, the regularization term is always considered. To this end, Vapnik [26] proposed the SRM method to balance the empirical risk of training data and generalization ability of learning machines. Later, the concept of universal consistency was introduced to demonstrate the consistency of many learning algorithms based on SRM method. For example, Steinwart [25] showed the universal consistency for SVMs and their different variants on a unified framework. Liu et al. [16] indicated that the extreme learning machine (ELM) was universally consistent for radial basis function networks, and pointed out the direction to select the optimal kernel functions in ELM application. Dumpert and Christmann [7] concerned the universal consistency of localized kernel based methods. Gyorfi et al. [12] shared the universal consistent results for the nearest-neighbor prototype algorithm in the multi-class classification setting, and the convergence rate was also conducted based on the universal consistency. In summary, universal consistency has been studied deeply in many problem settings. Here, the present paper focuses on universal consistency for binary classification problems.

SVMs are a type of powerful tools for binary classification problems. The key idea is to construct two parallel hyper-planes such that the positive and negative classes are separated well, and then maximize the margin between the two parallel hyper-planes, resulting in the minimization of the regularization term. SVMs were widely applied to many practical problems, such as text classification [15], face recognition [23] and bioinformatics [9] etc. Though successful in these applications, SVMs still have some difficulties, since they deal only with small sample problems. It would take very expensive computational cost for large scale sample problems.

In order to reduce the computational cost of SVMs, many extensions to SVM were proposed and studied. For instance, a generalized eigenvalue proximal support vector machine (GEPSVM) [18] was such an extension, which constructs two non-parallel hyper-planes, so that each hyper-plane is closest to one of the two classes and is also as far away from the other class as possible. Based on SVMs and GEPSVM, Jayadeva et al. [13] established another extension to SVMs, that is, twin support vector machine (TWSVM). The main idea of TWSVM is similar to that of GEPSVM, while the formulation is entirely different from that of GEPSVM. In fact, it derives a pair of quadratic programming problems (QPPs) for TWSVM, and the formulation of each QPP is similar to that of SVMs, except that only one class of the training samples appears in the constraints of each QPP. In brief, the computational cost of TWSVM is only one fourth of that of SVMs.

Similar to SVMs, TWSVM also has many variants for binary classification problem. For instance, smooth TWSVM algorithm [14] approximated a smoothing function $\rho(x, \eta)$ to the plus function $(x)_+$ in the process of solving the QPPs, such that the learnt classifier was more smoother than before. Twin bounded SVM (TBSVM) algorithm [21] embedded a regularization term to TWSVM, according to the SRM principle, and thus improved the classification performance. Weighted linear loss TWSVM algorithm [22] was constructed to adjust the impact of each point on the hyperplane, and thus the weights for the slack variables were given. Least squares TBSVM algorithm based on L1-norm distance metric [32] was a least square version of TBSVM firstly, and then substituted L1-norm for L2-norm to enhance the robustness. Besides, there are many other algorithms based on the same idea as TWSVM, like robust TWSVM algorithm [19], margin-based TWSVM with unity norm hyperplanes [20], fuzzy TWSVM algorithm [11] etc. Though these TWSVM variants are formulated well and applied to different practical problems, up to now their universal consistency has not been studied.

In this study, we address the universal consistency of TWSVM and its variants. Since it is very cumbersome to analyze the universal consistency of the TWSVM variants one by one, we suggest to do this work under a general framework. However, there is no unified framework for all TWSVM variants in the literature. So we first try to construct a general framework for TWSVM variants. Furthermore, as not all the variants are based on the same idea, it is very difficult to find a general framework fit to all variants, we determine to construct a general framework for most of the variants based on the idea of TWSVM, and formulate it as a general optimization problem. Concretely, the optimization problem consists of two minimization problems, each of which contains two terms: the first term measures the average of losses for both the positive and negative data, and the second term is expressed by a regularization term for maximizing some margin.

With the general framework, we then study the universal consistency of the general optimization problem. We first introduce the definition of universal consistency for the general optimization problem, and then show in what conditions, the universal consistency is valid. When introducing the definition of universal consistency, risk $R_P(f)$ and Bayes risk R_P for the general optimization problem are related, and thus are redefined here. When showing the validity of universal consistency in some conditions, an assertion is necessary and thus is proposed. Since the definitions like regularized L_1-risk $R^{reg}_{1,P,c_1}(f_1)$ and regularized L_2-risk $R^{reg}_{2,P,c_2}(f_2)$ are very important to describe the assertion, the detailed definitions are given before the assertion. The assertion is then described centered on a pair of concentration inequalities. Under three different conditions based on covering number, localized covering number and stability respectively, it derives three different pairs of concentration inequalities, and thus it concludes three different results for universal consistency.

The rest of the paper is organized as follows: Section 2 gives the preliminaries. Section 3 derives the general framework for most variants of TWSVM and formulates it as a general optimization problem. Section 4 introduces the definitions of risk, Bayes risk and universal consistency for this optimization problem and proposes an assertion to theoretically support the universal consistency. Section 5 presents the theoretical results about the assertion. Finally, Sect. 6 concludes this paper.

2 Preliminaries

Here, we give some notations and concepts that would be used in the following sections. Denote $\mathbb{R} = (-\infty, +\infty)$, $\mathbb{R}^+ = [0, +\infty)$. Suppose X is a compact metric space, and $k : X \times X \mapsto \mathbb{R}$ is a positive semi-definite kernel. We define a quantity K

$$K = \sup\left\{\sqrt{k(x,x)}, x \in X\right\}.$$

Let H be the reproducing kernel Hilbert space (RKHS) with respect to kernel k. Reminder that there is a mapping $\Phi : X \mapsto H$ satisfying the reproducing property, that is,

$$k(x_1,x_2) =< \Phi(x_1), \Phi(x_2) >, \quad x_1, x_2 \in X,$$

Suppose the kernel k is continuous, then the element of H is also continuous on X. In this situation, there is a mapping $I : H \mapsto \mathcal{C}(X)$, which continuously embeds the RKHS H into the space of all continuous functions $\mathcal{C}(X)$

$$I_f =< f, \Phi(x) >_H, \quad f \in H,$$

We say k is a universal kernel, if the mapping I is dense.

Let $\Omega : \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}^+$ be an non-descending function denoted by $\Omega(c,t)$. This function is continuous in 0 with respect to the variable c, and is unbounded when the variable t tends to infinity. We introduce the following definition [25] to explain in which case, Ω is a regularization function.

Definition 1 Given a function $\Omega(c,t)$, assume there exists $t > 0$ satisfying the inequality $\Omega(c,t) < \infty$ for any $c > 0$. Then, Ω is a regularization function, if for all $c > 0$, $s,t \in \mathbb{R}^+$, and for all sequences $(t_n) \subset \mathbb{R}^+$ with $t_n \to t$ and $\Omega(c,t_n) < \infty$, we have $\Omega(c,0) = \Omega(0,s)$ and $\Omega(c,t_n) \to \Omega(c,t)$.

For any given loss function $L : Y \times \mathbb{R} \times \mathbb{R}^+ \to \mathbb{R}^+$, denote

$$C(\alpha,t,\lambda) = \alpha L(1,t,\lambda) + (1-\alpha)L(-1,t,\lambda),$$

where $\alpha \in [0,1], t \in \mathbb{R}, \lambda \in \mathbb{R}^+$. Let

$$t_\alpha = \arg\min_{t\in\mathbb{R}} C(\alpha,t,\lambda),$$
$$M(\alpha,\lambda) = C(\alpha,t_\alpha,\lambda).$$

Then L is an admissible loss function [25], redefined as follows:

Definition 2 Given a continuous function $L : Y \times \mathbb{R} \times \mathbb{R}^+ \to \mathbb{R}^+$. Then L is an admissible loss function, if for any $\alpha \in [0,1/2)$ we have $t_\alpha < 0$, and if for any $\alpha \in (1/2,1]$ we have $t_\alpha > 0$.

3 A general framework for twin support vector machines

Denote by $X \subseteq \mathbb{R}^d$ the input space for instances and by $Y = \{-1,1\}$ the output space for labels. Let S be the training set belonging to the space $X \times Y$ with m samples $(x_i,y_i), i = 1,\dots,m$. Suppose S consists of m_1 positive samples relabeled by $(x^1_i,1), i = 1,\dots,m_1$, and m_2 negative samples relabeled by $(x^2_j,-1), j = 1,\dots,m_2$ with $m = m_1 + m_2$. Assume the data of the training set S are sampled from an unknown distribution P on $X \times Y$, and they are independent and identical distributed (*i.i.d.*). Denote matrices A, B and D as follows

$$A^T = (x^1_1,\dots,x^1_{m_1}),\ B^T = (x^2_1,\dots,x^2_{m_2}),\ D^T = (A^T,B^T).$$

We give a general framework to cover most of the variants of TWSVM, which is formulated as follows:

$$\begin{aligned}\min_{f_1\in H} \sum_{i=1}^{m_1} \ell_1(1,f_1(x^1_i)) + \sum_{j=1}^{m_2} \ell_2(-1,f_1(x^2_j),\lambda_1)\\ + \Omega^*(c_1,||f_1||_H),\end{aligned} \qquad (1)$$

$$\begin{aligned}\min_{f_2 \in H} \sum_{j=1}^{m_2} \ell_1(-1, f_2(x_j^2)) + \sum_{i=1}^{m_1} \ell_2(1, f_2(x_i^1), \lambda_2) \\ + \Omega^*(c_2, ||f_2||_H),\end{aligned} \tag{2}$$

where c_1, c_2, λ_1, λ_2 are all trade-off parameters, f_1, f_2 are the hyper-planes corresponding to the positive and negative classes, respectively, and H is RKHS. Here, ℓ_1 is one loss function measuring the squared distance from the data of one class to the corresponding hyper-plane. ℓ_2 is another loss function measuring the slack variable such that the distance from the data of the other class to the same hyper-plane is no smaller than 1. Ω^* is a regularization term for maximizing some margin.

Note that in the minimization problem Eq. (1), ℓ_1 measures the loss for one positive sample and ℓ_2 measures the loss for one negative sample. The sum of the first two terms is the total loss for all the positive and negative samples. Now we want to combine the two loss functions into one revised loss function such that it could measure the loss for any positive or negative sample about hyper-plane f_1. Given a function $p(y)$

$$p(y) = \begin{cases} 1, & y = 1, \\ 0, & y = -1, \end{cases}$$

the revised loss function L_1 can be defined as

$$\begin{aligned} L_1(y, f_1(x), \lambda_1) &= p(y)\ell_1(y, f_1(x)) \\ &\quad + (1 - p(y))\ell_2(y, f_1(x), \lambda_1) \\ &= \begin{cases} \ell_1(1, f_1(x)), & y = 1, \\ \ell_2(-1, f_1(x), \lambda_1), & y = -1. \end{cases} \end{aligned}$$

Analogously, a revised loss function L_2 is defined as

$$\begin{aligned} L_2(y, f_2(x), \lambda_2) &= (1 - p(y))\ell_1(y, f_2(x)) \\ &\quad + p(y)\ell_2(y, f_2(x), \lambda_2) \\ &= \begin{cases} \ell_2(1, f_2(x), \lambda_2), & y = 1, \\ \ell_1(-1, f_2(x)), & y = -1, \end{cases} \end{aligned}$$

in the minimization problem Eq. (2). It measures the loss for any positive or negative sample about hyper-plane f_2. Therefore, the general framework can be rewritten as

$$\begin{aligned} &\min_{f_1 \in H} \sum_{i=1}^{m} L_1(y, f_1(x), \lambda_1) + \Omega^*(c_1, ||f_1||_H), \\ &\min_{f_2 \in H} \sum_{i=1}^{m} L_2(y, f_2(x), \lambda_2) + \Omega^*(c_2, ||f_2||_H), \end{aligned}$$

Obviously, it is equivalent to the optimization problem

$$\begin{aligned} &\min_{f_1 \in H} \frac{1}{m} \sum_{i=1}^{m} L_1(y, f_1(x), \lambda_1) + \Omega(c_1, ||f_1||_H), \\ &\min_{f_2 \in H} \frac{1}{m} \sum_{i=1}^{m} L_2(y, f_2(x), \lambda_2) + \Omega(c_2, ||f_2||_H), \end{aligned} \tag{3}$$

where $\Omega(\cdot, \cdot) = \frac{1}{m}\Omega^*(\cdot, \cdot)$ is also a regularization function. To sum up, this optimization problem Eq. (3) is a unified framework for most of TWSVM variants considered in the paper.

Note, in linear case, the two non-parallel hyper-planes are conducted as $f_1(x) = x^T w_1 + b_1$ for positive samples and $f_2(x) = x^T w_2 + b_2$ for negative samples, respectively. Similarly in nonlinear case, they are formulated as $f_1(x) = k(x^T, D^T)u_1 + b_1$ for positive samples and $f_2(x) = k(x^T, D^T)u_2 + b_2$ for negative samples, where k is a kernel function. The final classifier $f : X \to \mathbb{R}$ for both linear and non-linear cases could be expressed as $f(x) = |f_2(x)| - |f_1(x)|$. Given a new sample x, the predicted label of x is

$$y = \begin{cases} 1, & f(x) > 0, \\ -1, & \text{otherwise}. \end{cases}$$

Let ξ, η are two slack variables, and e_1, e_2 are two vectors whose elements are all 1's and whose dimensions are m_1 and m_2, respectively. Below, we give several examples that can be expressed by the unified framework.

Example 1 (TWSVM [13]) For linear TWSVM, the optimization problem is formulated as follows:

$$\begin{aligned} &\min_{w_1, b_1, \xi} \frac{1}{2}\left\|A^T w_1 + e_1 b_1\right\|^2 + \lambda_1 e_2^T \xi,\ s.t. - (B^T w_1 + e_2 b_1) \\ &\quad + \xi \geq e_2, \xi \geq 0, \\ &\min_{w_2, b_2, \eta} \frac{1}{2}||B^T w_2 + e_2 b_2||^2 + \lambda_2 e_1^T \eta,\ s.t.(A^T w_2 + e_1 b_2) \\ &\quad + \eta \geq e_1, \eta \geq 0, \end{aligned}$$

For nonlinear TWSVM, the optimization problem is formulated as follows:

$$\begin{aligned} &\min_{u_1, b_1, \xi} \frac{1}{2}||k(A^T, D^T)u_1 + e_1 b_1||^2 + \lambda_1 e_2^T \xi, \\ &\qquad s.t. - (k(B^T, D^T)u_1 + e_2 b_1) + \xi \geq e_2, \xi \geq 0, \\ &\min_{u_2, b_2, \eta} \frac{1}{2}||k(B^T, D^T)u_2 + e_2 b_2||^2 + \lambda_2 e_1^T \eta, \\ &\qquad s.t.(k(A^T, D^T)u_2 + e_1 b_2) + \eta \geq e_1, \eta \geq 0. \end{aligned}$$

Let $\Omega^*(c_i, ||f_i||_H) = 0, i = 1, 2$. Let

$$\begin{aligned} &\ell_1(1, f_1(x_i^1)) = \frac{1}{2} f_1^2(x_i^1), i = 1, \dots, m_1, \\ &\ell_1(-1, f_2(x_j^2)) = \frac{1}{2} f_2^2(x_j^2), j = 1, \dots, m_2, \end{aligned} \tag{4}$$

and let

$$\begin{aligned} &\ell_2(1, f_2(x_i^1), \lambda_2) = \lambda_2 \eta_i = \lambda_2 \max(0, (1 - f_2(x_i^1))), \\ &\quad i = 1, \dots, m_1, \\ &\ell_2(-1, f_1(x_j^2), \lambda_1) = \lambda_1 \xi_j = \lambda_1 \max(0, (1 + f_1(x_j^2))), \\ &\quad j = 1, \dots, m_2. \end{aligned} \tag{5}$$

Then, the optimization problems of linear and nonlinear case of TWSVM could be both converted to the general framework Eq. (3).

Example 2 (TBSVM [21]) For linear TBSVM, the optimization problem is formulated as follows:

$$\begin{aligned} &\min_{w_1, b_1, \xi} \frac{1}{2} ||A^T w_1 + e_1 b_1||^2 + \lambda_1 e_2^T \xi + \frac{1}{2} c_1 (||w_1||^2 + b_1^2), \\ &\qquad s.t. - (B^T w_1 + e_2 b_1) + \xi \geq e_2, \xi \geq 0, \\ &\min_{w_2, b_2, \eta} \frac{1}{2} ||B^T w_2 + e_2 b_2||^2 + \lambda_2 e_1^T \eta + \frac{1}{2} c_2 (||w_2||^2 + b_2^2), \\ &\qquad s.t. (A^T w_2 + e_1 b_2) + \eta \geq e_1, \eta \geq 0. \end{aligned}$$

For nonlinear TBSVM, the optimization problem is formulated as follows:

$$\begin{aligned} &\min_{u_1, b_1, \xi} \frac{1}{2} ||k(A^T, D^T) u_1 + e_1 b_1||^2 + \lambda_1 e_2^T \xi \\ &\qquad + \frac{1}{2} c_1 (||u_1||^2 + b_1^2), \\ &\qquad s.t. - (k(B^T, D^T) u_1 + e_2 b_1) + \xi \geq e_2, \xi \geq 0, \\ &\min_{u_2, b_2, \eta} \frac{1}{2} ||k(B^T, D^T) u_2 + e_2 b_2||^2 + \lambda_2 e_1^T \eta \\ &\qquad + \frac{1}{2} c_2 (||u_2||^2 + b_2^2), \\ &\qquad s.t. (k(A^T, D^T) u_2 + e_1 b_2) + \eta \geq e_1, \eta \geq 0. \end{aligned}$$

Let

$$\begin{aligned} &\Omega^*(c_1, ||f_1||_H) = \frac{1}{2} c_1 (||w_1||^2 + b_1^2), \\ &\Omega^*(c_2, ||f_2||_H) = \frac{1}{2} c_2 (||w_2||^2 + b_2^2). \end{aligned} \tag{6}$$

Connected with Eqs. (4) and (5), the optimization problems of linear and nonlinear case of TBSVM could be both converted to the general framework Eq. (3).

Example 3 (Improved LSTSVM [31]) For linear improved LSTSVM, the optimization problem is formulated as follows:

$$\begin{aligned} &\min_{w_1, b_1, \xi} \frac{1}{2} ||A^T w_1 + e_1 b_1||^2 + \lambda_1 \xi^T \xi + \frac{1}{2} c_1 (||w_1||^2 + b_1^2), \\ &\qquad s.t. - (B^T w_1 + e_2 b_1) + \xi \geq e_2, \xi \geq 0, \\ &\min_{w_2, b_2, \eta} \frac{1}{2} ||B^T w_2 + e_2 b_2||^2 + \lambda_2 \eta^T \eta + \frac{1}{2} c_2 (||w_2||^2 + b_2^2), \\ &\qquad s.t. (A^T w_2 + e_1 b_2) + \eta \geq e_1, \eta \geq 0. \end{aligned}$$

For nonlinear improved LSTSVM, the optimization problem is formulated as follows:

$$\begin{aligned} &\min_{u_1, b_1, \xi} \frac{1}{2} \left\| k(A^T, D^T) u_1 + e_1 b_1 \right\|^2 + \lambda_1 \xi^T \xi \\ &\qquad + \frac{1}{2} c_1 (||u_1||^2 + b_1^2), \\ &\qquad s.t. - (k(B^T, D^T) u_1 + e_2 b_1) + \xi \geq e_2, \xi \geq 0, \\ &\min_{u_2, b_2, \eta} \frac{1}{2} \left\| k(B^T, D^T) u_2 + e_2 b_2 \right\|^2 + \lambda_2 \eta^T \eta \\ &\qquad + \frac{1}{2} c_2 \left(||u_2||^2 + b_2^2 \right), \\ &\qquad s.t. (k(A^T, D^T) u_2 + e_1 b_2) + \eta \geq e_1, \eta \geq 0. \end{aligned}$$

Let

$$\begin{aligned} &\ell_2(1, f_2(x_i^1), \lambda_2) = \lambda_2 \eta_i^2 = \lambda_2 (\max\{0, 1 - f_2(x_i^1)\})^2, \\ &\quad i = 1, \dots, m_1, \\ &\ell_2(-1, f_1(x_j^2), \lambda_1) = \lambda_1 \xi_j^2 = \lambda_1 (\max\{0, 1 + f_1(x_j^2)\})^2, \\ &\quad j = 1, \dots, m_2. \end{aligned}$$

Considering Eqs. (4) and (6), the optimization problems of linear and nonlinear case of improved LSTSVM could be both converted to the general framework Eq. (3).

4 Universal consistency of TWSVMs

Since TWSVM and its variants are all built based on SRM principle, we study the universal consistency of TWSVMs in a general framework, that is, the universal consistency of the optimization problem (3).

4.1 Definitions

Given the training set S, let $f_S : X \to \mathbb{R}$ be the classifier of the optimization problem Eq. (3) learned from the training set S, where $f_S(\cdot) = |f_{2,S}(\cdot)| - |f_{1,S}(\cdot)|$ and $f_{1,S}, f_{2,S} : X \to \mathbb{R}$ are measurable functions corresponding to the positive and negative hyper-planes, respectively. In order to make the classifier work well, we need to make sure it is as small as possible for the wrongly classified probability of a novel data (x, y) drown from P independently to S. The wrong classification represents that $\text{sign}(f_S(x)) \neq y$. In what follows, it is necessary to redefine *risk*, *Bayes risk* and *universal consistency* for the optimization problem Eq. (3).

Definition 3 (*Risk*)) Given a measurable function $f : X \to \mathbb{R}$, the *risk* of f is the wrongly classified probability of data (x, y) drawn from P independently to S, i.e.,

$$R_P(f) = P\{(x, y) : \text{sign}(f(x)) \neq y\},$$

where $f(\cdot) = |f_2(\cdot)| - |f_1(\cdot)|$, and $f_1, f_2 : X \to \mathbb{R}$ are both measurable functions corresponding to the positive and negative hyper-planes, respectively.

Definition 4 (*Bayes Risk*) The *Bayes risk* with respect to distribution P, denoted by R_P, is the smallest achievable risk

$$R_P = \inf\{R_P(f) | f : X \to \mathbb{R} \text{ is measurable}\},$$

where $f(\cdot) = |f_2(\cdot)| - |f_1(\cdot)|$, and $f_1, f_2 : X \to \mathbb{R}$ are both measurable functions corresponding to the positive and negative hyper-planes, respectively.

Bayes risk is the minimal value of *risk* $R_P(f)$, and thus is the minimal true risk with respect to distribution P on space $X \times Y$. Given the training set S, to make the wrongly classified probability as small as possible, we must make the risk $R_P(f_S)$ of the classifier f_S infinitely close to the minimal true risk. Thereby, the universal consistency is defined in the following:

Definition 5 (*Universal consistency*) The classifier f_S of optimization problem Eq. (3) is universally consistent, if the following equation

$$\lim_{m\to\infty} R_P(f_S) = R_P \tag{7}$$

is valid in probability with respect to distribution P on space $X \times Y$, where $f(\cdot) = |f_2(\cdot)| - |f_1(\cdot)|$, and $f_1, f_2 : X \to \mathbb{R}$ are both measurable functions corresponding to the positive and negative hyper-planes, respectively. Furthermore, if Eq. (7) is valid almost surely, the classifier f_S is strongly universally consistent.

Universal consistency is a key point to explain the success for the optimization problem Eq. (3), and to provide the solid theoretical basis. Thus in the below, we begin to discuss under what conditions, the universal consistency is guaranteed for the optimization problem Eq. (3).

4.2 Assertion

Though researchers have developed lots of variants based on the idea of TWSVM in the literature, there is still no theoretical study for the universal consistency of any variant. Also, no existing technique could be regarded as a reference in analyzing the universal consistency of the general optimization problem Eq. (3). Nevertheless, there is one specific technique for investigating the universal consistency of SVMs [25], and this technique can be applied to study the universal consistency of the problem Eq. (3). The reason is that TWSVMs are the extensions to SVMs, and the difference of the two problems is that TWSVMs formulate two QPPs, while SVMs formulate only one. The QPPs for TWSVMs and the QPP for SVMs are constructed in a similar way, both with two terms: loss function term and regularization function term. Therefore, in order to tackle the universal consistency of the optimization problem Eq. (3), a similar assertion like that in [25] is necessary to pay attention to. Before this, we give some definitions which would be used in the assertion.

Definition 6 (L_1*- and* L_2*-Risks*) Given two loss functions L_1, L_2 and a probability distribution P, L_1-risk is the expectation of loss function L_1 corresponding to the hyper-plane f_1, and L_2-risk is the expectation of L_2 corresponding to f_2, both defined as follows:

$$R_{1,P}(f_1) = E_{(x,y)\sim P} L_1(y, f_1(x), \lambda_1),\ R_{2,P}(f_2) = E_{(x,y)\sim P} L_2(y, f_2(x), \lambda_2).$$

Definition 7 (*Minimal* L_1*- and* L_2*-Risks*) The minimal L_1- and L_2-risks on the probability distribution P are the smallest achievable L_1-risk and L_2-risk, respectively, both defined as follows:

$$R_{1,P} = \inf\{R_{1,P}(f_1) | f_1 : X \to \mathbb{R} \text{ is measurable}\},$$
$$R_{2,P} = \inf\{R_{2,P}(f_2) | f_2 : X \to \mathbb{R} \text{ is measurable}\}.$$

Definition 8 (*Regularized* L_1*- and* L_2*-Risks*) Given the regularization function Ω and a RKHS H, the regularized L_1- and L_2-risks are defined by

$$R^{reg}_{1,P,c_1}(f_1) = R_{1,P}(f_1) + \Omega(c_1, ||f_1||_H),\ R^{reg}_{2,P,c_2}(f_2) = R_{2,P}(f_2) + \Omega(c_2, ||f_2||_H),$$

respectively for all $c_1, c_2 > 0$ and for all $f_1, f_2 \in H$.

Note, if the distribution P is an empirical measure on the training set S, we rewrite them as $R_{1,S}(f_1)$, $R_{2,S}(f_2)$, $R_{1,S}$, $R_{2,S}$, $R^{reg}_{1,S,c_1}(f_1)$ and $R^{reg}_{2,S,c_2}(f_2)$, respectively. We can see that $R^{reg}_{1,S,c_1}(f_1)$ and $R^{reg}_{2,S,c_2}(f_2)$ are just the two objective functions in Eq. (3). Then, the assertion is concluded in four steps as follows:

Step 1: Show that there exist two elements $f_{1,P,c_1}, f_{2,P,c_2} \in H$ minimizing the regularized L_1-risk and regularized L_2-risk, respectively,

$$f_{1,P,c_1} = \arg\min_{f_1 \in H} R^{reg}_{1,P,c_1}(f_1),\ \ f_{2,P,c_2} = \arg\min_{f_2 \in H} R^{reg}_{2,P,c_2}(f_2).$$

Step 2: Show that the minimal L_1-risk $R_{1,P}$ could be achieved at the element f_{1,P,c_1} by the regularized L_1-risk with c_1 tending to 0, and the minimal L_2-risk $R_{2,P}$ could be achieved at f_{2,P,c_2} by the regularized L_2-risk with c_2 tending to 0.

$$\lim_{c_1 \to 0} R^{reg}_{1,P,c_1}(f_{1,P,c_1}) = R_{1,P},\ \ \lim_{c_2 \to 0} R^{reg}_{2,P,c_2}(f_{2,P,c_2}) = R_{2,P}.$$

Step 3: For sequences of measurable functions $f_{1,m}$,$f_{2,m} : X \to \mathbb{R}$, $f_m(\cdot) = |f_{2,m}(\cdot)| - |f_{1,m}(\cdot)|$, and for admissible loss functions L_1,L_2, show that the Bayes risk R_P could be achieved at f_m with m tending to infinity, if the following equations hold true

$$\lim_{m \to \infty} R_{1,P}(f_{1,m}) = R_{1,P},\ \ \lim_{m \to \infty} R_{2,P}(f_{2,m}) = R_{2,P}.$$

Step 4: Find a pair of concentration inequalities, where one concentration inequality of them relates the L_1-risk with the empirical L_1-risk at point f_{1,S,c_1}, and the other relates the L_2-risk with the empirical L_2-risk at point f_{2,S,c_2}. The universal consistency of optimization problem Eq. (3) could be conducted from the pair of concentration inequalities.

Assume that the assertion holds true, now we can see how to demonstrate the universal consistency of optimization problem Eq. (3). First, we could find two elements

$$f_{1,S,c_1} = \arg\min_{f_1 \in H} R^{reg}_{1,S,c_1}(f_1),\ \ f_{2,S,c_2} = \arg\min_{f_2 \in H} R^{reg}_{2,S,c_2}(f_2),$$

and show the existence by Step 1, since S is the empirical measure of P. Next, the upper bounds for the probabilities of the events $|R_{1,S}(f_{1,S,c_1}) - |R_{1,P}(f_{1,S,c_1})| \geq \epsilon$ and $|R_{2,S}(f_{2,S,c_2}) - |R_{2,P}(f_{2,S,c_2})| \geq \epsilon$ occurring could be derived according to Hoeffding's inequalities [6], respectively. Furthermore, the two derived inequalities about upper bounds are exactly the pair of concentration inequalities we want in Step 4. Then, we force the two upper bounds both to tend to 0, by setting two sequences $c_1(m)$ and $c_2(m)$ both to tend to 0 when m tends to infinity. Following Step 2, we have two measurable sequences $f_{1,S,c_1(m)}$ and $f_{2,S,c_2(m)}$ such that

$$\lim_{m \to \infty} R_{1,P}(f_{1,S,c_1(m)}) = R_{1,P},\ \ \lim_{m \to \infty} R_{2,P}(f_{2,S,c_2(m)}) = R_{2,P}.$$

Naturally, the conditions for Step 3 are valid. Finally, the universal consistency is guaranteed by virtue of Step 3.

Note that, the assertion's validity is just some kind of assumption without any proof up to now. Thus, in the rest of the paper, we would complete the assertion by testifying it's validity.

5 Theoretical results

Here, we investigate the assertion and use some theorems to support the validity of the assertion in each step. Three theorems are first given to show Steps 1–3 of the assertion, respectively. In the next three subsections, three pairs of concentration inequalities are derived for Step 4 based on different conditions, including covering number, localized covering number and stability, respectively. Theorems are given to show that the universal consistency is valid based on different concentration inequalities. The proofs follow the idea in [25]. Because of space limit for the paper, the detailed proofs of Theorems and Lemmas are put to the supplementary material. Readers can see the supplementary material for the proofs.

Let k be a positive semi-definite kernel, L_1, L_2 be admissible loss functions, and Ω be a regularization function. Given the parameters λ_1, λ_2, define for $c_1, c_2 > 0$ that

$$\delta_{c_1} = \sup\{t : \Omega(c_1, t) \leq L_1(1, 0, \lambda_1) + L_1(-1, 0, \lambda_1)\},$$
$$\delta_{c_2} = \sup\{t : \Omega(c_2, t) \leq L_2(1, 0, \lambda_2) + L_2(-1, 0, \lambda_2)\},$$
$$L_{1,c_1} = L_{1|Y \times [-\delta_{c_1}K, \delta_{c_1}K] \times \lambda_1},\ \ L_{2,c_2} = L_{2|Y \times [-\delta_{c_2}K, \delta_{c_2}K] \times \lambda_2}.$$

Note that $0 < \delta_{c_1}, \delta_{c_2} < \infty$, and we have

$$\hat{\delta}_1 = \inf\{\delta_{c_1} : c_1 \in (0, 1]\} > 0,\ \hat{\delta}_2 = \inf\{\delta_{c_2} : c_2 \in (0, 1]\} > 0.$$

For the loss function $L_{i,c_i}, i = 1, 2$, denote by $|\cdot|_1$ the supremum

$$|L_{i,c_i}|_1 = \sup\left\{ \frac{|L_i(y, t', \lambda_i) - L_i(y, t'', \lambda_i)|}{|t' - t''|} : y \in Y, t', t'' \in [-\delta_{c_i}K, \delta_{c_i}K], t' \neq t'' \right\}.$$

Theorem 1 *Assume k is a continuous kernel on X. Let L_1 and L_2 be two admissible loss functions, and Ω be a regularization function. There exist two elements f_{1,P,c_1}, $f_{2,P,c_2} \in H$ such that*

$$R^{reg}_{1,P,c_1}(f_{1,P,c_1}) = \inf_{f_1 \in H} R^{reg}_{1,P,c_1}(f_1),\ \ R^{reg}_{2,P,c_2}(f_{2,P,c_2}) = \inf_{f_2 \in H} R^{reg}_{2,P,c_2}(f_2),$$

for the probability distribution P and for any $c_1, c_2 > 0$. Furthermore, we have $\|f_{1,P,c_1}\|_H \leq \delta_{c_1}$ and $\|f_{2,P,c_2}\|_H \leq \delta_{c_2}$.

Theorem 1 corresponds to Step 1 of the assertion, and ensures the existence of two elements f_{1,P,c_1} and f_{2,P,c_2} which minimize the regularized L_1-risk and regularized L_2-risk, respectively. Here, δ_{c_1} and δ_{c_2} are two critical quantities and give upper bounds on the norm of the solutions to the optimization problem Eq. (3), respectively.

Lemma 1 *There exist two measurable functions* $f_1^*, f_2^* : [0,1] \to \mathbb{R}$ *such that* $M_1(\alpha, \lambda_1) = C(\alpha, f_1^*(\alpha), \lambda_1)$ *and* $M_2(\alpha, \lambda_2) = C(\alpha, f_2^*(\alpha), \lambda_2)$ *for any* $\alpha \in [0,1]$*. Then, we have*

$$R_{1,P} = \int_X M_1(f_1^*(P(1|x)), \lambda_1) P_X(dx),$$
$$R_{2,P} = \int_X M_2(f_2^*(P(1|x)), \lambda_2) P_X(dx),$$

where P_X *is the marginal distribution of probability distribution* P *on* X*.*

Lemma 1 converts the minimal L_1-risk and minimal L_2-risk to the expectations of M_1 and M_2, respectively. It is necessary to the proof of Step 2.

Theorem 2 *Assume k is a universal kernel on X. Let* L_1 *and* L_2 *be two admissible loss functions, and* Ω *be a regularization function. We have*

$$\lim_{c_1 \to 0} R_{1,P,c_1}^{reg}(f_{1,P,c_1}) = R_{1,P}, \ \lim_{c_2 \to 0} R_{2,P,c_2}^{reg}(f_{2,P,c_2}) = R_{2,P},$$

for the probability distribution P and for any $c_1, c_2 > 0$,

Theorem 2 corresponds to Step 2 of the assertion, and guarantees that the minimal L_1-risk and minimal L_2-risk could be achieved at f_{1,P,c_1} and f_{2,P,c_2}, respectively, with c_1, c_2 tending to 0.

Theorem 3 *Let* L_1 *and* L_2 *be two admissible loss functions, and* Ω *be a regularization function. If there exist* $\delta_1 > 0$ *and* $\delta_2 > 0$ *such that* $R_{1,P}(f_1) \leq R_{1,P} + \delta_1$ *and* $R_{2,P}(f_2) \leq R_{2,P} + \delta_2$ *for any two measurable functions* $f_1, f_2 : X \to \mathbb{R}$*, we have* $R_P(f) \leq R_P + \epsilon$ *for any* $\epsilon > 0$*, where* $f(\cdot) = |f_2(\cdot)| - |f_1(\cdot)|$.

Theorem 3 corresponds to Step 3 of the assertion, and explains the following relations

$$\begin{cases} \lim_{m\to\infty} R_{1,P}(f_{1,m}) = R_{1,P} \\ \lim_{m\to\infty} R_{2,P}(f_{2,m}) = R_{2,P} \end{cases} \Longrightarrow \lim_{m\to\infty} R_P(f_m) = R_P$$

are valid for all sequences of measurable functions $f_{1,m}$, $f_{2,m}$ and $f_m(\cdot) = |f_{2,m}(\cdot)| - |f_{1,m}(\cdot)|$.

5.1 Universal consistency based on covering number

Now we pay attention to Step 4 of the assertion, and want to find a pair of concentration inequality based on covering number [33].

Definition 9 Given a metric space $(\mathcal{M}, d)$, the covering number of $\mathcal{M}$ is defined as

$$\mathcal{N}((\mathcal{M}, d), \epsilon)$$
$$= \min\left\{ n \in \mathbb{N} | x_1, \dots, x_n \in \mathcal{M} \subseteq \bigcup_{i=1}^{n} B(x_i, \epsilon) \right\},$$

where $B(x, \epsilon)$ is a closed ball with the center at point x and with a radius $\epsilon > 0$.

Instead of using covering number directly, its logarithmic form is employed more frequently, which is denoted as $\mathcal{H}((\mathcal{M}, d), \epsilon) = \ln \mathcal{N}((\mathcal{M}, d), \epsilon)$.

In addition, we have to measure the continuity of a function. Given a loss function L_1, the modulus and inverted modulus of continuity [2] of the function are expressed as $w(L_1, \delta)$ and $w^{-1}(L_1, \epsilon)$, respectively,

$$w(L_1, \delta) = \sup_{y \in Y, \lambda_1 \in \mathbb{R}^+, t', t'' \in \mathbb{R}, |t' - t''| \leq \delta} |L_1(y, t', \lambda_1) - L_1(y, t'', \lambda_1)|,$$
$$w^{-1}(L_1, \epsilon) = \sup\{\delta > 0 : w(L_1, \delta) \leq \epsilon\}.$$

With these definitions, we begin to formulate the pair of concentration inequalities for the optimization problem Eq. (3) by the following lemma, and establish the consistency result by the following theorem.

Lemma 2 *Assume k is a continuous kernel on X. Let* L_1 *and* L_2 *be two admissible loss functions, and* Ω *be a regularization function. For the probability distribution P and for any* $m \geq 1, c_1, c_2 > 0, \epsilon > 0$*, we have*

$$Pr\{S : |R_{1,S}(f_{1,S,c_1}) - R_{1,P}(f_{1,S,c_1})| \geq \epsilon\}$$
$$\leq 2\exp\left\{ \mathcal{H}(\delta_{c_1} I, \omega^{-1}(L_{1,c_1}, \epsilon/3)) - \frac{2\epsilon^2 m}{9\|L_{1,c_1}\|_\infty^2} \right\},$$
$$Pr\{S : |R_{2,S}(f_{2,S,c_2}) - R_{2,P}(f_{2,S,c_2})| \geq \epsilon\}$$
$$\leq 2\exp\left\{ \mathcal{H}(\delta_{c_2} I, \omega^{-1}(L_{2,c_2}, \epsilon/3)) - \frac{2\epsilon^2 m}{9\|L_{2,c_2}\|_\infty^2} \right\},$$

where Pr *is the joint probability of data* $(x_1, y_1) \times (x_2, y_2) \times \ldots \times (x_m, y_m)$ *from the training set S.*

Note that each sample $(x_i, y_i) \in S \subseteq X \times Y, i = 1, \dots, m$ follows the probability distribution P, then $(x_1, y_1) \times (x_2, y_2) \times \ldots \times (x_m, y_m) \in (X \times Y)^m$ follows the probability distribution P^m.

Theorem 4 *Assume k is a universal kernel on X. Let* L_1 *and* L_2 *be two admissible loss functions, and* Ω *be a regularization function. Suppose there are two positive sequences* $(c_1(m)), (c_2(m))$ *with* $c_1(m) \to 0, c_2(m) \to 0$ *and*

$$\frac{\|L_{1,c_1(m)}\|_\infty^2}{m}\mathcal{H}(\delta_{c_1(m)}I, \omega^{-1}(L_{1,c_1(m)}, \epsilon)) \longrightarrow 0,$$
$$\frac{\|L_{2,c_2(m)}\|_\infty^2}{m}\mathcal{H}(\delta_{c_2(m)}I, \omega^{-1}(L_{2,c_2(m)}, \epsilon)) \longrightarrow 0,$$

for all $\epsilon > 0$, when m approaches infinity. Then the optimization problem Eq. (3) is universally consistent. If we have the additional conditions

$$\sum_{m=1}^{\infty} \exp\{-\epsilon m/\|L_{1,c_1(m)}\|_\infty^2\} < \infty,$$
$$\sum_{m=1}^{\infty} \exp\{-\epsilon m/\|L_{2,c_2(m)}\|_\infty^2\} < \infty, \tag{8}$$

for all $\epsilon > 0$, the optimization problem Eq. (3) is even strongly universally consistent.

Lemma 2 and Theorem 4 are corresponding to the fourth step of the assertion. By virtue of covering numbers of $\delta_{c_1}I$ and $\delta_{c_2}I$, the pair of concentration inequalities are derived first. Based on them, the universal consistency of TWSVMs is then conducted. Next we give an example to explain the case in detail.

Example 4 Given a TWSVM variant like TBSVM [21], we consider the nonlinear case and use the Gaussian kernel k on $X \subset \mathbb{R}^d$. There is an upper bound for the covering number of I in [33]

$$\mathcal{H}(I, \epsilon) \leq a(\ln\frac{1}{\epsilon})^{d+1}, \text{ for some positive constant } a.$$

Then we have

$$\frac{\|L_{1,c_1(m)}\|_\infty^2}{m}\mathcal{H}(\delta_{c_1(m)}I, \omega^{-1}(L_{1,c_1(m)}, \epsilon))$$
$$\leq \frac{a\|L_{1,c_1(m)}\|_\infty^2}{m}\left(\ln\frac{1}{\omega^{-1}(L_{1,c_1(m)}, \epsilon)}\right)^{d+1} \xrightarrow{m\to\infty} 0,$$
$$\frac{\|L_{2,c_2(m)}\|_\infty^2}{m}\mathcal{H}(\delta_{c_2(m)}I, \omega^{-1}(L_{2,c_2(m)}, \epsilon))$$
$$\leq \frac{a\|L_{2,c_2(m)}\|_\infty^2}{m}\left(\ln\frac{1}{\omega^{-1}(L_{2,c_2(m)}, \epsilon)}\right)^{d+1} \xrightarrow{m\to\infty} 0$$

The classifier is universally consistent by Theorem 4.

5.2 Universal consistency based on localized covering number

Sometimes, we suggest the pair of concentration inequalities based on the localized covering number, instead of covering number. Given a function set $\mathcal{F} = \{f : X \mapsto \mathbb{R}\}$, the localized covering number of $\mathcal{F}$ is

$$\mathcal{N}(\mathcal{F}, m, \epsilon) = \sup\{\mathcal{N}((\mathcal{F}_{|X_0}, \ell_\infty^{|X_0|}), \epsilon) : X_0 \subset X, |X_0| \leq m\},$$

for any $\epsilon > 0$ and $m \geq 1$, where $\ell_\infty^{|X_0|}$ is the space $\mathbb{R}^{|X_0|}$ with the maximum norm, and $\mathcal{F}_{|X_0} = \{f_{|X_0} : f \in \mathcal{F}\}$ could be regarded as a subset of $\ell_\infty^{|X_0|}$. The logarithm of localized covering number is $\mathcal{H}(\mathcal{F}, m, \epsilon) = \ln \mathcal{N}(\mathcal{F}, m, \epsilon)$. Now we start to obtain another pair of concentration inequalities for the optimization problem Eq. (3) according to the following lemma, and derive the universal consistency by the following theorem.

Lemma 3 *Assume k is a continuous kernel on X. Let L_1 and L_2 be two admissible loss functions, and Ω be a regularization function. For the probability distribution P and for any $m \geq 1, c_1, c_2 > 0, \epsilon > 0$, we have*

$$Pr\{S : |R_{1,S}(f_{1,S,c_1}) - R_{1,P}(f_{1,S,c_1})| \geq \epsilon\}$$
$$\leq 12m\exp\left\{\mathcal{H}(\delta_{c_1}I, 2m, \omega^{-1}(L_{1,c_1}, \epsilon/6)) - \frac{\epsilon^2 m}{36\|L_{1,c_1}\|_\infty^2}\right\},$$
$$Pr\{S : |R_{2,S}(f_{2,S,c_2}) - R_{2,P}(f_{2,S,c_2})| \geq \epsilon\}$$
$$\leq 12m\exp\left\{\mathcal{H}(\delta_{c_2}I, 2m, \omega^{-1}(L_{2,c_2}, \epsilon/6)) - \frac{\epsilon^2 m}{36\|L_{2,c_2}\|_\infty^2}\right\},$$

where Pr is the joint probability of data $(x_1, y_1) \times (x_2, y_2) \times \ldots \times (x_m, y_m)$ from the training set S.

Theorem 5 *Assume k is a universal kernel on X. Let L_1 and L_2 be two admissible loss functions, and Ω be a regularization function. Suppose there are two positive sequences $(c_1(m)), (c_2(m))$ with $c_1(m) \to 0, c_2(m) \to 0$ and*

$$\frac{\|L_{1,c_1(m)}\|_\infty^2}{m}\big(\log m + \mathcal{H}(\delta_{c_1(m)}I, 2m, \omega^{-1}(L_{1,c_1(m)}, \epsilon))\big) \longrightarrow 0,$$
$$\frac{\|L_{2,c_2(m)}\|_\infty^2}{m}\big(\log m + \mathcal{H}(\delta_{c_2(m)}I, 2m, \omega^{-1}(L_{2,c_2(m)}, \epsilon))\big) \longrightarrow 0,$$

for all $\epsilon > 0$, when m approaches infinity. Then the optimization problem Eq. (3) is universally consistent. Furthermore, if we have additional conditions Eq. (8) for all $\epsilon > 0$, the problem Eq. (3) is even strongly universally consistent.

Lemma 3 presents another pair of concentration inequalities based on the localized covering numbers of $\delta_{c_1}I$ and $\delta_{c_2}I$. And Theorem 5 discusses the conditions under which, the universal consistency is valid for TWSVMs. Thus, Lemma 3 and Theorem 5 imply the validation of Step 4 of the assertion. In the following, an example is illustrated to exhibit the universal consistency.

Example 5 Given a TWSVM variant, we study the nonlinear case. The universal kernel $k(x, x') = \Sigma_{n=0}^{\infty} a_n \Phi_n(x)\Phi_n(x')$ is advised here, where $a_n > 0, n = 0, 1, 2, \ldots,$ and $\Phi_n : X \to \mathbb{R}$

are all continuous functions uniformly bounded with $|| \cdot ||_\infty$-norm. The localized covering number of I is upper bounded [24],

$$\mathcal{H}(I, m, \epsilon) \leq b(\frac{\log m}{\epsilon^2})^\rho \text{ for some postive constant } b, 0 < \rho < 1,$$

which indicates that

$$\begin{aligned}
&\frac{\|L_{1,c_1(m)}\|_\infty^2}{m}\Big(\log m + \mathcal{H}(\delta_{c_1(m)}I, 2m, \omega^{-1}(L_{1,c_1(m)}, \epsilon))\Big) \\
\leq &\frac{\|L_{1,c_1(m)}\|_\infty^2}{m}\Big(\log m + b(\frac{\log(2m)}{(\omega^{-1}(L_{1,c_1(m)}, \epsilon))^2})^\rho\Big) \xrightarrow{m\to\infty} 0, \\
&\frac{\|L_{2,c_2(m)}\|_\infty^2}{m}\Big(\log m + \mathcal{H}(\delta_{c_2(m)}I, 2m, \omega^{-1}(L_{2,c_2(m)}, \epsilon))\Big) \\
\leq &\frac{\|L_{2,c_2(m)}\|_\infty^2}{m}\Big(\log m + b(\frac{\log(2m)}{(\omega^{-1}(L_{2,c_2(m)}, \epsilon))^2})^\rho\Big) \xrightarrow{m\to\infty} 0.
\end{aligned}$$

By virtue of Theorem 5, it is obvious that the classifier is universal consistent.

5.3 Universal consistency based on stability

For practical problems, the case for convex loss functions and for the regularization function $\Omega(c, t) = ct^2$ is often considered for the optimization problem Eq. (3). In this case, a stable classifier is always employed. Here, the property for stability [3] is redefined as follows:

Definition 10 Given a training set $S = \{(x_i, y_i) \in X \times Y, i = 1, \dots, m\}$, let $f_S(\cdot) = |f_{2,S}(\cdot)| - |f_{1,S}(\cdot)|$ be the classifier on set S for the optimization problem Eq. (3). Replace the i'th sample (x_i, y_i) with (x, y), and the new set is denoted by $S_{i,(x,y)}$. If there exist two sequences $(\beta_1(m))$ and $(\beta_2(m))$ such that the following inequalities

$$|L_1(y', f_{1,S}(x'), \lambda_1) - L_1(y', f_{1,S_{i,(x,y)}}(x'), \lambda_1)| \leq \beta_1(i),$$
$$|L_2(y', f_{2,S}(x'), \lambda_2) - L_2(y', f_{2,S_{i,(x,y)}}(x'), \lambda_2)| \leq \beta_2(i),$$

are valid for any $(x', y') \in X \times Y$, the classifier f_S is stable with respect to sequences $(\beta_1(m))$ and $(\beta_2(m))$.

Lemma 4 *Let L_1 and L_2 be two convex loss functions, $\Omega(c,f) = c\|f\|^2$, and $(c_1(m))$, $(c_2(m))$ be two sequences for the regularization function. The classifier is stable with respect to sequences $(\frac{2K^2|L_{1,c_1(m)}|_1^2}{mc_1(m)})$ and $(\frac{2K^2|L_{2,c_2(m)}|_1^2}{mc_2(m)})$.*

In Lemma 4, the classifier of the optimization problem Eq. (3) has shown to be stable for the convex loss functions L_1, L_2 and for the regularization function $\Omega(c, t) = ct^2$. Below, we verify Step 4 of the assertion.

Lemma 5 *Assume k is a continuous kernel on X. Let L_1 and L_2 be two admissible loss functions, and Ω be a regularization function. Let $\beta_1(m) = \frac{2K^2|L_{1,c_1(m)}|_1^2}{mc_1(m)}$ and $\beta_2(m) = \frac{2K^2|L_{2,c_2(m)}|_1^2}{mc_2(m)}$. Suppose the classifier is stable with respect to k, L_1, L_2, Ω, $(c_1(m)), (c_2(m))$, $\beta_1(m)$ and $\beta_2(m)$. Then, we have*

$$\begin{aligned}
&Pr\{S : |R_{1,S}(f_{1,S,c_1}) - R_{1,P}(f_{1,S,c_1})| > \epsilon + \beta_1(m)\} \\
&\qquad \leq 2\exp\left\{-\frac{\epsilon^2 m}{2(m\beta_1(m) + \|L_{1,c_1(m)}\|_\infty)^2}\right\}, \\
&Pr\{S : |R_{2,S}(f_{2,S,c_2}) - R_{2,P}(f_{2,S,c_2})| > \epsilon + \beta_2(m)\} \\
&\qquad \leq 2\exp\left\{-\frac{\epsilon^2 m}{2(m\beta_2(m) + \|L_{2,c_2(m)}\|_\infty)^2}\right\},
\end{aligned}$$

where Pr is the joint probability of data $(x_1, y_1) \times (x_2, y_2) \times \ldots \times (x_m, y_m)$ from the training set S.

Theorem 6 *Assume k is a universal kernel on X. Let L_1 and L_2 be two convex admissible loss functions, and $\Omega(c,f) = c\|f\|^2$. Suppose there are two positive sequences $(c_1(m))$ and $(c_2(m))$ with $c_1(m) \to 0$, $c_2(m) \to 0$, and*

$$\frac{mc_1(m)^2}{|L_{1,c_1(m)}|_1^4} \longrightarrow \infty, \quad \frac{mc_2(m)^2}{|L_{2,c_2(m)}|_1^4} \longrightarrow \infty,$$

when m tends to infinity. Then the optimization problem Eq. (3) is universally consistent.

In Lemma 5, a pair of concentration inequalities is conducted under the condition that the classifier is stable. With the inequalities, the universal consistency of the optimization problem Eq. (3) is then obtained in Theorem 6.

6 Conclusion

In this paper, the universal consistency of TWSVMs for binary classification is addressed. Since many variants of TWSVM have been proposed, we first summarize a general framework of TWSVMs, which covers most of the TWSVM variants. We then perform theoretical study on universal consistency of the general framework in detail by defining an assertion. This assertion consists of four steps. In the first three steps, the regularized L_1-risk and regularized L_2-risk are introduced to build connections with the Bayes risk. In the last step, some pairs of concentration inequalities are derived based on different conditions, including covering number, localized covering number and stability. Universal consistency in different situations is proved based on different pairs of concentration inequalities, respectively.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s13042-021-01281-0.

Acknowledgements This work was supported by National Natural Science Foundation of China (NSFC) under grant No. 61972100.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Berner J, Grohs P, Jentzen A (2020) Analysis of the generalization error: empirical risk minimization over deep artificial neural networks overcomes the curse of dimensionality in the numerical approximation of black-scholes partial differential equations. SIAM J Math Data Sci 2(3):631–657
2. Bloom WR, Elliott D (1981) The modulus of continuity of the remainder in the approximation of Lipschitz functions. J Approx Theory 31(1):59–66
3. Bousquet O, Elisseeff A (2000) Algorithmic stability and generalization performance. In: Leen TK, Dietterich TG, Tresp V (eds) Advances in neural information processing systems, vol 13. MIT Press, Denver, pp 196–202
4. Brownlees CT, Joly E, Lugosi G (2015) Empirical risk minimization for heavy-tailed losses. Ann Stat 43(6):2507–2536
5. Chen DR, Sun T (2006) Consistency of multiclass empirical risk minimization methods based on convex loss. J Mach Learn Res 7(11):2435–2447
6. Devroye L, Györfi L, Lugosi G (1996) A probabilistic theory of pattern recognition, stochastic modelling and applied probability, vol 31. Springer, Berlin
7. Dumpert F, Andreas C (2018) Universal consistency and robustness of localized support vector machines. Neurocomputing 315:96–106
8. Fathony R, Behpour S, Zhang X, Ziebart BD (2018) Efficient and consistent adversarial bipartite matching. In: Dy JG, Krause A (eds) Proceedings of the 35th international conference on machine learning, ICML 2018, vol 80. PMLR, Stockholm, pp 1457–1466
9. Fiasche M (2014) SVM tree for personalized transductive learning in bioinformatics classification problems. In: Bassis S, Esposito A, Morabito FC (eds) Recent advances of neural network models and applications, WIRN 2013, vol 26. Springer, Salerno, pp 223–231
10. Fisher RA (1992) On the mathematical foundations of theoretical statistics. In: Kotz S, Johnson NL (eds) Breakthroughs in statistics. Springer series in statistics (perspectives in statistics). Springer, Berlin, pp 11–44
11. Gupta D, Richhariya B, Borah P (2019) A fuzzy twin support vector machine based on information entropy for class imbalance learning. Neural Comput Appl 31(11):7153–7164
12. Györfi L, Weiss R (2020) Universal consistency and rates of convergence of multiclass prototype algorithms in metric spaces. arXiv:2010.00636
13. Khemchandani RJ, Chandra S (2007) Twin support vector machines for pattern classification. IEEE Trans Pattern Anal Mach Intell 29(5):905–910
14. Kumar MA, Gopal M (2008) Application of smoothing technique on twin support vector machines. Pattern Recognit Lett 29(13):1842–1848
15. Kumar MA, Gopal M (2010) A comparison study on multiple binary-class svm methods for unilabel text categorization. Pattern Recognit Lett 31(11):1437–1444
16. Liu X, Wan A (2015) Universal consistency of extreme learning machine for RBFNs case. Neurocomputing 168:1132–1137
17. Liu Y (2007) Fisher consistency of multicategory support vector machines. In: Meila M, Shen X(eds) Proceedings of the eleventh international conference on artificial intelligence and statistics, AISTATS 2007, vol 2, San Juan, Puerto Rico, pp 291–298
18. Mangasarian OL, Wild EW (2006) Multisurface proximal support vector machine classification via generalized eigenvalues. IEEE Trans Pattern Anal Mach Intell 28(1):69–74
19. Qi Z, Tian Y, Shi Y (2013) Robust twin support vector machine for pattern classification. Pattern Recognit 46(1):305–316
20. Shao Y, Deng N (2013) A novel margin-based twin support vector machine with unity norm hyperplanes. Neural Comput Appl 22(7):1627–1635
21. Shao Y, Zhang C, Wang X, Deng N (2011) Improvements on twin support vector machines. IEEE Trans Neural Netw 22(6):962–968
22. Shao YH, Chen WJ, Wang Z, Li CN, Deng NY (2015) Weighted linear loss twin support vector machine for large-scale classification. Knowl Based Syst 73:276–288
23. Singh G, Chhabra I (2018) Effective and fast face recognition system using complementary OC-LBP and HOG feature descriptors with SVM classifier. J Inf Technol Res 91–110
24. Steinwart I (2001) On the influence of the kernel on the consistency of support vector machines. J Mach Learn Res 2(1):67–93
25. Steinwart I (2005) Consistency of support vector machines and other regularized kernel classifiers. IEEE Trans Inf Theory 51(1):128–142
26. Vapnik VN (1991) Principles of risk minimization for learning theory. In: Moody JE, Hanson SJ, Lippmann R (eds) Advances in neural information processing systems, vol 4. Morgan Kaufmann, Denver, pp 831–838
27. Vapnik VN, Chervonenkis AY (1971) On the uniform convergence of relative frequencies of events to their probabilities. Theory Probab Appl 16(2):264–280
28. Vapnik VN, Chervonenkis AY (1982) Necessary and sufficient conditions for the uniform convergence of the means to their expectations. Theory Probab Appl 26(3):532–553
29. Vapnik VN, Chervonenkis AY (1991) The necessary and sufficient conditions for consistency in the empirical risk minimization method. Pattern Recognit Image Anal 1(3):283–305
30. Xu W, Huang D, Zhou S (2019) Statistical learning with group invariance: problem, method and consistency. Int J Mach Learn Cybern 10:1503–1511
31. Xu Y, Xi W, Lv X, Guo R (2012) An improved least squares twin support vector machine. J Inf Comput Sci 9(4):1063–1071
32. Yan H, Ye Q, Zhang T, Yu DJ, Yuan X, Xu Y, Fu L (2018) Least squares twin bounded support vector machines based on L1-norm distance metric for classification. Pattern Recognit 74:434–447
33. Zhou D (2002) The covering number in learning theory. J Complexity 18(3):739–767

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.