On ψ-Learning

Xiaotong SHEN, George C. TSENG, Xuegong ZHANG, and Wing Hung WONG

The concept of large margins have been recognized as an important principle in analyzing learning methodologies, including boosting, neural networks, and support vector machines (SVMs). However, this concept alone is not adequate for learning in nonseparable cases. We propose a learning methodology, called ψ-learning, that is derived from a direct consideration of generalization errors. We provide a theory for ψ-learning and show that it essentially attains the optimal rates of convergence in two learning examples. Finally, results from simulation studies and from breast cancer classification confirm the ability of ψ-learning to outperform SVM in generalization.

KEY WORDS: Classification; Generalization error; Margins; Machine learning; Metric entropy; Support vector machine.

1. INTRODUCTION

Recent advances in machine learning have been driven by the recognition of importance of large margins, for instance in analyzing generalization of boosting, neural networks, and support vector machines (SVMs). The concept of margins provides foundation of the development of SVMs. The SVM was first proposed by Boser, Guyon, and Vapnik (1992) and Cortes and Vapnik (1995), and has gained its popularity and attracted tremendous interests due to its theoretical merits and successes in real applications, ranging from hand-written digit recognition to gene classification (see Vapnik 1998, 1999).

The theory of SVM was well developed for separable cases based on an idea of hard margins. However, its foundation becomes much less solid when extended to nonseparable cases. In such situations, generalization errors that become much more important have not been fully taken into account in the SVM formulation. Here we develop a learning methodology called ψ-learning. While retaining the interpretation of large margins for separable cases, ψ-learning delivers improved performance for nonseparable cases by appropriately controlling the training errors.

In this article we investigate the generalization ability of ψ-learning both theoretically and numerically. A theory is developed to quantify its learning accuracy as a function of the size of the training sample and the class of candidate decision functions. This theory is an extension of our earlier results (Shen and Wong 1994; Wong and Shen 1995; Shen 1998) from function estimation to machine learning. The theory not only explains why ψ-learning is expected to deliver high-accuracy performance, it also has an added bonus in that it reveals the trade-off between the learning choice of a tuning parameter and the size of a candidate function class. As suggested by the present theory and simulations, ψ-learning indeed has a theoretical advantage.

Although we demonstrate that ψ-learning has the potential to deliver high performance, the computational aspect of ψ-learning requires special attention because the minimization involved is nonconvex. Further computational development based on recent advances in global optimization will be presented in a subsequent article.

This article is organized as follows. Section 2 introduces the framework of ψ-learning and explains its connection with SVM, and Section 3 presents a learning theory and illustrative examples. Section 4 is devoted to implementation and algorithms. Section 5 examines the performance of ψ-learning via simulation and demonstrates that ψ-learning is more accurate than SVM. This section also presents an application of ψ-learning to the Wisconsin Breast Cancer Data for cancer classification. Section 6 discusses our learning principle. The Appendix is devoted to proofs.

2. ψ-LEARNING

Machine learning comprises four key components: an input space S, an index (output) space O, a decision function f, and a training sample. This section considers a simple scenario, known as binary classification, in which O is dyadic, with 1 and -1 indicating positive and negative classes $\mathcal{A}_{\pm}$.

Typically, machine learning is performed by constructing f, mapping from $\mathcal{S} \subset \mathcal{R}^d$ to $\mathcal{R}^1$ such that its sign, $Sign(f)$, called a "classifier" in the sequel, decides the class assignment of an instance $x \in \mathcal{S}$. To train f, a sample $(X_i, Y_i)_{i=1}^n$ of n input/output pairs is given that is independently and identically distributed according to an unknown joint distribution $P(x, y)$.

To analyze the learning scenario, we examine learning accuracy on inputs outside the training sample. We do this through an error function that measures the generalization ability. Here the error function is the generalization error (GE), defined as $Err(f) = P(Yf(X) < 0) = \frac{1}{2}E(1 - Sign(Yf(X)))$, whose empirical version is $(2n)^{-1}\sum_{i=1}^n (1 - Sign(Y_i f(X_i)))$, called the empirical generalization error (EGE).

2.1 Framework

For motivation, first consider linear classification, where the decision functions $f(x) = w \cdot x + b$ are hyperplanes, defined by the inner product $w \cdot x$ in $\mathcal{R}^d$, with $w \in \mathcal{R}^d$ and $b \in \mathcal{R}^1$. The basic form of SVM originated from the optimal separating hyperplane in separable cases where there exists a hyperplane separating two classes. The SVM maximizes the separation margin $\frac{2}{\|w\|}$, defined by the Euclidean norm $\|w\|$ of w, subject to constraints $y_i f(x_i) \geq 1$, $i = 1, \ldots, n$. These constraints enforce "0" training error. In nonseparable cases, however, these constraints

Xiaotong Shen is Professor, The Ohio State University and University of Minnesota, School of Statistics, University of Minnesota, Minneapolis, MN 55455, USA (E-mail: *xshen@stat.umn.edu*). George Tseng is Assistant Professor, Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA 15261, USA. Xuegong Zhang is Professor, Department of Automation, Tsinghua University, Beijing 100084, China (E-mail: *zhangxg@tsinghua.edu.cn*). Wing Hung Wong is Professor, Department of Statistics and Biostatistics, Harvard School of Public Health, Boston, MA 02115, USA (E-mail: *wwong@stat.harvard.edu*). The authors would like to thank the editors, the associate editor, and referees for helpful comments. The authors are grateful to Grace Wahba for pointing out a relation between $(1 - Sign)/2$ and the hinge loss. This work is supported in part by National Science Foundation grant, DMS-0072635, IIS-0328802, DMS-0090166, and National Science Foundation of China grant 69885004.

Journal of the American Statistical Association
September 2003, Vol. 98, No. 463, Theory and Methods
DOI 10.1198/016214503000000639

are not attainable; thus slack variables $\{\zeta_i\}_{i=1}^n$ are introduced to define the so-called "l_1 soft margin" $\frac{1}{2}\|w\|^2 + C\sum_{i=1}^n \zeta_i$, subject to constraints $\zeta_i \geq 1 - y_i(w \cdot x_i + b)$ and $\zeta_i \geq 0$, $i = 1, \ldots, n$. The Kuhn–Tucker constraint optimization theory says that the solution of SVM satisfies the active constraints in that either $\zeta_i = (1 - y_i f(x_i)) \geq 0$ or $\zeta_i = 0$ if $0 > 1 - y_i f(x_i)$, $i = 1, \ldots, n$, yielding an equivalent unconstrained version of SVM,

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^n \psi_{\text{svm}}(y_i f(x_i)), \tag{1}$$

where $\psi_{\text{svm}}(x) = 0$ if $x \geq 1$ and $\psi_{\text{svm}}(x) = 1 - x$, otherwise.

In our learning framework, SVM included, the primary goal is to seek a classifier, $Sign(f)$, to maximize generalization accuracy. Often a cost function is minimized to obtain the optimal f and thus $Sign(f)$, which involves the training sample $\{X_i, Y_i\}_{i=1}^n$ and $Sign(f)$. In principle, one might choose the cost function to be the EGE and minimize it with respect to (w, b). However, EGE suffers from the difficulty of multiple minimizers and the danger of overfitting. In the area of nonparametric function estimation, such a difficulty is handled through penalization. Thus here we use penalization (regularization) with penalty $\frac{1}{2}\|w\|^2$. In separable cases, the penalty $\frac{2}{\|w\|^2}$ has been shown by Boser et al. (1992) to represent the maximal margin or the separation margin between $A_\pm$. Furthermore, it can be used to discriminate among minimizers when multiple minimizers occur. In nonseparable cases, the use of a penalty has been further emphasized by Wahba (1998) in a context of nonlinear SVM. The foregoing discussion yields the following penalization cost function:

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^n \big(1 - Sign(y_i f(x_i))\big), \tag{2}$$

where the size of C ($C > 0$) reflects the relative importance of the EGE and the separation margin.

Evidently, the discrete nature of $Sign$ makes it more difficult to optimize (2). It is interesting to note that the SVM cost function ψ_{svm} in (1) is a convex upper envelope of $\frac{1}{2}(1 - Sign)$. (We are grateful to Grace Wahba for pointing out this relation.) In recent years, dramatic increases in computing power and the development of global optimization techniques have made it possible to tackle the problem of (2) directly. This allows us to substantially improve the generalization ability in terms of the GE, because there is a difference between a convex upper envelope and $(1 - Sign)/2$ itself, especially for nonseparable cases. We illustrate this aspect with a linear example in Sections 3 and 5.

Surprisingly, (2) is an undesirable cost function, because the minimizer of (2) becomes the zero function when no constraint on the size of (w, b) is imposed. This is because any positive scaling transformation of f leaves its sign unchanged; that is, $Sign$ forces $\|w\|$ of the solution to be 0. To eliminate this scaling problem, we introduce a penalty function, ψ, to drive correctly specified instances away from the decision boundary. This ψ is required to satisfy the property

$$\begin{aligned} &U \geq \psi(x) > 0, && \text{if } x \in (0, \tau] \\ &\psi(x) = (1 - Sign(x)) && \text{otherwise,} \end{aligned} \tag{3}$$

where $0 < \tau \leq 1$ and $U > 0$ are some constants. Furthermore, it seems sensible that any correctly specified instance is penalized less than any wrongly specified instance, which implies $U = 2$ in (3). This eliminates the scaling problem, which circumvents the difficulty of $Sign$. For any instance x_i such that $y_i f(x_i) \geq 0$, ψ pushes it toward $y_i f(x_i) \geq \tau$, because ψ assigns a positive penalty to any value in the range of $(0, \tau]$. This modification yields our cost function of ψ-learning,

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^n \psi(y_i f(x_i)), \tag{4}$$

where $C > 0$ is a tuning parameter that should depend on n and be chosen from data in practice. Based on our limited experience, the choice of C seems important for both separable and nonseparable problems.

The basic idea for choosing ψ is that it should be as close to $1 - Sign$ as possible, while eliminating the scaling problem. In this article we use a simple linear function ψ_0 in implementation, where $\psi_0(x)$ is defined to be 0 if $x \geq 1$, $1 - x$ if $0 \leq x \leq 1$, and 2 otherwise. The graph of ψ_0 is displayed in Figure 1. Other choices of ψ are possible. For instance, we use a different ψ function, ψ_1, based on a computational consideration, defined as 0 if $z \geq 1$, $2(1 - z)$ if $0 \leq z \leq 1$, and 2 otherwise. This continuous ψ_1 permits difference convex programming for globally solving (4) as opposed to ψ_0; see Section 6 for a further discussion.

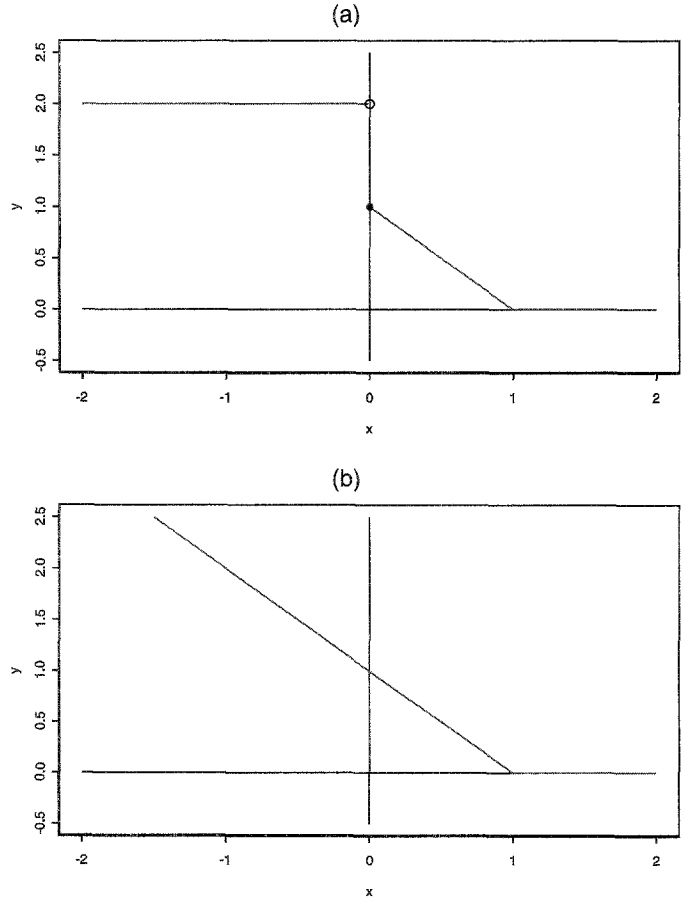

Figure 1. Plots of (a) ψ_0 and (b) ψ_{svm}.

For nonlinear problems, the decision function $f(x)$ is represented as $g(x) + b$ with $g(x) = \sum_{i=1}^{n} \alpha_i K(x, x_i)$, defined by a proper kernel $K(\cdot, \cdot)$ that maps from $S \times S$ to $\mathcal{R}$. This is equivalent to applying a certain nonlinear transform on S according to the theory of reproducing kernel Hilbert space (RKHS) (see Wahba 1990). Here $K(\cdot, \cdot)$ may be smooth or discontinuous and is required to satisfy Mercer's condition, which ensures that $\|g\|_K^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j K(x_i, x_j)$ is a proper norm. With this generalization, the concept of separability can be defined similarly by replacing the hyperplane representation by the kernel representation.

The kernel-based cost function of ψ-learning now becomes

$$\frac{1}{2}\|g\|_K^2 + C \sum_{i=1}^{n} \psi(y_i f(x_i)). \tag{5}$$

As an analogy, the penalty, $\|g\|_K^2$, of this form is often used to enforce smoothness of the regression function in nonparametric regression (see Wahba 1990). However, here $\|g\|_K^2$ may be induced by a nonsmooth kernel. For instance, the neural network step kernel (Mangasarian 2000) allows discontinuous jumps.

The classifier of ψ-learning then is $Sign(\hat{f})$, where $\hat{f}(x) = \hat{w} \cdot x + \hat{b}$ in (4) or $\hat{f}(x) = \sum_{i=1}^{n} \hat{\alpha}_i K(x, x_i) + \hat{b}$ in (5), which are defined by the minimizers $(\hat{w}, \hat{b}) = (\hat{w}_1, \ldots, \hat{w}_d, \hat{b})$ in (4) and $(\hat{\alpha}, \hat{b}) = (\hat{\alpha}_1, \ldots, \hat{\alpha}_n, \hat{b})$ in (5).

2.2 Properties of ψ

We now study properties of ψ to gain insight into the performance of ψ-learning.

Proposition 1. Let $\bar{f} = Sign(f^*)$, where $f^* = P(Y = 1|x) - 1/2$ is the Bayes decision function. Then, for any ψ satisfying (3), $\bar{f}$ minimizes $E\psi(Yf(X))$ and $E(1 - Sign(Yf(X)))$; that is, $E\psi(Yf(X)) \geq E\psi(Y\bar{f}(X)) = E(1 - Sign(Y\bar{f}(X))) \leq E(1 - Sign(Yf(X)))$, for any f. Furthermore, the minimizers for $E\psi(Yf(X))$ and $E(1 - Sign(Yf(X)))$ are not unique, for example, $c\bar{f}$ is also a minimizer for both quantities for any $c \geq 1$.

Proposition 1 says that the method of ψ-learning estimates the Bayes classifier $\bar{f} = Sign(f^*)$ rather than the Bayes decision function f^*. The Bayes classifier is the ideal optimal classifier, obtained by minimizing $E(1 - Sign(Yf(X)))$ for all f. This feature of ψ is essential, because the optimal performance of $\bar{f} = Sign(f^*)$ is realized by using the ψ function although it differs from $1 - Sign$.

Proposition 2. In separable cases, (4) and (1) yield the same solution when $C \to \infty$.

Proposition 2 says that ψ-learning with large C values is equivalent to the hard-margin SVM. When C is large, ψ enforces "0" training error in separable cases, which acts in a parallel fashion to the cost function of hard-margin SVM. Indeed, the two yield identical solutions for large C, as confirmed by simulation in Section 5.

In summary, ψ-learning and SVM are based on different principles. It is interesting to note that $\psi_{\text{svm}}(x) = \psi_0(x)$ when no training error is committed, such as in separable cases. In nonseparable cases, the advantage of ψ over ψ_{svm} is that it improves learning accuracy, and the disadvantage perhaps is that the optimization becomes nonconvex. Here we show that the potential gain in error rates offered by use of nonconvex cost functions is very substantial and suggests that research in optimization of cost functions of the form (4) should be a high priority. Comments on possible approaches to this optimization are provided in the beginning of Section 4.

3. LEARNING THEORY

To begin, we first consider the ideal optimal decision function (the Bayes decision function) f^* that minimizes $Err(f)$ in the sense that

$$\inf_{f \neq 0} Err(f) = Err(f^*) = E(1/2 - |f^*(X)|).$$

It is easy to show that $f^*(x) = P(Y = 1|X = x) - 1/2$. Here we measure the learning accuracy of f by

$$\begin{aligned} e(f, f^*) &= Err(f) - Err(f^*) \\ &= E|f^*(X)|\,|\,Sign(f(X)) - Sign(f^*(X))| \geq 0, \end{aligned} \tag{6}$$

the difference between the actual and ideal performances. As shown in Lemma 1, $e(f, f^*)$ reduces to $Err(f)$ in separable cases.

Next, we develop a learning theory to quantify the magnitude of $e(f, f^*)$ as a function of n, in terms of the value of C and the size of $G(\mathcal{F}) = \{G_f = \{x : f(x) \geq 0\} : f \in \mathcal{F}\}$. Here $G(\mathcal{F})$ is the class of candidate classification sets, induced by the class $\mathcal{F}$ comprising candidate decision functions. Similar to the theory of sieves (Shen and Wong 1994), $\mathcal{F}$ can depend on the sample size n, but this dependency is suppressed in our notation.

3.1 Theory

In this section we give an upper bound of the GE of ψ-learning in terms of the complexity and reveal the best tradeoff related to the choice of tuning parameter C. Our learning theory is formulated on the basis of the size of $G(\mathcal{F})$, measured by the metric entropy to be defined. For our theory, the ideal optimal classification set $G_{f^*} = \{x \in S : f^*(x) \geq 0\}$ is not required to belong to $G(\mathcal{F})$. Instead, it is assumed that $\bar{f} = Sign(f^*)$ can be well approximated by $\mathcal{F}$. To quantify this approximation, we consider $e_\psi(f, \bar{f}) = \frac{1}{2}(E\psi(Yf(X)) - E\psi(Y\bar{f}(X)))$, which measures the approximation error.

Let $J_0 = \max(J(f_0), 1)$. We make the following four assumptions.

Assumption A. For some positive sequence $s_n \to 0$ as $n \to \infty$, there exists $f_0 \in \mathcal{F}$ such that $e_\psi(f_0, \bar{f}) \leq s_n$. Equivalently, $\inf_{\{f \in \mathcal{F}\}} e_\psi(f, \bar{f}) \leq s_n$.

By Proposition 1, $e(f_0, f^*) \leq e_\psi(f_0, \bar{f}) \leq s_n$. If $\mathcal{F}$ and $\bar{f}$ are independent of n, then Assumption A means that $\inf_{\{f \in \mathcal{F}\}} e_\psi(f, \bar{f}) = 0$.

Assumption B. There exist some constants $0 < \alpha \leq +\infty$ and $c_1 > 0$ such that $P(x \in S : |f^*(x)| \leq \delta) \leq c_1 \delta^\alpha$ for any sufficiently small $\delta \geq 0$.

Assumption B is a Hölder type of condition that describes the behavior of f^* near the decision boundary $\{x : f^*(x) = 0\}$.

To specify Assumption C, we need to define the metric entropy for sets. For a given class $\mathcal{B}$ of subsets of S and any $\varepsilon > 0$, call $\{(G_1^l, G_1^u), \ldots, (G_m^l, G_m^u)\}$ an ε-bracketing set of

$\mathcal{B}$ if for any $G \in \mathcal{B}$ there is a j such that $G_j^l \subset G \subset G_j^u$ and $\max_{1\le j\le m} d(G_j^u, G_j^l) \le \varepsilon$, where $d(\cdot,\cdot)$ is a distance for any two sets $G_i \in \mathcal{S}$, defined as $d(G_1, G_2) = \int_{\{G_1 \Delta G_2\}} dP = P(G_1 \Delta G_2)$, and $G_1 \Delta G_2 = (G_1 \setminus G_2) \cup (G_2 \setminus G_1)$ is the set difference between G_i. Then the metric entropy $H(\varepsilon, \mathcal{B})$ of $\mathcal{B}$ with bracketing is defined as a logarithm of the cardinality of an ε-bracketing set of $\mathcal{B}$ of the smallest size.

Let

$$\mathcal{G}(k) = \{G_f = \{x : f(x) \ge 0\} : f \in \mathcal{F}, J(f) \le k\} \subset G(\mathcal{F}) = \{G_f = \{x : f(x) \ge 0\} : f \in \mathcal{F}, J(f) < +\infty\},$$

where $J(f)$ is $\frac{1}{2}\|w\|^2$ in (4) and is $\frac{1}{2}\|g\|_K^2$ in (5) for example.

Assumption C. For some positive constants c_i; $i = 2, \ldots, 4$, there exists some $\varepsilon_n > 0$ such that

$$\sup_{\{k \ge 1\}} \phi(\varepsilon_n, k) \le c_2 n^{1/2}, \tag{7}$$

where $\phi(\varepsilon_n, k) = \int_{c_4 L}^{c_3^{1/2} L^{\alpha/2(\alpha+1)}} H^{1/2}(u^2/2, \mathcal{G}(k))du/L$ and $L = L(\varepsilon_n, C, k) = \min(\varepsilon_n^2 + (Cn)^{-1} J_0(k/2 - 1), 1)$. For instance, $c_2 = 2^{-23/2}$, $c_3 = 2^{1/\alpha} U \max(4(2^{\alpha+2}c_1)^{1/(\alpha+2)} + 2), 8\max(U, 2))$, and $c_4 = 2^{-6}$.

Assumption D. For $x \in (0, \tau]$ with fixed constants $0 < \tau \le 1$ and U, $U \ge \psi(x) \ge (1 - Sign(x))$, and $\psi(x) = (1 - Sign(x))$ otherwise.

Theorem 1. Suppose that Assumptions A–D are met. Then, for any classifier of ψ-learning $Sign(\hat{f})$, there exists a constant $c_5 > 0$ such that

$$P\big(e(\hat{f}, \bar{f}) \ge \delta_n^2\big) \le 3.5 \exp\Big(-c_5 n (nC)^{-\frac{\alpha+2}{\alpha+1}} J_0^{\frac{\alpha+2}{\alpha+1}}\Big),$$

provided that $Cn \ge 2\delta_n^{-2} J_0$, where $\delta_n^2 = \min(\max(\varepsilon_n^2, 2s_n), 1)$.

Corollary 1. Under the assumptions of Theorem 1,

$$|e(\hat{f}, f^*)| = O_p(\delta_n^2), \qquad E|e(\hat{f}, f^*)| = O(\delta_n^2),$$

provided that $n^{-\frac{1}{\alpha+1}} (C^{-1} J_0)^{\frac{\alpha+2}{\alpha+1}}$ is bounded away from 0.

In any application, we need to verify that the assumptions are satisfied for $s_n \to 0$ and $\varepsilon_n \to 0$, then choose the optimal trade-off for δ_n. The optimal C that yields the best rate δ_n^2 for ψ-learning is determined by two inequalities: (1) $C/J_0 \ge 2n^{-1}\delta_n^{-2}$ and (2) C/J_0 is $O(n^{-\frac{1}{\alpha+2}})$. Usually, a good choice of C is of order of $n^{-1}\delta_n^{-2} J_0$. In this case, if $\alpha = \infty$, then the bound in Theorem 1 becomes $3.5\exp(-c_5 n \delta_n^2)$; if $\alpha \to 0$, then the bound there reduces to $3.5\exp(-c_5 n \delta_n^4)$, where δ_n is as defined in Theorem 1.

Remark 1. Although the present theory says that the foregoing result holds for any ψ satisfying Assumption D, it is important to note that the approximation error $e_\psi(f_0, \bar{f})$ can differ substantially depending on the choice of ψ. For instance, if $\psi(x) = 1/(1 - x)$ for any $0 < x \le 1$ and $\psi(x) = 1 - Sign(x)$ otherwise, then $e_\psi(f_0, \bar{f})$ could be much larger than $e_{\psi_0}(f_0, \bar{f})$. Additionally, the present theory allows f_0, f^*, and $\mathcal{F}$ to depend on n.

Remark 2. Let f_i; $i = 1, 2$, be the conditional probability densities of X given $Y = \pm 1$. By Bayes's rule,

$$P(Y = 1 | X = x) = \frac{\pi_1 f_1(x)}{\pi_1 f_1(x) + \pi_2 f_2(x)}, \tag{8}$$

where π_i; $i = 1, 2$, are the prior probabilities on $\mathcal{A}_\pm$. Assumption B is closely related to (4) of Mammen and Tsybakov (1999) but is different, in that this assumption is imposed only on $f^*(x)$ rather than on any $f \in \mathcal{F}$. A similar assumption was used by Lin (2000) to obtain error bounds for the difference between f^* and the SVM decision function. For Assumption B, the most interesting case is $\alpha = +\infty$. The case of $\alpha = 0$ is excluded. In any case, this is not of interest, because $\alpha = 0$ indicates that two classes are indistinguishable in an open neighborhood of the decision boundary.

Remark 3. The integral equation (7) of this form has been used to quantify rates of convergence of the maximum likelihood type of estimators (see, e.g., Van De Geer 1993; Birgé and Massart 1993; Wong and Shen 1995).

Remark 4. The result in Theorem 1 continues to hold if the "global" entropy is replaced by a corresponding "local" version (see, e.g., Van De Geer 1993). That is, $\mathcal{G}(k)$ is replaced by $\mathcal{G}_1(k) = \mathcal{G}(k) \cap \{e(f, f_0) \le 2u^2\}$. The proof requires only a trivial modification. The local entropy allows us to avoid the loss of a factor of $\log(n)$ for a linear problem, although it may not be useful for a nonlinear problem.

Theorem 1 and Corollary 1 provide probability and risk bounds for $e(\hat{f}, f^*)$, with the smallest ε_n satisfying (7) giving the best upper bound of the GE of ψ-learning. Indeed, there is a trade-off between the value of C and the performance; the best performance is realized when C gives the best balance between the size of $G(\mathcal{F})$ and n. Thus Theorem 1 provides guidance in the choice of C. As discussed in the beginning of Section 3, this aspect has not been revealed by the aforementioned first approach.

Finally, we compare Theorem 1 with learning theories in the literature. There are two main approaches to statistical learning theory for classification. The first approach is to bound the GE of a classifier in terms of the empirical training error and the Vapnik–Chervonenkis (VC) complexity of the class of candidate decision functions. In particular, this gives the following type of upper bound:

$$Err(\hat{f}) \le \inf_{\delta > 0} \left(n^{-1} \sum_{i=1}^{n} I(Y_i \hat{f}(X_i) \le \delta) + C(\mathcal{F}, \delta)/n^{1/2} \right),$$

where $C(\mathcal{F}, \delta)$ depends on some kind of entropy related to $\{(x, y) : yf(x) \le \delta\}$ for $f \in \mathcal{F}$. An upper bound of $Err(\hat{f})$ can then be obtained for a specific training sample by appropriately choosing δ depending on n (see, e.g., Vapnik 1998; Devroye, Györfi, and Lugosi 1996; Koltchinskii and Panchenko 2002). The second approach expresses an upper bound of the GE of a classifier in terms of the complexity of the class of candidate decision functions and the trade-off between the complexity and the training error (see Mammen and Tsybakov 1999; Lin 2000, 2002). The difference between these approaches have been discussed by Mammen and Tsybakov (1999, p. 1811). The two approaches are complementary with one another. The first approach is useful when we want a bound for the classification

error rate based on a particular observed dataset. But because this bound is random, it cannot be used to compare different classifiers a priori. For this latter purpose, we have to rely on the second approach. Although there are some results of the first approach in the machine learning literature (e.g., Koltchinskii and Panchenko 2002), to our knowledge the only published result of the second approach was given by Mammen and Tsybakov (1999). These authors' result is applicable only to classifiers that minimize the empirical training error over a compact class of sets (and therefore are not applicable to ψ-learning).

3.2 Illustrative Examples

In this section we apply the general theory to two specific learning examples.

3.2.1 Linear Classification. Linear classification uses a class of hyperplanes, $\mathcal{F} = \{x \in S : f(x) = w \cdot x + b : w \in \mathcal{R}^2\}$, where $S = \{(x_1, x_2) : x_1^2 + x_2^2 \le 1\}$ in $\mathcal{R}^2$ is a unit disk and the true decision function $f_t(x)$ is x_1 that yield the vertical line as the decision boundary. See Section 5 for more details about the setting.

With a choice of $f_0 = nf_t \in \mathcal{F}$, $e_\psi(f_0, \bar{f}) \le E(\psi(Yf_0(X)) - (1 - Y\,Sign(f^*(X))) \le s_n = c_1 n^{-1}$ for some constant $c_1 > 0$. Then Assumption A is met. It can be verified via (8) that $P(x \in S : |f^*(x)| \le \delta) = 0$ for any sufficiently small $\delta > 0$. Assumption B is satisfied with $\alpha = +\infty$. To check Assumption C, we compute the local entropy of $\mathcal{G}_1(k)$, defined in Remark 4. Note that $e(f, f_0) \le e_\psi(f, f_0) \le 2u^2$ implies that $\|w - w_0\| \le c'u^2$ for some $c' > 0$, where $f_0(x) = w_0 \cdot x$. Furthermore, $\min_{1\le i\le n} |(w - w_0) \cdot x_i| \le |b - b_0| \le \max_{1\le i\le n} |(w - w_0) \cdot x_i|$, because b minimizes $\sum_{i=1}^n \psi(y_i f(x_i))$ for any given w. Hence for any $f \in \mathcal{G}_1(k)$, $|b - b_0| \le \|w - w_0\|$ and $\|w\| \le (2k)^{1/2}$. Direct calculations yield that $H(u^2, \mathcal{G}_1(k)) \le O(\log(\min(k_1^{1/2}, c'u^2)/u^2))$ with $k_1 = (2k + \|w_0\|^2)^{1/2}$. Let $\phi_1(\varepsilon_n, k)$ be $(\log(\min(k_1^{1/2}, c'\varepsilon_n^2)/\varepsilon_n^2))^{1/2}/L^{1/2}$, where $L = \min(\varepsilon_n^2 + (Cn)^{-1}(k/2 - 1), 1)$. Easily, $\sup_{k\ge 1} \phi(\varepsilon_n, k) \le \phi_1(\varepsilon_n, 1) = c/\varepsilon_n$ for a constant $c > 0$. Solving (7) yields a rate $\varepsilon_n = n^{-1/2}$ when $C/\max(J(f_0), 1)$ is a sufficiently large constant. From Theorem 1, it follows that $e(\hat{f}, f^*) \le O(n^{-1}\log(1/\delta))$, except for a set of probability less than small $\delta > 0$. From Corollary 1, $Ee(\hat{f}, f^*) = O(n^{-1})$. This result holds generally for any ψ satisfying Assumption D, including the ψ_0 used in the simulations.

The foregoing rate expected to be optimal, in view of theorem 2.1 of Blumer, Ehrenfucht, Haussler, and Warmuth (1989). As indicated in theorems 4.5 and 4.6 of Bartlett and Shawe-Taylor (1999), the error rate of linear SVM is $n^{-1/2}$ in nonseparable cases and n^{-1} in separable cases. As suggested by Figure 2, the rate of $n^{-1/2}$ for linear SVM in the nonseparable cases cannot be improved further when $Err(f^*)$ is bounded away from 0. (Note that it is asymptotically separable if $Err(f^*) \to 0$ as $n \to \infty$.) In comparison, ψ-learning achieves a faster rate n^{-1} of convergence in the nonseparable cases, which agrees with our intuition discussed in Section 2.

3.2.2 Nonlinear Classification With Smoothness. Consider a function $\tilde{f}(x)$ with the bounded continuous pth derivative,

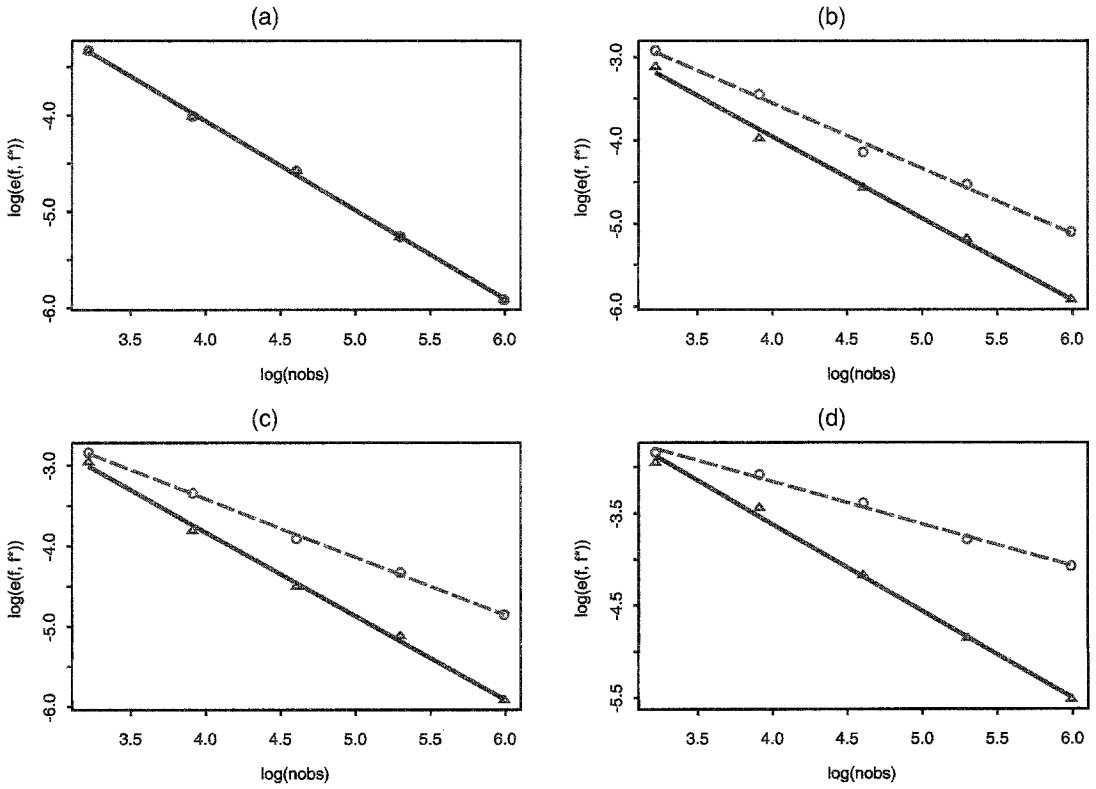

Figure 2. Plots of $\log(e(\hat{f}, f^*))$ as a Function of $\log n$ for ψ-Learning and SVM, Represented by the Vertical and Horizontal Lines. Here $C = 10^7$. (a) 0-flip (-- SVM: $Y = (-0.355) + (-0.926)X$; — ψL: $Y = (-0.355) + (-0.926)X$); (b) 1-flip (-- SVM: $Y = (-0.421) + (-0.783)X$; — ψL: $Y = (-0.0256) + (-0.983)X$); (c) 2-flip (-- SVM: $Y = (-0.531) + (-0.722)X$; — ψL: $Y = (0.345) + (-1.04)X$); (d) 10%-flip (-- SVM: $Y = (-1.34) + (-0.455)X$; — ψL: $Y = (0.159) + (-0.944)X$).

where the number of 0-crossing points of $\tilde{f}(x)$ in $[0, 1]$ is no more than $\tilde{p}$. A 0-crossing point x_0 of $\tilde{f}$ is defined as a point such that $\tilde{f}^{(i)}(x_0) = 0$ for $i = 0, \ldots, m$ with $m < p$, and $\tilde{f}^{(m)}(x_0^-)\tilde{f}^{(m)}(x_0^+) > 0$, where $\tilde{f}^{(i)}$ is the ith derivative of $\tilde{f}$ and $\tilde{f}^{(m)}(x_0^{\pm})$ are the corresponding left and right derivatives.

We assume that the joint distribution $P(\cdot, \cdot)$ of (X, Y) is induced by such a function $\tilde{f}(x)$. Specifically, the label variable Y is 1 if $\tilde{f}(x) \geq 0$ and is -1 otherwise. Let v be the total length of $G = \{x \in S : \tilde{f}(x) \geq 0\}$, and assume that $v \neq 1/3$.

For this example, $Err(f^*)$ may be easily computed. The conditional density of X given $Y = 1$ is $2/(1 + v)$ if $\tilde{f}(x) \geq 0$ and $1/(1 + v)$ if $\tilde{f}(x) < 0$. The conditional density of X given $Y = -1$ is $1/(2 - v)$ if $\tilde{f}(x) \geq 0$ and $2/(2 - v)$ if $\tilde{f}(x) < 0$. In this case, $P(Y = 1) = v$ and $Err(f^*) = \frac{|3v-1|}{2(1+v)} > 0$.

Nonlinear classification with smooth boundaries uses a kernel, $K(u, t)$. For a boundary with the degree p of smoothness on $S = [0, 1]$, K is the spline kernel of order p, defined by the Bernoulli polynomials in $\mathcal{R}^1$ (see Wahba 1990 for an expression). Here $p \geq 1$ is an integer. The spline kernel yields candidate decision functions with p derivatives belonging to $\mathcal{F}_n = \{f = g + b : g(x) = \sum_{i=1}^n w_i K(x, x_i), J(f) < \infty, f$ *has at most* $\tilde{p}$ *real roots*$\}$ with $J(f) = \int_0^1 [\frac{d^p g(x)}{dx^p}]^2 dx$ and $b \in \mathcal{R}^1$.

For Assumption A, it is easy to see that there exists a spline $f_t \in \mathcal{F}$ that interpolates $\tilde{f}$ and the corresponding derivatives at its 0-crossing points. This implies that $Sign(f_t(x)) = Sign(\tilde{f}(x)) = Sign(f^*(x))$. With a choice of $f_0 = \varepsilon_n^{-2} f_t$, $e_\psi(f_0, \tilde{f}) \leq s_n = c_1 \varepsilon_n^2$, where ε_n is defined below and $c_1 > 0$ is a constant. This implies Assumption A.

By assumptions and (8), $P(x \in S : |f^*(x)| \leq \delta) = 0$ for all sufficiently small $\delta > 0$, which implies Assumption B with $\alpha = +\infty$.

To verify Assumption C, we note that $H(u, \mathcal{G}(k)) \leq O(\log(1/u))$ for any k by Lemma 2. Let $\phi_1(\varepsilon, k) = c_3(\log(1/L^{1/2}))^{1/2}/L^{1/2}$, where $L = \min(\varepsilon^2 + (Cn)^{-1}(k/2 - 1), 1)$, which in turn yields $\sup_{k \geq 1} \phi(\varepsilon_n, k) \leq \phi_1(\varepsilon_n, 1) = c(\log(1/\varepsilon_n))^{1/2}/\varepsilon_n$ for some $c > 0$ and a rate $\varepsilon_n = (n^{-1} \log n)^{1/2}$ when $C/\max(J(f_0), 1) \sim \delta_n^{-2} n^{-1} \sim 1/\log n$. By Corollary 1, we conclude that $e(\hat{f}, f^*) \leq O(n^{-1} \log n)$ except for a set of probability tending to 0, and $Ee(\hat{f}, f^*) = O(n^{-1} \log n)$. This result holds generally for any ψ satisfying Assumption D.

This might be a somewhat surprising result, because the fast rate $n^{-1} \log n$ is nearly optimal, which is attainable by many classifiers only for linearly separable classification. Nevertheless, this rate has been attained by ψ-learning for nonlinear separable and nonseparable classification. As suggested by theorems 4.5 and 4.6 of Bartlett and Shawe-Taylor (1999), a rate faster than $n^{-1/2}$ is not generally expected for SVMs in nonseparable cases. A recent result of Lin (2000) also suggests this, although the lower bound of the error rate for SVMs is not yet available. As discussed previously, the GE of an SVM is determined by the estimation precision of f, which is a nonparametric rate for nonlinear learning.

4. IMPLEMENTATION

As discussed in Section 3, use of ψ may gain improved generalization ability, although the optimization involved in (4) or (5) becomes nonconvex. Fortunately, recent developments in deterministic global optimization techniques via concave programming have made it possible to tackle the optimization problem. The basic idea is to decompose (4) into a sum of a convex and concave functions. Then a sequence of outer approximations is constructed, for which a sequence of solutions, obtained via vertex enumeration of certain polytopes, converges to the global minimizer. Currently, we are developing global optimization routines based on the methods of An and Tao (1997) and Blanquero and Carrizosa (2000), and will report the results elsewhere.

In this article we pursuit a different strategy. The optimization in (4) and (5) involves unconstrained optimization, where the cost functions are piecewise quadratic in w and in α. For optimization of this type, a direct-search complex algorithm [available in (IMSL)] together with a good initial guess is applicable (see Gill, Murray, and Wright 1981 for the algorithm). The initial guess may be chosen as an estimate obtained from either an SVM or a stochastic search, such as a genetic algorithm. To prevent the algorithm from being trapped with a local optimizer, we recommend using multiple starting values.

Our limited experience suggests that this routine performs well for a low-dimensional problem, although it can not guarantee to obtain the global minimizer. In fact, no conventional deterministic optimization routines, using an initial guess, can guarantee to converge to the global optimizer, for a nonconvex objective function. Note that our theory extends to an ε_n-global minimizer of (4) whose objective function value is no greater than that of the global minimizer plus ε_n. This suggests that it is unnecessary to obtain the exact global minimizer as long as a reasonably good local minimizer can be found.

To apply a direct-search complex algorithm, we need to deal with b, because the choice of b that gives the minimal of (4) or (5) may be nonunique. For instance, in separable cases, any $b \in (b_{\min}, b_{\max})$ is a minimizer of (4) if $b_{\min}$ is the minimum of $w \cdot x$ for $\mathcal{A}_+$ and $b_{\max}$ is the maximum of $w \cdot x$ for $\mathcal{A}_-$, although $b = (b_{\min} + b_{\max})/2$ gives the optimal hyperplane when there is no a priori probability regarding $\mathcal{A}_\pm$. This difficulty yields estimation bias as well as a problem of convergence in optimization. To overcome the difficulty, we first identify the set of all minimizers b of the EGE given w or α, then compute the center of these minimizers. We give an efficient algorithm for performing this task.

To ensure fast function evaluation, we have developed an algorithm to express b as a function of w or α, as discussed in Section 2.1. For simplicity, we present only the algorithm for (4), which proceeds in three steps.

Algorithm:

Step 1: Initialization. Sort pairs $\{(t_i = wx_i, y_i)\}_{i=1}^n$ by t_i in an ascending order and let $\{(\tilde{t}_i, \tilde{y}_i)\}_{i=1}^n$ denote the sorted pairs.

Step 2: Counting and updating. Find b to minimize $g(b) = \sum_{i=1}^n \psi(y_i f(x_i))$ via dynamic programming, which is performed in three steps:

a. Compute A_i using a recursive formula: $A_i = A_{i-1} + \tilde{y}_i$; $i = 2, \ldots, n$, $A_1 = \#\{\tilde{y}_i = -1, i = 1, \ldots, n\}$, $\tilde{t}_0 = -\infty$, $\tilde{t}_{n+1} = \infty$.

b. Find the minimizer(s) of $\{A_1, \ldots, A_n\}$ and let $B = \{B_1 < B_2 < \cdots < B_m\}$ denote the collection of the distinct minimizer(s).

c. Compute

$$b^* = \begin{cases} -(\tilde{t}_{B_{m/2}} + \tilde{t}_{B_{m/2+1}})/2 & \text{if } m \text{ is even} \\ -(\tilde{t}_{B_{(m+1)/2}-1} + \tilde{t}_{B_{(m+1)/2}})/2 & \text{if } m \text{ is odd} \\ -(\tilde{t}_n + \epsilon) & \text{if } m = 1, B_1 = n+1 \\ -(\tilde{t}_1 - \epsilon) & \text{if } m = 1, B_1 = 1, \end{cases}$$

where ϵ is a small number, say the machine precision.

Step 3: Optimal b. Set $b = b^*$, if $\tilde{y}_i(\tilde{t}_i + b^*) \geq 1$; $i = 1, \ldots, n$, among all correctly specified instances such that $\tilde{y}_i(\tilde{t}_i + b^*) \geq 0$. Otherwise, compute the minimizer b^{**} of $g(b)$ using b^* as an initial value via a safeguarded quadratic interpolation algorithm for one-dimensional optimization. Then set $b = b^{**}$.

The computation complexity for this algorithm is $O(n)$ in step 2 and $O(n^2)$ in step 3. In most cases, step 2 suffices, yielding $O(n)$ computation. In step 3 the one-dimensional optimization is relatively easy, because $g(b)$ is piecewise linear in b. Overall, computational speed is not a concern; slow convergence of the optimization algorithm is offset by fast function evaluation.

5. NUMERICAL EXAMPLE

5.1 Simulation

We now examine the effectiveness of ψ-learning via simulation. We consider two-dimensional linear classification, in which $f(x) = \sum_{i=1}^{2} w_i x_i + b$. A random training sample $\{X_{i1}, X_{i2}, Y_i\}_{i=1}^n$ is generated as follows. First, $\{(X_{i1}, X_{i2})\}_{i=1}^n$ are sampled from the uniform distribution over the unit disk $\{(x_1, x_2) : x_1^2 + x_2^2 \leq 1\}$, and Y_i is assigned to 1 if $X_{i1} \geq 0$ and -1 otherwise. Then random selected labels $\{Y_i\}_{i=1}^n$ are flipped, which generates a random sample for nonseparable cases.

Simulations are conducted via IMSL optimization routines (see the IMSL manual for a description of optimization routines in the IMSL Math Library). Five different levels of contamination are considered: 0-flip, 1-flip, 2-flip, and 10%-flip, each with three different values, $C = 10, 10^3, 10^7$. For each simulation, the values of $e(f, f^*)$ for SVM and ψ-learning are computed, as reported in Table 1. To assure computation accuracy, our IMSL version of SVM is cross-examined by a popular version of SVM software SVMTorch (see Collobert and Bengio 2001).

An expression of $e(f, f^*)$ is obtained for fast evaluation, where $e(f, f^*)$, after adjusted for the ideal performance, becomes a basis for comparison. Direct calculations yield that $Err(f) = k/n + (1 - 2k/n)A$, $Err(f^*) = k/n$, and $e(f, f^*) = (1 - 2k/n)A$, where k is the number of flips; πA, given here, is the area between the vertical line and a decision line within the unit circle:

$$A = \begin{cases} \left(\frac{1}{2}|\theta_1 + \pi/2| + \frac{1}{2}|\theta_2 - \pi/2| + |b/2w_2| \right. \\ \quad \left. \times (|\cos(\theta_1)| - |\cos(\theta_2)|)\right)/\pi, & w_2 \neq 0 \\ \left(\pi/2 - \frac{1}{2}(\theta_2 - \theta_1 - \sin(\theta_2 - \theta_1))\right)/\pi, & w_2 = 0. \end{cases}$$

Here $\theta_1 = \Psi - \cos^{-1}(|b|/(w_1^2 + w_2^2)^{1/2})$, $\theta_2 = \Psi + \cos^{-1}(|b|/(w_1^2 + w_2^2)^{1/2})$, and $\Psi = \cos^{-1}(|w_1|/(w_1^2 + w_2^2)^{1/2})$. It can be verified that the bracketing functions of $\mathcal{F}$ also satisfy the foregoing inequality. This implies that $Err(f^*)$ is $1/n$ for 1-flip, $2/n$ for 2-flip, and .1 for 10%-flip.

Evidently, ψ-learning outperforms SVM in terms of generalization error in all the nonseparable and separable cases, except that they yield the identical results for large C in all the separable cases (0-flip). The result for the separable cases agrees with our theoretical result in Proposition 2. For the nonseparable cases, the improvements of ψ-learning over SVM increase with n and become rather significantly for large n and C, with the largest improvement of 425% (4.25 = .017/.0040), in the case of 10%-flip with $n = 400$ and $C = 10^7$. Overall, the amount of improvement is substantial. As C increases, the performances of ψ-learning and SVM usually do not deteriorate.

From Table 1, we see that even in the separable case (i.e., 0-flip), ψ-learning has the advantage of its performance being less sensitive to the choice of C. This is important when the methodology is applied to the nonlinear case when the choice of C is critical. In the nonseparable case (i.e., when the class labels are contaminated by noise), ψ-learning has smaller GE for all $C = 10^1, 10^3, 10^7$. Furthermore, the degree of improvement over SVM becomes more dramatic as the sample size increases. Thus ψ-learning has more robust performance both with respect to the choice of C and with respect to noisy data.

The simulations also indicate the large-sample performance of ψ-learning and SVM in terms of the sample size n. As suggested by the slopes of the least squares lines in Figure 2, the

Table 1. Average $e(\hat{f}, f^*)$ and the Standard Deviation (in parentheses) for SVM and ψ-Learning Over 100 Simulation Runs

C		0-flip	1-flip	2-flip	10%-flip
$n = 25$					
10^1	SVM	.0521(.0297)	.0586(.0336)	.0598(.0350)	.0598(.0350)
	ψL	.0464(.0293)	.0496(.0344)	.0569(.0403)	.0569(.0403)
10^3	SVM	.0356(.0229)	.0534(.0359)	.0584(.0319)	.0584(.0319)
	ψL	.0356(.0229)	.0417(.0339)	.0522(.0461)	.0522(.0461)
10^7	SVM	.0358(.0228)	.0537(.0370)	.0582(.0322)	.0582(.0322)
	ψL	.0358(.0228)	.0444(.0378)	.0520(.0453)	.0520(.0453)
$n = 50$					
10^1	SVM	.0327(.0180)	.0340(.0174)	.0394(.0187)	.0472(.0258)
	ψL	.0279(.0171)	.0285(.0186)	.0298(.0188)	.0314(.0219)
10^3	SVM	.0181(.0121)	.0317(.0176)	.0353(.0196)	.0457(.0263)
	ψL	.0179(.0122)	.0193(.0161)	.0239(.0206)	.0320(.0271)
10^7	SVM	.0181(.0123)	.0318(.0177)	.0353(.0196)	.0457(.0263)
	ψL	.0181(.0123)	.0188(.0142)	.0223(.0165)	.0321(.0260)
$n = 100$					
10^1	SVM	.0181(.0091)	.0195(.0102)	.0216(.0110)	.0344(.0171)
	ψL	.0132(.0090)	.0130(.0084)	.0141(.0093)	.0159(.0144)
10^3	SVM	.0108(.0073)	.0163(.0097)	.0201(.0099)	.0338(.0169)
	ψL	.0105(.0077)	.0105(.0074)	.0109(.0079)	.0148(.0139)
10^7	SVM	.0103(.0078)	.0159(.0100)	.0201(.0099)	.0337(.0169)
	ψL	.0103(.0078)	.0104(.0078)	.0111(.0091)	.0154(.0129)
$n = 200$					
10^1	SVM	.0133(.0071)	.0138(.0079)	.0143(.0085)	.0227(.0120)
	ψL	.0091(.0074)	.0090(.0073)	.0095(.0077)	.0107(.0094)
10^3	SVM	.0067(.0043)	.0109(.0061)	.0132(.0073)	.0228(.0120)
	ψL	.0064(.0046)	.0064(.0046)	.0066(.0047)	.0079(.0063)
10^7	SVM	.0052(.0043)	.0108(.0066)	.0132(.0071)	.0228(.0120)
	ψL	.0052(.0043)	.0056(.0051)	.0060(.0062)	.0078(.0087)
$n = 400$					
10^1	SVM	.0087(.0045)	.0087(.0048)	.0092(.0050)	.0172(.0094)
	ψL	.0053(.0038)	.0049(.0036)	.0049(.0037)	.0053(.0041)
10^3	SVM	.0038(.0021)	.0063(.0036)	.0079(.0042)	.0170(.0094)
	ψL	.0033(.0023)	.0033(.0024)	.0035(.0028)	.0039(.0027)
10^7	SVM	.0027(.0021)	.0061(.0037)	.0078(.0041)	.0170(.0094)
	ψL	.0027(.0021)	.0027(.0021)	.0027(.0021)	.0040(.0039)

NOTE: In computation, the percent of flips is rounded down to the number of flips.

error rate of ψ-learning is of order n^{-1} in all of the cases. In contrast, the error rate of SVM is somewhere between $n^{-1/2}$ and $n^{-3/4}$ for the nonseparable cases. This finding suggests that learning accuracy deteriorates greatly when generalizing from separable to nonseparable cases, which agrees with our theoretical analysis in Section 3 and the discussion in Section 2.

Finally, we scrutinize how SVM and ψ-learning perform on one randomly selected training sample of size 25. As illustrated in Figure 3, the decision function of SVM gives three training errors and the decision function of ψ-learning give, two training errors. To see why this occurs, note that SVM needs to compensate for the left- and right-contaminated instances far from the decision line by reducing the value of the second term of (1). Because of this, SVM must sacrifice learning accuracy.

In summary, ψ-learning has delivered the optimal performance in linear classification and substantially outperforms SVM. The simulations also confirm that ψ-learning has better generalization ability than SVM when the size of training sample is reasonably large, except in separable cases, where they have essentially the same performance.

5.2 Application to Breast Cancer Classification

The Wisconsin Breast Cancer database (WBCD), collected at University of Wisconsin Hospitals, concerns visually assessed nuclear features of fine-needle aspirates taken from patients' breasts. Each sample was assigned to a nine-dimensional vector of diagnostic characteristics, with each component being in the interval 1–10, with 1 corresponding to a normal state and 10 corresponding to a most abnormal state. The goal is to determine whether a sample is benign or malignant, as found on biopsy and examination. The data were described in detail by Wolberg and Mangasarian (1990).

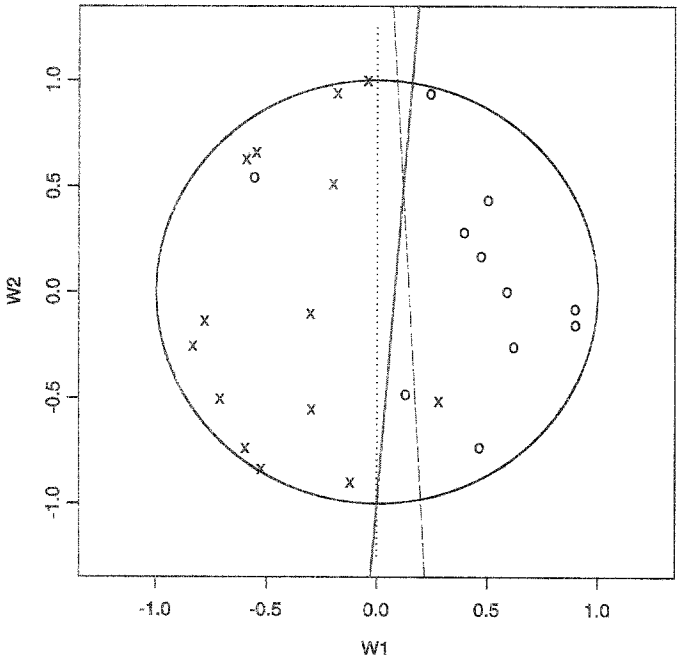

Figure 3. Plot of the Decision Functions of ψ-Learning and SVM, as Well as the True Decision Functions, Represented by the Orange (solid), Green (dashed), and Vertical (dotted) Lines, Respectively. The circled and crossed points are positive and negative instances. The training sample size is 25.

Table 2. Error Rates (in percentages) for WBCD

	SVM		ψL		Improvement (%)
Case	Testing	Training	Testing	Training	Testing
1	3.23%	2.35%	2.64%	1.47%	18.18%
2	2.64%	2.35%	2.64%	1.17%	0%
3	2.64%	2.35%	2.35%	1.17%	11.11%
4	2.05%	3.23%	1.76%	1.76%	14.29%
5	3.52%	1.47%	3.23%	1.17%	8.33%
6	1.76%	3.52%	1.47%	2.05%	16.67%
7	2.05%	2.64%	1.76%	1.47%	14.29%
8	1.76%	2.05%	1.17%	2.64%	33.33%
9	2.64%	2.93%	2.05%	1.76%	22.22%
10	2.35%	2.64%	1.76%	2.35%	25.00%

NOTE: For each case, a randomly selected subset of size 341 from a total of 682 samples is used to train, while the remaining samples are tested.

The WBCD has been used as a benchmark learning example, with a training error of 2.3% for linear machines when the 682 samples are used. (The details can be found at *ftp://ftp.cs.wisc.edu/math-prog/cpo-dataset/machine-learn/cancer1/*.) Here we apply ψ-learning to WBCD and compare its performance with SVM in the setting where $f(x) = w \cdot x + b$, where w and $x = (x_1, \ldots, x_9)$ are nine-dimensional vectors and $y = \pm 1$ indicate benign and malignancy.

To examine generalization ability, we randomly divide the 682 samples into two halves for testing and training. We then apply ψ-learning and SVM to the same randomly selected training and testing sets. Because the tuning parameter C may affect the performance of SVM and ψ-learning, we seek the best performance of ψ-learning and SVM with respect to a set of discretized C values in an interval $[10^{-3}, 10^3]$. The smallest testing errors and the corresponding training errors for SVM and ψ-learning based on 10 random selected partitions of WBCD are reported in Table 2. As suggested in the table, in terms of testing, ψ-learning outperforms SVM in 9 out 10 cases, all except for 1 case in which ψ-learning and SVM give the same performance. The percent improvement of ψ-learning over SVM range from 0 to 33.3%, which is not small in view of the high accuracy of SVM. For this application, the best performance of SVM is usually realized for a narrow range of small values of C, whereas that of ψ-learning is typically achieved for a much wider range of larger values of C.

6. DISCUSSION

This article has proposed a machine learning methodology, called ψ-learning, that is applicable to any problem of learning to classify data from examples. The theoretical and numerical analyses in this article show that ψ-learning has good generalization properties and can outperform popular SVMs. Although the computational complexity of ψ-learning is substantially higher than that of SVMs, we believe that the significant theoretical advantages, as suggested by the results presented herein, will make it worthwhile to aggressively pursue further computational developments of the ψ-learning methodology. Further developments of the theory are necessary to better understand how the performance of ψ-learning is related to the choice of $G(\mathcal{F})$. We hope that this article will serve to stimulate interest in these directions.

APPENDIX: PROOFS

Proof of Proposition 1. First, we prove that $\bar{f}$ minimizes $E(1 - Sign(Yf(X)))$. To this end, we note that $E((1 - Sign(Yf(X)))|X = x) = (1 - Sign(f(X))P(X) + (1 - Sign(-f(X))(1 - P(X)) = 1 - 2f^*(x)\, Sign(f)$, where $P(x) = P(Y = 1|x)$. Consequently, $E(1 - Sign(Yf(X)))$ is minimized when $f = \bar{f}$. Of course, the minimizer is not unique as f^* is also a minimizer. Furthermore, because $\psi(x) \geq (1 - Sign(x))$ and $\psi(y\, Sign(\bar{f}(x))) = 1 - y\, Sign(\bar{f}(x))$. The result then follows.

Proof of Proposition 2. We prove only the linear case. For large C, $\sum_{i=1}^n \psi(y_i f(x_i)) = 0$ is attained, implying that $y_i f(x_i) \geq 1; i = 1, \ldots, n$. This yields the desired result.

Lemma A.1. If two classes are separable, then $Err(f^*) = 0$.

Proof. We now prove a slightly stronger result. Let f_i; $i = 1, 2$, be the conditional probability densities for $\mathcal{A}_{\pm}$. By (8), $Err(f^*)$ can be written as

$$\int_{\{x: f_1/f_2 < \pi_1/\pi_2\}} \frac{\pi_1 f_1(x)}{\pi_1 f_1(x) + \pi_2 f_2(x)}\, dx + \int_{\{x: f_1/f_2 \geq \pi_1/\pi_2\}} \frac{\pi_2 f_2(x)}{\pi_1 f_1(x) + \pi_2 f_2(x)}\, dx.$$

By the definition of separability, $f_1(x) f_2(x) = 0$ for any $x \in S$, which implies the desired result.

Before proving Theorem 1, we need to define the L_2-metric entropy with bracketing for a function class $\mathcal{F}$. For any $\varepsilon > 0$, call $\{l_1^l, l_1^u, \ldots, l_m^l, l_m^u\}$ an ε-bracketing function if for any $f \in \mathcal{F}$ there is a j such that $l_j^l \leq l(f, \cdot) \leq l_j^u$ and $\max_{\{1 \leq j \leq m\}} \|l_j^u - l_j^l\|_2 \leq \varepsilon$, where $\|\cdot\|_2$ is the usual L_2-norm, defined as $\|l\|_2^2 = \int l^2 dP$. Then the L_2 metric entropy of $\mathcal{F}$ with bracketing $H_B(\varepsilon, \mathcal{F})$ is defined as a logarithm of the cardinality of the ε-bracketing of the smallest size.

Proof of Theorem 1. Before proceeding, we introduce some notation to be used later. Let $\tilde{l}_\psi(f, Z_i) = l_\psi(f, Z_i) + \lambda J(f)$ be the cost function to be minimized, as in (4), where $l_\psi(f, Z_i) = \psi(Y_i f(X_i))$ and $\lambda = 1/(Cn)$. Let $\tilde{l}(f, Z_i) = l(f, Z_i) + \lambda J(f)$ be the corresponding cost function defined by $Sign(\cdot)$, as in (2), where $l(f, Z_i) = (1 - Sign(Y_i f(X_i)))$. Define the scaled empirical process, $E_n(\tilde{l}(f, Z) - \tilde{l}_\psi(f_0, Z))$, as

$$n^{-1} \sum_{i=1}^n \big(\tilde{l}(f, Z_i) - \tilde{l}_\psi(f_0, Z_i) - E(\tilde{l}(f, Z_i) - \tilde{l}_\psi(f_0, Z_i))\big) = E_n\big(l(f, Z) - l_\psi(f_0, Z)\big),$$

where $Z = (X, Y)$. Let

$$A_{i,j} = \big\{f \in \mathcal{F} : 2^{i-1}\delta_n^2 \leq e(f, \bar{f}) < 2^i \delta_n^2, 2^{j-1} \max(J(f_0), 1) \leq J(f) < 2^j \max(J(f_0), 1)\big\},$$

and let

$$A_{i,0} = \big\{f \in \mathcal{F} : 2^{i-1}\delta_n^2 \leq e(f, \bar{f}) < 2^i \delta_n^2, J(f) < \max(J(f_0), 1)\big\},$$

for $j = 1, 2, \ldots$, and $i = 1, 2, \ldots$. Without loss of generality, we assume that $J(f_0) \geq 1$ in the sequel. Because $e(f, \bar{f}) \leq 1/2$ for any f, we assume, without loss of generality, that $\max(\varepsilon_n^2, 2s_n) < 1$.

The treatment used here is essentially the same as that of Wong and Shen (1995) and Shen (1998). Our approach for bounding $P(e(\hat{f}, \bar{f}) \geq \delta_n^2)$ is to reduce the problem of bounding a sequence of empirical processes that are induced by the cost function $\tilde{l}$. Specifically, we apply theorem 3 of Shen and Wong (1994), a large deviation inequality for empirical processes, to bound $P(A_{ij})$; $i, j = 1, \ldots, n$, by controlling the mean and variance, defined by $l(f, Z_i)$ and penalty λ. This yields an inequality for the sequence of empirical processes and thus for $e(\hat{f}, \bar{f})$.

First, we establish the connection between $e(\hat{f}, \bar{f})$ and the empirical processes. By Assumption D,

$$\tilde{l}_\psi(f_0, Z_i) - \tilde{l}(f, Z_i) \geq \tilde{l}_\psi(f_0, Z_i) - \tilde{l}_\psi(f, Z_i), \qquad i = 1, \ldots, n. \quad \text{(A.1)}$$

It follows from the fact that $\hat{f}$ is the maximizer of $-n^{-1}\sum_{i=1}^n \tilde{l}_\psi(f, Z_i)$, $e(f_0, \bar{f}) \leq e_\psi(f_0, \bar{f}) \leq \delta_n^2$, and (A.1) that

$$\big\{e(\hat{f}, \bar{f}) \geq \delta_n^2\big\} \subset \left\{\sup_{\{f \in \mathcal{F}: e(f, \bar{f}) \geq \delta_n^2\}} n^{-1} \sum_{i=1}^n (\tilde{l}_\psi(f_0, Z_i) - \tilde{l}_\psi(f, Z_i)) \geq 0\right\} \subset \left\{\sup_{\{f \in \mathcal{F}: e(f, \bar{f}) \geq \delta_n^2\}} n^{-1} \sum_{i=1}^n (\tilde{l}_\psi(f_0, Z_i) - \tilde{l}(f, Z_i)) \geq 0\right\}.$$

Hence

$$P\big(e(\hat{f}, \bar{f}) \geq \delta_n^2\big) \leq P^*\left(\sup_{\{f \in \mathcal{F}: e(f, \bar{f}) \geq \delta_n^2\}} n^{-1} \times \sum_{i=1}^n (\tilde{l}_\psi(f_0, Z_i) - \tilde{l}(f, Z_i)) \geq 0\right) = I,$$

where P^* denotes the outer probability measure.

To bound I, it suffices to bound the corresponding probability over A_{ij}, for each $i, j = 1, \ldots$. To this end, we need some inequalities regarding the first and second moments of $\tilde{l}(f, Z) - \tilde{l}_\psi(f_0, Z)$ for $f \in A_{ij}$.

For the first moment, note that $E(l(f, Z) - l_\psi(f_0, Z)) = E(l(f, Z) - l_\psi(\bar{f}, Z)) - E(l_\psi(f_0, Z)) - l_\psi(\bar{f}, Z))$, which is equal to $2(e(f, \bar{f}) - e_\psi(f_0, \bar{f}))$, because $El_\psi(\bar{f}, Z) = El(\bar{f}, Z)$ (Prop. 1). By Assumption A, $2e_\psi(f_0, \bar{f}) \leq 2s_n \leq \delta_n^2$. Then, using the assumption that $\max(J(f_0), 1)\lambda \leq \delta_n^2/2$, we have, for any integers $i, j \geq 1$,

$$\inf_{A_{ij}} E\big(\tilde{l}(f, Z) - \tilde{l}_\psi(f_0, Z)\big) \geq M(i, j) = (2^{i-1}\delta_n^2) + \lambda(2^{j-1} - 1)J(f_0) \quad \text{(A.2)}$$

and

$$\inf_{A_{i,0}} E(\tilde{l}(f, Z) - \tilde{l}_\psi(f_0, Z)) \geq (2^{i-1} - 1/2)\delta_n^2 \geq M(i, 0) = 2^{i-2}\delta_n^2, \quad \text{(A.3)}$$

where the fact that $2^i - 1 \geq 2^{i-1}$ has been used.

For the second moment, it follows from (6) and Assumption B that for any $f \in \mathcal{F}$,

$$\begin{aligned} e(f, \bar{f}) &= E|f^*(X)|\big|Sign(Y\bar{f}(X)) - Sign(Yf(X))\big| \\ &\geq \delta E\big|Sign(Y\bar{f}(X)) - Sign(Yf(X))\big| I(|f^*(X)| \geq \delta) \\ &\geq \delta\big(E|\, Sign(Y\bar{f}(X)) - Sign(Yf(X))| - 2c_1\delta^\alpha\big) \\ &\geq 2^{-1}(4c_1)^{-1/\alpha}\big(E|\, Sign(Y\bar{f}(X)) - Sign(Yf(X))|\big)^{(\alpha+1)/\alpha}, \end{aligned}$$

with a choice of $\delta = (E|\, Sign(Y\bar{f}(X)) - Sign(Yf(X))|/4c_1)^{1/\alpha}$. Now we establish a connection between the first and second moments. By Proposition 1, $E(\psi(Y\bar{f}(X)) - (1 - Sign(Y\bar{f}(X))) = 0$. Note that $\psi(x) \geq (1 - Sign(x))$ for any x, $E|\psi(Yf_0(X)) - (1 - Sign(Yf_0(X))| = E(\psi(Yf_0(X)) - (1 - Sign(Yf_0(X))) \leq e_\psi(f_0, \bar{f})$. Therefore, by the triangular inequality,

$$\begin{aligned} E(l(f, Z) - l_\psi(f_0, Z))^2 &\leq UE|(1 - Sign(Yf(X))) - \psi(Yf_0(X))| \\ &\leq U(E|\, Sign(Y\bar{f}(X)) - Sign(Yf(X))| \\ &\quad + E|\, Sign(Y\bar{f}(X)) - Sign(Yf_0(X))| \\ &\quad + e_\psi(f_0, \bar{f})). \end{aligned}$$

For any $f \in A_{i,j}$, $e(f,\bar{f})^{\frac{\alpha}{\alpha+1}} \geq (2^{-1}\delta_n^2)^{\frac{\alpha}{\alpha+1}} \geq 2^{-1}\delta_n^2 \geq e_\psi(f_0,\bar{f})$ and $e(f,\bar{f}) \geq e(f_0,\bar{f})$, implying that

$$\begin{aligned} &E\big(l(f,Z) - l_\psi(f_0,Z)\big)^2 \\ &\quad \leq U\big(2(4c_1)^{1/\alpha}\big(e(f,\bar{f})^{\frac{\alpha}{\alpha+1}} + e(f_0,\bar{f})^{\frac{\alpha}{\alpha+1}}\big) + e_\psi(f_0,\bar{f})\big) \\ &\quad \leq c_3(e(f,\bar{f})/2)^{\frac{\alpha}{\alpha+1}}, \end{aligned}$$

where $c_3 = 2^{1/\alpha}U\max(4(2^{\alpha+2}c_1)^{1/(\alpha+1)}+2), 8\max(U,2))$. Consequently,

$$\sup_{A_{i,j}} E\big(l_\psi(f_0,Z) - l(f,Z)\big)^2 \leq v^2(i,j) = c_3 M(i,j)^{\frac{\alpha}{\alpha+1}};$$
$$i = 1,\ldots,\ j = 0,\ldots.$$

We are now ready to bound I. Using the assumption that $\max(J(f_0),1)\lambda \leq \delta_n^2/2$, (A.2), and (A.3), we have

$$\begin{aligned} I &\leq \sum_{i,j} P^*\Big(\sup_{A_{i,j}} E_n(l_\psi(f_0,Z) - l(f,Z)) \geq M(i,j)\Big) \\ &\quad + \sum_i P^*\Big(\sup_{A_{i,0}} E_n(l_\psi(f_0,Y) - l(f,Y)) \geq M(i,0)\Big) \\ &= I_1 + I_2, \end{aligned}$$

Next we proceed to bound I_i separately. For I_1, we verify the required conditions (4.5)–(4.7) in theorem 3 of Shen and Wong (1994). To compute the metric entropy in that (4.7), we now define a bracketing function for $l_\psi(f_0,Z) - l(f,Z)$. Denote an ε-bracketing set for $\{G_f : G_f = \{x \in S : f(x) \geq 0\}, f \in A_{i,j}\}$ to be $\{(G_1^l, G_1^u),\ldots,(G_m^l, G_m^u)\}$. Let $s_j^l(x)$ be -1 if $x \in G_j^u$ and 1 otherwise, and $s_j^u(x)$ be -1 if $x \in G_j^l$ and 1 otherwise; $j = 1,\ldots,m$. Then, $\{(s_1^l, s_1^u),\ldots,(s_m^l, s_m^u)\}$ forms an ε-bracketing function of $-Sign(f)$ for $f \in A_{i,j}$. This implies that for any $\varepsilon \geq M(i,j)$ and $f \in A_{i,j}$, there exists a j $(1 \leq j \leq m)$ such that $l_j^l(z) \leq l(f,z) - l_\psi(f_0,z) \leq l_j^u(z)$ for any $z = (y,x)$, where

$$\begin{aligned} l_j^u(z) &= 1 + (s_j^u(x)(1+y)/2 - s_j^l(x)(1-y)/2) - l_\psi(f_0,z), \\ l_j^l(z) &= 1 + (s_j^l(x)(1+y)/2 - s_j^u(x)(1-y)/2) - l_\psi(f_0,z), \end{aligned}$$

and $(E(l_j^u - l_j^l)^2)^{1/2} = (E(s_j^u(x) - s_j^l(x))^2)^{1/2} \leq 2^{1/2}\varepsilon^{1/2}$. Hence $(E(l_j^u - l_j^l)^2)^{1/2} \leq \min((2\varepsilon)^{1/2}, 2^{1/2})$. Consequently $H_B(\varepsilon, \mathcal{F}(2^j)) \leq H(\varepsilon^2/2, \mathcal{G}(2^j))$ for any $\varepsilon > 0$ and $j = 0,\ldots,$ where $\mathcal{F}(2^j) = \{l(f,z) - l_\psi(f,z) : f \in \mathcal{F}, J(f) \leq 2^j\}$. Using the fact that $\int_{aM(i,j)}^{v(i,j)} H^{1/2}(u^2/2, \mathcal{G}(2^j))du/M(i,j)$ is nonincreasing in i and $M(i,j)$; $i = 1,\ldots,$ we have

$$\begin{aligned} &\int_{aM(i,j)}^{v(i,j)} H^{1/2}(u^2/2, \mathcal{G}(2^j))du/M(i,j) \\ &\quad \leq \int_{aM(1,j)}^{c_3^{1/2}M(1,j)^{\alpha/2(\alpha+1)}} H^{1/2}(u^2/2, \mathcal{G}(2^j))du/M(1,j) \leq \phi(\varepsilon_n, 2^j), \end{aligned}$$

where $a = \varepsilon/32$. Then Assumption C implies (4.7) with a choice of $\varepsilon = 1/2$ and c_i; $i = 3,4$. Finally, it is easy to see that (4.5)–(4.6) are satisfied with $\varepsilon = 1/2$ with the choice of $M(i,j)$, $v(i,j)$, and $T = \max(U,2)$. More specifically, $M(i,j)/v^2(i,j) \leq 1/8\max(U,2)$ implies (4.6), and (4.7) implies (4.5).

Note that $0 < \delta_n \leq 1$ and $\lambda\max(J(f_0),1) \leq \delta_n^2/2$. An application of theorem 3 of Shen and Wong (1994) with $M = n^{1/2}M(i,j)$, $v = v^2(i,j)$, $\varepsilon = 1/2$, and $T = \max(U,2)$ yields that

$$I_1 \leq \sum_{j=1}^{\infty}\sum_{i=!}^{\infty} 3\exp\Big(-\frac{(1-\varepsilon)nM(i,j)^2}{2(4v^2(i,j) + M(i,j)T/3)}\Big)$$

$$\begin{aligned} &\leq \sum_{j=1}^{\infty}\sum_{i=1}^{\infty} 3\exp\big(-c_5 nM(i,j)^{\frac{\alpha+2}{\alpha+1}}\big) \\ &\leq \sum_{j=1}^{\infty}\sum_{i=1}^{\infty} 3\exp\big(-c_5 n[2^{i-1}\delta_n^2 + (2^{j-1}-1)\lambda J(f_0)]^{\frac{\alpha+2}{\alpha+1}}\big) \\ &\leq \sum_{j=1}^{\infty}\sum_{i=1}^{\infty} 3\exp\big(-c_5 n[(2^{i-1}\delta_n^2)^{\frac{\alpha+2}{\alpha+1}} + ((2^{j-1}-1)\lambda J(f_0))^{\frac{\alpha+2}{\alpha+1}}]\big) \\ &\leq 3\exp\big(-c_5 n(\lambda J(f_0))^{\frac{\alpha+2}{\alpha+1}}\big)/\big[(1-\exp(-c_5 n(\lambda J(f_0))^{\frac{\alpha+2}{\alpha+1}}))\big]^2. \end{aligned}$$

Here and in the sequel, c_5 is a positive generic constant. Similarly, I_2 can be bounded.

Finally,

$$I \leq 6\exp\big(-c_5 n(\lambda J(f_0))^{\frac{\alpha+2}{\alpha+1}}\big)/\big[(1-\exp(-c_5 n(\lambda J(f_0))^{\frac{\alpha+2}{\alpha+1}}))\big]^2.$$

This implies that $I^{1/2} \leq (5/2 + I^{1/2})\exp(-c_5 n(\lambda J(f_0))^{\frac{\alpha+2}{\alpha+1}})$. The result then follows from the fact $I \leq I^{1/2} \leq 1$.

Proof of Corollary 1. The result then follows from the exponential inequality established in Theorem 1.

Lemma A.2: Metric Entropy of Example 3.2.2. Under the assumption in the nonlinear classification example in Section 3.2.2, we have

$$H(\varepsilon, G(\mathcal{F})) \leq O(\log(1/\varepsilon)).$$

Proof. First, we define a set of bracketing functions. Let $W = \{0, n^{-1}, 2n^{-1},\ldots,1\}$, where $n = [2\tilde{p}/\varepsilon]$ and $[x]$ is the integer part of x. Let $K(\mathcal{G},p^*)$ be $\{v : v = \bigcup_{i=1}^{p^*}[\tilde{l}_i,\tilde{u}_i), [\tilde{l}_i,\tilde{u}_i) \cap [\tilde{l}_j,\tilde{u}_j) = \Phi, i \neq j\}$, where Φ denotes the empty set.

Define $A(\mathcal{G},p^*) = \{u : u = \bigcup_{i=1}^{p^*}[l_i,u_i), [l_i,u_i) \cap [l_j,u_j) = \Phi, i \neq j\}$. Clearly, for any $u \in A(\mathcal{G},p^*)$, there exists $v \in K(\mathcal{G},p^*)$ such that $d(u,v) \leq 2\frac{\varepsilon}{2\tilde{p}}p^* \leq \varepsilon$. This is because for any given l_i, u_i there exists $\tilde{l}_i, \tilde{u}_i \in W$ such that $\max(|l_i - \tilde{l}_i|, |u_i - \tilde{u}_i|) \leq \frac{\varepsilon}{2\tilde{p}}$.

Note that $\mathcal{G} \subset \bigcup_{p^*=1}^{\tilde{p}} K(\mathcal{G},p^*)$. Then it suffices to bound the capacity $|K(\mathcal{G},p^*)|$, which is upper bounded by $C_{2p^*}^n < n^{2p^*}$. Hence $\sum_{p^*=1}^{\tilde{p}} K(\mathcal{G},p^*) \leq \sum_{p^*=1}^{\tilde{p}} n^{2p^*} < \tilde{p}n^{2\tilde{p}} = \tilde{p}(2\tilde{p}/\varepsilon)^{2\tilde{p}}$. The desired result then follows.

[Received November 2001. Revised April 2003.]

REFERENCES

An, L. T. H., and Tao, P. D. (1997), "Solving a Class of Linearly Constrained Indefinite Quadratic Problems by DC Algorithms," *Journal of Global Optimization*, 11, 253–285.

Bartlett, P. L., and Shawe-Taylor, J. (1999), Generalization Performance of Support Vector Machines and Other Pattern Classifiers," in *Advances in Kernel Methods: Support Vector Learning*, eds. B. Schokkopf, C. J. C. Burges, and A. J. Smola, Cambridge, MA: MIT Press, pp. 43–54.

Birgé, L. and Massart, P. (1993), "Rates of Convergence for Minimum Contrast Estimators," *Probability Theory and Related Fields*, 97, 113–150.

Blanquero, R., and Carrizosa, E. (2000), "On Covering Methods for DC Optimization," *Journal of Global Optimization*, 18, 265–274.

Blumer, A., Ehrenfucht, A., Haussler, D., and Warmuth, M. K. (1989), "Learnability and the Vapnik–Chervonenkis Dimension," *J. Assoc. Comput. Mach.*, 36, 929–965.

Boser, B., Guyon, I., and Vapnik, V. N. (1992), "A Training Algorithm for Optimal Margin Classifiers," in *Proceedings of the Fifth Annual Conference on Computational Learning Theory*, ACM, pp. 142–152.

Collobert, R., and Bengio, S. (2001), "SVMTorch: Support Vector Machines for Large-Scale Regression Problems," *Journal of Machine Learning Research*, 1, 143–160.

Cortes, C., and Vapnik, V. (1995), "Support-Vector Networks," *Machine Learning*, 20, 273–297.

Cristianini, N., and Shawe-Taylor, J. (1999), *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*, Cambridge, U.K.: Cambridge University Press.

Devroye, L., Györfi, L., and Lugosi, G. (1996), *A Probabilistic Theory of Pattern Recognition*, New York: Springer-Verlag.

Gill, P. E., Murray, W., and Wright, M. H. (1981), *Practical Optimization*, London: Academic Press.

Koltchinskii, V., and Panchenko, D. (2002), "Empirical Margin Distributions and Bounding the Generalization Error of Combined Classifier," *The Annals of Statistics*, 30, 1–50.

Lin, Y. (2000), "Some Asymptotic Properties of the Support Vector Machine," Technical report 1029, University of Wisconsin-Madison, Dept. of Statistics.

——— (2002), "A Note on Margin-Based Loss Functions in Classification," Technical Report 1043, University of Wisconsin-Madison, Dept. of Statistics.

Mammen, E., and Tsybakov, A. B. (1999), "Smooth Discrimination Analysis," *The Annals of Statistics*, 27, 1808–1829.

Mangasarian, O. L. (2000), "Generalized Support Vector Machine," in *Advances in Large Margin Classifiers*. Cambridge, MA: MIT Press.

Shen, X. (1998), "On the Method of Penalization," *Statistica Sinica*, 8, 337–357.

Shen, X., and Wong, W. H. (1994), "Convergence Rate of Sieve Estimates," *The Annals of Statistics*, 22, 580–615.

Van De Geer, S. (1993), "Hellinger-Consistency of Certain Nonparametric Maximum Likelihood Estimators," *The Annals of Statistics*, 21, 14–44.

Vapnik, V. (1998), *Statistical Learning Theory*, Chichester, UK: Wiley.

——— (1999), *The Nature of Statistical Learning Theory* (2nd ed.), New York: Springer-Verlag.

Wahba, G. (1998), "Support Vector Machines, Reproducing Kernel Hilbert Spaces and the Randomized GACV," Technical report 984, University of Wisconsin, Dept. of Statistics.

——— (1990), *Spline Models for Observational Data*, Regional Conference, Philadelphia: CBMS-NSF.

Wolberg, W. H., and Mangasarian, O. L. (1990), "Multisurface Method of Pattern Separation for Medical Diagnosis Applied to Breast Cytology," *Proceedings of the Nat. Academy of Sciences, USA*, 87, 9193–9196.

Wong, W. H., and Shen, X. (1995), "Probability Inequalities for Likelihood Ratios and Convergence Rates for Sieve MLEs," *The Annals of Statistics*, 23, 339–362.