# Multiclass Classification with Potential Function Rules: Margin Distribution and Generalization

Fei Teng[a,c], Yixin Chen[a,c,1], Xin Dang[b,c]

[a] *Department of Computer and Information Science,*
[b] *Department of Mathematics,*
[c] *The University of Mississippi, University, MS 38677, USA.*

**Abstract**

Motivated by the potential field of static electricity, a binary potential function classifier views each training sample as an electrical charge, positive or negative according to its class label. The resulting potential field divides the feature space into two decision regions based on the polarity of the potential. In this paper, we revisit potential function classifiers in their original form and reveal their connections with other well-known results in the literature. We derive a bound on the generalization performance of multi-class potential function classifiers based on the observed margin distribution of the training data. A new model selection criterion using a normalized margin distribution is then proposed to learn "good" potential function classifiers in practice.

*Keywords:* Multiclass classification, consistent classification rules, potential function rules, kernel rules, margin distribution, large margin classifiers, generalization bounds, model selection

## 1. Introduction

For thousands of years, various civilizations have observed "static electricity" where pieces of small objects with the same kind of electricity repelled each other and pieces with the opposite kind attracted each other. In pattern recognition and machine learning, *potential function rules* were motivated from the underlying property of static electricity to predict the

[1] To whom all correspondences should be addressed. Tel. 1-662-915-7438; Fax. 1-662-915-5623. ychen@cs.olemiss.edu

unknown binary nature of an observation, a problem commonly known as *binary classification*. Potential function rules were originally studied by Aizerman, Braverman, Rozonoer, and several other researchers in the 1960's ([1, 2, 3, 4, 10, 17, 18]). In its simplest form, a potential function rule puts a unit of positive electrical charge at every positive observation and a unit of negative electrical charge at every negative observation. The resulting potential field defines an intuitively appealing classifier: a new observation is predicted positive if the potential at that location is positive, and negative if its potential is negative.

In this paper, we revisit potential function rules (PFRs) in their original form and reveal their connections with other well-known results in the literature. We derive a bound on the generalization performance of potential function classifiers based on the observed margin distribution of the training data. A new model selection criterion using a normalized margin distribution is then proposed to learn "good" potential function classifiers in practice. There is an abundance of prior work in the field of pattern recognition and machine learning. It is beyond the scope of this article to supply a complete review of the area (for more comprehensive surveys on various subjects, the reader is referred to Devroye et al. [30], Duda et al. [35], Bishop [14] for patter recognition, to Schölkopf and Smola [58], Shawe-Taylor and Cristianini [59] for kernel methods, to Anthony and Biggs [5], Kearns and Vazirani [45] for computational learning theory, and to Mitchell [53], Hastie et al. [42], Vapnik [65] for machine/statistical learning). Nevertheless, a brief synopsis of some of the main findings most related to this article will serve to provide a rationale for the use of PFRs in pattern recognition and machine learning applications.

### *1.1. The Bayesian Decision Theory and Plug-In Decisions*

A multiclass classification problem aims at foretelling the unknown nature of an observation. More formally, an observation is a $d$-dimensional vector of numerical measurements denoted as $\mathbf{x} \in \mathbb{R}^d$. The unknown nature of the observation, $z$, takes values in a finite set $\mathbb{K} = \{1, 2, \ldots, K\}$, the set of *class labels*. A mapping $f : \mathbb{R}^d \to \mathbb{K}$, which is named a *classifier*, predicts the class label of an observation.

Does there exist an "optimal" classifier for a given classification task? Under a probabilistic setting, the Bayesian decision theory [11, 13] gives an affirmative answer – the *Bayes decision rule* (or called the Bayes classifier). If the pair of observation and its nature, $(\mathbf{x}, z)$, is a random variable with a joint

probability distribution $p(\mathbf{x}, z)$, the Bayes classifier, $f^*$, selects the class label for an observation $\mathbf{x}$ as $f^*(\mathbf{x}) = \text{argmax}_{z \in \mathbb{K}} \Pr(z|\mathbf{x}) = \text{argmax}_{z \in \mathbb{K}} p(\mathbf{x}, z)$. The optimality of $f^*$ is defined by the minimum probability of error, i.e., $\Pr[f^*(\mathbf{x}) \neq z] \leq \Pr[f(\mathbf{x}) \neq z]$ for any $f : \mathbb{R}^d \to \mathbb{K}$, which is well-known as the *Bayesian probability of error*. This probability measures the 'hardness' of a classification problem. It can theoretically be evaluated if the joint distribution is known, but the calculation may be (and usually is) intractable in practice due to the min operator inside the integral. Several tight bounds were proposed in the literature for computational approximations of the Bayesian probability of error [25, 41, 6].

The crux of the Bayesian approach is the difficulty of determining the joint distribution. Plug-in decision [30] is a natural way of applying the Bayesian classification in practice, where an approximated Bayes classifier is constructed using an estimated joint distribution. Depending upon the way in which the joint distribution is estimated, plug-in decision rules fall roughly into the parametric approaches and the nonparametric approaches.

In a parametric approach, the unknown joint distribution is described by a set of parameters based on certain structural assumptions, e.g., conditional independence of attributes within each class [49, 34, 15, 50], mixture of Gaussians [46, 61, 19], and mixture of Bernoullis [61]. The values of the parameters are obtained by optimizing a loss function, e.g., a likelihood function. In many applications, a parametric approach presents an efficient means of incorporating prior knowledge about the data. For example, Hofmann et al. [43] used a latent variable model (*aspect model*) to remove the statistical dependence among words in a document for textual data. Barnard et al. [7] explored several generative models to describe statistical relevance between images regions and associated texts. Veeramachaneni and Nagy [66] studied the interpattern dependence, named *style context*, for Optical Character Recognition. Intraclass style (statistical dependence between patterns of the same class in a field) and interclass style (statistical dependence between patterns of different classes in the same field) were formalized to derive style-constrained Bayesian classification.

The performance of a plug-in decision rule is determined by the quality of the estimated joint distribution. Ben-Bassat et al. [12] analyzed the sensitivity of Bayesian classification under multiplicative perturbation on the joint distribution. Devroye [26] presented a more general result showing that if the estimated posterior probability is close to the true posterior probability in $L_1$-sense, the error probability of the plug-in decision rule is near

the Bayesian probability of error. Nevertheless, does the error probability converge to the Bayesian probability of error if more training samples are obtained to approximate an arbitrary joint distribution? This is a question regarding the universal consistency of a classification rule. Loosely speaking, a *universally consistent rule* [30] guarantees us that taking more samples suffices to roughly reconstruct an arbitrary, fixed, but unknown distribution, hence to asymptotically achieve the optimality. While parametric approaches are efficient, in general they are not universally consistent.

In 1977, Stone proved the existence of a universally consistent rule [60]. He showed that any $k$-nearest neighbor classifier is universally consistent if $k$ is allowed to grow with $n$, the sample size, at a speed slower than that of $n$. Since then several rules have been shown to be universally consistent including histogram rules [37] and kernel rules [30]. We put these approaches under the category of nonparametric plug-in decisions because of the underlying nonparametric estimation of joint distributions. Representing all the data with a nonparametric model is sometimes preferred over summarizing it with a parametric model because of the rich detail held by very large data sources [40].

### *1.2. Classifier Selection and Complexity Regularization*

Universal consistency describes the asymptotic behavior of a classifier, i.e., the number of training samples goes to infinity. For real-life problems, however, the size of a training set is finite and, usually, fixed. This leads to a basic question in classifier design: how do we select a classifier, which performs well on future examples, from a given set of classifiers based on a given finite training set? Two basic principles were investigated in the literature for classifier selection: empirical risk minimization [63] and complexity regularization [51].

In order to achieve good generalization performance, the empirical risk minimization principle seeks for a classifier that minimizes the training error (empirical risk). Vapnik and Chervonenkis presented a theoretical ground for empirical risk minimization [63]. It was shown that if the 'capacity' of $\mathbb{C}$, the set of classifiers to choose from, is sufficiently restricted, minimizing the empirical risk guarantees a classifier whose performance is close to that of the best classifier in $\mathbb{C}$. Here the capacity of $\mathbb{C}$ is defined by the VC-dimension of $\mathbb{C}$. The above result reveals two competing factors in classifier selection. On one hand, a low capacity model set may not contain any classifier that generalizes well. On the other hand, too much freedom may over fit the

data resulting a model behaving like a refined look-up-table: perfect for the training data but poor on generalization.

This suggests that a classifier, built on a finite training set, generalizes the best if the right tradeoff is found between the training accuracy and the capacity of the model set. Complexity regularization applies the above idea to search for a classifier that minimizes the sum of empirical risk and a term penalizing the complexity [64, 8, 51, 9, 16]. Amongst various definitions of the penalty term, margin-based approaches received broad attention in the literature. A series of results were obtained that exhibit the intrinsic connection between generalization and different measures of margin distribution (e.g., maximal margin, margin percentile, soft margin) [65, 48, 59, 55, 39]. These theoretical results led to the discovery of new learning algorithms (e.g., support vector machines [65], margin distribution optimization [36], large margin multiple-instance learning [21], margin trees [62], large margin semi-supervised learning [67], dissimilarity-based learning [54], similarity-based learning [52, 22], large margin nearest neighbor classification [69, 44]), large margin conditional random fields [47], and new interpretations of known learning algorithms (e.g., boosting [57, 56], additive fuzzy systems [20]).

Classifiers derived from complexity regularization are not necessarily consistent. Lugosi and Zeger [51] presented a sufficient condition for universal consistency of a particular method of complexity regularization, structural risk minimization, using Vapnik-Chervonenkis complexity classes [65]. As discussed in Section 1.1, several nonparametric plug-in decision rules are universally consistent. A question arises naturally: Can we combine complexity regularization with a universally consistent nonparametric plug-in decision rule to improve the generalization performance thereof? This is the question that we attempt to address in this article. In particular, we focus on the regularization of multiclass PFRs based on a margin distribution.

### *1.3. A Historical Timeline of Potential Function Method*

Potential function method was invented by Russian researchers in the 1960's to model a general function reconstruction problem [1, 2, 3, 4, 10, 17, 18]. Various algorithms were proposed to tackle the reconstruction problem for specialized purposes. In a series of three articles, Aizerman, Braverman, and Rozonoer proposed algorithms using a potential function for binary classification [1], estimation of the posterior probability of binary classes [2], and approximating a functional relationship [3].

Under the specific problem settings, various convergence results were obtained [1, 17, 18, 4]. In [1], a learning method for potential function classifier was shown to converge in finite number of iterations when the two classes are separable by the given family of potential functions. Under the same assumption, a probabilistic bound was derived in [4] that can be used as a termination condition for the training process. The bound is essentially a sample complexity bound in computational learning theory [45]. Braverman and Pyatnitskii [17] established criteria for the choice of potential function used in convergent learning algorithms. In a later work, they estimated the rate of convergence [18].

In the simplest form, a binary PFR views each training sample as an electrical charge, positive or negative according to its class label. The resulting potential field divides the feature space into two decision regions based on the polarity of the potential. In the original formulation of PFR [1, 17], the electrical charge of each training sample is assigned a weight coefficient. In this article, we consider a simpler form with uniform weights. The basic idea of binary PFRs can be generalized to the multiclass scenario, in which a potential function is defined for each class using the training observations within that class. A new observation is then assigned a label corresponding to the class of the highest potential value. In addition to the intuitive appealingness, good scalability is a main advantage of PFRs in applications of a 'dynamic' nature, i.e., the structure of the problem may vary over time. Adding new classes does not affect the existing potential functions. Removing or merging classes influence only the potential functions of the classes involved in the operation.

### *1.4. An Overview of the Article*

The contributions of this article are given as follows:

- *Connections of PFRs with the Bayes decision theory.* Given charge density functions a priori, we present conditions under which a PFR is essentially optimal under the framework of the Bayesian decision theory. We then look into a more practical scenario where a PFR is built from a given set of training observations with unknown but fixed charge density functions. We show that a PFR is, in this case, equivalent to a plug-in decision rule using kernel density estimation, hence universally consistent.

- *A new generalization bound for PFRs.* We discuss the classifier selection for PFRs using complexity regularization. An upper bound on the generalization performance for PFRs is derived using a margin distribution.

- *A simple classifier selection method for PFRs.* Motivated by the above generalization bound, we propose a simple kernel selection method using a normalized margin distribution. Extensive experimental results on artificial data and real applications demonstrate the competitive performance of the proposed framework.

The remainder of the paper is organized as follows: Section 2 discusses PFRs from the perspective of the Bayesian decision theory. Conditions under which PFRs are equivalent to Bayes decision rules are presented. Section 3 shows a connection between a PFR and a plug-in decision rule using kernel density estimation. Section 4 presents a generalization bound for PFRs based on a margin distribution. Motivated by this bound on the generalization performance of PFRs, we propose in Section 5 a model selection method using a normalized margin distribution. In Section 6, we explain the experimental studies conducted and demonstrate the results. We conclude in Section 7.

## 2. Potential Function Rules and The Bayes Decision Theory

We start with a brief review of electrostatic potential functions [38]. We then introduce the general form of binary potential function classifiers. Finally, we demonstrate connections between PFRs and the Bayes classifiers.

### *2.1. Potential Function Rules*

Given a positive point charge at location $\mathbf{y}$, the electrostatic potential at location $\mathbf{x}$ is proportional to $\frac{1}{\|\mathbf{x}-\mathbf{y}\|}$, which is called the electrostatic point potential function. For a 'cloud' of positive charges with density $\rho_+$ over a space $\mathbb{X}$, the electrostatic potential function $\Phi$ is, modular a constant scale factor,

$$\Phi(\mathbf{x}) = \int_{\mathbb{X}} \frac{\rho_+(\mathbf{y})}{\|\mathbf{x}-\mathbf{y}\|} d\mathbf{y} \ .$$

Therefore, if $\rho_+$ and $\rho_-$ are respectively the charge density of positive and negative charges over $\mathbb{X}$, the electrostatic potential function $\Phi$ is, modular a

constant scale factor, defined as

$$\Phi(\mathbf{x}) = \int_{\mathbb{X}} \frac{\rho_+(\mathbf{y})}{\|\mathbf{x}-\mathbf{y}\|} d\mathbf{y} - \int_{\mathbb{X}} \frac{\rho_-(\mathbf{y})}{\|\mathbf{x}-\mathbf{y}\|} d\mathbf{y} \ .$$

The above electrostatic potential function can be generalized by replacing the electrostatic point potential function with a general *point potential function* $\psi : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$:

$$\Phi(\mathbf{x}) = \int_{\mathbb{X}} \rho_+(\mathbf{y})\psi(\mathbf{x},\mathbf{y}) d\mathbf{y} - \int_{\mathbb{X}} \rho_-(\mathbf{y})\psi(\mathbf{x},\mathbf{y}) d\mathbf{y} \ . \tag{1}$$

Note that the electrostatic potential at a location $\mathbf{x}$ is not well defined if $\mathbf{x}$ falls in the support of $\rho_+$ or $\rho_-$ due to the fact that $\frac{1}{\|\mathbf{x}-\mathbf{y}\|}$ is $\infty$ when $\mathbf{x} = \mathbf{y}$. This limitation, however, can be avoided by a general potential function (1) with a proper choice of the point potential function $\psi$.

Given $\rho_+$ and $\rho_-$, let $Q_+$ and $Q_-$ be the total positive charge and negative charge, respectively:

$$Q_+ = \int_{\mathbb{X}} \rho_+(\mathbf{x}) d\mathbf{x}, \quad Q_- = \int_{\mathbb{X}} \rho_-(\mathbf{x}) d\mathbf{x}.$$

We normalize the potential function (1) by the sum of the total positve and total negative charges:

$$\frac{\Phi(\mathbf{x})}{Q_+ + Q_-} = \frac{Q_+}{Q_+ + Q_-} \int_{\mathbb{X}} \frac{\rho_+(\mathbf{y})}{Q_+} \psi(\mathbf{x},\mathbf{y}) d\mathbf{y} - \frac{Q_-}{Q_+ + Q_-} \int_{\mathbb{X}} \frac{\rho_-(\mathbf{y})}{Q_-} \psi(\mathbf{x},\mathbf{y}) d\mathbf{y} \ .$$

It is not difficult to check that $\frac{\rho_+(\mathbf{x})}{Q_+}$ and $\frac{\rho_-(\mathbf{x})}{Q_-}$ can be viewed as probability density functions because they are nonnegative over $\mathbb{X}$ and $\int_{\mathbb{X}} \frac{\rho_+(\mathbf{x})}{Q_+} d\mathbf{x} = \int_{\mathbb{X}} \frac{\rho_-(\mathbf{x})}{Q_-} d\mathbf{x} = 1$, i.e., normalized charge densities are probability density functions. Therefore, we define conditional probability density functions as

$$p(\mathbf{x}|+) = \frac{\rho_+(\mathbf{x})}{Q_+}, \quad p(\mathbf{x}|-) = \frac{\rho_-(\mathbf{x})}{Q_-}, \tag{2}$$

and prior probability as

$$\Pr(+) = \frac{Q_+}{Q_+ + Q_-}, \quad \Pr(-) = \frac{Q_-}{Q_+ + Q_-}. \tag{3}$$

Consequently, the above normalized potential function is rewritten in terms of (2) and (3) as

$$\frac{\Phi(\mathbf{x})}{Q_+ + Q_-} = \Pr(+) \int_{\mathbb{X}} p(\mathbf{y}|+)\psi(\mathbf{x},\mathbf{y})d\mathbf{y} - \Pr(-) \int_{\mathbb{X}} p(\mathbf{y}|-)\psi(\mathbf{x},\mathbf{y})d\mathbf{y} \ .$$

Hence a binary potential function classifier is defined as

$$\begin{aligned} f(\mathbf{x}) &= \operatorname{sign}\left(\Phi(\mathbf{x})\right) \\ &= \operatorname{sign}\left(\Pr(+) \int_{\mathbb{X}} p(\mathbf{y}|+)\psi(\mathbf{x},\mathbf{y})d\mathbf{y} - \Pr(-) \int_{\mathbb{X}} p(\mathbf{y}|-)\psi(\mathbf{x},\mathbf{y})d\mathbf{y}\right), \quad (4) \end{aligned}$$

i.e., the polarity of the potential determines the class label.

*2.2. PFRs and Bayes Classifiers*

Next, we present a Bayesian interpretation of the above potential function classifier. In particular, we show that with a proper choice of $\psi$, the decision boundary of (4) is identical to that of the optimal Bayes classifier. Our first choice of $\psi$ is the Dirac delta function which is zero everywhere except at the origin, where it is infinite,

$$\delta(\mathbf{x}) = \begin{cases} +\infty & \mathbf{x} = \mathbf{0} \\ 0 & \mathbf{x} \neq \mathbf{0} \end{cases}$$

and which also satisfies the identity

$$\int_{-\infty}^{\infty} \delta(\mathbf{x})d\mathbf{x} = 1.$$

**Theorem 1.** *Let $\rho_+$ and $\rho_-$ be the charge densities; $p(\mathbf{x}|+)$, $p(\mathbf{x}|-)$, $\Pr(+)$, and $\Pr(-)$ be defined by (2) and (3). If we choose $\psi(\mathbf{x},\mathbf{y}) = \delta(\mathbf{x}-\mathbf{y})$, the decision boundary of the potential function classifier (4) is equivalent to that of the Bayes classifier for conditional probability distributions $p(\mathbf{x}|+)$ and $p(\mathbf{x}|-)$, and class prior probabilities $\Pr(+)$ and $\Pr(-)$.*

A proof of Theorem 1 is given in the Appendix. We may interpret the above theorem from the perspective of Fourier analysis. Specifically, for a translation invariant point potential function, i.e., $\psi(\mathbf{x},\mathbf{y}) = \psi(\mathbf{x}-\mathbf{y})$, the evaluation of $\int_{\mathbb{X}} p(\mathbf{y}|+)\psi(\mathbf{x}-\mathbf{y})d\mathbf{y}$ is essentially the convolution of $p(\mathbf{x}|+)$ and $\psi(\mathbf{x})$, which is equivalent to computing the inverse Fourier transform of

the product of the Fourier transforms of $p(\mathbf{x}|+)$ and $\psi(\mathbf{x})$. When $\psi$ is the Dirac delta function, potential function classifiers are equivalent to Bayes classifiers because the Fourier transform of the Dirac delta function is the constant 1.

Theorem 1 holds independent of the specific forms of the charge densities, i.e., it is distribution free. Nevertheless, the unboundedness of the Dirac delta function makes it a poor choice in numerical implementations. Next, by assuming that the Fourier transform of the charge densities have finite support, we extend the conclusion of Theorem 1 to a wider class of translation invariant point potential functions.

**Theorem 2.** *Let $\rho_+$ and $\rho_-$ be the charge densities; $p(\mathbf{x}|+)$, $p(\mathbf{x}|-)$, $\Pr(+)$, and $\Pr(-)$ be defined by (2) and (3). Let $\hat{p}_+(\boldsymbol{\omega})$ and $\hat{p}_-(\boldsymbol{\omega})$ be the Fourier transform of $p(\mathbf{x}|+)$ and $p(\mathbf{x}|-)$, respectively, i.e.,*

$$\begin{aligned} \hat{p}_+(\boldsymbol{\omega}) &= \int_{\mathbb{X}} p(\mathbf{x}|+) e^{-2\pi i \boldsymbol{\omega}^T \mathbf{x}} d\mathbf{x}, \\ \hat{p}_-(\boldsymbol{\omega}) &= \int_{\mathbb{X}} p(\mathbf{x}|-) e^{-2\pi i \boldsymbol{\omega}^T \mathbf{x}} d\mathbf{x}, \end{aligned}$$

*where $i$ is the complex number $\sqrt{-1}$. We assume that $\hat{p}_+$ and $\hat{p}_-$ have finite support, namely, there exist constants $s_+$ and $s_-$ such that $\hat{p}_+(\boldsymbol{\omega}) = 0$ for $\|\boldsymbol{\omega}\| \geq s_+$ and $\hat{p}_-(\boldsymbol{\omega}) = 0$ for $\|\boldsymbol{\omega}\| \geq s_-$. For any translation invariant point potential function $\psi(\mathbf{x}, \mathbf{y}) = \psi(\mathbf{x} - \mathbf{y})$, if its Fourier transform satisfies that $\Psi(\boldsymbol{\omega}) = 1$ for $\|\boldsymbol{\omega}\| < s = \max(s_+, s_-)$, the decision boundary of the potential function classifier (4) is identical to that of the Bayes classifier using conditional probability distributions $p(\mathbf{x}|+)$ and $p(\mathbf{x}|-)$, and class prior probabilities $\Pr(+)$ and $\Pr(-)$.*

A proof of Theorem 2 is given in the Appendix. The above theorem states that if the charge densities are 'band limited' (i.e., its Fourier transform is zero everywhere outside a hyperball of finite radius $s$) and the point potential function has value 1 over the support of charge densities in the frequency domain, the potential function conveys the same information as the class conditional density. In the one dimensional case, a possible choice of $\psi$ is a sinc function,

$$\psi(x, y) = \frac{\sin[2\pi s(x - y)]}{\pi(x - y)} = 2s \cdot \mathrm{sinc}[2s(x - y)],$$

whose Fourier transform is a rectangular window function

$$\Psi(\omega) = \begin{cases} 1 & |\omega| \leq s \\ 0 & |\omega| > s \end{cases} = \text{rect}\left(\frac{\omega}{2s}\right) \ .$$

This choice of $\Psi$ can be generalized to higher dimensional spaces: for a hyper-rectangular window function in a $d$-dimensional frequency domain,

$$\Psi(\boldsymbol{\omega}) = \begin{cases} 1 & |\omega_i| \leq s, \forall i \in [1,d] \\ 0 & |\omega_i| > s, \exists i \in [1,d] \end{cases} = \prod_{i=1}^{d} \text{rect}\left(\frac{\omega_i}{2s}\right) \ ,$$

the corresponding point potential function is

$$\psi(\mathbf{x}, \mathbf{y}) = (2s)^d \prod_{i=1}^{d} \text{sinc}[2s(x_i - y_i)] \ . \tag{5}$$

Theorem 2 has implications on the practical design of potential function classifiers using a finite training set. This will be discussed in the next section.

In statistics, an idea similar to Theorem 2 had been explored in nonparametric density estimation. Watson and Leadbetter [68] discussed the $L_2$ error of kernel density estimates and related it to the spectral property of the density function (i.e., characteristic function). They concluded that the form of the optimal estimate depends critically on the tail of the characteristic function. Davis [23, 24] showed that using a sinc kernel, with a carefully chosen scale factor, the Fourier integral error estimate is asymptotically optimal within a constant factor for all densities. The analyses in [68, 23, 24] were performed under the $L_2$ measure. Devroye [27, 28, 29] developed a series of asymptotic performance bounds for kernel estimates using the $L_1$ measure. Compared with the $L_2$ error, the $L_1$ error has a clearer physical interpretation. As consistent density estimates yield consistent classifiers, a plug-in decision rule using these kernel density estimates is naturally consistent.

## 3. Potential Function Rules as Plug-in Decision Rules

The main difficulty of using the potential function classifier (4) in practice is that charge densities are usually unknown. An approximation method is therefore presented in this section. Next, we first generalize the above binary potential function classifier to multiple classes. All the results discussed in Section 2 can be extended to the multi-class scenario. We then present an approximation on PFR and a discussion on its connection with plug-in decision rules.

*3.1. An Approximation on Multi-class Potential Function classifiers*

Let $z \in \mathbb{K} = \{1, \ldots, K\}$ be the class label of observation $\mathbf{x} \in \mathbb{X}$. The observation-label pair $(\mathbf{x}, z)$ is generated by a distribution $F$, which is a mixture of $K$ unknown distributions $F_1, \ldots, F_K$,

$$F = \sum_{k=1}^{K} P_k F_k,$$

where $P_k$ is the marginal probability of label $k$, i.e., $P_k = \Pr(z = k)$; $F_k$ is the cumulative distribution function of $\mathbf{x}$ conditioned on $z = k$. Analogous to (1), (2), and (3), we define $\Phi_k$ as a class potential function - the potential with respect to $P_k F_k$:

$$\Phi_k(\mathbf{x}) = P_k \int_{\mathbb{X}} \psi(\mathbf{x}, \mathbf{y}) dF_k(\mathbf{y}) \ . \tag{6}$$

A multi-class potential classifier is defined as

$$f(\mathbf{x}) = \operatorname*{argmax}_k \Phi_k(\mathbf{x}) \ . \tag{7}$$

Note that the class potential (6) is the product of $P_k$ and the expectation of the point potential function $\psi$ with respect to $F_k$, i.e.,

$$\Phi_k(\mathbf{x}) = P_k \mathbb{E}_{\mathbf{y} \sim F_k}[\psi(\mathbf{x}, \mathbf{y})] \ .$$

Although $F$ is unknown in most applications, a training set is usually given. Therefore, we approximate the above expectation by the sample mean. Let $\mathcal{S} = \{(\mathbf{x}_1, z_1), \ldots, (\mathbf{x}_\ell, z_\ell)\} \subset \mathbb{X} \times \mathbb{K}$ be the training set, a random i.i.d. sample from $F$.

**Definition 1 (Sample Class Potential Function).** *Given a point potential function $\psi : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$, we define the sample class potential of an observation $\mathbf{x}$ with respect to class $k$ and sample $\mathcal{S}$ as*

$$\phi_k(\mathbf{x}, \mathcal{S}) = \frac{1}{|\mathcal{S}|} \sum_{z_i = k} \psi(\mathbf{x}, \mathbf{x}_i). \tag{8}$$

A multi-class sample potential classifier is then defined using sample class potential functions as follows.

**Definition 2 (A Multi-class Sample Potential Function Classifier).** *Given $\mathcal{S}$, a set of i.i.d. training samples generated by an unknown distribution $F$ on $\mathbb{X} \times \mathbb{K}$, we define a potential classifier $f_{\mathcal{S}} : \mathbb{X} \to \mathbb{K}$ as*

$$f_{\mathcal{S}}(\mathbf{x}) = \operatorname*{argmax}_k \phi_k(\mathbf{x}, \mathcal{S}). \tag{9}$$

Clearly, the sample class potential (8) can be written as

$$\phi_k(\mathbf{x}, \mathcal{S}) = \frac{|\mathcal{S}_k|}{|\mathcal{S}|} \frac{1}{|\mathcal{S}_k|} \sum_{z_i=k} \psi(\mathbf{x}, \mathbf{x}_i) \ ,$$

where $\mathcal{S}_k = \{(\mathbf{x}, z) \in \mathcal{S} : z = k\}$. It is not difficult to observe that $\frac{|\mathcal{S}_k|}{|\mathcal{S}|}$ is an estimate of the marginal probability $P_k$. Furthermore, if we restrict $\psi$ to be a nonnegative translation invariant function and $\int_{\mathbb{X}} \psi(\mathbf{x}) d\mathbf{x} = c < \infty$, it is straightforward to show that $\frac{1}{c|\mathcal{S}_k|} \sum_{z_i=k} \psi(\mathbf{x}, \mathbf{x}_i)$ is an estimate of the probability density of $F_k$ at location $\mathbf{x}$ using the kernel density estimation ($\psi$ is the kernel function). Hence, for any given $\mathbf{x}$, $\phi_k(\mathbf{x}, \mathcal{S})$ is proportional to an estimation of the posterior probability $\Pr(z = k|\mathbf{x})$.

This implies that the family of multi-class potential function classifiers (9) includes those plug-in decision Bayes classifiers that use kernel density estimation. Therefore, if $\psi$ is chosen from regular kernels, the universal consistency of PFRs follows from the universal consistency of kernel rules (Devroye et al. [30]). Universal consistency characterizes an asymptotic property of a decision rule - a decision rule converges to the optimal solution as the number of training sample is sufficiently large. For kernel rules, universal consistency requires the 'width' of the kernel to decrease to 0 as the sample size increases to infinity. Next, we show that under the conditions of Theorem 2, for a fixed width of $\psi$ (i.e., $\frac{1}{s}$ in (5) is fixed), with high probability the prediction of a sample PFR converges to that of the Bayes classifier for any given input.

It is worth noting that when the point potential function $\frac{1}{\|\mathbf{x}-\mathbf{y}\|}$ is used, the above PFR is called Hilbert rule. Hilbert rules were investigated by Devroye, Krzyżak, and Györfi for density estimation, regression and classification [31, 32, 33]. An interesting property of Hilbert rules is that it does not have a smoothing parameter. Consistency theory of Hilbert rules was developed in [31, 32, 33].

### *3.2. The Potential Gap and the Generalization Performance*

For a set of numbers $a_1, \ldots, a_K$, the $k$-th smallest number is denoted by $a_{(k)}$, i.e., $a_{(1)} \leq a_{(2)} \leq \cdots \leq a_{(K)}$. We define the *potential gap* of a multi-class

classifier $f$ given in (7) on an observation $\mathbf{x}$ by

$$\Gamma(\mathbf{x}) = \Phi_{f(\mathbf{x})}(\mathbf{x}) - \Phi_{(K-1)}(\mathbf{x}), \tag{10}$$

which is the difference between the largest class potential and the second largest class potential at $\mathbf{x}$. It should be clear that $\Gamma(\mathbf{x}) \geq 0$. The following theorem demonstrates that under the conditions of Theorem 2 (class conditional densities are band limited), the performance of a sample potential classifier (9) is closely related to the potential gap.

**Theorem 3.** *Let $\mathcal{S} = \{(\mathbf{x}_1, z_1), \ldots, (\mathbf{x}_\ell, z_\ell)\} \subset \mathbb{R}^d \times \mathbb{K}$ be a random i.i.d. sample from $F$, a mixture of $K$ distributions $F_1, \ldots, F_K$: $F = \sum_{k=1}^{K} P_k F_k$, where $P_k$ is the marginal probability of class $k$; $F_k$, defined by a density function $p_k(\mathbf{x})$, is the distribution of $\mathbf{x}$ for class $k$. The conditional density functions are band limited, i.e., there exists $s > 0$ such that $\hat{p}_k(\boldsymbol{\omega}) = 0$ when $\|\boldsymbol{\omega}\| \geq s$ for all $k = 1, \ldots, K$, where $\hat{p}_k(\boldsymbol{\omega})$ is the Fourier transform of $p_k(\mathbf{x})$. For any $\mathbf{x} \in \mathbb{R}^d$ the following inequality holds:*

$$\Pr[f_{\mathcal{S}}(\mathbf{x}) \neq f^*(\mathbf{x})] \leq 2Ke^{-\frac{\ell\Gamma(\mathbf{x})^2}{2(2s)^{2d}}} , \tag{11}$$

*where $f_{\mathcal{S}}(\mathbf{x})$ is the sample potential function classifier given in Definition 2 with (5) being the point potential function, $f^*(\mathbf{x})$ the Bayes classifier, and $\Gamma(\mathbf{x})$ the potential gap.*

A proof of Theorem 3 is given in the Appendix.

Theorem 3 suggests that for any given $\mathbf{x}$ and a band limited joint probability density function, the probability that the sample potential function classifier behaves differently from the Bayes classifier depends on two parameters: the potential gap $\Gamma(\mathbf{x})$ and the sample size $\ell$. The larger the potential gap and the sample size, the more likely that the sample potential function classifier makes the optimal prediction. In this sense, the generalization performance of $f_{\mathcal{S}}$ depends on the potential gap. Nevertheless, Theorem 3 does not tell us how to pick a sample size $\ell$, neither could we compute the right hand side of the inequality (11), because the potential gap is unknown in practice. Motivated by the potential gap, we present, in the next section, a probabilistic bound on the generalization performance of a sample potential function classifier based on the margin of $f_{\mathcal{S}}$, which is closely related to the sample version of the potential gap,.

## 4. A Generalization Bound for Potential Function Classifiers

As indicated in Definition 1, the sample class potential $\phi_k(\mathbf{x}, \mathcal{S})$ is an estimate of the class potential $\Phi_k(\mathbf{x})$. Analogous to the potential gap, we define the *margin* of $f_{\mathcal{S}}$ on an observation $(\mathbf{x}, z) \in \mathbb{R}^d \times \mathbb{K}$ as

$$\gamma(\mathbf{x}, z, \mathcal{S}) = \phi_z(\mathbf{x}, \mathcal{S}) - \phi_{(K-1)}(\mathbf{x}, \mathcal{S}) \ . \tag{12}$$

Given a classifier $f_{\mathcal{S}}$ and a desired margin $\alpha > 0$, we denote by $\xi$ the bounded amount by which $f_{\mathcal{S}}$ fails to achieve the desired margin $\alpha$ on sample $(\mathbf{x}, z)$,

$$\xi = \min\{\alpha, [\alpha - \gamma(\mathbf{x}, z, \mathcal{S})]_+\} \ ,$$

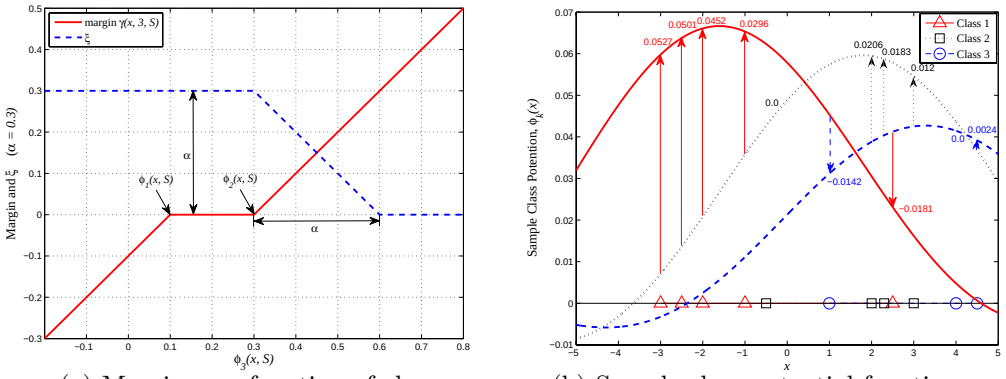
where $[x]_+ = x$ if $x \geq 0$ and 0 otherwise. For an observation $(\mathbf{x}_i, z_i) \in \mathcal{S}$, we define its margin shortage, $\xi_i$, as

$$\xi_i = \min\{\alpha, [\alpha - \gamma(\mathbf{x}_i, z_i, \mathcal{S}(i))]_+\} \ , \tag{13}$$

where $\mathcal{S}(i) = \mathcal{S} - \{(\mathbf{x}_i, z_i)\}$. Note that both $\xi$ and $\xi_i \in [0, \alpha]$.

We illustrate the concepts of margin and $\xi$ in Figure 1 under a 3-class scenario. The solid curve in Figure 1(a) shows the variations of the margin for an observation, $(\mathbf{x}, 3)$, as a function of its sample class potential $\phi_3(\mathbf{x}, \mathcal{S})$. The sample class potentials of $\mathbf{x}$ with respect to class 1 and class 2, i.e, $\phi_1(\mathbf{x}, \mathcal{S})$ and $\phi_2(\mathbf{x}, \mathcal{S})$, are fixed. For a desired margin $\alpha = 0.3$, the dashed curve represents the value of $\xi$: the bounded amount by which the margin is less than $\alpha$. Figure 1(b) shows three sample class potential functions constructed from 12 training observations. Each class is associated with a distinct marker: circle, triangle, or square. The point potential function defined in (5) with $s = 0.1$ is used in the evaluation of the sample class potential functions. We visualize the margin for each training observation using an arrow where the margin is computed as the difference between the vertical coordinate of the tip of the arrow ($\phi_z(\mathbf{x}, \mathcal{S})$) and that of the end of the arrow ($\phi_{(2)}(\mathbf{x}, \mathcal{S})$). The numerical value of a margin is also listed along with the arrow. For observations $(-0.5, 2)$ and $(4, 3)$, the arrows are absent because their margins are 0.

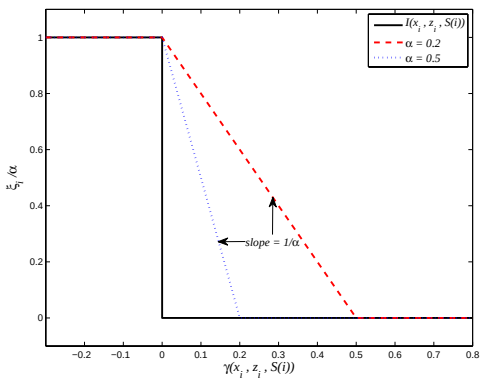
It is not difficult to relate margins to classification errors. Positive margins suggest correct classifications. Negative margins imply mis-classifications. There are only two scenarios that result in the 0 margin: $\phi_z(\mathbf{x}, \mathcal{S}) = \phi_{(K)}(\mathbf{x}, \mathcal{S}) = \phi_{(K-1)}(\mathbf{x}, \mathcal{S})$ or $\phi_{(K)}(\mathbf{x}, \mathcal{S}) > \phi_z(\mathbf{x}, \mathcal{S}) = \phi_{(K-1)}(\mathbf{x}, \mathcal{S})$. In the former case, which is rare in practice, the correctness of the classification depends on the



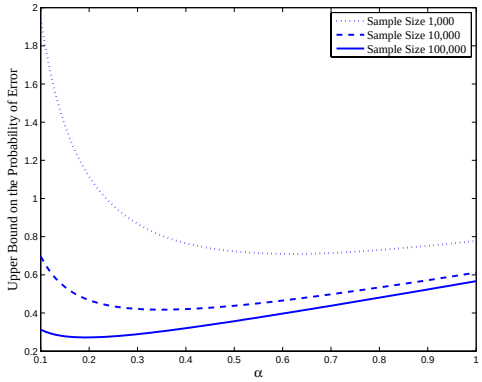

(a) Margin as a function of class potential.

(b) Sample class potential functions and margins.

Figure 1: Sample class potential functions and margins under a 3-class scenario. (a) The solid curve describes the variation of margin $\gamma(\mathbf{x}, 3, \mathcal{S})$ with respect to the sample class potential $\phi_3(\mathbf{x}, \mathcal{S})$ when the sample class potential $\phi_1(\mathbf{x}, \mathcal{S})$ and $\phi_2(\mathbf{x}, \mathcal{S})$ are fixed. The dashed curve represents $\xi$, the bounded amount by which the margin is less than $\alpha = 0.3$. (b) The three curves represent sample class potential functions built upon 12 training observations (denoted by the markers on the horizontal axis) using a 1-d sinc point potential function with $s = 0.1$. Each arrow corresponds to a margin, which is computed as the difference between the vertical coordinate of the tip of the arrow and that of the end of the arrow. The numeric value of the margin is given along with the arrow. The arrow is absent if the margin is 0.




(a) Bounded margin shortage.




(b) An upper bound on generalization performance.

Figure 2: (a) Plots of $\xi_i$, the bounded margin shortage, as a function of the margin $\gamma(\mathbf{x}_i, z_i, \mathcal{S}(i))$. When $\alpha$ approaches 0, $\frac{\xi_i}{\alpha}$ converges to the indicator function $I(\mathbf{x}_i, z_i, \mathcal{S}(i))$. (b) Plots of the upper bound on the probability of error in (16) as a function of the desired margin $\alpha$.

tie breaking strategy, which is usually random. The second case is more common, for example the two 0 margins in Figure 1(b). It leads to mis-classifications. If we introduce the following indicator function

$$I(\mathbf{x}, z, \mathcal{S}) = \begin{cases} 1 & \gamma(\mathbf{x}, z, \mathcal{S}) \leq 0 \\ 0 & \text{otherwise} \end{cases} , \tag{14}$$

$\sum_{i=1}^{\ell} I(\mathbf{x}_i, z_i, \mathcal{S}(i))$ is an upper bound on the number of mis-classified observations in a leave-one-out evaluation.

The connection between the bounded margin shortage $\xi_i$, which is defined in (13), and a classification error is more subtle. If we divide $\xi_i$ by $\alpha$, we have

$$\frac{\xi_i}{\alpha} = \begin{cases} 1 & \gamma(\mathbf{x}_i, z_i, \mathcal{S}(i)) \leq 0 \\ 1 - \frac{\gamma(\mathbf{x}_i, z_i, \mathcal{S}(i))}{\alpha} & 0 < \gamma(\mathbf{x}_i, z_i, \mathcal{S}(i)) \leq \alpha \\ 0 & 0 < \gamma(\mathbf{x}_i, z_i, \mathcal{S}(i)) \end{cases} . \tag{15}$$

Figure 2(a) compares $\frac{\xi_i}{\alpha}$ with $I(\mathbf{x}_i, z_i, \mathcal{S}(i))$ as a function of $\gamma(\mathbf{x}_i, z_i, \mathcal{S}(i))$. It is clear that $\frac{\xi_i}{\alpha}$ is always greater than or equal to $I(\mathbf{x}_i, z_i, \mathcal{S}(i))$. Therefore, $\sum_{i=1}^{\ell} \frac{\xi_i}{\alpha}$ is an upper bound on the number of mis-classified observations in a leave-one-out evaluation. Next, we present a generalization bound based on the desired margin $\alpha$ and the bounded margin shortage $\xi_i$ for any given

bounded point potential function $\psi$. Without loss of generality, we assume that $\psi : \mathbb{X} \times \mathbb{X} \to [0, 1]$.

**Theorem 4.** *Let $\mathcal{S} = \{(\mathbf{x}_1, z_1), \ldots, (\mathbf{x}_\ell, z_\ell)\} \subset \mathbb{X} \times \mathbb{K}$ be a random i.i.d. sample from an unknown distribution $F$, and $f_\mathcal{S} : \mathbb{X} \to \mathbb{K}$ a sample potential function classifier defined according to (9) using a given point potential function $\psi : \mathbb{X} \times \mathbb{X} \to [0, 1]$. For a fixed $\delta \in (0, 1)$, a desired margin $\alpha > 0$, and a new random sample $(\mathbf{x}, z)$ generated from $F$, the following bound holds with probability at least $1 - \delta$ over $\mathcal{S}$:*

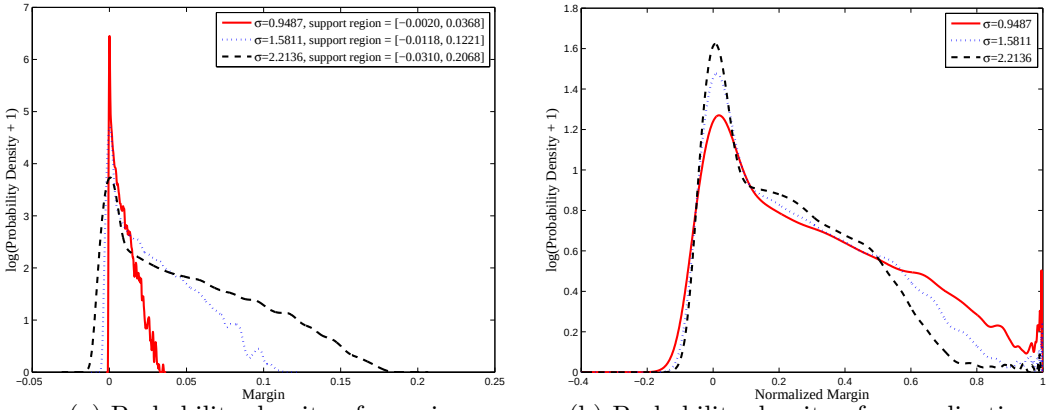
$$\Pr_F \left[ z \neq f_\mathcal{S}(\mathbf{x}) | \mathcal{S} \right] \leq \frac{1}{\ell} \sum_{i=1}^{\ell} \frac{\xi_i}{\alpha} + \frac{2}{\ell\alpha} + \left( 1 + \frac{4}{\alpha} \right) \sqrt{\frac{\ln(2/\delta)}{2\ell}} \,. \tag{16}$$

*where $\xi_i$ is defined in (13).*

A proof of Theorem 4 is given in the Appendix. It is worthwhile to note that there are two sources of randomness in the above inequality: the random sample $\mathcal{S}$ and the random observation $(\mathbf{x}, z)$. For a specific $\mathcal{S}$, the above bound is either true of false, i.e., it is not random. For a random sample $\mathcal{S}$, the probability that the bound is true is at least $1 - \delta$. The inequality shows that the error probability, $\Pr_F \left[ z \neq f_\mathcal{S}(\mathbf{x}) | \mathcal{S} \right]$, of a sample potential function classifier depends on three terms. The first term, $\frac{1}{\ell} \sum_{i=1}^{\ell} \frac{\xi_i}{\alpha}$, is an upper bound on the leave-one-out training error. The second and the third terms are determined by the training sample size $\ell$, the desired margin $\alpha$, and the confidence parameter $\delta$. In general, for fixed $\ell$ and $\delta$, the generalization performance of $f_\mathcal{S}$ is a trade-off between training error and the desired margin $\alpha$. On one hand, a smaller $\alpha$ produces a tighter bound on the training error, but larger values for the second and the third term. On the other hand, a larger $\alpha$ can reduce the values of the second and the third term, but makes the first term a looser bound on the training error. This is illustrated in Figure 2(b) using margins generated from a uniform distribution on $[-0.1, 1]$. The values of the upper bound are shown as a function of the desired margin $\alpha$. In the next section, we discuss classifier selection methods motivated by the above bound on the generalization performance.

## 5. Margin Distributions and Classifier Selection

The learning of a potential function classifier is essentially the selection of a point potential function (or its parameters). Figure 2(b) shows that




(a) Probability density of margin.

(b) Probability density of normalization margin.

Figure 3: Distributions of margin and normalized margin under a Gaussian point potential function $e^{-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{\sigma^2}}$ with different values of $\sigma$.

given $\ell$ and $\delta$, the upper bound on the probability of error has a minimum. Hence it is tempting to choose a classifier that minimizes the upper bound in (16). Unfortunately, this is not an effective approach in practice because the bound is usually loose even for large training sets with $50,000$–$100,000$ observations.

As we discussed in Section 4, the desired margin $\alpha$ plays a key role in estimating the generalization performance. If we define

$$i^* = \underset{i=1,\ldots,\ell,\gamma(\mathbf{x}_i,z_i,\mathcal{S}(i))>0}{\operatorname{argmin}} \gamma(\mathbf{x}_i, z_i, \mathcal{S}(i)),$$

it is clear from Figure 2(a) that $\frac{1}{\ell}\sum_{i=1}^{\ell}\frac{\xi_i}{\alpha}$ achieves the minimum (which is equal to the training error) when $0 < \alpha \leq \gamma(\mathbf{x}_{i^*}, z_{i^*}, \mathcal{S}(i^*))$ [2]. Although a larger value of $\alpha$ decreases the values of the last two terms in (16), it also increases the value of $\frac{1}{\ell}\sum_{i=1}^{\ell}\frac{\xi_i}{\alpha}$. However, for a fixed value of $\alpha$, the bound is tigher if the margins are concentrated more towards the positive end than towards the negative end. This suggest that we may select classifiers based on the distribution of margins.

However, a direct comparison of margin distributions may not be meaningful because the support region of a margin distribution largely depends

[2]This is because there will be no observations whose margin falls into the sloped region. Hence $\frac{1}{\ell}\sum_{i=1}^{\ell}\frac{\xi_i}{\alpha} = \frac{1}{\ell}\sum_{i=1}^{\ell} I(\mathbf{x}_i, z_i, \mathcal{S}(i))$, which is the leave-one-out training error.

on the selected point potential function $\psi$ and its parameters. For example, Figure 3(a) shows the probability distributions of margins under a Gaussian point potential function (i.e., $\psi(\mathbf{x}, \mathbf{y}) = e^{-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{\sigma^2}}$) using the MAGIC dataset from UCI Machine Learning Repository (details of the dataset are given in Section 6). The support region of the margin distribution varies significantly with the values of $\sigma$.

To make margins comparable under different point potential functions or different parameter values, we propose the following normalization procedure. For any given $\mathbf{x} \in \mathbb{R}^d$, we define a normalized sample class potential, $\hat{\phi}_k(\mathbf{x}, \mathcal{S})$, as

$$\hat{\phi}_k(\mathbf{x}, \mathcal{S}) = \frac{\phi_k(\mathbf{x}, \mathcal{S})}{\sum_{i=1}^{K} |\phi_i(\mathbf{x}, \mathcal{S})|}.$$

Clearly, the above normalization does not change the order of sample class potentials, hence the classification decisions. The normalized margin of $f_{\mathcal{S}}$ on an observation $(\mathbf{x}, z) \in \mathbb{R}^d \times \mathbb{K}$ is then defined as

$$\hat{\gamma}(\mathbf{x}, z, \mathcal{S}) = \hat{\phi}_z(\mathbf{x}, \mathcal{S}) - \hat{\phi}_{(K-1)}(\mathbf{x}, \mathcal{S}) \ .$$

Figure 3(b) shows the probability densities of the margins after normalization. In both figures, the densities are shown under a log transformation. As we discussed in Section 3.1, if $\psi$ is a nonnegative translation invariant function that is integrable over $\mathbb{X}$, $\phi_k(\mathbf{x}, \mathcal{S})$ is *proportional to* an estimation of the posterior probability $\Pr(z = k|\mathbf{x})$. The normalized class potential, $\hat{\phi}_k(\mathbf{x}, \mathcal{S})$, is an estimate of the posterior probability $\Pr(z = k|\mathbf{x})$. Hence the normalized margin $\hat{\gamma}$ can be viewed as an estimation on the posterior probability gap.

In classifier selection, we would like to choose a classifier whose margins concentrate towards the positive end. In terms of normalized margin, this suggests that $\hat{\gamma}$ should concentrate towards 1. We propose the following metric:

$$h(f_{\mathcal{S}}) = \mathrm{var}_{\hat{\gamma}} - \mathrm{mean}_{\hat{\gamma}} \tag{17}$$

where $\mathrm{mean}_{\hat{\gamma}} = \frac{1}{\ell} \sum_{i=1}^{\ell} \hat{\gamma}(\mathbf{x}_i, z_i, \mathcal{S}(i))$ and $\mathrm{var}_{\hat{\gamma}} = \frac{1}{\ell-1} \sum_{i=1}^{\ell} [\hat{\gamma}(\mathbf{x}_i, z_i, \mathcal{S}(i)) - \mathrm{mean}_{\hat{\gamma}}]^2$. Clearly, the desired normalized margins should have large mean and small variance, i.e., we select a classifier that minimizes $h$.

## 6. Experimental Results

We compare the proposed model selection method using normalized margin distribution with a traditional approach using the leave-one-out training

Table 1: The comparison results of model selection using leave-one-out error and the margin distribution metric defined in (17). $\ell_{train}$: the size of training set; $\ell_{test}$: the size of test set; $d$: feature dimension; $K$: the number of classes; $n_B$: the number of experiments in which the proposed method performs better than leave-one-out model selection; $n_E$: the number of experiments in which the proposed method and leave-one-out model selection have same test error; $n_W$: the number of experiments in which the proposed method performs worse than leave-one-out model selection.

| **Dataset** | $\ell_{train}$ | $\ell_{test}$ | $d$ | $K$ | $n_B$ | $n_E$ | $n_W$ |
|---|---|---|---|---|---|---|---|
| **Balancescale** | 570 | 55 | 4 | 2 | **7** | **37** | **6** |
| **Bloodtransfusion** | 600 | 148 | 4 | 2 | **29** | **13** | **8** |
| **Breastcancer** | 600 | 83 | 9 | 2 | **10** | **39** | **1** |
| **Ecoli** | 200 | 136 | 7 | 8 | **15** | **22** | **13** |
| **Glass** | 150 | 64 | 9 | 6 | **22** | **21** | **7** |
| **Imageseg** | 2100 | 210 | 19 | 7 | **15** | **22** | **13** |
| Ionosphere | 320 | 31 | 34 | 2 | 5 | 29 | 16 |
| **Letter** | 18000 | 2000 | 16 | 26 | **32** | **7** | **11** |
| **Liver** | 300 | 45 | 6 | 2 | **23** | **14** | **13** |
| **Magic** | 10000 | 9020 | 10 | 2 | **44** | **4** | **2** |
| **Multi-Feature1** | 1800 | 200 | 216 | 10 | **9** | **34** | **7** |
| **Multi-Feature2** | 1800 | 200 | 64 | 10 | **7** | **43** | **0** |
| **Multi-Feature3** | 1800 | 200 | 240 | 10 | **9** | **39** | **2** |
| **Satimage** | 5835 | 600 | 36 | 6 | **16** | **23** | **11** |
| Sonar | 150 | 58 | 60 | 2 | 15 | 20 | 15 |
| **Spectfheart** | 200 | 67 | 44 | 2 | **3** | **37** | **0** |
| **Survival** | 206 | 100 | 3 | 2 | **22** | **8** | **20** |
| Vehicle | 800 | 46 | 18 | 4 | 8 | 31 | 11 |
| Vowel | 890 | 100 | 10 | 11 | 3 | 33 | 14 |
| Winequality | 6000 | 497 | 11 | 7 | 17 | 0 | 33 |

error. The experiments were conducted on 20 datasets from UCI Machine Learning Repository. Each dataset is randomly divided into a training set and a test set. We built a potential function classifier with Gaussian point potential function for each dataset. The bandwidth parameter $\sigma$ of the point potential function is determined from 20 different values (0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 2.0, 3.0, and 4.0), using two strategies: (1) minimizing the leave-one-out training error; (2) minimizing the margin distribution metric defined in (17). The above procedure was repeated for 50 runs. In each run, test errors were recorded. In Table 1, we list the names of the datasets, the sizes of training and test sets, the dimension of the feature space, the number of categories, and the number of runs in which the proposed model selection method outperformed

($n_B$), tied with ($n_E$), and underperformed ($n_W$) the leave-one-out approach. Among the 20 datasets, the proposed method outperformed the leave-one-out model selection on 15 datasets (i.e. $n_B > n_W$), which are highlighted in Table 1. The two approaches tied on 1 dataset (Sonar). This suggests a very competitive performance of the proposed method.

## 7. Conclusions

In this paper, we revisited potential function rules (PFRs) in their original form and reveal their connections with other well-known results in the literature. We derive a bound on the generalization performance of potential function classifiers based on the observed margin distribution of the training data. A new model selection criterion using a normalized margin distribution is then proposed to learn "good" potential function classifiers in practice. We evaluated the proposed model selection method over 20 UCI data sets. In comparison with the traditional model selection using leave-one-out training error, the margin distribution based metric demonstrates very competitive performance.

## Acknowledgments

This work was supported by the US National Science Foundation under award number MCB-1027989 and EPS-0903787.

## Appendix A.

**Proof of Theorem 1:** Because $\delta(\cdot)$ is a Dirac delta function, it follows that

$$\int_{\mathbb{X}} p(\mathbf{y}|+)\delta(\mathbf{x}-\mathbf{y})d\mathbf{y} = p(\mathbf{x}|+), \int_{\mathbb{X}} p(\mathbf{y}|-)\delta(\mathbf{x}-\mathbf{y})d\mathbf{y} = p(\mathbf{x}|-).$$

Therefore,

$$\begin{aligned} \Pr(+)\int_{\mathbb{X}} p(\mathbf{y}|+)\delta(\mathbf{x}-\mathbf{y})d\mathbf{y} &\propto \Pr(+|\mathbf{x}) \\ \Pr(-)\int_{\mathbb{X}} p(\mathbf{y}|-)\delta(\mathbf{x}-\mathbf{y})d\mathbf{y} &\propto \Pr(-|\mathbf{x}), \end{aligned}$$

i.e., the potential of the positive (negative) class is proportional to the posterior probability of the positive (negative) class. Hence the decision boundary of (4) is identical to that of the Bayes classifier. □

**Proof of Theorem 2:** For a translation invariant $\psi$,

$$\int_{\mathbb{X}} p(\mathbf{y}|+)\psi(\mathbf{x},\mathbf{y})d\mathbf{y} = \int_{\mathbb{X}} p(\mathbf{y}|+)\psi(\mathbf{x}-\mathbf{y})d\mathbf{y} = \mathcal{F}^{-1}[\hat{p}_+(\boldsymbol{\omega})\Psi(\boldsymbol{\omega})]$$

where $\mathcal{F}^{-1}$ is the inverse Fourier transform. Because $\hat{p}_+(\boldsymbol{\omega}) = 0$ for $\|\boldsymbol{\omega}\| \geq s$ and $\Psi(\boldsymbol{\omega}) = 1$ for $\|\boldsymbol{\omega}\| \leq s$, we have $\hat{p}_+(\boldsymbol{\omega})\Psi(\boldsymbol{\omega}) = \hat{p}_+(\boldsymbol{\omega})$. It follows that

$$\Pr(+)\int_{\mathbb{X}} p(\mathbf{y}|+)\psi(\mathbf{x},\mathbf{y})d\mathbf{y} = \Pr(+)p(\mathbf{x}|+) \propto \Pr(+|\mathbf{x}).$$

Similarly,

$$\Pr(-)\int_{\mathbb{X}} p(\mathbf{y}|-)\psi(\mathbf{x},\mathbf{y})d\mathbf{y} = \Pr(-)p(\mathbf{x}|-) \propto \Pr(-|\mathbf{x}).$$

The potential of the positive (negative) class hence is proportional to the posterior probability of the positive (negative) class. Therefore the decision boundary of (4) is identical to that of the Bayes classifier. □

We need the following Lemma to prove Theorem 3.

**Lemma 1.** *For any $a_1, a_2, \ldots, a_K \in \mathbb{R}$ and $b_1, b_2, \ldots, b_K \in \mathbb{R}$, if $|a_k - b_k| \leq \epsilon$ for all $k \in \mathbb{K}$, we have $|a_{(j)} - b_{(j)}| \leq \epsilon$ for all $j \in \mathbb{K}$.*

**Proof:** For any $j \in \mathbb{K}$,

$$a_{(j)} - \epsilon \leq a_{(j+1)} - \epsilon \leq \cdots \leq a_{(K)} - \epsilon.$$

Because $b_k \geq a_k - \epsilon$ for all $k \in \mathbb{K}$, the number of $b_k$'s that are greater than or equal to $a_{(j)} - \epsilon$ is at least $K - j + 1$. Therefore $b_{(j)} \geq a_{(j)} - \epsilon$. Similarly, for any $j \in \mathbb{K}$,

$$a_{(1)} + \epsilon \leq a_{(2)} + \epsilon \leq \cdots \leq a_{(j)} + \epsilon.$$

Because $b_k \leq a_k + \epsilon$ for all $k \in \mathbb{K}$, the number of $b_k$'s that are less than or equal to $a_{(j)} + \epsilon$ is at least $j$. Therefore $b_{(j)} \leq a_{(j)} + \epsilon$. This completes the proof. □

**Proof of Theorem 3:** We introduce a new random variable $\mathbf{S}_k = |\{(\mathbf{x}, z) \in \mathcal{S} : z = k\}|$. For any given $\mathbf{x}$,

$$\begin{aligned}
\mathbb{E}_F[\phi_k(\mathbf{x},\mathcal{S})] &= \mathbb{E}_{\mathbf{S}_k}\left\{\mathbb{E}_{F|\mathbf{S}_k}\left[\frac{\mathbf{S}_k}{\ell}\frac{1}{\mathbf{S}_k}\sum_{z_i=k}\psi(\mathbf{x},\mathbf{x}_i)\right]\right\} \\
&= \mathbb{E}_{\mathbf{S}_k}\left\{\frac{\mathbf{S}_k}{\ell}\mathbb{E}_{\mathbf{y}\sim F_k}[\psi(\mathbf{x},\mathbf{y})]\right\} = P_k\mathbb{E}_{\mathbf{y}\sim F_k}[\psi(\mathbf{x},\mathbf{y})] = \Phi_k(\mathbf{x}).
\end{aligned}$$

We rewrite $\phi_k(\mathbf{x},\mathcal{S})$ as $\phi_k(\mathbf{x},\mathcal{S}) = \frac{1}{\ell}\sum_{i=1}^{\ell} I(z_i = k)\psi(\mathbf{x},\mathbf{x}_i)$ where the indicator function $I(z_i = k) = 1$ if $z_i = k$, $I(z_i = k) = 0$ otherwise. Because $(\mathbf{x}_i, z_i)$'s are i.i.d., so are $I(z_i = k)\psi(\mathbf{x},\mathbf{x}_i)$. In addition, from (5) it is clear that $|I(z_i = k)\psi(\mathbf{x},\mathbf{x}_i)| \leq (2s)^d$ . It follows from Hoeffding's inequality that for any given $\mathbf{x}$, $\epsilon > 0$, and $k = 1, \ldots, K$,

$$\Pr[|\phi_k(\mathbf{x},\mathcal{S}) - \Phi_k(\mathbf{x})| \geq \epsilon] \leq 2e^{-\frac{2\ell\epsilon^2}{(2s)^{2d}}} . \tag{A.1}$$

Because the conditional densities are band limited, it follows from the proof of Theorem 2 that

$$\Phi_k(\mathbf{x}) = P_k \mathbb{E}_{\mathbf{y}\sim F_k}[\psi(\mathbf{x},\mathbf{y})] \propto \Pr(z=k|\mathbf{x}) \ .$$

Hence we have $f^*(\mathbf{x}) = \text{argmax}_k \, \Phi_k(\mathbf{x})$. From Lemma 1, we know that if $|\phi_k(\mathbf{x},\mathcal{S}) - \Phi_k(\mathbf{x})| \leq \frac{\Gamma(\mathbf{x})}{2}$ for $k = 1,\ldots,K$, $|\phi_{(K)}(\mathbf{x},\mathcal{S}) - \Phi_{(K)}(\mathbf{x})| \leq \frac{\Gamma(\mathbf{x})}{2}$. Combining this with the facts that

$$\phi_{f_{\mathcal{S}}(\mathbf{x})}(\mathbf{x},\mathcal{S}) = \phi_{(K)}(\mathbf{x},\mathcal{S}) \text{ and } \Phi_{f^*(\mathbf{x})}(\mathbf{x}) = \Phi_{(K)}(\mathbf{x}) \ ,$$

it is straightforward to derive that $f_{\mathcal{S}}(\mathbf{x}) = f^*(\mathbf{x})$. Therefore,

$$\Pr\left[|\phi_k(\mathbf{x},\mathcal{S}) - \Phi_k(\mathbf{x})| < \frac{\Gamma(\mathbf{x})}{2}, \ \forall \ k = 1,\ldots,K\right] \leq \Pr[f_{\mathcal{S}}(\mathbf{x}) = f^*(\mathbf{x})] \ . \tag{A.2}$$

Let $\epsilon = \frac{\Gamma(\mathbf{x})}{2}$. Using (A.1), (A.2), and the union bound, we have

$$\Pr[f_{\mathcal{S}}(\mathbf{x}) \neq f^*(\mathbf{x})] \leq \Pr\left[\exists k, \ |\phi_k(\mathbf{x},\mathcal{S}) - \Phi_k(\mathbf{x})| \geq \frac{\Gamma(\mathbf{x})}{2}\right] \leq 2Ke^{-\frac{\ell\Gamma(\mathbf{x})^2}{2(2s)^{2d}}} \ .$$

This completes the proof. □

In order to prove the upper bound on generalization of sample potential function classifiers in Theorem 4, we need the following Lemma and an inequality attributed to McDiarmid.

**Lemma 2.** *Let $\mathcal{S}(i) = \mathcal{S} - \{(\mathbf{x}_i, z_i)\}$. For a change of one $(\mathbf{x}_t, z_t)$ to $(\hat{\mathbf{x}}_t, \hat{z}_t)$, denote*

$$\hat{\mathcal{S}}_t = \{(\mathbf{x}_1,z_1),\ldots,(\mathbf{x}_{t-1},z_{t-1}),(\hat{\mathbf{x}}_t,\hat{z}_t),(\mathbf{x}_{t+1},z_{t+1}),\ldots,(\mathbf{x}_\ell,z_\ell)\}.$$

*We define $\hat{\mathcal{S}}_t(i) = \hat{\mathcal{S}}_t - \{(\mathbf{x}_i, z_i)\}$, hence $\hat{\mathcal{S}}_t(t) = \mathcal{S}(t)$. Let $\mathbf{x} \in \mathbb{X}$ be any observation in $\mathbb{X}$ and $z \in \mathbb{K}$ a class label. The following inequalities hold for any point potential function $\psi : \mathbb{X}\times\mathbb{X} \to [0,1]$:*

$$|\gamma(\mathbf{x},z,\mathcal{S}) - \gamma(\mathbf{x},z,\mathcal{S}(i))| \ \leq \ \frac{2}{\ell}, \tag{A.3}$$

$$\left|\gamma(\mathbf{x},z,\mathcal{S}(i)) - \gamma(\mathbf{x},z,\hat{\mathcal{S}}_t(i))\right| \ \leq \ \frac{2}{\ell-1}, \tag{A.4}$$

$$\left|\gamma(\mathbf{x}_i,z_i,\mathcal{S}(i)) - \gamma(\mathbf{x}_i,z_i,\hat{\mathcal{S}}_t(i))\right| \ \leq \ \frac{2}{\ell-1}, \text{ if } i \neq t. \tag{A.5}$$

**Proof:** It is readily checked that for any $z \in \mathbb{K}$,

$$\begin{aligned} |\phi_z(\mathbf{x},\mathcal{S}) - \phi_z(\mathbf{x},\mathcal{S}(i))| \ &= \ \begin{cases} \left|\frac{1}{\ell}\sum_{z_j=z}\psi(\mathbf{x},\mathbf{x}_j) - \frac{1}{\ell-1}\sum_{z_j=z,j\neq i}\psi(\mathbf{x},\mathbf{x}_j)\right| & \text{if } z = z_i \\ \left|\frac{1}{\ell}\sum_{z_j=z}\psi(\mathbf{x},\mathbf{x}_j) - \frac{1}{\ell-1}\sum_{z_j=z}\psi(\mathbf{x},\mathbf{x}_j)\right| & \text{if } z \neq z_i \end{cases} \\ &= \ \begin{cases} \left|\frac{1}{\ell}\psi(\mathbf{x},\mathbf{x}_i) - \frac{1}{\ell(\ell-1)}\sum_{z_j=z,j\neq i}\psi(\mathbf{x},\mathbf{x}_j)\right| \leq \frac{1}{\ell} & \text{if } z = z_i \\ \left|\frac{1}{\ell(\ell-1)}\sum_{z_j=z}\psi(\mathbf{x},\mathbf{x}_j)\right| \leq \frac{1}{\ell} & \text{if } z \neq z_i \end{cases} . \end{aligned}$$

From (12) we have

$$\begin{aligned}|\gamma(\mathbf{x},z,\mathcal{S})-\gamma(\mathbf{x},z,\mathcal{S}(i))| &= \left|\phi_z(\mathbf{x},\mathcal{S})-\phi_{(K-1)}(\mathbf{x},\mathcal{S})-\phi_z(\mathbf{x},\mathcal{S}(i))+\phi_{(K-1)}(\mathbf{x},\mathcal{S}(i))\right| \\ &\leq \frac{1}{\ell}+\left|\phi_{(K-1)}(\mathbf{x},\mathcal{S})-\phi_{(K-1)}(\mathbf{x},\mathcal{S}(i))\right| \leq \frac{2}{\ell},\end{aligned}$$

where the last step is based on Lemma 1.

It is not difficult to show that for any $z \in \mathbb{K}$, $\left|\phi_z(\mathbf{x},\mathcal{S}(i))-\phi_z(\mathbf{x},\hat{\mathcal{S}}_t(i))\right| = 0$ when $i=t$, otherwise,

$$\left|\phi_z(\mathbf{x},\mathcal{S}(i))-\phi_z(\mathbf{x},\hat{\mathcal{S}}_t(i))\right| = \begin{cases} 0 & \text{if } z_t \neq z,\ \hat{z}_t \neq z \\ \left|\frac{1}{\ell-1}\psi(\mathbf{x},\hat{\mathbf{x}}_t)\right| & \text{if } z_t \neq z,\ \hat{z}_t = z \\ \left|\frac{1}{\ell-1}\psi(\mathbf{x},\mathbf{x}_t)-\frac{1}{\ell-1}\psi(\mathbf{x},\hat{\mathbf{x}}_t)\right| & \text{if } z_t = z,\ \hat{z}_t = z \\ \left|\frac{1}{\ell-1}\psi(\mathbf{x},\mathbf{x}_t)\right| & \text{if } z_t = z,\ \hat{z}_t \neq z \end{cases} \leq \frac{1}{\ell-1}.$$

Therefore,

$$\begin{aligned}&\left|\gamma(\mathbf{x},z,\mathcal{S}(i))-\gamma(\mathbf{x},z,\hat{\mathcal{S}}_t(i))\right| \\ &\leq \left|\phi_z(\mathbf{x},\mathcal{S}(i))-\phi_z(\mathbf{x},\hat{\mathcal{S}}_t(i))\right|+\left|\phi_{(K-1)}(\mathbf{x},\mathcal{S}(i))-\phi_{(K-1)}(\mathbf{x},\hat{\mathcal{S}}_t(i))\right| \\ &\leq \frac{2}{\ell-1}.\end{aligned}$$

Finally, for $i \neq t$ and any $z_i \in \mathbb{K}$,

$$\begin{aligned}&\left|\phi_{z_i}(\mathbf{x}_i,\mathcal{S}(i))-\phi_{z_i}(\mathbf{x}_i,\hat{\mathcal{S}}_t(i))\right| \\ &= \begin{cases} 0 & \text{if } z_t \neq z_i,\ \hat{z}_t \neq z_i \\ \left|\frac{1}{\ell-1}\psi(\mathbf{x}_i,\hat{\mathbf{x}}_t)\right| & \text{if } z_t \neq z_i,\ \hat{z}_t = z_i \\ \left|\frac{1}{\ell-1}\psi(\mathbf{x}_i,\mathbf{x}_t)-\frac{1}{\ell-1}\psi(\mathbf{x}_i,\hat{\mathbf{x}}_t)\right| & \text{if } z_t = z_i,\ \hat{z}_t = z_i \\ \left|\frac{1}{\ell-1}\psi(\mathbf{x}_i,\mathbf{x}_t)\right| & \text{if } z_t = z_i,\ \hat{z}_t \neq z_i \end{cases} \leq \frac{1}{\ell-1}.\end{aligned}$$

Therefore,

$$\begin{aligned}&\left|\gamma(\mathbf{x}_i,z_i,\mathcal{S}(i))-\gamma(\mathbf{x}_i,z_i,\hat{\mathcal{S}}_t(i))\right| \\ &\leq \left|\phi_{z_i}(\mathbf{x}_i,\mathcal{S}(i))-\phi_{z_i}(\mathbf{x}_i,\hat{\mathcal{S}}_t(i))\right|+\left|\phi_{(K-1)}(\mathbf{x}_i,\mathcal{S}(i))-\phi_{(K-1)}(\mathbf{x}_i,\hat{\mathcal{S}}_t(i))\right| \leq \frac{2}{\ell-1}.\end{aligned}$$

This completes the proof. □

**Lemma 3 (McDiarmid's Inequality).** *Let $X_1, X_2, \ldots, X_n$ be independent random variables taking values in a set $\mathbb{X}$. Suppose that $f : \mathbb{X}^n \to \mathbb{R}$ satisfies*

$$\sup_{\mathbf{x}_1,\ldots,\mathbf{x}_n,\hat{\mathbf{x}}_j \in \mathbb{X}} |f(\mathbf{x}_1,\ldots,\mathbf{x}_n)-f(\mathbf{x}_1,\ldots,\hat{\mathbf{x}}_j,\ldots,\mathbf{x}_n)| \leq c_j$$

*for constants* $c_j, 1 \le j \le n$. *Then for every* $\epsilon > 0$,

$$\Pr[f(X_1, \ldots, X_n) - \mathbb{E}f \ge \epsilon] \le \exp\left(\frac{-2\epsilon^2}{\sum_{j=1}^n c_j^2}\right).$$

**Proof of Theorem 4:** Consider the loss function

$$g(\mathbf{x}, z, \mathcal{S}) = \begin{cases} 1, & \text{if } \gamma(\mathbf{x}, z, \mathcal{S}) \le 0, \\ \frac{\alpha - \gamma(\mathbf{x}, z, \mathcal{S})}{\alpha}, & \text{if } 0 < \gamma(\mathbf{x}, z, \mathcal{S}) \le \alpha, \\ 0, & \text{otherwise.} \end{cases}$$

It is not difficult to show that

$$\Pr_F\left[z \ne f_{\mathcal{S}}(\mathbf{x}) | \mathcal{S}\right] \le \mathbb{E}_{F|\mathcal{S}}\left[g(\mathbf{x}, z, \mathcal{S})\right] \ ,$$

where the equality holds when $\alpha = 0$. Hence it suffices to show that $\mathbb{E}_{F|\mathcal{S}}\left[g(\mathbf{x}, z, \mathcal{S})\right]$ is bounded by the right side of (16).

We break $\mathbb{E}_{F|\mathcal{S}}[g(\mathbf{x}, z, \mathcal{S})] - \frac{1}{\ell\alpha}\sum_{i=1}^{\ell}\xi_i = \mathbb{E}_{F|\mathcal{S}}[g(\mathbf{x}, z, \mathcal{S})] - \frac{1}{\ell}\sum_{i=1}^{\ell} g(\mathbf{x}_i, z_i, \mathcal{S}(i))$ into $A + B + C$:

$$\begin{aligned} A &= \mathbb{E}_{F|\mathcal{S}}[g(\mathbf{x}, z, \mathcal{S})] - \mathbb{E}_{F|\mathcal{S}}\left[\frac{1}{\ell}\sum_{i=1}^{\ell} g(\mathbf{x}, z, \mathcal{S}(i))\right], \\ B &= \mathbb{E}_{F|\mathcal{S}}\left[\frac{1}{\ell}\sum_{i=1}^{\ell} g(\mathbf{x}, z, \mathcal{S}(i))\right] - \mathbb{E}_F[g(\mathbf{x}_j, z_j, \mathcal{S}(j))], \\ C &= \mathbb{E}_F[g(\mathbf{x}_j, z_j, \mathcal{S}(j))] - \frac{1}{\ell}\sum_{i=1}^{\ell} g(\mathbf{x}_i, z_i, \mathcal{S}(i)), \end{aligned}$$

where $(\mathbf{x}_j, z_j)$ is any fixed sample in $\mathcal{S}$.

We first look at $A$. It is straightforward to show that

$$|g(\mathbf{x}, z, \mathcal{S}) - g(\mathbf{x}, z, \mathcal{S}(i))| \le \frac{1}{\alpha}|\gamma(\mathbf{x}, z, \mathcal{S}) - \gamma(\mathbf{x}, z, \mathcal{S}(i))| \le \frac{2}{\ell\alpha},$$

where the last inequality is based on (A.3). Therefore

$$\begin{aligned} A &= \mathbb{E}_{F|\mathcal{S}}\left\{\frac{1}{\ell}\sum_{i=1}^{\ell}\left[g(\mathbf{x}, z, \mathcal{S}) - g(\mathbf{x}, z, \mathcal{S}(i))\right]\right\} \\ &\le \mathbb{E}_{F|\mathcal{S}}\left|\frac{1}{\ell}\sum_{i=1}^{\ell}\left[g(\mathbf{x}, z, \mathcal{S}) - g(\mathbf{x}, z, \mathcal{S}(i))\right]\right| \le \frac{2}{\ell\alpha}. \end{aligned} \tag{A.6}$$

Next, we look at $B$. It is not difficult to verify that

$$\mathbb{E}_F\left\{\mathbb{E}_{F|\mathcal{S}}\left[\frac{1}{\ell}\sum_{i=1}^{\ell} g(\mathbf{x}, z, \mathcal{S}(i))\right]\right\} = \mathbb{E}_F[g(\mathbf{x}_j, z_j, \mathcal{S}(j))].$$

For a change of one $(\mathbf{x}_t, z_t)$ to $(\hat{\mathbf{x}}_t, \hat{z}_t)$, we denote

$$\hat{\mathcal{S}}_t = \{(\mathbf{x}_1, z_1), \ldots, (\mathbf{x}_{t-1}, z_{t-1}), (\hat{\mathbf{x}}_t, \hat{z}_t), (\mathbf{x}_{t+1}, z_{t+1}), \ldots, (\mathbf{x}_\ell, z_\ell)\}.$$

From (A.4) we have for any $z \in \mathbb{K}$

$$\left| g(\mathbf{x}, z, \mathcal{S}(i)) - g(\mathbf{x}, z, \hat{\mathcal{S}}_t(i)) \right| \leq \frac{1}{\alpha} \left| \gamma(\mathbf{x}, z, \mathcal{S}(i)) - \gamma(\mathbf{x}, z, \hat{\mathcal{S}}_t(i)) \right| \leq \frac{2}{\alpha(\ell - 1)}.$$

Therefore,

$$\begin{aligned}
&\sup_{(\mathbf{x}_1, z_1), \ldots, (\mathbf{x}_\ell, z_\ell), (\hat{\mathbf{x}}_t, \hat{z}_t)} \left| \mathbb{E}_{F|\mathcal{S}} \left[ \frac{1}{\ell} \sum_{i=1}^{\ell} g(\mathbf{x}, z, \mathcal{S}(i)) \right] - \mathbb{E}_{F|\hat{\mathcal{S}}_t} \left[ \frac{1}{\ell} \sum_{i=1}^{\ell} g(\mathbf{x}, z, \hat{\mathcal{S}}_t(i)) \right] \right| \\
= &\sup_{(\mathbf{x}_1, z_1), \ldots, (\mathbf{x}_\ell, z_\ell), (\hat{\mathbf{x}}_t, \hat{z}_t)} \frac{1}{\ell} \left| \sum_{i=1}^{\ell} \mathbb{E}_{F|\mathcal{S}, \hat{\mathcal{S}}_t} \left[ g(\mathbf{x}, z, \mathcal{S}(i)) - g(\mathbf{x}, z, \hat{\mathcal{S}}_t(i)) \right] \right| \\
\leq &\sup_{(\mathbf{x}_1, z_1), \ldots, (\mathbf{x}_\ell, z_\ell), (\hat{\mathbf{x}}_t, \hat{z}_t)} \frac{1}{\ell} \sum_{i=1}^{\ell} \mathbb{E}_{F|\mathcal{S}, \hat{\mathcal{S}}_t} \left| g(\mathbf{x}, z, \mathcal{S}(i)) - g(\mathbf{x}, z, \hat{\mathcal{S}}_t(i)) \right| \leq \frac{2}{\alpha \ell}. \quad \text{(A.7)}
\end{aligned}$$

By (A.7), we apply the McDiamid's inequality to get

$$\Pr(B > \epsilon_1) \leq \exp\left( \frac{-\alpha^2 \ell \epsilon_1^2}{2} \right). \quad \text{(A.8)}$$

Next, we look at $C$. It is clear that

$$\mathbb{E}_F \left[ \frac{1}{\ell} \sum_{i=1}^{\ell} g(\mathbf{x}_i, z_i, \mathcal{S}(i)) \right] = \mathbb{E}_F[g(\mathbf{x}_j, z_j, \mathcal{S}(j))].$$

Let $\bar{g}(\mathcal{S}) = \frac{1}{\ell} \sum_{i=1}^{\ell} g(\mathbf{x}_i, z_i, \mathcal{S}(i))$. For a change of one $(\mathbf{x}_t, z_t)$ to $(\hat{\mathbf{x}}_t, \hat{z}_t)$, denote

$$\hat{\mathcal{S}}_t = \{(\mathbf{x}_1, z_1), \ldots, (\mathbf{x}_{t-1}, z_{t-1}), (\hat{\mathbf{x}}_t, \hat{z}_t), (\mathbf{x}_{t+1}, z_{t+1}), \ldots, (\mathbf{x}_\ell, z_\ell)\}.$$

For any $i \neq t$, it follows from (A.5) that for any $z_i \in \mathbb{K}$

$$\left| g(\mathbf{x}_i, z_i, \mathcal{S}(i)) - g(\mathbf{x}_i, z_i, \hat{\mathcal{S}}_t(i)) \right| \leq \frac{1}{\alpha} \left| \gamma(\mathbf{x}_i, z_i, \mathcal{S}(i)) - \gamma(\mathbf{x}_i, z_i, \hat{\mathcal{S}}_t(i)) \right| \leq \frac{2}{\alpha(\ell - 1)}.$$

Therefore,

$$\begin{aligned}
&\sup_{(\mathbf{x}_1, z_1), \ldots, (\mathbf{x}_\ell, z_\ell), (\hat{\mathbf{x}}_t, \hat{z}_t)} \left| \bar{g}(\mathcal{S}) - \bar{g}(\hat{\mathcal{S}}_t) \right| \\
= &\sup_{(\mathbf{x}_1, z_1), \ldots, (\mathbf{x}_\ell, z_\ell), (\hat{\mathbf{x}}_t, \hat{z}_t)} \frac{1}{\ell} \left| \sum_{i=1}^{\ell} g(\mathbf{x}_i, z_i, \mathcal{S}(i)) - \left[ g(\hat{\mathbf{x}}_t, \hat{z}_t, \hat{\mathcal{S}}_t(t)) + \sum_{i=1, i \neq t}^{\ell} g(\mathbf{x}_i, z_i, \hat{\mathcal{S}}_t(i)) \right] \right| \\
\leq &\frac{1}{\ell} + \sup_{(\mathbf{x}_1, z_1), \ldots, (\mathbf{x}_\ell, z_\ell), (\hat{\mathbf{x}}_t, \hat{z}_t)} \frac{1}{\ell} \left| \sum_{i=1, i \neq t}^{\ell} \left[ g(\mathbf{x}_i, z_i, \mathcal{S}(i)) - g(\mathbf{x}_i, z_i, \hat{\mathcal{S}}_t(i)) \right] \right| \leq \frac{1}{\ell} + \frac{2}{\alpha \ell}. \quad \text{(A.9)}
\end{aligned}$$

By (A.9), we apply the McDiarmid's inequality to get

$$\Pr(C > \epsilon_2) \leq \exp\left(\frac{-2\ell\epsilon_2^2}{\left(1+\frac{2}{\alpha}\right)^2}\right). \tag{A.10}$$

Finally, setting

$$\exp\left(\frac{-\gamma^2\ell\epsilon_1^2}{2}\right) = \exp\left(\frac{-2\ell\epsilon_2^2}{\left(1+\frac{2}{\alpha}\right)^2}\right) = \frac{\delta}{2}$$

and solving for $\epsilon_1$ and $\epsilon_2$, we obtain

$$\epsilon_1 = \frac{1}{\alpha}\sqrt{\frac{2\ln(2/\delta)}{\ell}}, \;\; \epsilon_2 = \left(\frac{1}{2}+\frac{1}{\alpha}\right)\sqrt{\frac{2\ln(2/\delta)}{\ell}}\,.$$

Because $B + C > \epsilon_1 + \epsilon_2$ implies $B > \epsilon_1$ or $C > \epsilon_2$,

$$\Pr(B + C > \epsilon_1 + \epsilon_2) \leq \Pr(B > \epsilon_1 \text{ or } C > \epsilon_2) \leq \Pr(B > \epsilon_1) + \Pr(C > \epsilon_2) \leq \delta.$$

So, with probability at least $1-\delta$, $B + C \leq \epsilon_1 + \epsilon_2$. Because $A \leq \frac{2}{\ell\alpha}$, $B + C \leq \epsilon_1 + \epsilon_2$ implies that $A + B + C \leq \epsilon_1 + \epsilon_2 + \frac{2}{\ell\alpha}$. Therefore, with probability at least $1-\delta$, $A + B + C \leq \epsilon_1 + \epsilon_2 + \frac{2}{\ell\alpha}$, i.e.

$$\mathbb{E}_{F|\mathcal{S}}[g(\mathbf{x}, z, \mathcal{S})] \leq \frac{1}{\ell}\sum_{i=1}^{\ell} g(\mathbf{x}_i, z_i, \mathcal{S}(i)) + \frac{2}{\ell\alpha} + \left(1+\frac{4}{\alpha}\right)\sqrt{\frac{\ln(2/\delta)}{2\ell}}\,.$$

It is easy to verify that $\frac{1}{\ell}\sum_{i=1}^{\ell} g(\mathbf{x}_i, z_i, \mathcal{S}(i)) = \frac{1}{\ell}\sum_{i=1}^{\ell}\frac{\xi_i}{\alpha}$. Therefore, with probability at least $1-\delta$,

$$\mathbb{E}_{F|\mathcal{S}}[g(\mathbf{x}, z, \mathcal{S})] \leq \frac{1}{\ell}\sum_{i=1}^{\ell}\frac{\xi_i}{\alpha} + \frac{2}{\ell\alpha} + \left(1+\frac{4}{\alpha}\right)\sqrt{\frac{\ln(2/\delta)}{2\ell}}\,.$$

This completes the proof. □

## References


[1] M. A. Aizerman, E. M. Braverman, and L. I. Rozonoer, "Theoretical Fundations of the Potential Function Method in Pattern Recognition Learning," *Automation and Remote Control*, vol. 25, no. 6, pp. 917–936, 1964.

[2] M. A. Aizerman, E. M. Braverman, and L. I. Rozonoer, "The Probability Problem of Pattern Recognition Learning and The Method of Potential Functions," *Automation and Remote Control*, vol. 25, no. 9, pp. 1307–1323, 1964.

[3] M. A. Aizerman, E. M. Braverman, and L. I. Rozonoer, "The Method of Potential Functions for the Problem of Restoring the Characteristic of a Function Converter from Randomly Observed Points," *Automation and Remote Control*, vol. 25, no. 12, pp. 1705–1714, 1964.

[4] M. A. Aizerman, E. M. Braverman, and L. I. Rozonoer, "Extrapolative Problems in Automatic Control and the Method of Potential Functions," *American Mathematical Society Translations*, vol. 87, no. 2, pp. 281–303, 1970.

[5] M. Anthony and N. Biggs, *Computational Learning Theory*, Cambridge University Press, 1992.

[6] H. Avi-Itzhak and T. Diep, "Arbitrarily Tight Upper and Lower Bounds on the Bayesian Probability of Error," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 1, pp. 89–91, 1996.

[7] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. M. Blei, M. I. Jordan, "Matching Words and Pictures," *Journal of Machine Learning Research*, vol. 3, pp. 1107–1135, 2003.

[8] A. Barron, "Complexity Regularization with Application to Artificial Neural Networks," *Nonparametric Functional Estimation and Related Topics*, pp. 561–576, Kluwer Academic Publisher, 1991.

[9] P. L. Bartlett, "For Valid Generalization, the Size of the Weights is More Important Than the Size of the Network," *Advances in Neural Information Processing Systems 9*, pp. 134–140, 1997.

[10] O. A. Bashkirov, E. M. Braverman, and I. B. Muchnik, "Potential Function Algorithms for Pattern Recognition Learning Machines," *Automation and Remote Control*, vol. 25, no. 5, pp. 692–695, 1964.

[11] T. Bayes, "An Essay Towards Solving a Problem in the Doctrine of Chances," *The Philosophical Transactions*, vol. 53, pp. 370–418, 1763.

[12] M. Ben-Bassat, K. L. Klove, and M. H. Weil, "Sensitivity Analysis in Bayesian Classification Models: Multiplicative Deviations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 2, no. 3, pp. 261–266, 1980.

[13] J. O. Berger, *Statistical Decision Theory and Bayesian Analysis*, second edition, Springer, 1985.

[14] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[15] M. Boulle, "Compression-Based Averaging of Selective Naive Bayes Classifiers," *Journal of Machine Learning Research*, vol. 8, pp. 1659–1685, 2007.

[16] O. Bousquet and A. Elisseeff, "Stability and Generalization," *Journal of Machine Learning Research*,, vol. 2, pp. 499–526, 2002.

[17] E. M. Braverman, "On the Method of Potential Functions," *Automation and Remote Control*, vol. 26, no. 12, pp. 2205–2213, 1965.

[18] E. M. Braverman and E. S. Pyatnitskii, "Estimation of the Rate of Convergence of Algorithms Based on the Potential Functions Method," *Automation and Remote Control*, vol. 27, no. 1, pp. 95–112, 1966.

[19] P. Bruneau, M. Gelgon, and F. Picarougne, "Parsimonious Reduction of Gaussian Mixture Models with a Variational-Bayes Approach," *Pattern Recognition*, vol. 43, no. 3, pp. 850–858, 2010.

[20] Y. Chen and J. Z. Wang, "Support Vector Learning for Fuzzy Rule-Based Classification Systems," *IEEE Transactions on Fuzzy Systems*, vol. 11, no. 6, pp. 716–728, 2003.

[21] Y. Chen, J. Bi, and J. Z. Wang, "MILES: Multiple-Instance Learning via Embedded Instance Selection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 1931–1947, 2006.

[22] Y. Chen, E. K. Garcia, M. R. Gupta, A. Rahimi, and L. Cazzanti, "Similarity-based Classification: Concepts and Algorithms," *Journal of Machine Learning Research*, vol. 10, pp. 747–776, 2009.

[23] K. B. Davis, "Mean Square Error Properties of Density Estimates," *The Annals of Statistics*, vol. 3, no. 4, pp. 1025–1030, 1975.

[24] K. B. Davis, "Mean Integrated Square Error Properties of Density Estimates," *The Annals of Statistics*, vol. 5, no. 3, pp. 530–535, 1977.

[25] P. A. Devijver, "On a New Class of Bounds on Bayes Risk in Multi-Hypothesis Pattern Recognition," *IEEE Transactions on Computers*, vol. 23, no. 1, pp. 70–80, 1974.

[26] L. Devroye, "On the Asymptotic Probability of Error in Nonparametric Discrimination," *The Annals of Statistics*, vol. 9, no. 6, pp. 1320–1327, 1981.

[27] L. Devroye, "The Equivalence of Weak, Strong and Complete Convergence in $L_1$ for Kernel Density Estimates," *The Annals of Statics*, vol. 11, no. 3, pp. 896–904, 1983.

[28] L. Devroye, "Asymptotic Performance Bounds for the Kernel Estimate," *The Annals of Statistics*, vol. 16, no. 3, pp. 1162–1179, 1988.

[29] L. Devroye, "A Universal Lower Bound for the Kernel Estimate," *Statistics and Probability Letters*, vol. 8, pp. 419-423, 1989.

[30] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, Springer-Verlag New York, 1996.

[31] L. Devroye, L. Györfi, and A. Krzyżak, "The Hilbert Kernel Regression Estimate," *Journal of Multivariate Analysis*, vol. 65, pp. 209–227, 1998.

[32] L. Devroye and A. Krzyżak, "On the Hilbert Kernel Density Estimate," *Statistics and Probability Letters*, vol. 44, pp. 299–308, 1999.

[33] L. Devroye and A. Krzyżak, "New Multivariate Product Density Estimator," *Journal of Multivariate Analysis*, vol. 82, pp. 88–110, 2002.

[34] P. Domingos and M. J. Pazzani, "On the Optimality of the Simple Bayesian Classifier Under Zero-one Loss," *Machine Learning*, vol. 29, no. 2-3, pp. 103–130, 1997.

[35] R. O. Duda, P. E. Hart, D. G. Stork, *Pattern Classification*, Second Edition John Wiley & Sons, Inc., 2001.

[36] A. Garg and D. Roth, "Margin Distribution and Learning Algorithms," *Proc. of Twentieth International Conf. on Machine Learning*, pp. 210–217, 2003.

[37] L. Gordon and R. A. Olshen, "Asymptotically Efficient Solutions to the Classification Problem," *The Annals of Statistics*, vol. 6, no. 3, pp. 515–533, 1978.

[38] D. J. Griffiths, *Introduction to Electrodynamics*, Third Edition, Prentice Hall, 1998.

[39] Y. Guermeur, "VC Theory of Large Margin Multi-Category Classifiers," *Journal of Machine Learning Research*, vol. 8, pp. 2551–2594, 2007.

[40] A. Halevy, P. Norvig, and F. Pereira, "The Unreasonable Effectiveness of Data," *IEEE Intellegent Systems*, vol. 24, no. 2, pp. 8–12, 2009.

[41] W. A. Hashlamoun, P. K. Varshney, and V. N. S. Samarasooriya, "A Tight Upper Bound on the Bayesian Probability of Error," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 2, pp. 220–224, 1994.

[42] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, 2001.

[43] T. Hofmann, J. Puzicha, and M. I. Jordan, "Unsupervised Learning from Dyadic Data," *Advances in Neural Information Processing Systems 11*, pp. 466-472, 1999.

[44] X.-B. Jin, C.-L. Liu, and X. Hou, "Regularized Margin-based Conditional Log-likelihood Loss for Prototype Learning," *Pattern Recognition*, vol. 43, no. 7, pp. 428–2438, 2010.

[45] M. J. Kearns and U. V. Vazirani, *An Introduction to Computational Learning Theory*, The MIT Press, 1994.

[46] H.-C. Kim and Z. Ghahramani, "Bayesian Gaussian Process Classification with the EM-EP Algorithm," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 1948–1959, 2006.

[47] M. Kim, "Large Margin Cost-sensitive Learning of Conditional Random Fields," *Pattern Recognition*, vol. 43, no. 10, pp. 3683–3692, 2010.

[48] J. Langford and J. Shawe-Taylor, "PAC-Bayes and Margins," Advances in Neural Information Processing Systems 15, pp. 439–446, 2002.

[49] P. Langley, W. Iba, and K. Thompson, "An Analysis of Bayesian Classifiers," *Proc. of the Tenth National Conf. on Artificial Intelligence*, pp. 223-228, 1992.

[50] H. Langseth and T. D. Nielsen, "Latent Classification Models for Binary Data," *Pattern Recognition*, vol. 42, no. 11, pp. 2724–2736, 2009.

[51] G. Lugosi and K. Zeger, "Concept Learning Using Complexity Regularization," *IEEE Transactions on Information Theory*, vol. 42, no. 1, pp. 48–54, 1996.

[52] A. Maurer, "Learning Similarity with Operator-valued Large-margin Classifiers," *Journal of Machine Learning Research*, vol. 9, pp. 1049–1082, 2008.

[53] T. M. Mitchell, *Machine Learning*, McGraw-Hill Companies, Inc., 1997.

[54] E. Pekalska, P. Paclik and R. P.W. Duin, "A Generalized Kernel Approach to Dissimilarity-based Classification," *Journal of Machine Learning Research*, vol. 2, pp. 175–211, 2001.

[55] G. Rätsch and M. K. Warmuth, "Efficient Margin Maximizing with Boosting," *Journal of Machine Learning Research*, vol. 6, pp. 2131–2152, 2005.

[56] S. Rosset, J. Zhu, and T. Hastie, "Boosting as a Regularized Path to a Maximum Margin Classifier," *Journal of Machine Learning Research*, vol. 5, pp. 941–973, 2004.

[57] R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee, "Boosting the Margin: A New Explanation for the Effectiveness of Voting Methods," *The Annals of Statistics*, vol. 26, no. 5, pp. 1651–1686, 1998.

[58] B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, The MIT Press, 2002.

[59] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*, Cambridge University Press, 2004.

[60] C. J. Stone, "Consistent Nonparametric Regression," *The Annals of Statistics*, vol. 5, no. 4, pp. 595–620, 1977.

[61] J. Sung, Z. Ghahramani, and S.-Y. Bang, "Latent-Space Variational Bayes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 12, pp. 2236–2242, 2008.

[62] R. Tibshirani and T. Hastie, "Margin Trees for High-dimensional Classification," *Journal of Machine Learning Research*, vol. 8, pp. 637–652, 2007.

[63] V. N. Vapnik and A. Ya. Chervonenkis, "On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities," *Theory of Probabilities and Its Applications*, vol. 16, no. 2, pp. 264–280, 1971.

[64] V. N. Vapnik, *Estimation of Dependencies Based on Empirical Data*, Springer-Verlag, 1982.

[65] V. N. Vapnik, *Statistical Learning Theory*, John Wiley & Sons, Inc., 1998.

[66] S. Veeramachaneni and G. Nagy, "Analytical Results on Style-Constrained Bayesian Classification of Pattern Fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 29, no. 7, pp. 1280–1285, 2007.

[67] J. Wang and X. Shen, "Large Margin Semi-supervised Learning," *Journal of Machine Learning Research*, vol. 8, pp. 1867–1891, 2007.

[68] G. S. Watson and M. R. Leadbetter, "On the Estimation of the Probability Density, I," *The Annals of Mathematical Statistics*, vol. 34, no. 2, pp. 480–491, 1963.

[69] K. Q. Weinberger and L. K. Saul, "Distance Metric Learning for Large Margin Nearest Neighbor Classification," *Journal of Machine Learning Research*, vol. 10, pp. 207–244, 2009.