Journal of Machine Learning Research 13 (2012) 2705-2734 Submitted 4/11; Revised 2/12; Published 9/12

Coherence Functions with Applications in Large-Margin Classification Methods

Zhihua Zhang ZHZHANG@ZJU.EDU.CN
Dehua Liu DEHUALIU0427@GMAIL.COM
Guang Dai GUANG.GDAI@GMAIL.COM
College of Computer Science and Technology
Zhejiang University
Hangzhou, Zhejiang 310027, China

Michael I. Jordan JORDAN@CS.BERKELEY.EDU
Computer Science Division and Department of Statistics
University of California
Berkeley, CA 94720-1776, USA

Editor: Xiaotong Shen

Abstract

Support vector machines (SVMs) naturally embody sparseness due to their use of hinge loss functions. However, SVMs can not directly estimate conditional class probabilities. In this paper we propose and study a family of *coherence functions*, which are convex and differentiable, as surrogates of the hinge function. The coherence function is derived by using the maximum-entropy principle and is characterized by a temperature parameter. It bridges the hinge function and the logit function in logistic regression. The limit of the coherence function at zero temperature corresponds to the hinge function, and the limit of the minimizer of its expected error is the minimizer of the expected error of the hinge loss. We refer to the use of the coherence function in large-margin classification as "$\mathcal{C}$-*learning*," and we present efficient coordinate descent algorithms for the training of regularized $\mathcal{C}$-learning models.

Keywords: large-margin classifiers, hinge functions, logistic functions, coherence functions, $\mathcal{C}$-learning

1. Introduction

Large-margin classification methods have become increasingly popular since the advent of boosting (Freund, 1995), support vector machines (SVM) (Vapnik, 1998) and their variants such as ψ-learning (Shen et al., 2003). Large-margin classification methods are typically devised based on a majorization-minimization procedure, which approximately solves an otherwise intractable optimization problem defined with the 0-1 loss. For example, the conventional SVM employs a hinge loss, the AdaBoost algorithm employs the exponential loss, and ψ-learning employs a so-called ψ-loss, as majorizations of the 0-1 loss.

Large-margin classification methods can be unified using the tools of regularization theory; that is, they can be expressed as the form of "loss" + "penalty" (Hastie et al., 2001). Sparseness has also emerged as a significant theme generally associated with large-margin methods. Typical approaches for achieving sparseness are to use either a non-differentiable penalty or a non-differentiable loss.

Recent developments in the former vein focus on the use of the ℓ_1 penalty (Tibshirani, 1996) or the elastic-net penalty (a mixture of the ℓ_1 and ℓ_2 penalties) (Zou and Hastie, 2005) instead of the ℓ_2 penalty which is typically used in large-margin classification methods. As for non-differentiable losses, the paradigm case is the hinge loss function that is used for the SVM and which leads to a sparse expansion of the discriminant function.

Unfortunately, the conventional SVM does not directly estimate a conditional class probability. Thus, the conventional SVM is unable to provide estimates of uncertainty in its predictions—an important desideratum in real-world applications. Moreover, the non-differentiability of the hinge loss also makes it difficult to extend the conventional SVM to multi-class classification problems. Thus, one seemingly natural approach to constructing a classifier for the binary and multi-class problems is to consider a smooth loss function, while an appropriate penalty is employed to maintain the sparseness of the classifier. For example, regularized logistic regression models based on logit losses (Friedman et al., 2010) are competitive with SVMs.

Of crucial concern are the statistical properties (Lin, 2002; Bartlett et al., 2006; Zhang, 2004) of the majorization function for the original 0-1 loss function. In particular, we analyze the statistical properties of extant majorization functions, which are built on the exponential, logit and hinge functions. This analysis inspires us to propose a new majorization function, which we call a *coherence function* due to a connection with statistical mechanics. We also define a loss function that we refer to as $\mathcal{C}$-loss based on the coherence function.

The $\mathcal{C}$-loss is smooth and convex, and it satisfies a Fisher-consistency condition—a desirable statistical property (Bartlett et al., 2006; Zhang, 2004). The $\mathcal{C}$-loss has the advantage over the hinge loss that it provides an estimate of the conditional class probability, and over the logit loss that one limiting version of it is just the hinge loss. Thus, the $\mathcal{C}$-loss as well as the coherence function have several desirable properties in the context of large-margin classifiers.

In this paper we show how the coherence function can be used to develop an effective approach to estimating the class probability of the conventional binary SVM. Platt (1999) first exploited a sigmoid link function to map the SVM outputs into probabilities, while Sollich (2002) used logarithmic scoring rules (Bernardo and Smith, 1994) to transform the hinge loss into the negative of a conditional log-likelihood (i.e., a predictive class probability). Recently, Wang et al. (2008) developed an interval estimation method. Theoretically, Steinwart (2003) and Bartlett and Tewari (2007) showed that the class probability can be asymptotically estimated by replacing the hinge loss with a differentiable loss. Our approach also appeals to asymptotics to derive a method for estimating the class probability of the conventional binary SVM.

Using the $\mathcal{C}$-loss, we devise new large-margin classifiers which we refer to as *$\mathcal{C}$-learning*. To maintain sparseness, we use the elastic-net penalty in addition to $\mathcal{C}$-learning. We in particular propose two versions. The first version is based on reproducing kernel Hilbert spaces (RKHSs) and it can automatically select the number of support vectors via penalization. The second version focuses on the selection of features again via penalization. The classifiers are trained by coordinate descent algorithms developed by Friedman et al. (2010) for generalized linear models.

The rest of this paper is organized as follows. In Section 2 we summarize the fundamental basis of large-margin classification. Section 3 presents $\mathcal{C}$-loss functions and their mathematical properties. Section 4 gives a method for class probability estimation of the conventional SVM outputs and Section 5 studies our $\mathcal{C}$-learning algorithms. We conduct an experimental analysis in Section 6 and conclude our work in Section 7. All proofs are deferred to the appendix.

2. Large-Margin Classifiers

We consider a *binary* classification problem with a set of training data $\mathcal{T} = \{\mathbf{x}_i, y_i\}_1^n$, where $\mathbf{x}_i \in \mathcal{X} \subset \mathbb{R}^d$ is an input vector and $y_i \in \mathcal{Y} = \{1, -1\}$ is the corresponding class label. Our goal is to find a decision function $f(\mathbf{x})$ over a measurable function class $\mathcal{F}$. Once such an $f(\mathbf{x})$ is obtained, the classification rule is $y = \mathrm{sign}(f(\mathbf{x}))$ where $\mathrm{sign}(a) = 1, 0, -1$ according to $a > 0$, $a = 0$ or $a < 0$. Thus, we have that $\mathbf{x}$ is misclassified if and only if $yf(\mathbf{x}) \leq 0$ (here we ignore the case that $f(\mathbf{x}) = 0$).

Let $\eta(\mathbf{x}) = \Pr(Y = 1|X = \mathbf{x})$ be the conditional probability of class 1 given $\mathbf{x}$ and let $P(X,Y)$ be the probability distribution over $\mathcal{X} \times \mathcal{Y}$. For a measurable decision function $f(\mathbf{x}) : \mathcal{X} \to \mathbb{R}$, the expected error at $\mathbf{x}$ is then defined by

$$\Psi(f(\mathbf{x})) = \mathbb{E}(I_{[Yf(X)\leq 0]}|X = \mathbf{x}) = I_{[f(\mathbf{x})\leq 0]}\eta(\mathbf{x}) + I_{[f(\mathbf{x})>0]}(1-\eta(\mathbf{x})),$$

where $I_{[\#]} = 1$ if # is true and 0 otherwise. The generalization error is

$$\Psi_f = \mathbb{E}_P I_{[Yf(X)\leq 0]} = \mathbb{E}_X\left[I_{[f(X)\leq 0]}\eta(X) + I_{[f(X)>0]}(1-\eta(X))\right],$$

where the expectation $\mathbb{E}_P$ is taken with respect to the distribution $P(X,Y)$ and $\mathbb{E}_X$ denotes the expectation over the input data $\mathcal{X}$. The optimal Bayes error is $\hat{\Psi} = \mathbb{E}_P I_{[Y(2\eta(X)-1)\leq 0]}$, which is the minimum of Ψ_f with respect to measurable functions f.

A classifier is a classification algorithm which finds a measurable function $f_{\mathcal{T}} : \mathcal{X} \to \mathbb{R}$ based on the training data $\mathcal{T}$. We assume that the $(\mathbf{x}_i, y_i)$ in $\mathcal{T}$ are independent and identically distributed from $P(X,Y)$. A classifier is said to be *universally consistent* if

$$\lim_{n\to\infty} \Psi_{f_{\mathcal{T}}} = \hat{\Psi}$$

holds in probability for any distribution P on $\mathcal{X}\times\mathcal{Y}$. It is strongly universally consistent if the condition $\lim_{n\to\infty} \Psi_{f_{\mathcal{T}}} = \hat{\Psi}$ is satisfied almost surely (Steinwart, 2005).

The empirical generalization error on the training data $\mathcal{T}$ is given by

$$\Psi_{emp} = \frac{1}{n}\sum_{i=1}^{n} I_{[y_i f(\mathbf{x}_i)\leq 0]}.$$

Given that the empirical generalization error Ψ_{emp} is equal to its minimum value zero when all training data are correctly classified, we wish to use Ψ_{emp} as a basis for devising classification algorithms. However, the corresponding minimization problem is computationally intractable.

2.1 Surrogate Losses

A wide variety of classifiers can be understood as minimizers of a continuous *surrogate loss* function $\phi(yf(\mathbf{x}))$, which upper bounds the 0-1 loss $I_{[yf(\mathbf{x})\leq 0]}$. Corresponding to $\Psi(f(\mathbf{x}))$ and Ψ_f, we denote $R(f(\mathbf{x})) = \phi(f(\mathbf{x}))\eta(\mathbf{x}) + \phi(-f(\mathbf{x}))(1-\eta(\mathbf{x}))$ and

$$R_f = \mathbb{E}_P[\phi(Yf(X))] = \mathbb{E}_X\left[\phi(f(X))\eta(X) + \phi(-f(X))(1-\eta(X))\right].$$

For convenience, we assume that $\eta \in [0,1]$ and define the notation

$$R(\eta, f) = \eta\phi(f) + (1-\eta)\phi(-f).$$

Exponential Loss	Logit Loss	Hinge Loss	Squared Hinge Loss
$\exp[-yf(\mathbf{x})/2]$	$\log[1+\exp(-yf(\mathbf{x}))]$	$[1-yf(\mathbf{x})]_+$	$([1-yf(\mathbf{x})]_+)^2$

Table 1: Surrogate losses for margin-based classifier.

The surrogate ϕ is said to be Fisher consistent, if for every $\eta \in [0,1]$ the minimizer of $R(\eta, f)$ with respect to f exists and is unique and the minimizer (denoted $\hat{f}(\eta)$) satisfies $\text{sign}(\hat{f}(\eta)) = \text{sign}(\eta - 1/2)$ (Lin, 2002; Bartlett et al., 2006; Zhang, 2004). Since $\text{sign}(u) = 0$ is equivalent to $u = 0$, we have that $\hat{f}(1/2) = 0$. Substituting $\hat{f}(\eta)$ into $R(\eta, f)$, we also define the following notation:

$$\hat{R}(\eta) = \inf_f R(\eta, f) = R(\eta, \hat{f}(\eta)).$$

The difference between $R(\eta, f)$ and $\hat{R}(\eta)$ is

$$\triangle R(\eta, f) = R(\eta, f) - \hat{R}(\eta) = R(\eta, f) - R(\eta, \hat{f}(\eta)).$$

When regarding $f(\mathbf{x})$ and $\eta(\mathbf{x})$ as functions of $\mathbf{x}$, it is clear that $\hat{f}(\eta(\mathbf{x}))$ is the minimizer of $R(f(\mathbf{x}))$ among all measurable function class $\mathcal{F}$. That is,

$$\hat{f}(\eta(\mathbf{x})) = \operatorname*{argmin}_{f(\mathbf{x}) \in \mathcal{F}} R(f(\mathbf{x})).$$

In this setting, the difference between R_f and $\mathbb{E}_X[R(\hat{f}(\eta(X)))]$ (denoted $R_{\hat{f}}$) is given by

$$\triangle R_f = R_f - R_{\hat{f}} = \mathbb{E}_X \triangle R(\eta(X), f(X)).$$

If $\hat{f}(\eta)$ is invertible, then the inverse function $\hat{f}^{-1}(f(\mathbf{x}))$ over $\mathcal{F}$ can be regarded as a class-conditional probability estimate given that $\eta(\mathbf{x}) = \hat{f}^{-1}(\hat{f}(\mathbf{x}))$. Moreover, Zhang (2004) showed that $\triangle R_f$ is the expected distance between the conditional probability $\hat{f}^{-1}(f(\mathbf{x}))$ and the true conditional probability $\eta(\mathbf{x})$. Thus, minimizing R_f is equivalent to minimizing the expected distance between $\hat{f}^{-1}(f(\mathbf{x}))$ and $\eta(\mathbf{x})$.

Table 1 lists four common surrogate functions used in large-margin classifiers. Here $[u]_+ = \max\{u, 0\}$ is a so-called hinge function and $([u]_+)^2 = (\max\{u, 0\})^2$ is a squared hinge function which is used for developing the ℓ_2-SVM (Cristianini and Shawe-Taylor, 2000). Note that we typically scale the logit loss to equal 1 at $yf(\mathbf{x}) = 0$. These functions are convex and the upper bounds of $I_{[yf(\mathbf{x}) \leq 0]}$. Moreover, they are Fisher consistent. In particular, the following result has been established by Friedman et al. (2000) and Lin (2002).

Proposition 1 *Assume that* $0 < \eta(\mathbf{x}) < 1$ *and* $\eta(\mathbf{x}) \neq 1/2$. *Then, the minimizers of* $\mathbb{E}(\exp[-Yf(X)/2]|X = \mathbf{x})$ *and* $\mathbb{E}(\log[1 + \exp(-Yf(X))]|X = \mathbf{x})$ *are both* $\hat{f}(\mathbf{x}) = \log \frac{\eta(\mathbf{x})}{1-\eta(\mathbf{x})}$, *the minimizer of* $\mathbb{E}\big([1 - Yf(X)]_+|X = \mathbf{x}\big)$ *is* $\hat{f}(\mathbf{x}) = \text{sign}(\eta(\mathbf{x}) - 1/2)$, *and the minimizer of* $\mathbb{E}\big(([1 - Yf(X)]_+)^2|X = \mathbf{x}\big)$ *is* $\hat{f}(\mathbf{x}) = 2\eta(\mathbf{x}) - 1$.

When the exponential or logit loss function is used, $\hat{f}^{-1}(f(\mathbf{x}))$ exists. It is clear that $\eta(\mathbf{x}) = \hat{f}^{-1}(\hat{f}(\mathbf{x}))$. For any $f(\mathbf{x}) \in \mathcal{F}$, we denote the inverse function by $\tilde{\eta}(\mathbf{x})$, which is

$$\tilde{\eta}(\mathbf{x}) = \hat{f}^{-1}(f(\mathbf{x})) = \frac{1}{1 + \exp(-f(\mathbf{x}))}.$$

Unfortunately, the minimization of the hinge loss (which is the basis of the SVM) does not yield a class probability estimate (Lin et al., 2002).

2.2 The Regularization Approach

Given a surrogate loss function ϕ, a large-margin classifier typically solves the following optimization problem:

$$\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \phi(y_i f(\mathbf{x}_i)) + \gamma J(h),$$

where $f(\mathbf{x}) = \alpha + h(\mathbf{x})$, $J(h)$ is a regularization term to penalize model complexity and γ is the degree of penalization.

Suppose that $f = \alpha + h \in (\{1\} + \mathcal{H}_K)$ where $\mathcal{H}_K$ is a reproducing kernel Hilbert space (RKHS) (Wahba, 1990) induced by a reproducing kernel $K(\cdot,\cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Finding $f(\mathbf{x})$ is then formulated as a regularization problem of the form

$$\min_{f \in \mathcal{H}_K} \left\{ \frac{1}{n} \sum_{i=1}^{n} \phi(y_i f(\mathbf{x}_i)) + \frac{\gamma}{2} \|h\|^2_{\mathcal{H}_K} \right\}, \tag{1}$$

where $\|h\|^2_{\mathcal{H}_K}$ is the RKHS norm. By the representer theorem, the solution of (1) is of the form

$$f(\mathbf{x}_i) = \alpha + \sum_{j=1}^{n} \beta_j K(\mathbf{x}_i, \mathbf{x}_j) = \alpha + \mathbf{k}_i' \beta, \tag{2}$$

where $\beta = (\beta_1, \ldots, \beta_n)'$ and $\mathbf{k}_i = (K(\mathbf{x}_i, \mathbf{x}_1), \ldots, K(\mathbf{x}_i, \mathbf{x}_n))'$. Noticing that $\|h\|^2_{\mathcal{H}_K} = \sum_{i,j=1}^{n} K(\mathbf{x}_i, \mathbf{x}_j)\beta_i\beta_j$ and substituting (2) into (1), we obtain the minimization problem with respect to α and β as

$$\min_{\alpha, \beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \phi(y_i(\alpha + \mathbf{k}_i'\beta)) + \frac{\gamma}{2} \beta' \mathbf{K} \beta \right\},$$

where $\mathbf{K} = [\mathbf{k}_1, \ldots, \mathbf{k}_n]$ is the $n \times n$ kernel matrix. Since $\mathbf{K}$ is symmetric and positive semidefinite, the term $\beta'\mathbf{K}\beta$ is in fact an empirical RKHS norm on the training data.

In particular, the conventional SVM defines the surrogate $\phi(\cdot)$ as the hinge loss and solves the following optimization problem:

$$\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} [1 - y_i(\alpha + \mathbf{k}_i'\beta)]_+ + \frac{\gamma}{2} \beta' \mathbf{K} \beta. \tag{3}$$

In this paper, we are especially interested in *universal kernels*, namely, kernels whose induced RKHS is dense in the space of continuous functions over $\mathcal{X}$ (Steinwart, 2001). The Gaussian RBF kernel is such an example.

2.3 Methods for Class Probability Estimation of SVMs

Let $\hat{f}(\mathbf{x})$ be the solution of the SVM problem in (3). In an attempt to address the problem of class probability estimation for SVMs, Sollich (2002) proposed a class probability estimate

$$\hat{\eta}(\mathbf{x}) = \begin{cases} \frac{1}{1+\exp(-2\hat{f}(\mathbf{x}))} & \text{if } |\hat{f}(\mathbf{x})| \leq 1, \\ \frac{1}{1+\exp[-(\hat{f}(\mathbf{x})+\text{sign}(\hat{f}(\mathbf{x})))]} & \text{otherwise.} \end{cases}$$

This class probability was also used in the derivation of a so-called complete SVM studied by Mallick et al. (2005).

Another proposal for obtaining class probabilities from SVM outputs was developed by Platt (1999), who employed a post-processing procedure based on the parametric formula

$$\hat{\eta}(\mathbf{x}) = \frac{1}{1+\exp(A\hat{f}(\mathbf{x})+B)},$$

where the parameters A and B are estimated via the minimization of the empirical cross-entropy error over the training data set.

Wang et al. (2008) proposed a nonparametric form obtained from training a sequence of weighted classifiers:

$$\min_{f\in\mathcal{F}} \frac{1}{n}\Big\{(1-\pi_j)\sum_{y_i=1}[1-y_i f(\mathbf{x}_i)]_+ + \pi_j \sum_{y_i=-1}[1-y_i f(\mathbf{x}_i)]_+\Big\} + \gamma J(h) \tag{4}$$

for $j=1,\ldots,m{+}1$ such that $0=\pi_1<\cdots<\pi_{m+1}=1$. Let $\hat{f}_{\pi_j}(\mathbf{x})$ be the solution of (4). The estimated class probability is then $\hat{\eta}(\mathbf{x})=\frac{1}{2}(\pi_*+\pi^*)$ where $\pi_*=\min\{\pi_j : \mathrm{sign}(\hat{f}_{\pi_j}(\mathbf{x}))=-1\}$ and $\pi^*=\max\{\pi_j : \mathrm{sign}(\hat{f}_{\pi_j}(\mathbf{x}))=1\}$.

Additional contributions are due to Steinwart (2003) and Bartlett and Tewari (2007). These authors showed that the class probability can be asymptotically estimated by replacing the hinge loss with various differentiable losses.

3. Coherence Functions

In this section we present a smooth and Fisher-consistent majorization loss, which bridges the hinge loss and the logit loss. We will see that one limit of this loss is equal to the hinge loss. Thus, it is applicable to the asymptotical estimate of the class probability for the conventional SVM as well as the construction of margin-based classifiers, which will be presented in Section 4 and Section 5.

3.1 Definition

Under the $0-1$ loss the misclassification costs are specified to be one, but it is natural to set the misclassification costs to be a positive constant $u>0$. The empirical generalization error on the training data is given in this case by

$$\frac{1}{n}\sum_{i=1}^{n} u I_{[y_i f(\mathbf{x}_i)\leq 0]},$$

where $u>0$ is a constant that represents the misclassification cost. In this setting we can extend the hinge loss as

$$H_u(yf(\mathbf{x})) = [u{-}yf(\mathbf{x})]_+.$$

It is clear that $H_u(yf(\mathbf{x})) \geq uI_{[yf(\mathbf{x})\leq 0]}$. This implies that $H_u(yf(\mathbf{x}))$ is a majorization of $uI_{[yf(\mathbf{x})\leq 0]}$.

We apply the maximum entropy principle to develop a smooth surrogate of the hinge loss $[u{-}z]_+$. In particular, noting that $[u{-}z]_+=\max\{u{-}z,\, 0\}$, we maximize $w(u{-}z)$ with respect to $w\in(0,1)$ under the entropy constraint; that is,

$$\max_{w\in(0,1)} \Big\{F = w(u{-}z) - \rho\big[w\log w + (1-w)\log(1-w)\big]\Big\},$$

where $-[w\log w+(1{-}w)\log(1{-}w)]$ is the entropy and $\rho>0$, a Lagrange multiplier, plays the role of temperature in thermodynamics.

The maximum of F is

$$V_{\rho,u}(z) = \rho \log \big[1+\exp\frac{u-z}{\rho}\big] \tag{5}$$

at $w = \exp((u-z)/\rho)/[1+\exp((u-z)/\rho)]$. We refer to functions of this form as *coherence functions* because their properties (detailed in the next subsection) are similar to statistical mechanical properties of deterministic annealing (Rose et al., 1990).

We also consider a scaled variant of $V_{\rho,u}(z)$:

$$C_{\rho,u}(z) = \frac{u}{\log[1+\exp(u/\rho)]} \log \big[1+\exp\frac{u-z}{\rho}\big],\ \rho > 0,\ u > 0, \tag{6}$$

which has the property that $C_{\rho,u}(z) = u$ when $z = 0$. Recall that u as a misclassification cost should be specified as a positive value. However, both $C_{\rho,0}(z)$ and $V_{\rho,0}(z)$ are well defined mathematically. Since $C_{\rho,0}(z) = 0$ is a trivial case, we always assume that $u > 0$ for $C_{\rho,u}(z)$ here and later. In the binary classification problem, z is defined as $yf(\mathbf{x})$. In the special case that $u = 1$, $C_{\rho,1}(yf(\mathbf{x}))$ can be regarded as a smooth alternative to the SVM hinge loss $[1-yf(\mathbf{x})]_+$. We refer to $C_{\rho,u}(yf(\mathbf{x}))$ as $\mathcal{C}$-losses.

It is worth noting that $V_{1,0}(z)$ is the logistic function and $V_{\rho,0}(z)$ has been proposed by Zhang and Oles (2001) for binary logistic regression. We keep in mind that $u \geq 0$ for $V_{\rho,u}(z)$ through this paper.

3.2 Properties

It is obvious that $C_{\rho,u}(z)$ and $V_{\rho,u}$ are infinitely smooth with respect to z. Moreover, the first-order and second-order derivatives of $C_{\rho,u}(z)$ with respect to z are given as

$$C'_{\rho,u}(z) = -\frac{u}{\rho \log[1+\exp(u/\rho)]} \frac{\exp\frac{u-z}{\rho}}{1+\exp\frac{u-z}{\rho}},$$

$$C''_{\rho,u}(z) = \frac{u}{\rho^2 \log[1+\exp(u/\rho)]} \frac{\exp\frac{u-z}{\rho}}{(1+\exp\frac{u-z}{\rho})^2}.$$

Since $C''_{\rho,u}(z) > 0$ for any $z \in \mathbb{R}$, $C_{\rho,u}(z)$ as well as $V_{\rho,u}(z)$ are strictly convex in z, for fixed $\rho > 0$ and $u > 0$.

We now investigate relationships among the coherence functions and hinge losses. First, we have the following properties.

Proposition 2 *Let $V_{\rho,u}(z)$ and $C_{\rho,u}(z)$ be defined by (5) and (6). Then,*

(i) $u \times I_{[z \leq 0]} \leq [u-z]_+ \leq V_{\rho,u}(z) \leq \rho \log 2 + [u-z]_+$;

(ii) $\frac{1}{2}(u-z) \leq V_{\rho,u}(z) - \rho \log 2$;

(iii) $\lim_{\rho \to 0} V_{\rho,u}(z) = [u-z]_+$ *and* $\lim_{\rho \to \infty} V_{\rho,u}(z) - \rho \log 2 = \frac{1}{2}(u-z)$;

(iv) $u \times I_{[z \leq 0]} \leq C_{\rho,u}(z) \leq V_{\rho,u}(z)$;

(v) $\lim_{\rho \to 0} C_{\rho,u}(z) = [u-z]_+$ *and* $\lim_{\rho \to \infty} C_{\rho,u}(z) = u$*, for* $u > 0$.

As a special case of $u = 1$, we have $C_{\rho,1}(z) \geq I_{[z\leq 0]}$. Moreover, $C_{\rho,1}(z)$ approaches $(1-z)_+$ as $\rho \to 0$. Thus, $C_{\rho,1}(z)$ is a majorization of $I_{[z\leq 0]}$.

As we mentioned earlier, $V_{\rho,0}(z)$ are used to devise logistic regression models. We can see from Proposition 2 that $V_{\rho,0}(z) \geq [-z]_+$, which implies that a logistic regression model is possibly no longer a large-margin classifier. Interestingly, however, we consider a variant of $V_{\rho,u}(z)$ as

$$L_{\rho,u}(z) = \frac{1}{\log(1+\exp(u/\rho))} \log\left[1+\exp((u-z)/\rho)\right],\ \rho > 0,\ u \geq 0,$$

which always satisfies that $L_{\rho,u}(z) \geq I_{[z\leq 0]}$ and $L_{\rho,u}(0) = 1$, for any $u \geq 0$. Thus, the $L_{\rho,u}(z)$ for $\rho > 0$ and $u \geq 0$ are majorizations of $I_{[z\leq 0]}$. In particular, $L_{\rho,1}(z) = C_{\rho,1}(u)$ and $L_{1,0}(z)$ is the logit function.

In order to explore the relationship of $C_{\rho,u}(z)$ with $(u-z)_+$, we now consider some properties of $L_{\rho,u}(z)$ when regarding it respectively as a function of ρ and of u.

Proposition 3 *Assume* $\rho > 0$ *and* $u \geq 0$. *Then,*

(i) $L_{\rho,u}(z)$ *is a decreasing function in* ρ *if* $z < 0$, *and it is an increasing function in* ρ *if* $z \geq 0$;

(ii) $L_{\rho,u}(z)$ *is a decreasing function in* u *if* $z < 0$, *and it is an increasing function in* u *if* $z \geq 0$.

Results similar to those in Proposition 3-(i) also apply to $C_{\rho,u}(z)$ because of $C_{\rho,u}(z) = uL_{\rho,u}(z)$. Then, according to Proposition 2-(v), we have that $u = \lim_{\rho\to+\infty} C_{\rho,u}(z) \leq C_{\rho,u}(z) \leq \lim_{\rho\to 0} C_{\rho,u}(z) = (u-z)_+$ if $z < 0$ and $(u-z)_+ = \lim_{\rho\to 0} C_{\rho,u}(z) \leq C_{\rho,u}(z) \leq \lim_{\rho\to+\infty} C_{\rho,u}(z) = u$ if $z \geq 0$. It follows from Proposition 3-(ii) that $C_{\rho,1}(z) = L_{\rho,1}(z) \leq L_{\rho,0}(z)$ if $z < 0$ and $C_{\rho,1}(z) = L_{\rho,1}(z) \geq L_{\rho,0}(z)$ if $z \geq 0$. In addition, it is easily seen that $(1-z)_+ \geq ((1-z)_+)^2$ if $z \geq 0$ and $(1-z)_+ \leq ((1-z)_+)^2$ otherwise. We now obtain the following proposition:

Proposition 4 *Assume* $\rho > 0$. *Then,* $C_{\rho,1}(z) \leq \min\left\{L_{\rho,0}(z),\ [1-z]_+,\ ([1-z]_+)^2\right\}$ *if* $z < 0$, *and* $C_{\rho,1}(z) \geq \max\left\{L_{\rho,0}(z),\ [1-z]_+,\ ([1-z]_+)^2\right\}$ *if* $z \geq 0$.

This proposition is depicted in Figure 1. Owing to the relationships of the $\mathcal{C}$-loss $C_{\rho,1}(yf(x))$ with the hinge and logit losses, it is potentially useful in devising new large-margin classifiers.

We now turn to the derivatives of $C_{\rho,u}(z)$ and $(u-z)_+$. It is immediately verified that $-1 \leq C'_{\rho,u}(z) \leq 0$. Moreover, we have

$$\lim_{\rho\to 0} C'_{\rho,u}(z) = \lim_{\rho\to 0} V'_{\rho,u}(z) = \begin{cases} 0 & z > u, \\ -\frac{1}{2} & z = u, \\ -1 & z < u. \end{cases}$$

Note that $(u-z)'_+ = -1$ if $z < u$ and $(u-z)'_+ = 0$ if $z > u$. Furthermore, $\partial(u-z)_+|_{z=u} = [-1,0]$ where $\partial(u-z)_+|_{z=u}$ denotes the subdifferential of $(u-z)_+$ at $z = u$. Hence,

Proposition 5 *For a fixed* $u > 0$, *we have that* $\lim_{\rho\to 0} C'_{\rho,u}(z) \in \partial(u-z)_+$.

This proposition again establishes a connection of the hinge loss with the limit of $C_{\rho,u}(z)$ at $\rho = 0$. Furthermore, we obtain from Propositions 2 and 5 that $\partial(u-z)_+ = \partial \lim_{\rho\to 0} C_{\rho,u}(z) \ni \lim_{\rho\to 0} \partial C_{\rho,u}(z)$.

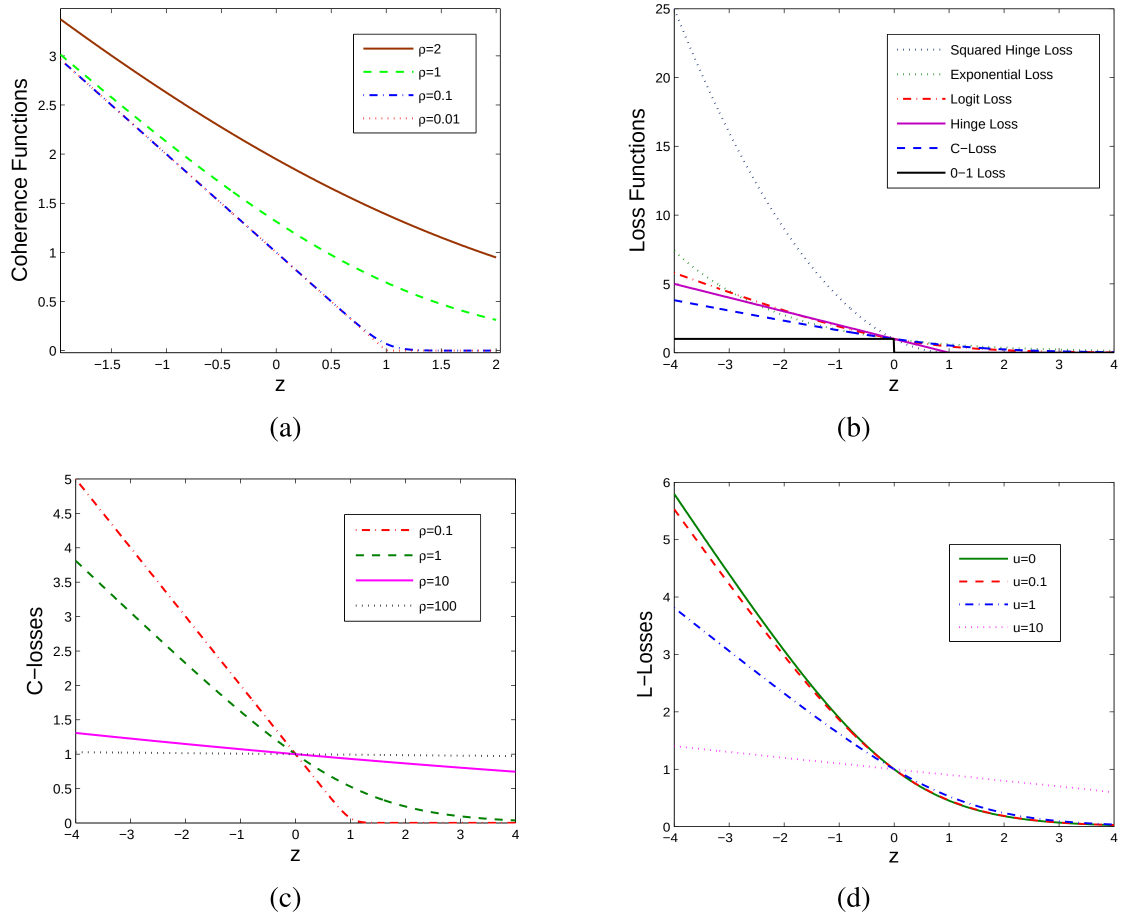


Figure 1: These functions are regarded as a function of $z = yf(\mathbf{x})$. (a) Coherence functions $V_{\rho,1}(z)$ with $\rho = 0.01$, $\rho = 0.1$, $\rho = 1$ and $\rho = 2$. (b) A variety of majorization loss functions, $\mathcal{C}$*-loss*: $C_{1,1}(z)$; *Logit loss*: $L_{1,0}(z)$; *Exponential loss*: $\exp(-z/2)$; *Hinge loss*: $[1-z]_+$; *Squared Hinge Loss*: $([1-z]_+)^2$. (c) $C_{\rho,1}(z)$ (or $L_{\rho,1}(z)$) with $\rho = 0.1$, $\rho = 1$, $\rho = 10$ and $\rho = 100$ (see Proposition 3-(i)). (d) $L_{1,u}(z)$ with $u = 0$, $u = 0.1$, $u = 1$ and $u = 10$ (see Proposition 3-(ii)).

3.3 Consistency in Classification Methods

We now apply the coherence function to the development of classification methods. Recall that $C'_{\rho,u}(0)$ exists and is negative. Thus, the $\mathcal{C}$-loss $C_{\rho,u}(yf(x))$ is Fisher-consistent (or classification calibrated) (Bartlett et al., 2006). In particular, we have the following theorem.

Theorem 6 *Assume* $0 < \eta < 1$ *and* $\eta \neq \frac{1}{2}$. *Consider the optimization problem*

$$\min_{f \in \mathbb{R}} R(f, \eta) := V_{\rho,u}(f)\eta + V_{\rho,u}(-f)(1-\eta)$$

for fixed $\rho > 0$ *and* $u \geq 0$. *Then, the minimizer is unique and is given by*

$$f_*(\eta) = \rho \log \frac{(2\eta - 1)\exp(\frac{u}{\rho}) + \sqrt{(1-2\eta)^2 \exp(\frac{2u}{\rho}) + 4\eta(1-\eta)}}{2(1-\eta)}. \tag{7}$$

Moreover, we have $f_* > 0$ *if and only if* $\eta > 1/2$. *Additionally, the inverse function* $f_*^{-1}(f)$ *exists and it is given by*

$$\tilde{\eta}(f) := f_*^{-1}(f) = \frac{1 + \exp(\frac{f-u}{\rho})}{1 + \exp(-\frac{u+f}{\rho}) + 1 + \exp(\frac{f-u}{\rho})}, \; for \; f \in \mathbb{R}. \tag{8}$$

The minimizer $f_*(\mathbf{x})$ of $R(f(\mathbf{x})) := \mathbb{E}(V_{\rho,u}(Yf(X))|X = \mathbf{x})$ and its inverse $\tilde{\eta}(\mathbf{x})$ are immediately obtained by replacing f with $f(\mathbf{x})$ in (7) and (8). Since for $u > 0$ the minimizers of $\mathbb{E}(C_{\rho,u}(Yf(X))|X = \mathbf{x})$ and $\mathbb{E}(V_{\rho,u}(Yf(X))|X = \mathbf{x})$ are the same, this theorem shows that $C_\rho(yf(\mathbf{x}), u)$ is also Fisher-consistent. We see from Theorem 6 that the explicit expressions of $f_*(\mathbf{x})$ and its inverse $\tilde{\eta}(\mathbf{x})$ exist. In the special case that $u = 0$, we have $f_*(\mathbf{x}) = \rho \log \frac{\eta(\mathbf{x})}{1-\eta(\mathbf{x})}$ and $\tilde{\eta}(\mathbf{x}) = \frac{1}{1+\exp(-f(\mathbf{x})/\rho)}$. Furthermore, when $\rho = 1$, as expected, we recover logistic regression. In other words, the result is identical with that in Proposition 1 for logistic regression.

We further consider properties of $f_*(\eta)$. In particular, we have the following proposition.

Proposition 7 *Let* $f_*(\eta)$ *be defined by (7). Then,*

(i) $\mathrm{sign}(f_*(\eta)) = \mathrm{sign}(\eta - 1/2)$.

(ii) $\lim_{\rho \to 0} f_*(\eta) = u \times \mathrm{sign}(\eta - 1/2)$.

(iii) $f_*'(\eta) = \frac{df_*(\eta)}{d\eta} \geq \frac{\rho}{\eta(1-\eta)}$ *with equality if and only if* $u = 0$.

Proposition 7-(i) shows that the classification rule with $f_*(\mathbf{x})$ is equivalent to the Bayes rule. In the special case that $u = 1$, we have from Proposition 7-(ii) that $\lim_{\rho \to 0} f_*(\mathbf{x}) = \mathrm{sign}(\eta(\mathbf{x}) - 1/2)$. This implies that the current $f_*(\mathbf{x})$ approaches the solution of $\mathbb{E}((1 - Yf(X))_+|X = \mathbf{x})$, which corresponds to the conventional SVM method (see Proposition 1).

We now treat $\tilde{\eta}(f)$ as a function of ρ. The following proposition is easily proven.

Proposition 8 *Let* $\tilde{\eta}(f)$ *be defined by (8). Then, for fixed* $f \in \mathbb{R}$ *and* $u > 0$, $\lim_{\rho \to \infty} \tilde{\eta}(f) = \frac{1}{2}$ *and*

$$\lim_{\rho \to 0} \tilde{\eta}(f) = \begin{cases} 1 & if\; f > u, \\ \frac{2}{3} & if\; f = u, \\ \frac{1}{2} & if\; -u < f < u, \\ \frac{1}{3} & if\; f = -u, \\ 0 & if\; f < -u. \end{cases}$$

As we discuss in the previous subsection, $V_{\rho,u}(z)$ is obtained when $w = \exp((u - z)/\rho)/(1 + \exp((u - z)/\rho))$ by using the maximum entropy principle. Let $z = yf(\mathbf{x})$. We further write w as $w_1(f) = 1/[1 + \exp((f - u)/\rho)]$ when $y = 1$ and w as $w_2(f) = 1/[1 + \exp(-(f + u)/\rho)]$ when $y = -1$.

We now explore the relationship of $\tilde{\eta}(f)$ with $w_1(f)$ and $w_2(f)$. Interestingly, we first find that

$$\tilde{\eta}(f) = \frac{w_2(f)}{w_1(f)+w_2(f)}.$$

It is easily proven that $w_1(f)+w_2(f) \geq 1$ with equality if and only if $u=0$. We thus have that $\tilde{\eta}(f) \leq w_2(f)$, with equality if and only if $u=0$; that is, the loss becomes logit function $V_{\rho,0}(z)$. Note that $w_2(f)$ represents the probability of the event $\{u+f>0\}$ and $\tilde{\eta}(f)$ represents the probability of the event $\{f>0\}$. Since the event $\{f>0\}$ is a subset of the event $\{u+f>0\}$, we have $\tilde{\eta}(f) \leq w_2(f)$. Furthermore, the statement that $\tilde{\eta}(f) = w_2(f)$ if and only if $u=0$ is equivalent to $\{u+f>0\} = \{f>0\}$ if and only if $u=0$. This implies that only the logit loss induces $\tilde{\eta}(f) = w_2(f)$.

As discussed in Section 2.1, $\tilde{\eta}(\mathbf{x})$ can be regarded as a reasonable estimate of the true class probability $\eta(\mathbf{x})$. Recall that $\triangle R(\eta, f) = R(\eta, f) - R(\eta, f_*(\eta))$ and $\triangle R_f = \mathbb{E}_X[\triangle R(\eta(X), f(X))]$ such that $\triangle R_f$ can be viewed as the expected distance between $\tilde{\eta}(\mathbf{x})$ and $\eta(\mathbf{x})$.

For an arbitrary fixed $f \in \mathbb{R}$, we have

$$\triangle R(\eta, f) = R(\eta, f) - R(\eta, f_*(\eta)) = \eta\rho \log \frac{1+\exp\frac{u-f}{\rho}}{1+\exp\frac{u-f_*(\eta)}{\rho}} + (1-\eta)\rho \log \frac{1+\exp\frac{u+f}{\rho}}{1+\exp\frac{u+f_*(\eta)}{\rho}}.$$

The first-order derivative of $\triangle R(\eta, f)$ with respect to η is

$$\frac{d\triangle R(\eta, f)}{d\eta} = \rho \log \frac{1+\exp\frac{u-f}{\rho}}{1+\exp\frac{u-f_*(\eta)}{\rho}} - \rho \log \frac{1+\exp\frac{u+f}{\rho}}{1+\exp\frac{u+f_*(\eta)}{\rho}}.$$

The Karush-Kuhn-Tucker (KKT) condition for the minimization problem is as follows:

$$\eta \frac{\exp\frac{u-f_*(\eta)}{\rho}}{1+\exp\frac{u-f_*(\eta)}{\rho}} + (1-\eta)\frac{\exp\frac{u+f_*(\eta)}{\rho}}{1+\exp\frac{u+f_*(\eta)}{\rho}} = 0,$$

and the second-order derivative of $\triangle R(\eta, f)$ with respect to η is given by

$$\frac{d^2\triangle R(\eta, f)}{d\eta^2} = \Big(\frac{1}{1+\exp(-\frac{u-f_*(\eta)}{\rho})} + \frac{1}{1+\exp(-\frac{u+f_*(\eta)}{\rho})}\Big) f_*'(\eta) = \big[w_1(f_*(\eta)) + w_2(f_*(\eta))\big] f_*'(\eta).$$

According to Proposition 7-(iii) and using $w_1(f_*(\eta)) + w_2(f_*(\eta)) \geq 1$, we have

$$\frac{d^2\triangle R(\eta, f)}{d\eta^2} \geq \frac{\rho}{\eta(1-\eta)},$$

with equality if and only if $u=0$. This implies $\frac{d^2\triangle R(\eta, f)}{d\eta^2} > 0$. Thus, for a fixed f, $\triangle R(\eta, f)$ is strictly convex in η. Subsequently, we have that $\triangle R(\eta, f) \geq 0$ with equality $\eta = \tilde{\eta}$, or equivalently, $f = f_*$.

Using the Taylor expansion of $\triangle R(\eta, f)$ at $\tilde{\eta} := \tilde{\eta}(f) = f_*^{-1}(f)$, we thus obtain a lower bound for $\triangle R(\eta, f)$; namely,

$$\begin{aligned}\triangle R(\eta, f) &= \triangle R(\tilde{\eta}, f) - \frac{d\triangle R(\tilde{\eta}, f)}{d\eta}(\eta - \tilde{\eta}) + \frac{1}{2}\frac{d^2\triangle R(\bar{\eta}, f)}{d\eta^2}(\eta - \tilde{\eta})^2 \\ &= \frac{1}{2}\frac{d^2\triangle R(\bar{\eta}, f)}{d\eta^2}(\eta - \tilde{\eta})^2 \geq \frac{\rho}{2\bar{\eta}(1-\bar{\eta})}(\eta - \tilde{\eta})^2 \geq 2\rho(\eta - \tilde{\eta})^2,\end{aligned}$$

where $\bar{\eta} \in (\tilde{\eta}, \eta) \subset [0, 1]$. In particular, we have that $\triangle R(\eta, 0) \geq 2\rho(\eta - 0.5)^2$. According to Theorem 2.1 and Corollary 3.1 in Zhang (2004), the following theorem is immediately established.

Theorem 9 *Let $\varepsilon_1 = \inf_{f(\cdot)\in\mathcal{F}} \mathbb{E}_X[\triangle R(\eta(X), f(X))]$, and let $f_*(\mathbf{x}) \in \mathcal{F}$ such that*

$$\mathbb{E}_X[R(\eta(X), f_*(X))] \leq \inf_{f(\cdot)\in\mathcal{F}} \mathbb{E}_X[R(\eta(X), f(X))] + \varepsilon_2$$

for $\varepsilon_2 \geq 0$. Then for $\varepsilon = \varepsilon_1 + \varepsilon_2$,

$$\triangle R_{f_*} = \mathbb{E}_X[\triangle R(\eta(X), f_*(X))] \leq \varepsilon$$

and

$$\Psi_{f_*} \leq \hat{\Psi} + \big(2\varepsilon/\rho\big)^{1/2},$$

where $\Psi_{f_*} = \mathbb{E}_P I_{[Y f_*(X)\leq 0]}$, and $\hat{\Psi} = \mathbb{E}_P I_{[Y(2\eta(X)-1)\leq 0]}$ is the optimal Bayes error.

3.4 Analysis

For notational simplicity, we will use $C_\rho(z)$ for $C_{\rho,1}(z)$. Considering $f(\mathbf{x}) = \alpha + \beta'\mathbf{k}$, we define a regularized optimization problem of the form

$$\min_{\alpha,\beta}\Big\{\frac{1}{n}\sum_{i=1}^{n} C_\rho(y_i f(\mathbf{x}_i)) + \frac{\gamma_n}{2}\beta'\mathbf{K}\beta\Big\}. \tag{9}$$

Here we assume that the regularization parameter γ relies on the number n of training data points, thus we denote it by γ_n.

Since the optimization problem (9) is convex with respect to α and β, the solution exists and is unique. Moreover, since C_ρ is infinitely smooth, we can resort to the Newton-Raphson method to solve (9).

Proposition 10 *Assume that γ_n in (9) and γ in (3) are same. Then the minimizer of (9) approaches the minimizer of (3) as $\rho \to 0$.*

This proposition is obtained directly from Proposition 5. For a fixed ρ, we are also concerned with the universal consistency of the classifier based on (9) with and without the offset term α.

Theorem 11 *Let $K(\cdot,\cdot)$ be a universal kernel on $\mathcal{X}\times\mathcal{X}$. Suppose we are given such a positive sequence $\{\gamma_n\}$ that $\gamma_n \to 0$. If*

$$n\gamma_n^2/\log n \to \infty,$$

then the classifier based on (9) is strongly universally consistent. If

$$n\gamma_n^2 \to \infty,$$

then the classifier based on (9) with $\alpha = 0$ is universally consistent.

Finally, we explore the convergence rate of the classification model in (9). It is worth pointing out that Lin (2002) studied the convergence rate of the ℓ_2-SVM based on Theorem 4.1 in Lin (2000). Additionally, Lin (2002) argued that it is harder for the standard SVM to establish a similar convergence rate result due to two main reasons. The first one is that $(1-z)_+$ is not differentiable and the second one is that the target function $\mathrm{sign}(\eta-1/2)$ is not in the assumed RKHS. We note that Theorem 4.1 in Lin (2000) is elaborated only for the least squares problem. It is also difficult to apply this theorem to our case, although the above two reasons no longer exist in our case. In Section 6.1 we illustrate how our classification model in (9) approaches the corresponding target function given in (7) using the same simulation data set as in Lin (2002). Moreover, under certain conditions, we can have the following theorem about convergence rate.

Theorem 12 *Suppose* $\mathbf{x}$ *takes values in a finite region with density* $p(\mathbf{x})$ *and* $f(\mathbf{x}) = \alpha + h(\mathbf{x})$ *with* α *taking values in a finite open interval. Assume that* h *belongs to a bounded open set of RKHS* $\mathcal{H}_K$ *with positive definite kernel* $K(\cdot,\cdot)$ *and that there exists an* $M_0 > 0$ *such that* $|K(\mathbf{x}_1,\mathbf{x}_2)| < M_0$ *if* $(\mathbf{x}_1,\mathbf{x}_2)$ *are bounded. Let* $\phi(z)$ *be a strictly convex function and twice continuously differentiable. We define*

$$\begin{aligned} f^* &= \operatorname*{argmin}_f \int \phi(yf(\mathbf{x}))dF(\mathbf{x},y), \\ \hat{f}_n &= \operatorname*{argmin}_f \int \phi(yf(\mathbf{x}))dF_n(\mathbf{x},y) + \frac{\gamma}{2}\|h\|^2_{\mathcal{H}_K}. \end{aligned}$$

where $F(\mathbf{x},y)$ *is the distribution of* $(\mathbf{x},y)$ *and* $F_n(\mathbf{x},y)$ *is the empirical distribution of* $(\mathbf{x},y)$*. Then we have*

$$\int |\hat{f}_n(\mathbf{x}) - f^*(\mathbf{x})| p(\mathbf{x}) d\mathbf{x} = O(\gamma) + O_p(\frac{1}{\sqrt{n}}).$$

Obviously, the Gaussian RBF kernel satisfies the condition in Theorem 12. Since $C_\rho(z)$ is strictly convex and infinitely smooth, we can directly apply Theorem 12 to the classification model in (9). In particular, let $\hat{f}_n$ be the minimizer of Problem (9) and $f_*(\eta)$ be defined in (7). Under the conditions in Theorem 12, we have

$$\int \big|\hat{f}_n(\mathbf{x}) - f_*(\eta(\mathbf{x}))\big| p(\mathbf{x}) d\mathbf{x} = O(\gamma) + O_p(\frac{1}{\sqrt{n}}).$$

4. Class Probability Estimation of SVM Outputs

As discussed earlier, the limit of the coherence function, $V_{\rho,1}(yf(\mathbf{x}))$, at $\rho = 0$ is just the hinge loss. Moreover, Proposition 7 shows that the minimizer of $V_{\rho,1}(f)\eta + V_{\rho,1}(-f)(1-\eta)$ approaches that of $H(f)\eta + H(-f)(1-\eta)$ as $\rho \to 0$. Thus, Theorem 6 provides us with an approach to the estimation of the class probability for the conventional SVM.

In particular, let $\hat{f}(\mathbf{x})$ be the solution of the optimization problem (3) for the conventional SVM. In terms of Theorem 6, we suggest that the estimated class probability $\hat{\eta}(\mathbf{x})$ is defined as

$$\hat{\eta}(\mathbf{x}) = \frac{1+\exp(\frac{\hat{f}(\mathbf{x})-1}{\rho})}{1+\exp(-\frac{1+\hat{f}(\mathbf{x})}{\rho}) + 1 + \exp(\frac{\hat{f}(\mathbf{x})-1}{\rho})}. \tag{10}$$

Proposition 7 would seem to motivate setting ρ to a very small value in (10). However, as shown in Proposition 8, the probabilistic outputs degenerate to 0, $1/3$, $1/2$, $2/3$ and 1 in this case. Additionally, the classification function $\hat{f}(\mathbf{x}) = \hat{\alpha} + \sum_{i=1}^{n} \hat{\beta}_i K(\mathbf{x}, \mathbf{x}_i)$ is obtained via fitting a conventional SVM model on the training data. Thus, rather than attempting to specify a fixed value of ρ via a theoretical argument, we instead view it as a hyperparameter to be fit empirically.

In particular, we fit ρ by minimizing the generalized Kullback-Leibler divergence (or cross-entropy error) between $\hat{\eta}(X)$ and $\eta(X)$, which is given by

$$\mathrm{GKL}(\eta, \hat{\eta}) = \mathbb{E}_X \Big[\eta(X) \log \frac{\eta(X)}{\hat{\eta}(X)} + (1-\eta(X)) \log \frac{1-\eta(X)}{1-\hat{\eta}(X)} \Big].$$

Alternatively, we formulate the optimization problem for obtaining ρ as

$$\min_{\rho > 0} \ \mathrm{EKL}(\hat{\eta}) := -\frac{1}{n} \sum_{i=1}^{n} \Big\{ \frac{1}{2}(y_i + 1) \log \hat{\eta}(\mathbf{x}_i) + \frac{1}{2}(1 - y_i) \log(1 - \hat{\eta}(\mathbf{x}_i)) \Big\}. \tag{11}$$

The problem can be solved by the Newton method. In summary, one first obtains $\hat{f}(\mathbf{x}) = \hat{\alpha} + \sum_{i=1}^{n} \hat{\beta}_i K(\mathbf{x}, \mathbf{x}_i)$ via the conventional SVM model, and estimates ρ via the optimization problem in (11) based on the training data; one then uses the formula in (10) to estimate the class probabilities for the training samples as well as the test samples.

5. $\mathcal{C}$-Learning

Focusing on the relationships of the $\mathcal{C}$-loss $C_\rho(yf(x))$ (i.e., $C_{\rho,1}(yf(x))$) with the hinge and logit losses, we illustrate its application in the construction of large-margin classifiers. Since $C_\rho(yf(\mathbf{x}))$ is smooth, it does not tend to yield a sparse classifier. However, we can employ a sparsification penalty $J(h)$ to achieve sparseness. We use the elastic-net penalty of Zou and Hastie (2005) for the experiments in this section. Additionally, we study two forms of $f(\mathbf{x})$: kernel expansion and feature expansion. Built on these two expansions, sparseness can subserve the selection of support vectors and the selection of features, respectively. The resulting classifiers are called *$\mathcal{C}$-learning*.

5.1 The Kernel Expansion

In the kernel expansion approach, given a reproducing kernel $K(\cdot, \cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, we define the kernel expansion as $f(\mathbf{x}) = \alpha + \sum_{i=1}^{n} \beta_i K(\mathbf{x}_i, \mathbf{x})$ and solve the following optimization problem:

$$\min_{\alpha, \beta} \frac{1}{n} \sum_{i=1}^{n} C_\rho(y_i f(\mathbf{x}_i)) + \gamma \Big((1-\omega) \frac{1}{2} \beta' \mathbf{K} \beta + \omega \|\beta\|_1 \Big), \tag{12}$$

where $\mathbf{K} = [K(\mathbf{x}_i, \mathbf{x}_j)]$ is the $n \times n$ kernel matrix.

It is worth pointing out that the current penalty is slightly different from the conventional elastic-net penalty, which is $(1-\omega)\frac{1}{2}\beta'\beta + \omega\|\beta\|_1$. In fact, the optimization problem (12) can be viewed equivalently as the optimization problem

$$\min_{\alpha, \beta} \frac{1}{n} \sum_{i=1}^{n} C_\rho(y_i f(\mathbf{x}_i)) + \frac{\gamma}{2} \beta' \mathbf{K} \beta \tag{13}$$

under the ℓ_1 penalty $\|\beta\|_1$. Thus, the method derived from (12) enjoys the generalization ability of the conventional kernel supervised learning method derived from (13) but also the sparsity of the ℓ_1 penalty.

Recently, Friedman et al. (2010) devised a pathwise coordinate descent algorithm for regularized logistic regression problems in which the elastic-net penalty is used. In order to solve the optimization problem in (12), we employ this pathwise coordinate descent algorithm.

Let the current estimates of α and β be $\hat{\alpha}$ and $\hat{\beta}$. We first form a quadratic approximation to $\frac{1}{n}\sum_{i=1}^{n} C_\rho(y_i f(\mathbf{x}_i))$, which is

$$Q(\alpha,\beta) = \frac{1}{2n\rho}\sum_{i=1}^{n} q(\mathbf{x}_i)(1-q(\mathbf{x}_i))\big(\alpha+\mathbf{k}_i'\beta - z_i\big)^2 + \text{Const},$$

where

$$\begin{aligned} z_i &= \hat{\alpha}+\mathbf{k}_i'\hat{\beta} + \frac{\rho}{y_i(1-q(\mathbf{x}_i))}, \\ q(\mathbf{x}_i) &= \frac{\exp[(1-y_i(\hat{\alpha}+\mathbf{k}_i'\hat{\beta})/\rho]}{1+\exp[(1-y_i(\hat{\alpha}+\mathbf{k}_i'\hat{\beta}))/\rho]}, \\ \mathbf{k}_i &= (K(\mathbf{x}_1,\mathbf{x}_i),\ldots,K(\mathbf{x}_n,\mathbf{x}_i))'. \end{aligned}$$

We then employ coordinate descent to solve the weighted least-squares problem as follows:

$$\min_{\alpha,\beta}\; G(\alpha,\beta) := Q(\alpha,\beta) + \gamma\Big((1-\omega)\frac{1}{2}\beta'\mathbf{K}\beta + \omega\|\beta\|_1\Big). \tag{14}$$

Assume that we have estimated $\tilde{\beta}$ for β using $G(\alpha,\beta)$. We now set $\frac{\partial G(\alpha,\tilde{\beta})}{\partial\alpha} = 0$ to find the new estimate of α:

$$\tilde{\alpha} = \frac{\sum_{i=1}^{n} q(\mathbf{x}_i)(1-q(\mathbf{x}_i))\big(z_i - \mathbf{k}_i'\tilde{\beta}\big)}{\sum_{i=1}^{n} q(\mathbf{x}_i)(1-q(\mathbf{x}_i))}. \tag{15}$$

On the other hand, assume that we have estimated $\tilde{\alpha}$ for α and $\tilde{\beta}_l$ for β_l $(l=1,\ldots,n, l\neq j)$. We now optimize β_j. In particular, we only consider the gradient at $\beta_j \neq 0$. If $\beta_j > 0$, we have

$$\frac{\partial G(\tilde{\alpha},\tilde{\beta})}{\partial\beta_j} = \frac{1}{n\rho}\sum_{i=1}^{n} K_{ij} q(\mathbf{x}_i)(1-q(\mathbf{x}_i))\big(\alpha+\mathbf{k}_i'\tilde{\beta}-z_i\big) + \gamma(1-\omega)(K_{jj}\beta_j+\mathbf{k}_j\check{\beta}) + \gamma\omega$$

and, hence,

$$\tilde{\beta}_j = \frac{S(t-\gamma(1-\omega)\mathbf{k}_j'\check{\beta}, \gamma\omega)}{\frac{1}{n\rho}\sum_{i=1}^{n} K_{ij}^2 q(\mathbf{x}_i)(1-q(\mathbf{x}_i)) + \gamma(1-\omega)K_{jj}}, \tag{16}$$

where $t = \frac{1}{n\rho}\sum_{i=1}^{n} K_{ij} q(\mathbf{x}_i)(1-q(\mathbf{x}_i))\big(z_i - \tilde{\alpha} - \mathbf{k}_i'\check{\beta}\big)$, $\check{\beta} = (\tilde{\beta}_1,\ldots,\tilde{\beta}_{j-1},0,\tilde{\beta}_{j+1},\ldots,\tilde{\beta}_n)'$, $K_{ij} = K(\mathbf{x}_i,\mathbf{x}_j)$, and $S(\mu,\nu)$ is the soft-thresholding operator:

$$\begin{aligned} S(\mu,\nu) &= \text{sign}(\mu)(|\mu|-\nu)_+ \\ &= \begin{cases} \mu-\nu & \text{if } \mu>0 \text{ and } \mu<|\nu| \\ \mu+\nu & \text{if } \mu<0 \text{ and } \mu<|\nu| \\ 0 & \text{if } \mu>|\nu|. \end{cases} \end{aligned}$$

Algorithm 1 summarizes the coordinate descent algorithm for binary $\mathcal{C}$-learning.

Algorithm 1 The coordinate descent algorithm for binary $\mathcal{C}$-learning

Input: $\mathcal{T} = \{\mathbf{x}_i, y_i\}_{i=1}^n$, γ, ω, ε_m, ε_i, ρ;
Initialize: $\tilde{\alpha} = \alpha_0$, $\tilde{\beta} = \beta_0$
repeat
  Calculate $G(\tilde{\alpha}, \tilde{\beta})$ using (14);
  $\alpha^\star \leftarrow \tilde{\alpha}$;
  $\beta^\star \leftarrow \tilde{\beta}$;
  repeat
    $\bar{\alpha} \leftarrow \tilde{\alpha}$;
    $\bar{\beta} \leftarrow \tilde{\beta}$;
    Calculate $\tilde{\alpha}$ using (15);
    for $j = 1$ **to** n **do**
      Calculate $\tilde{\beta}_j$ using (16);
    end for
  until $\|\tilde{\alpha} - \bar{\alpha}\| + \|\tilde{\beta} - \bar{\beta}\| < \varepsilon_i$
until $\|\tilde{\alpha} - \alpha^\star\| + \|\tilde{\beta} - \beta^\star\| < \varepsilon_m$
Output: $\tilde{\alpha}$, $\tilde{\beta}$, and $f(\mathbf{x}) = \tilde{\alpha} + \sum_{i=1}^n K(\mathbf{x}_i, \mathbf{x})\tilde{\beta}_i$.

5.2 The Linear Feature Expansion

In the linear feature expansion approach, we let $f(\mathbf{x}) = a + \mathbf{x}'\mathbf{b}$, and pose the following optimization problem:

$$\min_{a,\mathbf{b}} \frac{1}{n} \sum_{i=1}^n C_\rho(y_i f(\mathbf{x}_i)) + \gamma J_\omega(\mathbf{b}), \tag{17}$$

where for $\omega \in [0,1]$

$$J_\omega(\mathbf{b}) = (1-\omega)\frac{1}{2}\|\mathbf{b}\|_2^2 + \omega\|\mathbf{b}\|_1 = \sum_{j=1}^d \left[\frac{1}{2}(1-\omega)b_j^2 + \omega|b_j|\right].$$

The elastic-net penalty maintains the sparsity of the ℓ_1 penalty, but the number of variables to be selected is no longer bounded by n. Moreover, this penalty tends to generate similar coefficients for highly-correlated variables. We also use a coordinate descent algorithm to solve the optimization problem (17). The algorithm is similar to that for the kernel expansion and the details are omitted here.

6. Experimental Results

In Section 6.1 we illustrate convergence analysis of our classification method based on the $\mathcal{C}$-loss. In Section 6.2 we report the results of experimental evaluations of our method for class probability estimation of the conventional SVM given in Section 4. In Section 6.3 we present results for the $\mathcal{C}$-learning method given in Section 5.

6.1 Simulation Analysis for Convergence of $\mathcal{C}-$Learning

In Section 3.4 we have presented a theoretical analysis for convergence rate of the classification model defined in (9). We now conduct an empirical analysis to illustration how our model in (9) approaches the target function $f_*(\eta)$ given in (7) and how the class probability estimate $\tilde{\eta}(\hat{f}_n(\mathbf{x}))$ in (8) approaches the underlying class probability $\eta(\mathbf{x})$.

For the purpose of visualization, we employ the same simulation data set as in Lin (2002). In particular, we take n equidistant points on the interval $[0,1]$; that is, $x_i = (i-1)/n$ for $i = 1,\ldots,n$. Let $\eta(x) = \Pr(Y = 1|X = x) = 1-|1-2x|$. Then the target function $f_*(\eta(x))$ for our model is computed by (7). We randomly generate y_i to be 1 or -1 with probability $\eta(x_i)$ and $1-\eta(x_i)$, and form a training data set $\{(x_i,y_i); i = 1\ldots,n\}$. Following the setting in Lin (2002), we implement our simulations on RKHS $\mathcal{H}^m([0,1])$ (i.e., the Sobolev Hilbert space with order m of univariate functions on domain $[0,1]$) and four data sets with size $n = 33$, 65, 129 and 257. The parameters γ and ρ are chosen to minimize the generalized KL (GKL) divergence by the grid search. The implementation is based on the $\mathcal{C}$-learning algorithm in Section 5 where the hyperparameter ω is approximately set as 0.

Figures 2 and 3 respectively depict the solutions $\hat{f}_n$ to the regularization problem in (9) and the class probability estimates $\tilde{\eta}(\hat{f}_n)$ given in (8), when the sample size is $n = 33$, 65, 129, and 257. We can see that the solution $\hat{f}_n$ is close the target function $f_*(\eta)$ and that the class probability estimate $\tilde{\eta}(\hat{f}_n)$ is close the underlying class probability η, as n increases. Thus, our simulation example shows that our method based on $\mathcal{C}$-loss not only can approach its corresponding target function $f_*(\eta)$ but also can estimate the underlying class probability $\eta(\mathbf{x})$. It should be pointed out that the similar experimental results can be found on the Gaussian RBF kernel, so here we do not include the results with the Gaussian RBF kernel.

6.2 Simulation for Class Probability Estimation of SVM Outputs

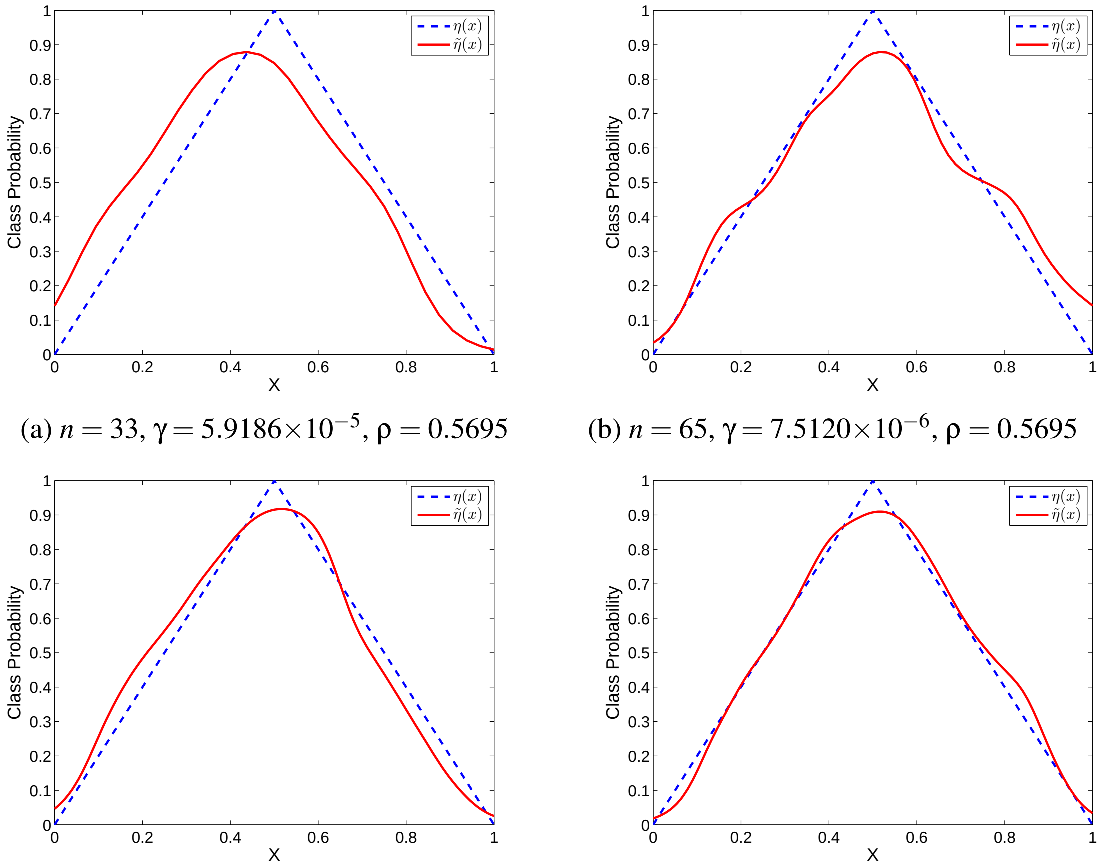
We validate our estimation method for the class probability of SVM outputs ("Ours for SVM"), comparing it with several alternatives: Platt's method (Platt, 1999), Sollich's method (Sollich, 2002), and the method of Wang et al. (2008) (WSL's). Since penalized (or regularized) logistic regression (PLR) and $\mathcal{C}$-learning can directly calculate class probability, we also implement them. Especially, the class probability of $\mathcal{C}$-learning outputs is based on (8) where we set $\rho = 1$ and $u = 1$ since $\mathcal{C}$-learning itself employs the same setting.

We conducted our analysis over two simulation data sets which were used by Wang et al. (2008). The first simulation data set, $\{(x_{i1},x_{i2};y_i)\}_{i=1}^{1000}$, was generated as follows. The $\{(x_{i1},x_{i2})\}_{i=1}^{1000}$ were uniformly sampled from a unit disk $\{(x_1,x_2) : x_1^2+x_2^2 \leq 1\}$. Next, we set $y_i = 1$ if $x_{i1} \geq 0$ and $y_i = -1$ otherwise, $i = 1,\ldots,1000$. Finally, we randomly chose 20% of the samples and flipped their labels. Thus, the true class probability $\eta(Y_i = 1|x_{i1},x_{i2})$ was either 0.8 or 0.2.

The second data set, $\{(x_{i1},x_{i2};y_i)\}_{i=1}^{1000}$, was generated as follows. First, we randomly assigned 1 or -1 to y_i for $i = 1,\ldots,1000$ with equal probability. Next, we generated x_{i1} from the uniform distribution over $[0,2\pi]$, and set $x_{i2} = y_i(\sin(x_{i1})+\varepsilon_i)$ where $\varepsilon_i \sim N(\varepsilon_i|1,0.01)$. For the data, the true class probability of $Y = 1$ was given by

$$\eta(Y = 1|x_1,x_2) = \frac{N(x_2|\sin(x_1)+1,0.01)}{N(x_2|\sin(x_1)+1,0.01)+N(x_2|-\sin(x_1)-1,0.01)}.$$

The simulation followed the same setting as that in Wang et al. (2008). That is, we randomly selected 100 samples for training and the remaining 900 samples for test. We did 100 replications for each data set. The values of generalized Kullback-Leibler loss (GKL) and classification error rate (CER) on the test sets were averaged over these 100 simulation replications. Additionally, we



(a) $n = 33$, $\gamma = 5.9186 \times 10^{-5}$, $\rho = 0.5695$

(b) $n = 65$, $\gamma = 7.5120 \times 10^{-6}$, $\rho = 0.5695$

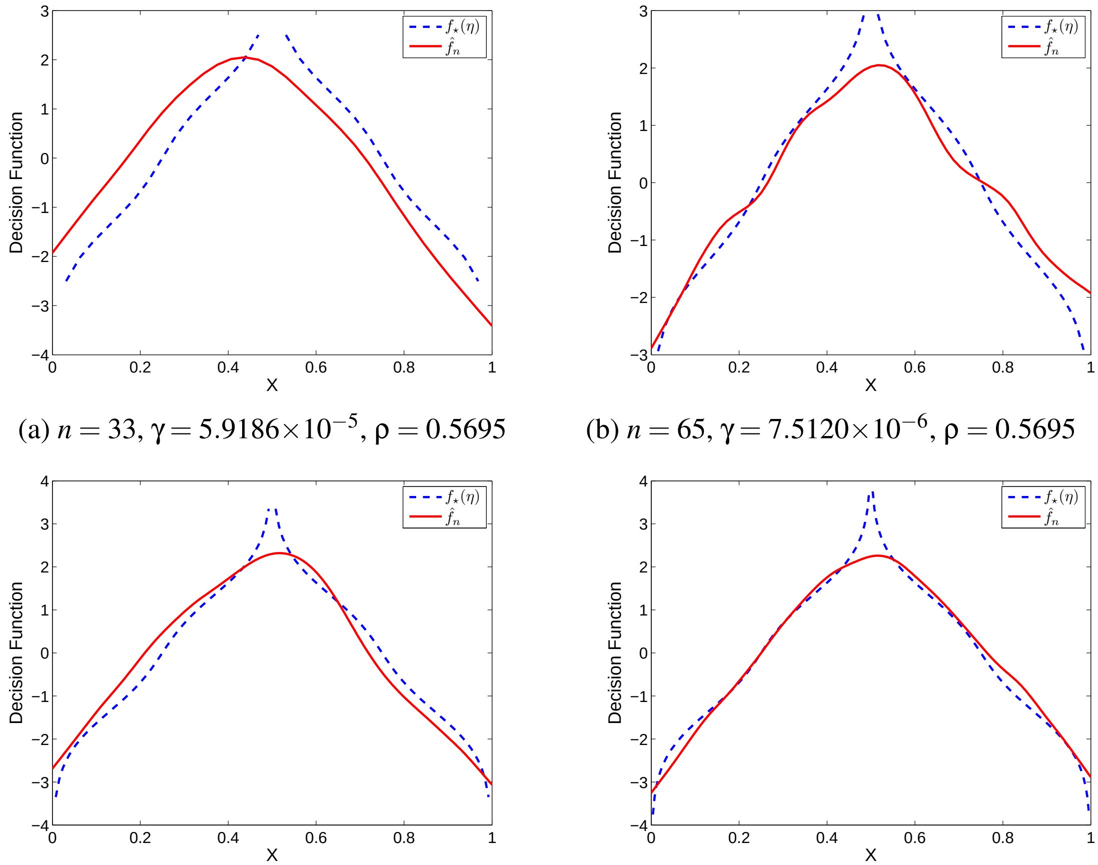
(c) $n = 129$, $\gamma = 1.5141 \times 10^{-5}$, $\rho = 0.5695$

(d) $n = 257$, $\gamma = 3.7999 \times 10^{-6}$, $\rho = 0.5695$

Figure 2: The underlying class probabilities $\eta(x)$ ("blue + dashed line") and estimated class probabilities $\tilde{\eta}(x) = \tilde{\eta}(\hat{f}_n(x))$ ("red + solid line") on RKHS $\mathcal{H}^m([0,1])$ and the simulation data sets with the size $n = 33$, 65, 129, and 257. Here the values of parameters γ and ρ in each data set are obtained by minimizing the GKL divergence.

employed a Gaussian RBF kernel $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/\sigma^2)$ where the parameter σ was set as the median distance between the positive and negative classes. We reported GKL and CER as well as the corresponding standard deviations in Tables 2 and 3 in which the results with the PLR method, the tuned Platt method and the WSL method are directly cited from Wang et al. (2008).

Note that the results with PLR were averaged only over 66 nondegenerate replications (Wang et al., 2008). Based on GKL and CER, the performance of $\mathcal{C}$-learning is the best in these two simulations. With regard to GKL, our method for SVM outperforms the original and tuned versions of Platt's method as well as the method of Wang et al. (2008). Since our estimation method is based on the $\hat{\eta}(\mathbf{x})$ in (10), the CER with this class probability $\hat{\eta}(\mathbf{x})$ is identical to that with the conventional SVM. This also applies to Sollich's method, thus we did not include the CER of this method. However, Table 3 shows that this does not necessarily hold for Platt's method for SVM



(a) $n = 33$, $\gamma = 5.9186\times 10^{-5}$, $\rho = 0.5695$ (b) $n = 65$, $\gamma = 7.5120\times 10^{-6}$, $\rho = 0.5695$

(c) $n = 129$, $\gamma = 1.5141\times 10^{-5}$, $\rho = 0.5695$ (d) $n = 257$, $\gamma = 3.7999\times 10^{-6}$, $\rho = 0.5695$

Figure 3: The target decision function $f_*(\eta)$ ("blue + dashed line") and estimated decision functions $\hat{f}_n$ ("red + solid line") on RKHS $\mathcal{H}^m([0,1])$ and the toy data sets with the size $n = 33$, 65, 129, and 257. The parameters are identical to the settings in Figure 2.

probability outputs. In other words, $\hat{\eta}(\mathbf{x}) > 1/2$ is not equivalent to $\hat{f}(\mathbf{x}) > 0$ for Platt's method. In fact, Platt (1999) used the sigmoid-like function to improve the classification accuracy of the conventional SVM. As for the method of Wang et al. (2008) which is built on a sequence of weighted classifiers, the CERs of the method should be different from those of the original SVM. With regard to CER, the performance of PLR is the worst in most cases.

6.3 The Performance Analysis of $\mathcal{C}$-Learning

To evaluate the performance of our $\mathcal{C}$-learning method, we further conducted empirical studies on several benchmark data sets and compared $\mathcal{C}$-learning with two closely related classification methods: the hybrid huberized SVM (HHSVM) of Wang et al. (2007) and the regularized logistic regression model (RLRM) of Friedman et al. (2010), both with the elastic-net penalty. All the three classification methods were implemented in both the feature and kernel expansion settings.

	PLR	Platt's	Tuned Platt	WSL's	Sollich's	Ours for SVM	$\mathcal{C}$-learning
Data 1	0.579 (±0.0021)	0.582 (±0.0035)	0.569 (±0.0015)	0.566 (±0.0014)	0.566 (±0.0021)	0.558 (±0.0015)	0.549 (±0.0016)
Data 2	0.138 (±0.0024)	0.163 (±0.0018)	0.153 (±0.0013)	0.153 (±0.0010)	0.155 (±0.0017)	0.142 (±0.0016)	0.134 (±0.0014)

Table 2: Values of GKL over the two simulation test sets (standard deviations are shown in parentheses).

	PLR	Platt's	WSL's	Ours for SVM	$\mathcal{C}$-learning
Data 1	0.258 (±0.0053)	0.234 (±0.0026)	0.217 (±0.0021)	0.219 (±0.0021)	0.214 (±0.0015)
Data 2	0.075 (±0.0018)	0.077 (±0.0024)	0.069 (±0.0014)	0.065 (±0.0015)	0.061 (±0.0019)

Table 3: Values of CER over the two simulation test sets (standard deviations are shown in parentheses).

In the experiments we used 11 binary classification data sets. Table 4 gives a summary of these benchmark data sets. The seven binary data sets of digits were obtained from the publicly available USPS data set of handwritten digits as follows. The first six data sets were generated from the digit pairs $\{$(1, 7), (2, 3), (2, 7), (3, 8), (4, 7), (6,9)$\}$, and 200 digits were chosen within each class of each data set. The USPS (odd vs. even) data set consisted of the first 80 images per digit in the USPS training set.

The two binary artificial data sets of "g241c" and "g241d" were generated via the setup presented by Chapelle et al. (2006). Each class of these two data sets consisted of 750 samples.

The two binary gene data sets of "colon" and "leukemia" were also used in our experiments. The "colon" data set, consisting of 40 colon tumor samples and 22 normal colon tissue samples with 2,000 dimensions, was obtained by employing an Affymetrix oligonucleotide array to analyze more than 6,500 human genes expressed in sequence tags (Alon et al., 1999). The "leukemia" data set is of the same type as the "colon" cancer data set (Golub et al., 1999), and it was obtained with respect to two variants of leukemia, that is, acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL). It initially contained expression levels of 7129 genes taken over 72 samples (AML, 25 samples, or ALL, 47 samples), and then it was pre-feature selected, leading to a feature space with 3571 dimensions.

In our experiments, each data set was randomly partitioned into two disjoint subsets as the training and test, with the percentage of the training data samples also given in Table 4. Twenty random partitions were chosen for each data set, and the average and standard deviation of their classification error rates over the test data were reported.

Although we can seek an optimum ρ using computationally intensive methods such as cross-validation, the experiments showed that when ρ takes a value in $[0.1, 2]$, our method is always able to obtain promising performance. Here our reported results are based on the setting of $\rho = 1$, due to

Data Set	m	d	k	n/k
USPS (1 vs. 7)	2	256	400	3%
USPS (2 vs. 3)	2	256	400	3%
USPS (2 vs. 7)	2	256	400	3%
USPS (3 vs. 8)	2	256	400	3%
USPS (4 vs. 7)	2	256	400	3%
USPS (6 vs. 9)	2	256	400	3%
USPS (Odd vs. Even)	2	256	800	3%
g241c	2	241	1500	10%
g241d	2	241	1500	10%
colon	2	2000	62	25.8%
leukemia	2	3571	72	27.8%

Table 4: Summary of the benchmark data sets: m—the number of classes; d—the dimension of the input vector; k—the size of the data set; n—the number of the training data.

Data Set	HHSVM	RLRM	$\mathcal{C}$-learning
(1 vs. 7)	2.29±1.17	2.06±1.21	**1.60±0.93**
(2 vs. 3)	**8.13±2.02**	8.29±2.76	8.32±2.73
(2 vs. 7)	5.82±2.59	6.04±2.60	**5.64±2.44**
(3 vs. 8)	12.46±2.90	**10.77±2.72**	11.74±2.83
(4 vs. 7)	7.35±2.89	6.91±2.72	**6.68±3.53**
(6 vs. 9)	2.32±1.65	2.15±1.43	**2.09±1.41**
(Odd vs. Even)	20.94±2.02	19.83±2.82	**19.74±2.81**
g241c	22.30±1.30	21.38±1.12	**21.34±1.11**
g241d	24.32±1.53	**23.81±1.65**	23.85±1.69
colon	14.57±1.86	14.47±2.02	**12.34±1.48**
leukemia	4.06±2.31	4.43±1.65	**3.21±1.08**

Table 5: Classification error rates (%) and standard deviations on the 11 data sets for the feature expansion setting.

the relationship of the $\mathcal{C}$-loss $C(z)$ with the hinge loss $(1-z)_+$ and the logit loss $\log(1+\exp(-z))$ (see our analysis in Section 3 and Figure 1).

As for the parameters γ and ω, they were selected by cross-validation for all the classification methods. In the kernel expansion, the RBF kernel $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/\sigma^2)$ was employed, and σ was set to the mean Euclidean distance among the input samples. For $\mathcal{C}$-learning, the other parameters were set as follows: $\varepsilon_m = \varepsilon_i = 10^{-5}$.

Tables 5 and 6 show the test results corresponding to the linear feature expansion and RBF kernel expansion, respectively. From the tables, we can see that for the overall performance of $\mathcal{C}$-learning is slightly better than the two competing methods in the feature and kernel settings generally.

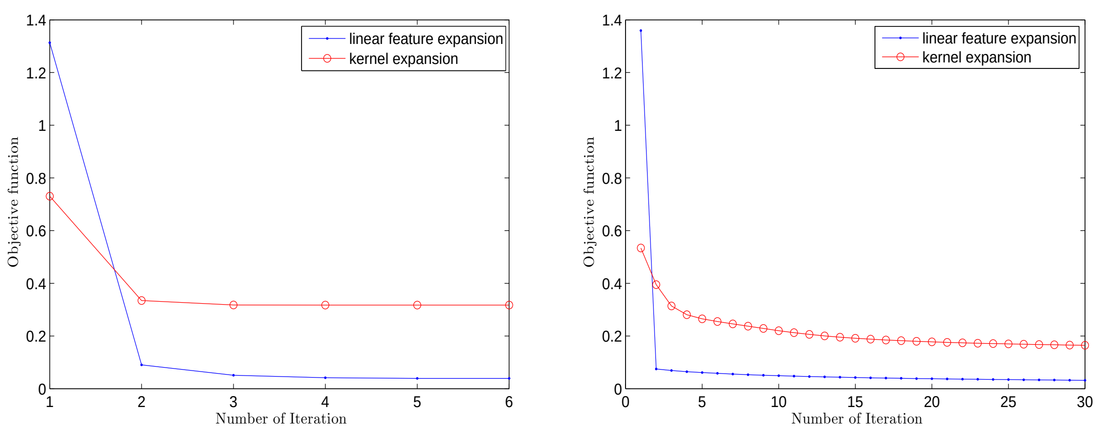
Figure 4 reveals that the values of the objective functions for the linear feature and RBF kernel versions in the outer and inner iterations tend to be significantly reduced as the number of iterations in the coordinate descent procedure increases. Although we report only the change of the values of

Data Set	HHSVM	RLRM	$\mathcal{C}$-learning
(1 vs. 7)	1.73±1.64	1.39±0.64	**1.37±0.65**
(2 vs. 3)	8.55±3.36	8.45±3.38	**8.00±3.32**
(2 vs. 7)	5.09±2.10	4.02±1.81	**3.90±1.79**
(3 vs. 8)	12.09±3.78	10.58±3.50	**10.36±3.52**
(4 vs. 7)	6.74±3.39	6.92±3.37	**6.55±3.28**
(6 vs. 9)	2.12±0.91	1.74±1.04	**1.65±0.99**
(Odd vs. Even)	28.38±10.51	26.92±6.52	**26.29±6.45**
g241c	**21.38±1.45**	21.55±1.42	21.62±1.35
g241d	25.89±2.15	22.34±1.27	**20.37±1.20**
colon	14.26±2.66	14.79±2.80	**13.94±2.44**
leukemia	2.77±0.97	2.74±0.96	**2.55±0.92**

Table 6: Classification error rates (%) and standard deviations on the 11 data sets for the RBF kernel setting.

the objective function for the data set USPS (1 vs. 7) similar results were found on all other data sets. This shows that the coordinate descent algorithm is very efficient.

We also conducted a systematic study of sparseness from the elastic-net penalty. Indeed, the elastic-net penalty does give rise to sparse solutions for our $\mathcal{C}$-learning methods. Moreover, we found that similar to other methods the sparseness of the solution is dependent on the parameters γ and ω that were set to different values for different data sets using cross validation.



(a) (b)

Figure 4: Change of the value of the objective function for the $\mathcal{C}$-learning as the number of iterations in the coordinate descent procedure increases in the linear feature and RBF kernel cases on the data set USPS (1 vs. 7): (a) the values of the objective function (12) in the outer iteration; (b) the objective function values $G(\alpha,\beta)$ for the fature and RBF kernel cases in the inner iteration.

7. Conclusions

In this paper we have studied a family of coherence functions and considered the relationship between coherence functions and hinge functions. In particular, we have established some important properties of these functions, which lead us to a feasible approach for class probability estimation in the conventional SVM. Moreover, we have proposed large-margin classification methods using the $\mathcal{C}$-loss function and the elastic-net penalty, and developed pathwise coordinate descent algorithms for parameter estimation. We have theoretically established the Fisher-consistency of our classification methods and empirically tested the classification performance on several benchmark data sets. Our approach establishes an interesting link between SVMs and logistic regression models due to the relationship of the $\mathcal{C}$-loss with the hinge and logit losses.

Acknowledgments

The authors would like to thank the Action Editor and three anonymous referees for their constructive comments and suggestions on the original version of this paper. This work has supported in part by the Natural Science Foundations of China (No. 61070239), the US ARL and the US ARO under contract/grant number W911NF-11-1-0391.

Appendix A. The Proof of Proposition 2

First, we have

$$\rho \log 2 + [u-z]_+ - V_{\rho,u}(z) = \rho \log \frac{2 \exp \frac{[u-z]_+}{\rho}}{1+\exp \frac{u-z}{\rho}} \geq 0.$$

Second, note that

$$\begin{aligned} \rho \log 2 + \frac{u-z}{2} - V_{\rho,u}(z) &= \rho \log \frac{2 \exp \frac{1}{2} \frac{u-z}{\rho}}{1+\exp \frac{u-z}{\rho}} \\ &\leq \rho \log \frac{\exp \frac{u-z}{\rho}}{1+\exp \frac{u-z}{\rho}} \leq 0, \end{aligned}$$

where we use the fact that $\exp(\cdot)$ is convex.

Third, it immediately follows from Proposition (i) that $\lim_{\rho \to 0} V_{\rho,u}(z) = [u-z]_+$. Moreover, it is easily obtained that

$$\begin{aligned} \lim_{\rho \to \infty} V_{\rho,u}(z) - \rho \log 2 &= \lim_{\rho \to \infty} \frac{\log \frac{1+\exp \frac{u-z}{\rho}}{2}}{\frac{1}{\rho}} = \lim_{\alpha \to 0} \frac{\log \frac{1+\exp \alpha(u-z)}{2}}{\alpha} \\ &= \lim_{\alpha \to 0} \frac{\frac{1}{2}[u-z] \exp[\alpha(u-z)]}{\frac{1+\exp \alpha(u-z)}{2}} = \frac{1}{2}(u-z). \end{aligned}$$

Since $\log(1+a) \geq \log(a)$ for $a > 0$, we have

$$\frac{u}{\log[1+\exp(u/\rho)]} \log\left[1+\exp \frac{u-z}{\rho}\right] \leq \frac{u}{u/\rho} \log\left[1+\exp \frac{u-z}{\rho}\right] = \rho \log\left[1+\exp \frac{u-z}{\rho}\right].$$

We now consider that

$$\lim_{\rho\to\infty} C_{\rho,u}(z) = u \lim_{\rho\to\infty} \frac{\log\left[1+\exp\frac{u-z}{\rho}\right]}{\log[1+\exp(u/\rho)]} = u.$$

Finally, since

$$\lim_{\alpha\to\infty} \frac{\log[1+\exp(u\alpha)]}{\alpha u} = \lim_{\alpha\to\infty} \frac{\exp(u\alpha)]}{1+\exp(u\alpha)} = 1 \ \text{ for } \ u > 0,$$

we obtain $\lim_{\rho\to 0} C_{\rho,u}(z) = [u-z]_+$.

Appendix B. The Proof of Proposition 3

Before we prove Proposition 3, we establish the following lemma.

Lemma 13 *Assume that* $x > 0$*, then* $f_1(x) = \frac{x}{1+x}\frac{\log x}{\log(1+x)}$ *and* $f_2(x) = \frac{x}{1+x}\frac{1}{\log(1+x)}$ *are increasing and decreasing, respectively.*

Proof The first derivatives of $f_1(x)$ and $f_2(x)$ are

$$f_1'(x) = \frac{1}{(1+x)^2\log^2(1+x)}\Big[\log x\log(1+x) + \log(1+x) + x\log(1+x) - x\log x\Big]$$

$$f_2'(x) = \frac{1}{(1+x)^2\log^2(1+x)}[\log(1+x) - x] \le 0.$$

This implies that $f_2(x)$ is decreasing. If $\log x \ge 0$, we have $x\log(1+x) - x\log x \ge 0$. Otherwise, if $\log x < 0$, we have $\log x[\log(1+x) - x] \ge 0$. This implies that $f_1'(x) \ge 0$ is always satisfied. Thus, $f_1(x)$ is increasing. ∎

Let $\alpha = 1/\rho$ and use $h_1(\alpha)$ for $L_{\rho,u}(z)$ to view it as a function of α. We now compute the derivative of $h_1(\alpha)$ w.r.t. α:

$$\begin{aligned} h_1'(\alpha) &= \frac{\log[1+\exp(\alpha(u-z))]}{\log[1+\exp(u\alpha)]}\times \\ &\quad \Big[\frac{\exp(\alpha(u-z))}{1+\exp(\alpha(u-z))}\frac{u-z}{\log[1+\exp(\alpha(u-z))]} - \frac{\exp(\alpha u)}{1+\exp(\alpha u)}\frac{u}{\log[1+\exp(\alpha u)]}\Big] \\ &= \frac{\log[1+\exp(\alpha(u-z))]}{\alpha\log[1+\exp(u\alpha)]}\times \\ &\quad \Big[\frac{\exp(\alpha(u-z))}{1+\exp(\alpha(u-z))}\frac{\log\exp(\alpha(u-z))}{\log[1+\exp(\alpha(u-z))]} - \frac{\exp(\alpha u)}{1+\exp(\alpha u)}\frac{\log\exp(\alpha u)}{\log[1+\exp(\alpha u)]}\Big]. \end{aligned}$$

When $z < 0$, we have $\exp(\alpha(u-z)) > \exp(\alpha u)$. It then follows from Lemma 13 that $h_1'(\alpha) \ge 0$. When $z \ge 0$, we have $h_1'(\alpha) \le 0$ due to $\exp(\alpha(u-z)) \le \exp(\alpha u)$. The proof of (i) is completed.

To prove part (ii), we regard $L_{\rho,u}(z)$ as a function of u and denote it with $h_2(u)$. The first derivative $h_2'(u)$ is given by

$$\begin{aligned} h_2'(u) &= \frac{\alpha\log[1+\exp(\alpha(u-z))]}{\log[1+\exp(u\alpha)]}\times \\ &\quad \Big[\frac{\exp(\alpha(u-z))}{1+\exp(\alpha(u-z))}\frac{1}{\log[1+\exp(\alpha(u-z))]} - \frac{\exp(\alpha u)}{1+\exp(\alpha u)}\frac{1}{\log[1+\exp(\alpha u)]}\Big]. \end{aligned}$$

Using Lemma 13, we immediately obtain part (ii).

Appendix C. The Proof of Theorem 6

We write the objective function as

$$\begin{aligned} L(f) &= V_{\rho,u}(f)\eta + V_{\rho,u}(-f)(1-\eta) \\ &= \rho \log\big[1+\exp\frac{u-f}{\rho}\big]\eta + \rho\log\big[1+\exp\frac{u+f}{\rho}\big](1-\eta). \end{aligned}$$

The first-order and second-order derivatives of L w.r.t. f are given by

$$\begin{aligned} \frac{dL}{df} &= -\eta \frac{\exp\frac{u-f}{\rho}}{1+\exp\frac{u-f}{\rho}} + (1-\eta)\frac{\exp\frac{u+f}{\rho}}{1+\exp\frac{u+f}{\rho}}, \\ \frac{d^2L}{df^2} &= \frac{\eta}{\rho}\frac{\exp\frac{u-f}{\rho}}{1+\exp\frac{u-f}{\rho}}\frac{1}{1+\exp\frac{u-f}{\rho}} + \frac{1-\eta}{\rho}\frac{\exp\frac{u+f}{\rho}}{1+\exp\frac{u+f}{\rho}}\frac{1}{1+\exp\frac{u+f}{\rho}}. \end{aligned}$$

Since $\frac{d^2L}{df^2} > 0$, the minimum of L is unique. Moreover, letting $\frac{dL}{df} = 0$ yields (7).

Appendix D. The Proof of Proposition 7

First, if $\eta > 1/2$, we have $4\eta(1-\eta) > 4(1-\eta)^2$ and $(2\eta-1)\exp(u/\rho) > 0$. This implies $f_* > 0$. When $\eta < 1/2$, we have $(2\eta-1)\exp(u/\rho) > 0$. In this case, since

$$(1-2\eta)^2\exp(2u/\rho) + 4\eta(1-\eta) < (1-2\eta)^2\exp(2u/\rho) + 4(1-\eta)^2 + 4(1-\eta)(1-2\eta)\exp(u/\rho),$$

we obtain $f_* < 0$.

Second, letting $\alpha = 1/\rho$, we express f_* as

$$\begin{aligned} f_* &= \frac{1}{\alpha}\log\frac{(2\eta-1)\exp(u\alpha) + \sqrt{(1-2\eta)^2\exp(2u\alpha) + 4\eta(1-\eta)}}{2(1-\eta)} \\ &= \frac{1}{\alpha}\log\frac{\frac{(2\eta-1)}{|2\eta-1|} + \sqrt{1+\frac{4\eta(1-\eta)}{(1-2\eta)^2\exp(2u\alpha)}}}{2(1-\eta)\exp(-u\alpha)/|2\eta-1|} \\ &= \frac{1}{\alpha}\log\left[\frac{(2\eta-1)}{|2\eta-1|} + \sqrt{1+\frac{4\eta(1-\eta)}{(1-2\eta)^2\exp(2u\alpha)}}\right] - \frac{1}{\alpha}\log\Big[\frac{2(1-\eta)}{|2\eta-1|}\Big] + u. \end{aligned}$$

Thus, if $\eta > 1/2$, it is clear that $\lim_{\alpha\to\infty} f_* = u$. In the case that $\eta < 1/2$, we have

$$\begin{aligned} \lim_{\alpha\to\infty} f_* &= u - u\lim_{\alpha\to\infty}\frac{1}{-1+\sqrt{1+\frac{4\eta(1-\eta)}{(1-2\eta)^2\exp(2u\alpha)}}}\frac{1}{\sqrt{1+\frac{4\eta(1-\eta)}{(1-2\eta)^2\exp(2u\alpha)}}}\frac{4\eta(1-\eta)}{(1-2\eta)^2}\exp(-2u\alpha) \\ &= u - \frac{4\eta(1-\eta)u}{(1-2\eta)^2}\lim_{\alpha\to\infty}\frac{\exp(-2u\alpha)}{-1+\sqrt{1+\frac{4\eta(1-\eta)}{(1-2\eta)^2\exp(2u\alpha)}}} \\ &= u - 2u\lim_{\alpha\to\infty}\sqrt{1+\frac{4\eta(1-\eta)}{(1-2\eta)^2\exp(2u\alpha)}} \\ &= -u. \end{aligned}$$

Here we use l'Hôpital's rule in calculating limits.

Third, let $\alpha = \exp(u/\rho)$. It is then immediately calculated that

$$f_*'(\eta) = 2\rho \frac{\alpha + \frac{(1-2\eta)(1-\alpha^2)}{\sqrt{(1-2\eta)^2\alpha^2+4\eta(1-\eta)}}}{(2\eta-1)\alpha + \sqrt{(1-2\eta)^2\alpha^2+4\eta(1-\eta)}} + \frac{\rho}{1-\eta}.$$

Consider that

$$\begin{aligned} A &= \frac{2\alpha + \frac{2(1-2\eta)(1-\alpha^2)}{\sqrt{(1-2\eta)^2\alpha^2+4\eta(1-\eta)}}}{(2\eta-1)\alpha + \sqrt{(1-2\eta)^2\alpha^2+4\eta(1-\eta)}} - \frac{1}{\eta} \\ &= \frac{\alpha - \frac{2\eta+(1-2\eta)\alpha^2}{\sqrt{(1-2\eta)^2\alpha^2+4\eta(1-\eta)}}}{\eta(2\eta-1)\alpha + \eta\sqrt{(1-2\eta)^2\alpha^2+4\eta(1-\eta)}}. \end{aligned}$$

It suffices for $f_*'(\eta) \geq \frac{\rho}{\eta(1-\eta)}$ to show $A \geq 0$. Note that

$$\frac{(2\eta+(1-2\eta)\alpha^2)^2}{(1-2\eta)^2\alpha^2+4\eta(1-\eta)} - \alpha^2 = \frac{4\eta^2(1-\alpha^2)}{(1-2\eta)^2\alpha^2+4\eta(1-\eta)} \leq 0$$

due to $\alpha \geq 1$, with equality when and only when $\alpha = 1$ or, equivalently, $u = 0$. Accordingly, we have $\alpha - \frac{2\eta+(1-2\eta)\alpha^2}{\sqrt{(1-2\eta)^2\alpha^2+4\eta(1-\eta)}} \geq 0$.

Appendix E. The Proof of Theorem 11

In order to prove the theorem, we define

$$\delta_\gamma := \sup\{t : \gamma t^2 \leq 2V_\rho(0)\} = \sqrt{2/\gamma}$$

for $\gamma > 0$ and let $V_\rho^{(\gamma)}(yf)$ be the coherence function $V_\rho(yf)$ restricted to $\mathcal{Y} \times [-\delta_\gamma k_{\max}, \delta_\gamma k_{\max}]$, where $k_{\max} = \max_{\mathbf{x}\in\mathcal{X}} K(\mathbf{x},\mathbf{x})$. For the Gaussian RBF kernel, we have $k_{\max} = 1$.

It is clear that

$$\|V_\rho^{(\gamma)}\|_\infty := \sup\left\{V_\rho^{(\gamma)}(yf), (y,f) \in \mathcal{Y} \times [-\delta_\gamma k_{\max}, \delta_\gamma k_{\max}]\right\} = \rho \log\left(1 + \exp\frac{u+k_{\max}\sqrt{2/\gamma}}{\rho}\right).$$

Considering that

$$\lim_{\gamma\to 0} \frac{\|V_\rho^{(\gamma)}\|_\infty}{k_{\max}\sqrt{2/\gamma}} = \lim_{\alpha\to\infty} \frac{\exp\frac{u+\alpha}{\rho}}{1+\exp\frac{u+\alpha}{\rho}} = 1,$$

we have $\lim_{\gamma\to 0} \|V_\rho^{(\gamma)}\|_\infty / \sqrt{1/\gamma} = \sqrt{2}k_{\max}$. Hence, we have $\|V_\rho^{(\gamma)}\|_\infty \sim \sqrt{1/\gamma}$.

On the other hand, since

$$V_\rho^{(\gamma)}(yf) = V_\rho^{(\gamma)}(yf_1) - \frac{\partial V_\rho^{(\gamma)}(yf_2)}{\partial f}(f - f_1),$$

where $f_2 \in [f, f_1] \subseteq [-\delta_\gamma k_{\max}, \delta_\gamma k_{\max}]$, we have

$$
\begin{aligned}
|V_\rho^{(\gamma)}|_1 &:= \sup\left\{ \frac{|V_\rho^{(\gamma)}(yf) - V_\rho^{(\gamma)|}(yf_1)}{|f-f_1|},\ y \in \mathcal{Y}, f, f_1 \in [-\delta_\gamma k_{\max}, \delta_\gamma k_{\max}], f \neq f_1 \right\} \\
&= \sup\left\{ \Big|\frac{\partial V_\rho^{(\gamma)}(yf_2)}{\partial f}\Big|,\ y \in \mathcal{Y}, f_2 \in [-\delta_\gamma k_{\max}, \delta_\gamma k_{\max}] \right\} \\
&= \frac{\exp\frac{u + k_{\max}\sqrt{2/\gamma}}{\rho}}{1 + \exp\frac{u + k_{\max}\sqrt{2/\gamma}}{\rho}}.
\end{aligned}
$$

In this case, we have $\lim_{\gamma \to 0} |V_\rho^{(\gamma)}|_1 = 1$, which implies that $|V_\rho^{(\gamma)}|_1 \sim 1$.

We now immediately conclude Theorem 11 from Corollary 3.19 and Theorem 3.20 of Steinwart (2005).

Appendix F. The Proof of Theorem 12

Note that the condition for function in the theorem implies that $\phi''(z) > 0$ in $\mathbb{R}$, then it follows that $\phi''(z) > \delta$ for some positive δ in a compact region (and certainly also holds in any bounded region). We also denote

$$\bar{f}_n = \operatorname*{argmin}_f \int \phi(yf(\mathbf{x}))dF_n(\mathbf{x}, y)$$

and

$$
\begin{aligned}
L(f) &= \int \phi(yf(\mathbf{x}))dF(\mathbf{x}, y), \\
L_{n1}(f) &= \int \phi(yf(\mathbf{x}))dF_n(\mathbf{x}, y) + \frac{\gamma}{2}\|h\|^2_{\mathcal{H}_K}, \\
L_{n2}(f) &= \int \phi(yf(\mathbf{x}))dF_n(\mathbf{x}, y).
\end{aligned}
$$

We have $f^*, \bar{f}_n, \hat{f}_n$ are all unique, because the corresponding objective functions are strictly convex.

Taking the derivative of the functional $L(f)$ w.r.t. f yields

$$\int yv\phi'(yf^*)dF(\mathbf{x}, y) = 0 \quad \text{for any } v \in \mathcal{H}_K. \tag{18}$$

Differentiating the functional $L_{n2}(f)$ w.r.t. f, we have

$$\int yv\phi'(y\bar{f}_n)dF_n(\mathbf{x}, y) = 0 \quad \text{for any } v \in \mathcal{H}_K. \tag{19}$$

It follows from the derivative of the functional $L_{n1}(f)$ w.r.t. h that

$$\int yv\phi'(y\hat{f}_n)dF_n(\mathbf{x}, y) + \gamma\langle \hat{h}_n, v\rangle = 0 \quad \text{for any } v \in \mathcal{H}_K \tag{20}$$

with $\bar{f}_n = \bar{h}_n + \bar{\alpha}_n$ and $\hat{f}_n = \hat{\alpha}_n + \hat{h}_n$. Since $\{\hat{h}_n\}$ is uniformly bounded (the condition of the theorem), and from the property $|\hat{h}_n(\mathbf{x})| \leq \|\hat{h}_n\| K(\mathbf{x}, \mathbf{x})^{1/2}$, we have $\sup_{n,\mathbf{x}} |\hat{f}_n(\mathbf{x})| < M_1$ for some $M_1 > 0$. Similarly, $\sup_{n,\mathbf{x}} |\bar{f}_n(\mathbf{x})| < M_2$ for some $M_2 > 0$.

The left-hand side of (20) is a linear functional, so it has a representer which is of the form $G_1(\hat{h}_n, \hat{\alpha}_n, \gamma) = \psi(\hat{\alpha} + \hat{h}_n) + \gamma\hat{h}_n$. Hence, the Equation (20) can be written as

$$\langle \psi(\hat{\alpha} + \hat{h}_n) + \gamma\hat{h}_n, v\rangle = 0$$

for any v. This implies $G_1(\hat{h}_n, \hat{\alpha}_n, \gamma) = 0$.

Differentiating $L_{n1}(f)$ w.r.t. α, we get $\int y\phi'(y\hat{f}_n(\mathbf{x}))dF_n(\mathbf{x}, y) = 0$ which is of the form $G_2(\hat{h}_n, \hat{\alpha}_n, \gamma) = 0$. Note that the second-order derivative of $L_{n1}(f)$ w.r.t. h and α (simply denoted as D^2) is bounded and bounded away from zero, because $|y_i\hat{f}_n(\mathbf{x}_i)|$ is uniformly bounded and $\phi''(z)$ satisfies the conditions in the theorem. Based on the implicit function theorem, $(\hat{h}_n, \hat{\alpha}_n)$ can be represented as $(\hat{h}_n, \hat{\alpha}_n)^T = \zeta(\gamma)$ for some function $\zeta(\cdot)$ (here we omit the other arguments), and $\zeta'(\gamma) = -(D^2)^{-1}(\hat{h}_n, 0)^T$. As a result, $|\zeta'(\gamma)| \leq M_3$ for some $M_3 > 0$. Then we have

$$\sup_n\{\|\hat{h}_n - \bar{h}_n\| + |\hat{\alpha}_n - \bar{\alpha}_n|\} \to 0 \text{ if } \gamma \to 0.$$

We need to estimate $\hat{f}_n - f^*$ which is represented as the sum of $\bar{f}_n - f^*$ and $\hat{f}_n - \bar{f}_n$. Expanding the function $\phi'(y\hat{f}_n)$ as

$$\phi'(y\hat{f}_n) = \phi'(y\bar{f}_n) + y\phi''(y\bar{f}_n)(\hat{f}_n - \bar{f}_n) + o(\hat{f}_n - \bar{f}_n)$$

and substituting them in Equation (20), we obtain (also keep in mind that (19) holds)

$$\int v\Big[\phi''(y\bar{f}_n) + o(1)\Big](\hat{f}_n - \bar{f}_n)dF_n(\mathbf{x}, y) = -\gamma\langle\hat{h}_n, v\rangle,$$

which is just

$$\frac{1}{n}\sum_{i=1}^{n}[v\phi''(y_i\bar{f}_n(\mathbf{x}_i)) + o(1)](\hat{f}_n - \bar{f}_n) = -\gamma\langle\hat{h}_n, v\rangle.$$

Note that we can chose v to be bounded, say, $\|v\| \leq 1$. From the property of $\phi''(z)$ and uniformly boundedness of $\{y_i\bar{f}_n(\mathbf{x}_i)\}$, we find that $v\phi''(y_i\bar{f}_n(\mathbf{x}_i)) + o(1)$ is uniformly bounded away from 0 if γ is sufficiently small. Hence we get $\sup_{n,\mathbf{x}}|\hat{f}_n(\mathbf{x}) - \bar{f}_n(\mathbf{x})| = O(\gamma)$, which implies $\int|\hat{f}_n(\mathbf{x}) - \bar{f}_n(\mathbf{x})|p(\mathbf{x})d\mathbf{x} = O(\gamma)$.

The next step is to estimate $\int|\bar{f}_n(\mathbf{x}) - f^*(\mathbf{x})|p(\mathbf{x})d\mathbf{x}$. Note that

$$\mathbb{E}(vy\phi'((y\bar{f}_n(\mathbf{x})) = \int vy\phi'(y\bar{f}_n(\mathbf{x}))dF(\mathbf{x}, y).$$

Based on the central limit theorem, we have

$$0 = \int vy\phi'(y\bar{f}_n(\mathbf{x})))dF_n(\mathbf{x}, y) = \int vy\phi'(y\bar{f}_n(\mathbf{x})))dF(\mathbf{x}, y) + O_p(\frac{1}{\sqrt{n}}).$$

Together with Equation (18), we have

$$\int vy[\phi'(y\bar{f}_n(\mathbf{x})) - \phi'(yf^*(\mathbf{x}))]dF(\mathbf{x}, y) = O_p(\frac{1}{\sqrt{n}}),$$

$$\Big|\int v\phi''(yf^*(\mathbf{x}))(\bar{f}_n(\mathbf{x}) - f^*(\mathbf{x}))dF(\mathbf{x}, y)\Big| = O_p(\frac{1}{\sqrt{n}}). \quad (21)$$

If $\bar{f}_n(\mathbf{x}) - f^*(\mathbf{x}) \neq 0$, we substitute $v(\mathbf{x}) = \frac{\bar{f}_n(\mathbf{x})-f^*(\mathbf{x})}{|\bar{f}_n(\mathbf{x})-f^*(\mathbf{x})|}$ into the Equation (21). From the boundedness of $f^*(\mathbf{x})$ and the property stated at the beginning of the proof, we conclude

$$\int \big|\bar{f}_n(\mathbf{x}) - f^*(\mathbf{x})\big| p(\mathbf{x}) d\mathbf{x} = O_p(\frac{1}{\sqrt{n}}).$$

Summarily, we have

$$\int |\hat{f}_n(\mathbf{x}) - f^*(\mathbf{x})| p(\mathbf{x}) d\mathbf{x} = O(\gamma) + O_p(\frac{1}{\sqrt{n}}).$$

References

U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences (PNAS)*, 96 (12):6745–6750, 1999.

P. Bartlett and A. Tewari. Sparseness vs estimating conditional probabilities: some asymptotic results. *Journal of Machine Learning Research*, 8:775–790, 2007.

P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

J. M. Bernardo and A. F. M. Smith. *Bayesian Theory*. John Wiley and Sons, New York, 1994.

O. Chapelle, B. Schölkopf, and A. Zien, editors. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006.

N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, Cambridge, U.K., 2000.

Y. Freund. Boosting a weak learning algorithm by majority. *Information and Computation*, 21: 256–285, 1995.

J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

J. H. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: A statistical view of boosting. *Annals of Statistics*, 28(2):337–374, 2000.

T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classitcation of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286: 531–537, 1999.

T. Hastie, R. Tishirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, New York, 2001.

Y. Lin. Tensor product space ANOVA models. *The Annals of Statistics*, 28(3):734–755, 2000.

Y. Lin. Support vector machines and the Bayes rule in classification. *Data Mining and Knowledge Discovery*, 6:259–275, 2002.

Y. Lin, G. Wahba, H. Zhang, and Y. Lee. Statistical properties and adaptive tuning of support vector machines. *Machine Learning*, 48:115–136, 2002.

B. K. Mallick, D. Ghosh, and M. Ghosh. Bayesian classification of tumours by using gene expression data. *Journal of the Royal Statistical Society Series B*, 67:219–234, 2005.

J. C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pages 61–74, Cambridge, MA, 1999. MIT Press.

K. Rose, E. Gurewitz, and G. C. Fox. Statistical mechanics and phase transitions in clustering. *Physics Review Letter*, 65:945–948, 1990.

X. Shen, G. C. Tseng, X. Zhang, and W. H. Wong. On ψ-learning. *Journal of the American Statistical Association*, 98:724–734, 2003.

P. Sollich. Bayesian methods for support vector machines: evidence and predictive class probabilities. *Machine Learning*, 46:21–52, 2002.

I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2:67 – 93, 2001.

I. Steinwart. Sparseness of support vector machines. *Journal of Machine Learning Research*, 4: 1071 – 1105, 2003.

I. Steinwart. Consistency of support vector machines and other regularized kernel classifiers. *IEEE Transactions on Information Theory*, 51(1):128 – 142, 2005.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1996.

V. Vapnik. *Statistical Learning Theory*. John Wiley and Sons, New York, 1998.

G. Wahba. *Spline Models for Observational Data*. SIAM, Philadelphia, 1990.

J. Wang, X. Shen, and Y. Liu. Probability estimation for large-margin classifiers. *Biometrika*, 95 (1):149–167, 2008.

L. Wang, J. Zhu, and H. Zou. Hybrid huberized support vector machines for microarray classification. In *Proceedings of the 24th International Conference on Machine Learning (ICML)*, pages 983–990, 2007.

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

T. Zhang and F. Oles. Text categorization based on regularized linear classification methods. *Information Retrieval*, 4:5–31, 2001.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society B*, 67:301–320, 2005.