GENERALIZATION ERROR RATES FOR MARGIN-BASED CLASSIFIERS

DISSERTATION

Presented in Partial Fulfillment of the Requirements for
the Degree Doctor of Philosophy in the
Graduate School of The Ohio State University

By

Changyi Park, M.S.

* * * * *

The Ohio State University

2005

Dissertation Committee:

Xiaotong Shen, Adviser

Prem K. Goel

Joseph S. Verducci

Yoonkyung Lee

Approved by

Adviser
Department of Statistics

ABSTRACT

Margin-based classifiers defined by functional margins are generally believed to yield high performance in classification. In this thesis, a general theory that quantifies the size of generalization error of a margin classifier is presented. The trade-off between geometric margins and training errors is captured, in addition to the complexity of a classification problem. The theory permits an investigation of the generalization ability of convex and nonconvex margin classifiers, including support vector machines (SVM), kernel logistic regression (KLR), and ψ-learning. Our theory indicates that the generalization ability of a certain class of nonconvex losses may be substantially faster than those for convex losses. Illustrative examples for both linear and nonlinear classification are provided.

To God.

ACKNOWLEDGMENTS

I would like to thank Professor Xiaotong Shen, my advisor, for introducing me interesting topics in statistical learning theory. Without his guidance and constant support, I could not have made any progress at all. I am thankful to my committee members; Professor Prem K. Goel, Joseph S. Verducci, and Yoonkyung Lee, for helpful comments.

Yufeng Liu at the University of North Carolina and Sijin Liu supplied me with the preprints of their recent work with Professor Shen. This gave me a better understanding of perspective on my own research topic.

I am grateful to my family for their love and support. Also, I should mention my friends at the Ohio State University including members of the OSU tennis club and the Korean Catholic Community.

Because every first-born belongs to God, I dedicate my first intellectual product to God. Thanks be to God.

Columbus, Ohio

August 17, 2005 Changyi Park

VITA

February 3, 1972	Born - Seosan, Korea
1994	B.S. Computer Science and Statistics, Seoul National University, Korea.
1996	M.S. Statistics, Seoul National University, Korea.
2000-2001	Graduate Teaching Associate, The Ohio State University.
2002-present	Graduate Research Associate, The Ohio State University.

FIELDS OF STUDY

Major Field: Statistics

TABLE OF CONTENTS

Page

Chapters:

Appendices:

LIST OF TABLES

LIST OF FIGURES

Figure **Page**

CHAPTER 1

INTRODUCTION

Classification, as a tool to extract information from data, has played an important role in science and engineering. There have been various methods for classification in the literature including traditional ones such as discriminant analysis and logistic regression and more modern ones such as classification trees and aggregating classifiers. Hastie, Tibshirani, and Friedman (2001) provide a good introduction to different classification techniques.

Let us briefly review some of recent classification techniques first.

- Classification trees

  Classification trees; c.f., Breiman, Friedman, Olshen, and Stone (1984), are constructed through recursive partitioning: splitting the feature space into partitions (nodes), and then splitting it up further on each of the partitions. Class labels are assigned to each node. An attractive feature of a classification tree is that it can be readily displayed in graphics. However, classification trees have problems such as high variance and lack of smoothness in prediction.

- Aggregating classifiers

  Boosting produces a composite classifier by combining simple base classifiers in a greedy fashion; c.f., Freund and Schapire (1997). Recently, it was shown that boosting can overfit (Jiang, 2004). To prevent overfitting, early stopping of boosting is necessary. Zhang and Yu (2005) studied convergence and consistency of boosting with early stopping. Bagging is another aggregating method that reduces the variance by aggregating many trees via bootstrap; c.f., Breiman (1996).

- Margin-based classifiers

  Margin-based classification has seen significant developments in the past several years, including many well-known classifiers such as support vector machine (SVM, Cortes and Vapnik, 1995), kernel logistic regression (KLR, Zhu and Hastie, 2005), and ψ-learning (Shen, Tseng, Zhang, and Wong, 2003) among others.

In this thesis, we will focus on the generalization accuracy of margin-based classifiers. The central problem to be addressed is (1) the generalization accuracy of various margin classifiers, obtained by minimizing a penalized margin cost function, and (2) the optimal performance for any classifier.

There has been considerable interest on the generalization accuracy of margin classifiers, in particular those delivering good numerical results, in the literature. Zhang (2004a), and Lugosi and Vayatis (2004) obtained consistency for convex margin losses. Lin (2000) studied the rates of convergence for SVM, based on a formulation of the method of sieves, where the rates are the same as those in function estimation. Shen, Zhang, Tseng, and Wong (2003) derived a learning theory for nonconvex ψ-learning, where the rates are usually faster than those in function estimation. In fact, they show that a fast rate of n^{-1} is attainable by ψ-learning in a linear nonseparable example. Bartlett, Jordan, and McAuliffe (2003) obtained the rates of convergence for convex margin losses, but did not cover SVM. It seems that treating margin classification via penalization is most relevant to the present formulation.

Despite progress, several important issues remain yet unresolved. First, what is the size of the generalization error of a general margin classifier, convex or nonconvex? Second, what is the best performance that one anticipates for any classifier?

In this thesis, we derive a general upper bound theory for margin-based classifiers, convex or nonconvex, obtained by minimizing a certain cost function via penalization, in addition to a lower bound theory quantifying the optimal performance. Specifically, we derive some probability as well as risk upper bounds of the generalization error of a general loss. In nonseparable cases, a class of convex margin losses usually yields slow nonparametric function estimation rates, whereas a class of nonconvex margin losses, including ψ-learning losses, leads to sharper rates. In separable cases, both classes yield sharp rates. Most importantly, our lower bound theory provides an insight into the issue of attainment of the optimal rates by various classifiers. Through an application of the lower bound theory, we show that the optimal performances are achieved by the class of the nonconvex losses in classification examples.

This thesis is organized as follows. Chapter 2 discusses margin-based classifiers with respect to the choice of losses. Chapter 3 establishes a general upper bound theory concerning the generalization error of a general margin-based classifier, followed by a general lower bound theory for any classifier. Chapter 4 presents some numerical examples, followed by a discussion in Chapter 5. The appendix contains technical proofs.

CHAPTER 2

MARGIN-BASED CLASSIFICATION

Classification can be characterized by four key components: an input space $\mathcal{X}$, an output space $\mathcal{Y}$, a decision function f, and a training sample $(X_i, Y_i)_{i=1}^n$. Let us confine ourselves to binary classification only, i.e., $\mathcal{Y}$ is dyadic, with ± 1 indicating positive and negative classes $\mathcal{A}_\pm$. Classification is performed by constructing f, mapping from $\mathcal{X} \subset \mathbb{R}^d$ to $\mathbb{R}^1$, such that its sign, $Sign(f)$, called a classifier, decides the class assignment of an input $x \in \mathcal{X}$. Throughout the thesis, f is assumed to be measurable. A sample $(X_i, Y_i)_{i=1}^n$ of n input/output pairs is used to train f, which is independent and identically distributed according to an unknown joint probability $P(\cdot, \cdot)$ on $(\mathcal{X} \times \{-1, 1\}, \sigma(\mathcal{X}) \times 2^{\{-1,1\}})$ with $\sigma(\mathcal{X})$ a σ-field on $\mathcal{X}$.

2.1 Concept of Margins

The concept of margins is important in the accuracy of generalization. The *functional margin* of an instance (x_i, y_i) with respect to a decision function f is defined as

$$\gamma_i = y_i f(x_i),$$

indicating correct classification when $\gamma_i > 0$. This concept is directly related to generalization in that the ratio of the number of negative margins to n for sample

$(X_i, Y_i)_{i=1}^n$ is the training error. Consequently, $\{\gamma_i\}_{i=1}^n$ indicate the overall performance of classification with respect to f.

For linearly separable cases, SVM maximizes a separation margin or the *geometric margin* $2/\|w\|^2$ with respect to a linear decision function f subject to the constraints $y_i f(x_i) \geq 1$; $i = 1, \cdots, n$, enforcing zero training error where $f(x) = \langle w, x \rangle + b$ is a hyperplane with $\langle \cdot, \cdot \rangle$ the usual inner product in $\mathbb{R}^d$ and $b \in \mathbb{R}^1$. For non-separable cases, a soft-margin SVM is introduced to minimize $\frac{1}{2}\|w\|^2 + C\sum_{i=1}^n \zeta_i$ subject to the constraints $\zeta_i \geq 1 - y_i(\langle w, x_i \rangle + b)$ and $\zeta_i \geq 0$; $i = 1, \cdots, n$, where $\{\zeta_i\}_{i=1}^n$ are called the slack variables. The equivalent unconstrained minimization problem is

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^n [1 - y_i f(x_i)]_+ \tag{2.1}$$

with $[z]_+ = z$ if $z \geq 0$ and $[z]_+ = 0$ otherwise. This cost function is extended to a general class of margin-based loss functions via penalization.

For nonlinear classification, the geometric margin becomes $2/\|g\|_K^2$ when f has a kernel representation of $g(x) + b \equiv \sum_{i=1}^n \alpha_i K(x, x_i) + b$, where $\|g\|_K^2 = \sum_{i=1}^n \sum_{j=1}^n$ $\alpha_i\ \alpha_j\ K(x_i, x_j)$ is the reproducing kernel norm of g. Here $K(\cdot, \cdot)$ is a Mercer kernel (Mercer, 1909) that maps from $\mathcal{X} \times \mathcal{X}$ to $\mathbb{R}$ so that $\|g\|_K^2$ is a proper norm. Then the problem is minimizing

$$\frac{1}{2}\|g\|_K^2 + C\sum_{i=1}^n [1 - y_i f(x_i)]_+ \tag{2.2}$$

The penalty term $\|g\|_K^2$ is inversely proportional to the geometric margin.

2.2 Margin-based Losses

Note that direct minimization of the cost function defined by the misclassification (or $0-1$) loss, requiring set estimation, is feasible only for some special classes such as monotone classes; c.f., Mammen and Tsybakov (1999). Furthermore, the solution obtained by such a minimization is essentially zero in linear classification because *Sign* function is scaling invariant (Shen et al., 2003). In practice, a surrogate loss as an upper bound of the misclassification loss is often used in optimization.

The basic idea of margin classification is to construct a margin loss $V(\cdot)$ so that it is a function of the functional margin of (x, y), i.e., $V = V(yf(x))$. To obtain a classifier, it is natural to minimize a cost function $\sum_{i=1}^{n} V(Y_i f(X_i))$ defined by a training sample $(X_i, Y_i)_{i=1}^{n}$ with respect to $f \in \mathcal{F}$, a class of candidate decision functions. This mimics the minimization of $EV(Yf(X))$ with respect to $f \in \mathcal{F}$. To prevent overfitting from occurring, a nonnegative penalty $J[f]$ is added to yield a penalized margin cost function:

$$J[f] + C \sum_{i=1}^{n} V(Y_i f(X_i)), \tag{2.3}$$

where $C > 0$ is a tuning parameter and J penalizes some undesirable properties of f. Then C controls the trade-off between the training error and the penalty. The minimizer of (2.3) with respect to $f \in \mathcal{F}$ yields an estimated decision function $\hat{f}$, and hence the classifier $Sign(\hat{f})$. In the machine learning literature, J is often the inverse of the geometric margin or the conditional Fisher information (Corduneanu and Jaakola, 2003).

Different choices of V yield different classification methodologies. SVM uses the hinge loss defined as $V(z) = [1 - z]_+$. Its variants are in a more general form $V(z) = [1 - z]_+^q$ for $q \geq 1$; c.f., Lin (2000). The kernel logistic regression (KLR) adopts the logistic loss $V(z) = \log(1 + e^{-z})$, c.f., Zhu and Hastie (2005). The ψ-loss is of the form of $V(z) = \psi(z)$, defined as $\psi(z) = 0$ if $z \geq 1$, $\psi(z) = 1 - z$ if $0 \leq z \leq 1$, and 2 otherwise, c.f., Shen et al. (2003). The normalized sigmoid loss $V(z) = 1 - \tanh(cz)$

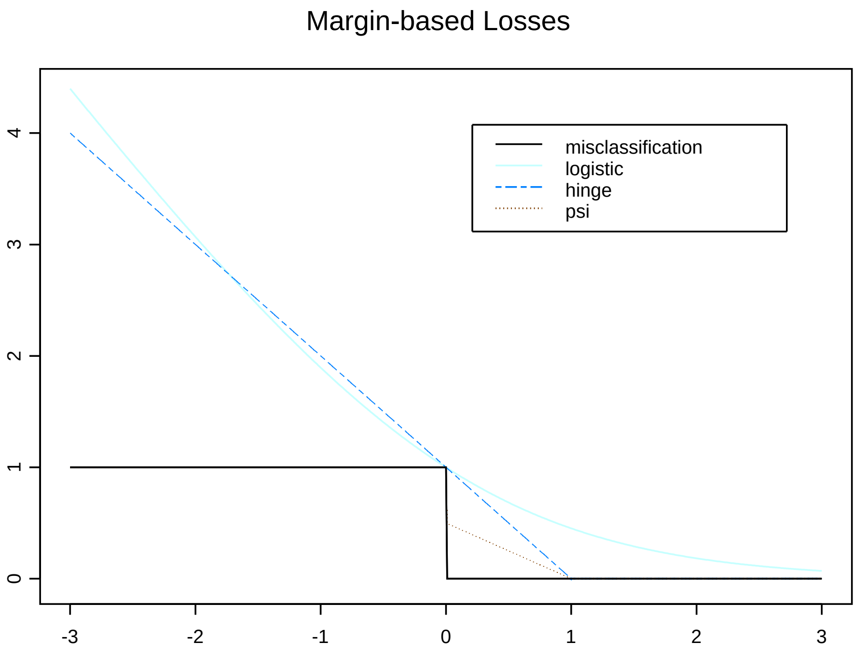

Figure 2.1: Examples of margin-based losses.

is also margin-based; c.f., Mason, Baxter, Bartlett, and Frean (2000). The weighted distance discriminant analysis uses a nonstandard loss function depending on margin. However, their loss may not be compared directly with other margin-based losses because their penalty term is different from the standard choice, c.f., Marron and Todd (2002). Examples of margin-based losses such as logistic, hinge, and ψ are given in Figure 2.1, where each loss is scaled so that it is the tight upper bound of the misclassification loss. Throughout the thesis, we consider scaled margin losses so that they are tight upper bounds of the misclassification loss.

CHAPTER 3

STATISTICAL LEARNING THEORY

3.1 Generalization and Surrogate Errors

Let L be the misclassification loss defined as $L(z) = \frac{1}{2}(1 - Sign(z))$. For classification, a specific loss V rather than L is often used. Note that the minimization of $EV(Yf(X))$, called the *surrogate risk*, over all measurable functions is not feasible in practice. Instead, the minimization of $EV(Yf(X))$ is performed over some class $\mathcal{F}$ with its minimizer f_0. Throughout the thesis, $\mathcal{F}$ is assumed to be a linear space. However, it should be noted that f_0 may not belong to $\mathcal{F}$. As is to be seen in illustrative examples in section 3.5, f_0 may not be close to f_g in terms of the surrogate risk. In this situation, $e_V(f, f_0) = EV(Yf(X)) - EV(Yf_0(X)) \geq 0$, called the *excess surrogate risk over* $\mathcal{F}$, is naturally introduced. Because f_0 is the feasible minimizer of $EV(Yf(X))$, a reasonable measure of performance for a classifier $Sign(f)$ is the *excess risk over* $\mathcal{F}$ defined as $|e(f, f_0)| = |EL(Yf(X)) - EL(Yf_0(X))|$. Consequently, we study the connection between $|e(f, f_0)|$ and $e_V(f, f_0)$ in what follows.

In the literature, the *excess risk* defined as $e(f, \bar{f}) = EL(Yf(X)) - EL(Y\bar{f}(X)) \geq 0$ has been commonly used to measure the performance of a classifier $Sign(f)$. Here $\bar{f} = Sign(f^*)$ is the *Bayes classifier* with $f^*(x) = p^*(x) - 1/2$, obtained by minimizing

$EL(Yf(X))$, called the *misclassification risk*, over all measurable functions where $p^*(x) = P(Y = 1|X = x)$ is the unknown conditional probability of the positive class given $X = x$. Bartlett et al. (2003) established the connection between $e(f, \bar{f})$ and $e_V(f, f_g)$ for convex losses where f_g denotes the minimizer of $EV(Yf(X))$ over all measurable functions. They obtained the rates of convergence in terms of $e(f, \bar{f})$. However, this formulation is appropriate only when the *approximation error* defined by $e_V(f_0, f_g) = EV(Yf_0(X)) - EV(Yf_g(X))$ is zero (or tends to zero for $\mathcal{F}$ depending on n.) The approximation error depends on the surrogate loss V, the size of $\mathcal{F}$, and the underlying distribution. For convex losses, the approximation may not be zero (or tends to zero) in general unless $\mathcal{F}$ is sufficiently large and the underlying distribution belongs to some restricted class of distributions.

The following conditions are used to characterize a general loss:

(C-1) (Behavior of loss) $V(z) < V(-z)$ for all $z > 0$.

(C-2) (Strictly convex loss) V is strictly convex.

(C-3) (Nonstrictly convex loss) $V(z) = [1 - z]_+^q$ for $q \geq 1$.

(C-4) (Nonconvex loss) V is nonincreasing, bounded, continuous on $(-\infty, 1)$, $V(-1) = V(0-) \geq L(0-)$, and $V \equiv 0$ on $[1, \infty)$.

Without loss of generality, we assume that V is nonnegative. Condition (C-1) says that a wrongly classified instance should have a penalty higher than its counterpart yielding a correct classification, which is commonly used; e.g., Lin (2002a), and Bartlett et al. (2003). Condition (C-2) is satisfied by strictly convex losses such as the logistic loss. Condition (C-3) specifies a general hinge loss, which is convex but not strictly convex. Condition (C-4) describes a class of bounded losses, typically nonconvex, and is satisfied by the ψ-losses.

Before proceeding, we need to clarify one important issue, that is, whether the cost function (2.3) estimates $\bar{f}$. For any specific loss V, it is evident that it targets directly at the excess risk when $f_g = \bar{f}$, and it estimates f_g when $f_g \neq \bar{f}$. This results in different types of classifiers behaving dramatically differently, due to the choice of V. The classifiers of the first type, usually strictly convex, correspond to the situation of $f_g \neq \bar{f}$. They yield correct classification when $Sign(f_g) = Sign(f^*)$ a.s. The classifiers of the second type, normally nonconvex, estimate $\bar{f}$ directly. A general hinge loss with $q \geq 2$ belongs to the first type. The hinge loss ($q = 1$) targets at $\bar{f}$, but usually estimates $f_0 \neq \bar{f}$ because the approximation error is not (or tends to) zero in general. In this sense, nonstrictly convex losses do not belong to any of these two types, which makes the situation complicated.

The following assumption, called the *low noise assumption*, is used in establishing the connection; c.f. Bartlett et al. (2003), and Mammen and Tsybakov (1999). Here the low noise assumption is for the class minimizer f_0. For the special case when the approximation error is zero (or tends to zero), this assumption can be imposed on f^* with noise level $0 < \alpha < \infty$ and constant $c_1 > 0$. The parameter β indicates the noise level, and reflects the difficulty of classification with $\beta = +\infty$ corresponding to the easiest classification. Other equivalent conditions of the low noise assumption can be found in Bousquet, Boucheron, and Lugosi (2004a). By restricting the class of distributions via the low noise assumption, fast rates of convergence to the optimal risk can be obtained; c.f., Mammen and Tsybakov (1999) and Shen et al. (2003). Without any assumption on underlying distribution, the best possible rate for any classifiers is $n^{-1/2}$ for nonseparable cases as is suggested in Yang (1998). Our theory in section 3.2 is based on empirical processes. The low noise assumption is also useful in bounding the second moment of an empirical process by some power of its first moment.

Assumption A. There exist some constant $0 < \beta < \infty$ and $c_1^* > 0$ such that $P(x \in \mathcal{X} : |f_0(x)| \leq \delta) \leq c_1^* \delta^\beta$ for sufficiently small $\delta > 0$.

Lemma 1 describes a general situation for the classifiers of the first type, whereas Lemma 2 concerns classifiers of the second type, targeting directly at the excess risk. Also, note that the class minimizer may not belong to the class $\mathcal{F}$. For the classifiers of the first type, it is reasonable to assume that the class minimizer is unique because the surrogate risk is strictly convex. However, for the second type or the hinge loss, the class minimizer may not be unique.

Lemma 1. *Assume that V satisfies (C-1) and (C-2). In addition, f_0 satisfies Asssumption A. If $|f_0| \leq a$ a.s. on $\mathcal{X}$ for some $a > 0$, then*

$$|e(f, f_0)| \leq c_{V^*} e_{V^*}(f, f_0)^{\frac{\beta}{\beta+2}} \tag{3.1}$$

for a positive constant c_{V^*} depending only on V^*, where V^* is a truncated surrogate loss at $0 \leq T_2 < T_1 < \infty$, defined as

$$V^*(z) = \begin{cases} T_1, & \textit{if } V(z) > T_1 \\ V(z), & \textit{if } T_2 \leq V(z) \leq T_1 \\ T_2, & \textit{otherwise.} \end{cases}$$

In Lemma 1, $e(f, f_0)$ may not be nonnegative because f_0 is the class minimizer of the surrogate risk not the misclassification risk, as noted in Bousquet, Boucheron, and Lugosi (2004b). Under an additional assumption of consistency, $e(f, f_0) = e(f, \bar{f}) \geq 0$.

Proposition 1. *In addition to the assumptions of Lemma 1, if* $e(f_0, \bar{f}) = 0$*,*

$$e(f, \bar{f}) \leq c_{V^*} e_{V^*}(f, f_0)^{\frac{\beta}{\beta+2}} \tag{3.2}$$

Lemma 1 yields exponent $\frac{\beta}{\beta+2}$ in (3.2), which leads to a higher generalization error, as compared to exponent 1 in (3.3). The consistency defined in the literature such as Lin (2002a), Zhang (2004a), or Bartlett et al. (2003) is basically Fisher-consistency implying that $e(f_g, \bar{f}) = 0$. This is appropriate only when $e(f_0, f_g) = 0$ because f_0 is the feasible minimizer of $EV(Yf(X))$.

The assumption below states that $e_V(f_0, f_g)$, the approximation error, can be arbitrarily small. With this assumption in place, the generalization error rate is not impeded if the approximation tends to zero sufficiently fast.

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

The following lemma follows from Proposition 1 in Shen et al. (2003) because $V^* = V$ for V satisfying (C-4). For V satisfying (C-3) with $q = 1$, the connection for V can be found in Zhang (2004) and Bartlett et al. (2003).

Lemma 2. *Suppose* V *satisfies either (C-3) with* $q = 1$ *or (C-4). Then,*

$$e(f, \bar{f}) \leq c_{V^*} e_{V^*}(f, \bar{f}) \tag{3.3}$$

for some constant $c_{V^*} > 0$ *depending on a truncated loss* V^**.*

A few remarks are necessary for Lemma 2. Any loss satisfying (C-4) is very close to the misclassification loss in that it yields $f_0 \approx \bar{f}$. However, Assumption B may not be satisfied by the hinge loss satisfying (C-3) with $q = 1$ unless the underlying distribution is restricted in some fashion and $\mathcal{F}$ is sufficiently large. Although the hinge loss targets at $\bar{f}$, it may not satisfy Assumption B. In that case, SVM estimates f_0 instead of $\bar{f}$.

3.2 Upper Bound Theory

We now develop a general theory for these two types of classifiers separately as they behave differently in terms of the excess risk (over $\mathcal{F}$). This, together with the results in section 3.1 and the lower bound result in section 3.3, provides an insight into why classifiers of the second type, defined by certain nonconvex losses, enable to achieve sharper rates of convergence.

Our learning theory is derived via the metric entropy-based complexity measure as well as the characteristics of V. In particular, the theory uses the metric entropy for a class of functions. Note that the metric entropy for sets, which results in sharper rates of convergence, can be applied only to the second type.

For a class of functions $\mathcal{F}$, we define the L_2-metric entropy with bracketing that measures the massiveness of $\mathcal{F}$. Given any $\varepsilon > 0$, the set $\{(f_k^l, f_k^u)\}_{k=1}^{n_c}$ is called an *ε-bracketing function* of $\mathcal{F}$ if for any $f \in \mathcal{F}$, there is a k such that $f_k^l \leq f \leq f_k^u$ and $\|f_k^u - f_k^l\|_2 \leq \varepsilon$ for all $k = 1, \cdots, n_c$ where $\|\cdot\|_2$ is the L_2-norm. The *L_2-metric entropy $H_B(\varepsilon, \mathcal{F})$ of $\mathcal{F}$ with bracketing* is defined as logarithm of the cardinality of ε-bracketing function of $\mathcal{F}$ of the smallest size.

Let $\mathcal{F}(k) = \{f \in \mathcal{F} : J[f] \leq k\} \subset \mathcal{F} = \{f \in \mathcal{F} : J[f] < \infty\}$ and $J_0 = \max\{J[f_0], 1\}$. Denote $\mathcal{F}_V(k) = \{l_V(f, z) - l_V(f_0, z) : f \in \mathcal{F}(k)\}$ where $l_V(f, Z) = V(Yf(X))$ and $Z = (X, Y)$ is an instance. The following assumption on the metric entropy is made.

Assumption C. For some positive constants c_2, c_3, and c_4, there exists some $\bar{\varepsilon}_n > 0$ such that

$$\sup_{k \geq 1} \phi(\bar{\varepsilon}_n, k) \leq c_2 n^{1/2}, \tag{3.4}$$

where $\phi(\bar{\varepsilon}_n, k) = \int_{c_4 D}^{c_3^{1/2} D^{1/2}} H_B^{1/2}(u, \mathcal{F}_V(k)) du/D$ and $D = D(\bar{\varepsilon}_n, C, k) = \min\{\bar{\varepsilon}_n^2 + (Cn)^{-1} J_0(k/2 - 1), 1\}$.

Two aspects govern the performance of a classifier. First, the size of $\mathcal{F}$, described by the metric entropy, determines $\bar{\varepsilon}_n$. Specifically, the smallest $\bar{\varepsilon}_n$ satisfying Assumption C yields the best upper bound of the generalization error rate for a classifier. Second, there is a trade-off between the geometric margin and the training error, which is controlled by the choice of C. The best error rate of a margin classifier is realized when C strikes the balance of the trade-off, which can be summarized in terms of the size of $\mathcal{F}$, n, and C.

Theorem 1 and Corollary 1 provide some probability and risk bounds for strictly convex losses in terms of V^*, a truncated version of V. Since the convergence is in V^*, the metric entropy is for $\mathcal{F}_{V^*}(k) = \{l_{V^*}(f, z) - l_{V^*}(f_0, z) : f \in \mathcal{F}(k)\}$. However, the metric entropy for $\mathcal{F}_{V^*}(k)$ can be replaced by that for $\mathcal{F}(k)$ because $H_B(u, \mathcal{F}_{V^*}(k)) \leq H_B(u, \mathcal{F}(k))$, which is shown in the proof.

Theorem 1. *Assume that V satisfies (C-1) and (C-2). Suppose that Assumption A and C are met with $\phi(\bar{\varepsilon}_n, k)$ defined by $\int_{c_4 D}^{c_3^{1/2} D^{1/2}} H_B^{1/2}(u, \mathcal{F}(k)) du/D$. For any margin classifier $Sign(\hat{f})$ defined in (2.3), there exists a constant $c_5 > 0$ such that*

$$P\left(|e(\hat{f}, f_0)| \geq \delta_n^{\frac{2\beta}{\beta+2}}\right) \leq 3.5 \exp\left(-c_5 n (nC)^{-1} J_0\right)$$

provided that $Cn \geq 2\delta_n^{-2} J_0$ where $\delta_n^2 = \min\{\bar{\varepsilon}_n^2, 1\}$.

Corollary 1. *Under the assumptions of Theorem 1,*

$$|e(\hat{f}, f_0)| = O_p(\delta_n^{\frac{2\beta}{\beta+2}}) \quad \textit{and} \quad E|e(\hat{f}, f_0)| = O(\delta_n^{\frac{2\beta}{\beta+2}}).$$

Theorem 2 and Corollary 2 yield probability and risk bounds for SVM and ψ-learning. The metric entropy for functions is adopted here. This result is useful for ψ-learning when it is difficult to compute the metric entropy for sets although there may be some loss of power in the rate. To simplify the metric entropy, we may assume that ψ-loss satisfies Lipschitz condition :

$$|\psi(z_1) - \psi(z_2)| \leq D|z_1 - z_2|, \tag{3.5}$$

where D is a positive constant. However, (3.5) is irrelevant to Theorem 3 adopting the metric entropy for sets.

Theorem 2. *In addition to Assumptions A-C with* $\phi(\bar{\varepsilon}_n, k)$ *defined by* $\int_{c_4 D}^{c_3^{1/2} D^{\frac{\alpha}{2(\alpha+1)}}} H_B^{1/2}$ $(u^2/2, \mathcal{F}(k))$ du/D, V *satisfies (C-3) with* $q = 1$ *or (C-4). For any classifier* $Sign(\hat{f})$ *defined in (2.3), there exists a constant* $c_5 > 0$ *such that*

$$P\left(e(\hat{f}, \bar{f}) \geq \delta_n^2\right) \leq 3.5 \exp\left(-c_5 n (nC)^{-\frac{\alpha+2}{\alpha+1}} J_0^{\frac{\alpha+2}{\alpha+1}}\right)$$

provided that $Cn \geq 2\delta_n^{-2} J_0$ *where* $\delta_n^2 = \min\{\max\{\bar{\varepsilon}_n^2, 2s_n\}, 1\}$.

Corollary 2. *Under the assumptions of Theorem 2,*

$$|e(\hat{f}, \bar{f})| = O_p(\delta_n^2) \quad \textit{and} \quad E|e(\hat{f}, \bar{f})| = O(\delta_n^2),$$

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

To develop our theory for nonconvex losses, we now define the metric entropy with bracketing for a class $\mathcal{G}$ of sets in $\mathcal{X}$. Given any $\varepsilon > 0$, the set $\{(G_k^l, G_k^u)\}_{k=1}^{n_c}$ is called an *ε-bracketing set* of $\mathcal{G}$ if for any $G \in \mathcal{G}$, there is a k such that $G_k^l \subset G \subset G_k^u$ and $d_\Delta(G_k^u, G_k^l) \leq \varepsilon$ for all $k = 1, \cdots, n_c$ where $d_\Delta(\cdot,\cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$ is a metric defined by $d_\Delta(G_1, G_2) = P(G_1 \Delta G_2)$ and $G_1 \Delta G_2$ is the symmetric difference set of G_1 and G_2. The *metric entropy $H_B(\varepsilon, \mathcal{G})$ of $\mathcal{G}$ with bracketing* is defined as logarithm of the cardinality of ε-bracketing set of $\mathcal{G}$ of the smallest size.

Let $\mathcal{G}(k) = \{G_f = \{x : f(x) \geq 0\} : f \in \mathcal{F}, J[f] \leq k\} \subset \mathcal{G}(\mathcal{F}) = \{G_f = \{x : f(x) \geq 0\} : f \in \mathcal{F}, J[f] < \infty\}$. Here the metric entropy for sets is adopted because the convergence is in terms of L. For the second type, this yields a little sharper rates than Theorem 2 and Corollary 2 using the metric entropy for functions.

Theorem 3. *In addition to Assumptions A-C with $\phi(\bar{\varepsilon}_n, k)$ defined by $\int_{c_4 D}^{c_3^{1/2} D^{\frac{\alpha}{2(\alpha+1)}}} H_B^{1/2}(u^2/2, \mathcal{G}(k))\, du/D$, V satisfies (C-4). For any classifier $Sign(\hat{f})$ defined in (2.3), there exists a constant $c_5 > 0$ such that*

$$P\left(e(\hat{f}, \bar{f}) \geq \delta_n^2\right) \leq 3.5 \exp\left(-c_5 n (nC)^{-\frac{\alpha+2}{\alpha+1}} J_0^{\frac{\alpha+2}{\alpha+1}}\right)$$

provided that $Cn \geq 2\delta_n^{-2} J_0$ where $\delta_n^2 = \min\{\max\{\bar{\varepsilon}_n^2, 2s_n\}, 1\}$.

Corollary 3. *Under the assumptions of Theorem 3,*

$$|e(\hat{f}, \bar{f})| = O_p(\delta_n^2) \quad \textit{and} \quad E|e(\hat{f}, \bar{f})| = O(\delta_n^2),$$

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

Theorem 3 holds for ψ-learning with $U \geq \psi(z) > 0$ for $z \in (0, \tau]$ and $\psi(z) = V(1 - Sign(z))$ for $z \notin (0, \tau]$, where $0 < \tau \leq 1$, $V \geq \max\{U, 1/2\}$ and $U > 0$. Similarly, the smallest $\bar{\varepsilon}_n$ satisfying (3.4) in Assumption C gives the best performance for $\hat{f}$, and C needs to be suitably chosen to yield the best trade-off of the training error and the margin.

Remark 1. The result in Theorems 1-3 continue to hold if the "global" entropy is replaced by a corresponding "local" version; c.f., Van de Geer (1993). That is, $\mathcal{F}(k)$ is replaced by $\mathcal{F}_1(k) = \mathcal{F}(k) \cap \{|e(f, f_0)| \leq 2\varepsilon_n\}$, and $\mathcal{G}(k)$ is replaced by $\mathcal{G}_1(k) = \mathcal{G}(k) \cap \{|e(f, f_0)| \leq 2\varepsilon_n^2\}$. The proof requires only a trivial modification. The local entropy allows us to avoid loss of factor of $\log(n)$ in a linear problem, although it may not be useful for a nonlinear problem.

Remark 2. To illustrate the calculation of upper rates using the upper risk bounds in Corollaries 1-3, suppose $H_B(\epsilon, \mathcal{F}(k)) \leq A_1\epsilon^{-\kappa_1}$ and $H_B(\epsilon, \mathcal{G}(k)) \leq A_2\epsilon^{-\kappa_2}$ for some $0 < \kappa_1, \kappa_2 < 1$ and positive constants A_1 and A_2. Intuitively, the metric entropy for sets may not be larger than the metric entropy for functions, i.e., $\kappa_2 \leq \kappa_1$, because $\mathcal{G}(k)$ is induced by $\mathcal{F}(k)$. In our linear and polynomial kernel examples of section 3.5, $\kappa_2 = \kappa_1$. The metric entropy for sets with smooth boundaries in Dudley (1974) is an example that $\kappa_2 < \kappa_1$. Let $s \in (0, 1]$ be the exponent of the s_n. By solving entropy equations using the metric entropy for functions, we have the rates $n^{-\frac{2\beta}{(\beta+2)(\kappa_1+2)}}$ and $n^{-\frac{\alpha+1}{\kappa_1(\alpha+1)+\alpha+2}}$ for strictly convex losses and nonconvex losses, respectively. For nonconvex losses, using $H_B(\epsilon, \mathcal{G}(k))$ yields the rate $n^{-\frac{\alpha+1}{\kappa_2(\alpha+1)+\alpha+2}}$, which may be faster than $n^{-\frac{\alpha+1}{\kappa_1(\alpha+1)+\alpha+2}}$. Usually, the rates for nonconvex losses are not impeded by the size of approximation error because $s = 1$ in many cases. If Assumption B is satisfied for the hinge loss, the rate for the hinge loss is given by $n^{-\min\{\frac{\alpha+1}{\kappa_1(\alpha+1)+\alpha+2}, s\}}$ where $0 < s < 1$. The rate for SVM is expected to be impeded by the size of approximation error in many cases.

3.3 Lower Bound Theory

This section develops our lower bound theory. Mammen and Tsybakov (1999) obtained a minimax lower bound for a class of boundary fragments with smooth boundaries and Scott and Nowak (2004) studied minimax rates of dyadic decision trees for the box counting class, a natural class for image analysis. However, a general lower bound for classification in terms of the excess risk has not been yet available in the statistics literature, although other types of lower bounds were studied in the machine learning literature. Most relevant work can be found in Yang (1999), where a minimax lower bound for the excess risk is derived using a class of conditional probability densities, yielding rates usually slower than $n^{-1/2}$.

Our formulation uses a class of decision functions $f \in \mathcal{F}$ whose sign yields classification, as opposed to a class of conditional probability densities. As is to be seen, the general lower bound for the excess risk can be faster than $n^{-1/2}$ under the low noise assumption.

To begin, let us first define ε-capacity of a class $\mathcal{G}$ of classification sets in metric d_Δ, which is more natural measure of complexity than the metric entropy in quantifying the lower bound. A finite subclass N of $\mathcal{G}$ is said to be *ε-separated* if

$$\inf_{G_i, G_j \in N; G_i \neq G_j} d_\Delta(G_i, G_j) \geq \varepsilon$$

for $\varepsilon > 0$. The *ε-capacity* $C(\varepsilon, \mathcal{G})$ of $\mathcal{G}$ is defined as the logarithm of the cardinality of the maximal ε-separated set.

Under the assumption that $\mathcal{G}$ is totally bounded, Theorem 4 of Kolmogorov and Tikhomirov (1959), that holds for a class of functions, can be extended to class of sets in metric d_Δ, to yield the following relationship with the metric entropy:

$$C(2\varepsilon, \mathcal{G}) \leq H(\varepsilon, \mathcal{G}) \leq C(\varepsilon, \mathcal{G}), \quad for \;\; any \; \varepsilon > 0, \tag{3.6}$$

where $H(\varepsilon, \mathcal{G})$ is the metric entropy without bracketing, which is no greater than the corresponding entropy with bracketing $H_B(\varepsilon, \mathcal{G})$. This relation says that $C(\varepsilon, \mathcal{G})$ and $H(\varepsilon, \mathcal{G})$ are of the same order in ε.

Let $G_{f^*} = \{x : f^*(x) \geq 0\}$ be the *Bayes classification set*, which may not belong to $\mathcal{G}(\mathcal{F})$. Define a local version of $\mathcal{G}(\mathcal{F})$ as $\mathcal{G}(\varepsilon, \mathcal{F}) = \{G_f : f \in \mathcal{F}, d_\Delta(G_f, G_{f^*}) \leq 2\varepsilon\}$ for any $\varepsilon > 0$.

Theorem 4 below provides a general lower bound for the excess risk, which can be used to compare the upper bound results derived in Theorem 3.

Theorem 4. *Assume that $\mathcal{G}(\varepsilon, \mathcal{F})$ has finite metric entropy. In addition, Assumption A is met with $\underline{\varepsilon}_n$ satisfying*

$$C(\underline{\varepsilon}_n^{\frac{2\alpha}{\alpha+1}}, \mathcal{G}(\underline{\varepsilon}_n, \mathcal{F})) \geq n\underline{\varepsilon}_n^2, \tag{3.7}$$

where α is defined in Assumption A. Then,

$$\sup_{f \in \mathcal{F}} P(e(f, \bar{f}) \geq \underline{\varepsilon}_n^2) \geq \frac{1}{4}.$$

Corollary 4. *If $\bar{\varepsilon}_n = \underline{\varepsilon}_n = \varepsilon_n$,*

$$\sup_{f \in \mathcal{F}} Ee(f, \bar{f}) = O(\varepsilon_n^2).$$

The lower bound theory is formulated on the basis of the local capacity, which allows to cover the result of n^{-1} in the linear case, which is in contrast to the upper bound results.

3.4 Attainment: Upper and Lower Bounds

For the comparison of rates, let us define a few notations. If $\limsup(a_n/b_n) < \infty$, then $a_n \ll b_n$. We will use $a_n \asymp b_n$ if $a_n \ll b_n$ and $a_n \gg b_n$.

Consider the optimal rate (lower bound) $\underline{\varepsilon}_n^2$ defined by the relation:

$$C(\underline{\varepsilon}_n^{\frac{2\alpha}{\alpha+1}}, \mathcal{G}(\underline{\varepsilon}_n, \mathcal{F})) \asymp H(\underline{\varepsilon}_n^{\frac{2\alpha}{\alpha+1}}, \mathcal{G}(\underline{\varepsilon}_n, \mathcal{F})) \asymp n\underline{\varepsilon}_n^2, \tag{3.8}$$

obtained from Theorem 4 and (3.6). In view of the corresponding best attainable upper bound for the second type of classifier, we note that

$$\int_{\bar{\varepsilon}_n^2}^{\bar{\varepsilon}_n^{\frac{\alpha}{\alpha+1}}} H_B^{1/2}(u^2, \mathcal{G}(\bar{\varepsilon}_n, \mathcal{F}))du \asymp n\bar{\varepsilon}_n^2, \tag{3.9}$$

obtained by suitably chosen C ($Cn \sim \delta_n^{-2}$) with $\delta_n^2 = \min\{\max\{\bar{\varepsilon}_n^2, 2s_n\}, 1\}$ as defined in Theorem 3. In contrast to the best attainable upper bound for the first type of classifier, the corresponding rate is obtained by

$$\int_{\bar{\varepsilon}_n^2}^{\bar{\varepsilon}_n} H_B^{1/2}(u, \mathcal{F}_1(\bar{\varepsilon}_n))du \asymp n\bar{\varepsilon}_n^2 \tag{3.10}$$

where $\mathcal{F}_1(\varepsilon) = \mathcal{F} \cap \{|e(f, f_0)| \leq 2\varepsilon\}$ for strictly convex losses. In this case, the rate is suboptimal.

When the size of $\mathcal{G}(\bar{\varepsilon}_n, \mathcal{F})$ is not large, $\int_{\bar{\varepsilon}_n^2}^{\bar{\varepsilon}_n^{\frac{\alpha}{\alpha+1}}} H_B^{1/2}(u^2, \mathcal{G}(\bar{\varepsilon}_n, \mathcal{F}))du \asymp H(\bar{\varepsilon}_n^{\frac{2\alpha}{\alpha+1}}, \mathcal{G}(\underline{\varepsilon}_n, \mathcal{F}))$ in (3.9), $\bar{\varepsilon}_n \asymp \underline{\varepsilon}_n$. This is illustrated in the linear example of section 3.5.1, where the optimal rate $\underline{\varepsilon}_n$ defined by (3.8) is achieved by the second class of nonconvex margin classifiers.

When the size of $\mathcal{G}(\bar{\varepsilon}_n, \mathcal{F})$ is large, (3.8) and (3.9) yield that $\bar{\varepsilon}_n \gg \underline{\varepsilon}_n$. This means that the margin classifiers of any type, as defined in section 3.3, fail to attain the optimal rates of convergence.

Theorem 4 and Corollary 4 say that the classification error rates, as measured by the excess risk, can be as fast as n^{-1}. Furthermore, an application of Theorem 3 to linear example of Shen et. al (2003) showed that the rate n^{-1} is achieved by ψ-learning, which is in fact optimal for both separable and nonseparable cases. More generally, Theorem 4 says that the rates obtained in Theorem 3 are optimal. In view of (3.10), classifiers of the first type converges slower than $n^{-1/2}$ in the nonseparable case.

3.5 Examples

This section applies our general theory to some specific learning examples to derive the rates of convergence in terms of the excess risk (over $\mathcal{F}$), for convex and nonconvex margin classifiers in linear and nonlinear cases. Throughout learning examples in this section, it is assumed that the underlying marginal distribution on $\mathcal{X}$ is uniform.

3.5.1 Linear classification : linear kernel

Consider classification with linear decision functions belonging to $\mathcal{F} = \{f(x) = \langle w, x \rangle + b : w \in \mathbb{R}^d\}$, with the input space $\mathcal{X} = \{(x_1, x_2, \cdots, x_d) : \sum_{j=1}^{d} x_j^2 \leq 1\}$ a unit sphere in $\mathbb{R}^d$. Suppose that $f_t(x) = x_1$ is the true decision function. This example is a generalization of the linear example in Shen et al. (2003) to a d-dimensional feature space.

Suppose that the positive class label $Y = +1$ is assigned if $x_1 \geq 0$ and the negative class label $Y = -1$ is assigned otherwise for any $x \in \mathcal{X}$. Assume that class labels are flipped at random with unknown probability $0 < r < \frac{1}{2}$, so that the Bayes risk is r.

For ψ-learning, we apply Corollary 3. The losses are the same as those in Shen et al. (2003) except for scaling constants. For a choice of $f_0 = nf_t$, $e_V(f_0, \bar{f}) \leq s_n = c_1 n^{-1}$ for some constant $c_1 > 0$. This implies Assumption B. For any sufficiently small $\delta > 0$, $P(x \in \mathcal{X} : |f^*(x)| \leq \delta) = 0$, implying Assumption A with $\alpha = +\infty$. To check Assumption C, let us compute the local metric entropy of $\mathcal{G}_1(k) = \mathcal{G}(k) \cap \{|e(f, f_0)| \leq 2u^2\}$; see Remark 1. Note that $|e(f, f_0)| \leq 2u^2$ implies that

$||w - w_0|| \leq c'u^2$ for some $c' > 0$ where $f_0(x) = \langle w_0, x\rangle$. Moreover, $\min_{1\leq i\leq n} |\langle w - w_0, x_i\rangle| \leq |b - b_0| \leq \max_{1\leq i\leq n} |\langle w - w_0, x_i\rangle|$ since b minimizes $\sum_{i=1}^n V(y_i f(x_i))$ for any given w. Hence $|b - b_0| \leq \|w - w_0\|$ and $\|w\| \leq (2k)^{1/2}$ for any $f \in \mathcal{G}_1(k)$. Then $H_B(u^2, \mathcal{G}_1(k)) \leq O(d \log(\min(k_1^{1/2}, c'u^2)/u^2))$ with $k_1 = (2k + \|w_0\|^2)^{1/2}$. The local metric entropy of $\mathcal{G}_1(k) = \mathcal{G}(k) \cap \{|e(f, f_0)| \leq 2u^2\}$ is given by $H_B(u^2, \mathcal{G}_1(k)) \leq O(d \log(\min(k_1^{1/2}, c'u^2)/u^2))$ with $k_1 = (2k + \|w_0\|^2)^{1/2}$.

Let $\phi(\bar\varepsilon_n, k)$ be $(\log(\min(k_1^{1/2}, c'\bar\varepsilon_n^2)/\bar\varepsilon_n^2))^{1/2}/D^{1/2}$ where $D = \min(\bar\varepsilon_n^2 + (Cn)^{-1}(k/2 - 1), 1)$. Easily, $\sup_{k\geq 1} \phi(\bar\varepsilon_n, k) \leq \phi(\bar\varepsilon_n, 1) = c/\bar\varepsilon_n$ for some constant $c > 0$. Solving (3.4) in Assumption C yields a rate $\bar\varepsilon_n = n^{-1/2}$ when $C/\max(J[f_0], 1)$ is sufficiently large. By Corollary 3, we have $e(\hat f, \bar f) \leq O(n^{-1} \log(1/\delta))$ except for a set of probability less than small $\delta > 0$ and $Ee(\hat f, \bar f) = O(n^{-1})$. To obtain the corresponding lower bound, we need to obtain a lower bound for $C(\varepsilon, \mathcal{G}(\varepsilon, \mathcal{F}))$. Because $\mathcal{G}(\varepsilon, \mathcal{F})$ sits within a d-dimensional cube, then $C(\varepsilon, \mathcal{G}(\varepsilon, \mathcal{F})) \geq c \log\left(\frac{(2\varepsilon)^d}{\varepsilon^d}\right)$ for some constant $c > 0$, which yields a lower bound of $\underline{\varepsilon}_n = n^{-1/2}$ by solving $\underline{\varepsilon}_n$ in (3.7). By Corollary 4, $\sup_{f\in\mathcal{F}} Ee(f, \bar f) = O(n^{-1})$. Consequently, the rate n^{-1} is optimal.

For the logistic loss, let $f_0 = nf_t$. It can be shown that $P(|f_0| \leq \delta) \leq c_1^*\delta$ for some $c_1^* > 0$. This implies Assumption A for f_0 with $\beta = 1$. To apply Corollary 1, we need to compute the metric entropy. By the relation $H_B(2u, \mathcal{F}_1(k)) \leq H_B(u^2, \mathcal{G}_1(k))$, we have $H_B(u, \mathcal{F}_1(k)) \leq O(d \log(\min(4k_1^{1/2}, c'u^2)/u^2))$. Similarly, let $\phi(\bar\varepsilon_n, k)$ be $(\log(\min(4k_1^{1/2},\ c'\bar\varepsilon_n^2)\ /\bar\varepsilon_n^2))^{1/2}\ /D^{1/2}$, where $D = \min(\bar\varepsilon_n^2 + (Cn)^{-1}(k/2 - 1), 1)$ and get $\bar\varepsilon_n = n^{-1/2}$, when $C\ /\max(J[f_0], 1)$ is sufficiently large. By Corollary 1, we have $|e(\hat f, f_0)| \leq O(n^{-1/3}\ \log\ (1/\delta))$ except for a set of probability less than $\delta > 0$ and $E|e(\hat f, f_0)| = O(n^{-1/3})$.

For the hinge loss satisfying (C-3) with $q = 1$, Assumption B is not satisfied. It can be easily checked that the minimizer of $EV(Yf(X))$ over $\mathcal{F}$ is $f_0 = \sqrt{\frac{1-r}{r}} f_t$. Since $e_V(f_0, \bar{f}) = \sqrt{r(1-r)} > 0$, Assumption B is not satisfied. Hence Corollary 2 is not applicable. Obviously, SVM estimates f_0 not $\bar{f}$ in this example. In Proposition 2, the rate $E|e(\hat{f}, f_0)| = O(n^{-1/2})$ is obtained via direct calculations.

3.5.2 Nonlinear classification : polynomial kernel

Let $K(x, y) = (\langle x, y\rangle + 1)^{m_p}$ for $x, y \in \mathcal{X}$ be a polynomial kernel of order m_p, where $\mathcal{X} = \{x \in \mathbb{R} : x_1^1 + \cdots + x_d^2 \leq 1\}$ for an integer $d > 1$. This kernel induces a RKHS $\mathcal{F}$ that consists of all polynomials of order at most m_p. Denote by $\mathcal{F}(k) = \{f \in \mathcal{F} : J[f] = \|f\|_K \leq k^2\}$ and by the corresponding class $\mathcal{G}(k)$ of classification sets induced by $\mathcal{F}(k)$.

Let the conditional densities of X given $Y = \pm 1$ be $\exp(p_i(x)) \,/ \int_{\mathcal{X}} \exp(p_i(x))dx$; $i = 1, 2$ where $p_i \in \mathcal{F}$; $i = 1, 2$ such that p_1, p_2, and $p_1 - p_2$ are polynomials of order at least 1. Suppose that $\frac{\pi_1}{\pi_2} = \frac{\int_{\mathcal{X}} \exp(p_1(x))dx}{\int_{\mathcal{X}} \exp(p_2(x))dx}$ where π_i; $i = 1, 2$ are the prior probabilities on $\mathcal{A}_\pm$, respectively. By Bayes' Theorem, $p^* = \dfrac{\exp(p_1 - p_2)}{1 + \exp(p_1 - p_2)}$. Denote true decision function by $f_t = p_1 - p_2$.

Consider ψ-learning with a class of ψ losses. For a choice of $f_0 = nf_t$, $e_V(f_0, \bar{f}) \leq s_n = c_1 n^{-1}$ for some constant $c_1 > 0$. This implies Assumption B. For any sufficiently small $\delta > 0$,

$$\begin{aligned} P(x \in \mathcal{X} : |f^*(x)| \leq \delta) &= P\left(x \in \mathcal{X} : |f_t(x)| \leq \log\left(\frac{1 + 2\delta}{1 - 2\delta}\right)\right) \\ &\leq P\left(x \in \mathcal{X} : |f_t(x)| \leq c\delta\right) \end{aligned}$$

for some $c > 0$ using Taylor series expansion. Because f_t is a non-constant polynomial, $P(|f_t| \leq c\delta) \leq c_1\delta$ for some $c_1 > 0$. This implies Assumption A with $\alpha = 1$. By

Lemma 4, the metric entropy for $\mathcal{G}(k)$ is $H_B(u, \mathcal{G}(u)) \leq O(\log(k/u))$. For Assumption C, let $\phi(\bar{\varepsilon}_n, k)$ be $(\log(k/\bar{\varepsilon}_n^2))^{1/2}/D^{3/4}$ where $D = \min(\bar{\varepsilon}_n^2 + (Cn)^{-1}(k/2-1), 1)$. Easily, $\sup_{k \geq 1} \phi(\bar{\varepsilon}_n, k) \leq \phi(\bar{\varepsilon}_n, 1) = \dfrac{c}{\bar{\varepsilon}_n^{3/2}} \left(\log \dfrac{1}{\bar{\varepsilon}_n^2}\right)^{1/2}$ for some constant $c > 0$. Solving (3.4) in Assumption C yields a rate $\bar{\varepsilon}_n = n^{-1/3}(\log n)^{1/3}$ when $C/\max(J[f_0], 1)$ is sufficiently large. By Corollary 3, we have $e(\hat{f}, \bar{f}) \leq O(n^{-2/3}(\log n)^{2/3} \log(1/\delta))$ except for a set of probability less than small $\delta > 0$ and $Ee(\hat{f}, \bar{f}) = O(n^{-2/3}\ (\log n)^{2/3})$.

Now let us consider the logistic loss. Since $f_g = f_t/2 \in \mathcal{F}$, let $f_0 = f_t/2$. Then Assumption A is satisfied for f_0 with $\beta = 1$ and $c_1^* = 2c$ because $P(x \in \mathcal{X} : |f_0(x)| \leq \delta) \leq 2c_1\delta$ for some $c_1 > 0$. Note that $H_B(u, \mathcal{F}(k)) \leq O(\log(2k/u))$. Similarly, let $\phi(\bar{\varepsilon}_n, k)$ be $(\log(2k/\bar{\varepsilon}_n))^{1/2}/D^{1/2}$, where $D = \min(\bar{\varepsilon}_n^2 + (Cn)^{-1}\ (k/2-1), 1)$. Then $\bar{\varepsilon}_n = n^{-1/2}(\log n)^{1/2}$, when $C/\max(J[f_0], 1)$ is sufficiently large. By Corollary 1, we have $|e(\hat{f}, f_0)| \leq O(n^{-1/3}(\log n)^{1/3}\ \log(1/\delta))$ except for a set of probability less than small $\delta > 0$ and $E|e(\hat{f}, f_0)| = O(n^{-1/3}(\log n)^{1/3})$.

For the hinge loss satisfying (C-3) with $q = 1$, Assumption B is not satisfied because $\bar{f}$ has a jump discontinuity at the decision boundary and $\mathcal{F}$ is the class of polynomial of fixed degree m_p. Hence, Corollary 2 is not applicable.

3.5.3 Nonlinear classification : gaussian kernel

Let $\mathcal{X}$ be the unit sphere in $\mathbb{R}^d$. Consider a Gaussian kernel defined as

$$K(x, y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right).$$

Let $\mathcal{F}$ be the RKHS induced by this kernel. Define $\mathcal{F}(k) = \{f \in \mathcal{F} : J[f] = \|f\|_K \leq k^2\}$. The metric entropy of $\mathcal{F}(k)$ in sup-norm is given by $H(\epsilon, \mathcal{F}(k)) \leq O((\log \frac{k}{\epsilon})^{d+1})$ by (4.8) of Zhou (2002). It is easy to show that $H_B(2\epsilon, \mathcal{F}(k)) \leq O((\log \frac{k}{\epsilon})^{d+1})$.

Assume that the underlying joint distribution $P(\cdot, \cdot)$ of (X, Y) is the mixture distribution of two normal distributions with mean vector μ_i; $i = 1, 2$ and covariance matrix $\sigma^2 I$, where $\mu_1 = (+1, 0, \cdots, 0)'$ and $\mu_2 = (-1, 0, \cdots, 0)'$. Let $\theta \in (0, 1)$ be the mixing parameter such that $\left|\log\left(\frac{\theta}{1-\theta}\right)\right| < \frac{2}{\sigma^2}$. By Bayes' Theorem, $p^* = \dfrac{\theta \exp(-2x_1/\sigma^2)}{1 - \theta + \theta \exp(-2x_1/\sigma^2)}$. Denote the true decision function as $f_t(x) = x_1 - x_1^*$ where $x_1^* = \frac{\sigma^2}{2} \ln\left(\frac{\theta}{1-\theta}\right)$.

First, consider ψ-learning with a class of ψ losses. For a choice of $f_0 = n f_t$, $e_V(f_0, \bar{f}) \leq s_n = c_1 n^{-1}$ for some constant $c_1 > 0$. This implies Assumption B. For any sufficiently small $\delta > 0$ and some $c_1 > 0$,

$$\begin{aligned} P(x \in \mathcal{X} : |f^*(x)| \leq \delta) &\leq P\left(x \in \mathcal{X} : \frac{\sigma^2}{2} |x_1 - x_1^*| \leq \ln\left(\frac{1+2\delta}{1-2\delta}\right)\right) \\ &\leq P\left(x \in \mathcal{X} : |x_1 - x_1^*| \leq c\delta\right) \leq c\delta \end{aligned}$$

for some generic $c > 0$ using Taylor series expansion. This implies Assumption A with $\alpha = 1$. For Assumption C, let $\phi(\bar{\varepsilon}_n, k)$ be $(\log(k/\bar{\varepsilon}_n^2))^{\frac{d+1}{2}}/D^{3/4}$ where $D = \min(\bar{\varepsilon}_n^2 + (Cn)^{-1}(k/2-1), 1)$. Easily, $\sup_{k \geq 1} \phi(\bar{\varepsilon}_n, k) \leq \phi(\bar{\varepsilon}_n, 1) = \frac{c}{\bar{\varepsilon}_n^{2/3}} \left(\log \frac{1}{\bar{\varepsilon}_n^2}\right)^{\frac{d+1}{2}}$ for some constant $c > 0$. Solving (3.4) in Assumption C yields a rate $\bar{\varepsilon}_n = n^{-1/3} (\log n)^{(d+1)/3}$ when $C/\max(J[f_0], 1)$ is sufficiently large. By Corollary 2, we have $e(\hat{f}, \bar{f}) \leq O(n^{-2/3}$ $(\log n)^{2(d+1)/3}$ $\log(1/\delta))$ except for a set of probability less than small $\delta > 0$ and $Ee(\hat{f}, \bar{f}) = O(n^{-2/3} (\log n)^{2(d+1)/3})$.

For the hinge loss satisfying (C-3) with $q = 1$, let $f_0 \in \mathcal{F}$ with $\|f_0 - f_1\|_\infty = \inf_{f \in \mathcal{F}} \|f - f_1\|_\infty$ and $|f_0| \leq 1$ where $f_1 = \tanh(n f_t) \in C^\infty$. Assumption B is met with f_0. Because $|f_0| \leq 1$, we have $e_V(f_0, \bar{f}) \leq E|V(Y f_0(X)) - V(Y \bar{f}(X))| = E|f_0(X) - \bar{f}(X)| \leq s_n = c_1 n^{-s}$ for some $0 < s < 1$. Although the size of the approximation error is not available, it is expected to impede the rate for the hinge loss. Hence the rate may be slower than $n^{-2/3}$ $(\log n)^{2(d+1)/3}$. In this example, we applied Corollary 2 instead of Corollary 3 for ψ-learning because it may not be easy to compute $H_B(u, \mathcal{G}(k))$. A tight upper bound of the metric entropy for sets may eliminate the extra power of $\log n$ factor in the rate for ψ-learning.

Let us consider any convex loss satisfying (C-1) and (C-2). For KLR, it is easily shown that $f_g = \frac{1}{\sigma^2} f_t \in \mathcal{F}$. Let $f_0 = \frac{1}{\sigma^2} f_t$. It can be checked that Assumption A is satisfied for f_0 with $\beta = 1$ because $P(x \in \mathcal{X} : |f_0(x)| \leq \delta) \leq c\delta$ for some $c > 0$. Note that $H_B(u, \mathcal{F}_{V^*}(k)) \leq O((\log(2k/u))^{d+1})$. Similarly, let $\phi(\bar{\varepsilon}_n, k)$ be $(\log(2k/\bar{\varepsilon}_n^2))^{(d+1)/2} / \ D^{1/2}$, where $D = \min(\bar{\varepsilon}_n^2 + (Cn)^{-1}(k/2 - 1), 1)$. Then $\bar{\varepsilon}_n = n^{-1/2}$ $(\log n)^{(d+1)/2}$, when $C / \max(J[f_0], 1)$ is sufficiently large. By Corollary 1, we have $|e(\hat{f}, f_0)| \leq O(n^{-1/3} \ (\log n)^{(d+1)/3} \ \log(1/\delta))$ except for a set of probability less than small $\delta > 0$ and $E|e(\hat{f}, f_0)| = O(n^{-1/3} \ (\log n)^{(d+1)/3})$.

It may be worthwhile to mention that Scovel and Steinwart (2004) obtained fast rates for SVM using Gaussian kernels. In addition to the low noise assumption, they imposed some kind of restriction, so-called the geometric noise condition, on the underlying distribution. The geometric noise condition describes the concentration of $|2p^* - 1| dP_X$ near the decision boundary where P_X denotes the marginal distribution of X. Their conditions seem to imply our Assumption B.

CHAPTER 4

NUMERICAL EXAMPLES

4.1 Optimization

4.1.1 Support vector machine

The solution of (2.2) has a finite dimensional representation $f(x) = g(x)+b$, where $g(x) = \sum_{i=1}^{n} \alpha_i K(x, x_i)$ by the property of RKHS (Wahba, 1990). The cost function (2.2) can be minimized via a constrained quadratic minimization of

$$\|g\|_K^2/2 + C\sum_{i=1}^{n} \xi_i \tag{4.1}$$

subject to the constraints

$$\xi_i \geq 1 - y_i f(x_i), \text{ and } \xi_i \geq 0; i = 1, \cdots, n.$$

The coefficients ζ_i's are determined by its dual problem

$$W(\zeta) = \frac{1}{2}\sum_{1\leq i,j\leq n} \zeta_i\zeta_j y_i y_j K(x_i, x_j) - \sum_{i=1}^{n} \zeta_i \tag{4.2}$$

subject to

$$0 \leq \zeta_i \leq C, i = 1, \cdots, n, \qquad \sum_{i=1}^{n} \zeta_i y_i = 0 \tag{4.3}$$

where $\zeta = (\zeta_1, \cdots, \zeta_n)$. By minimizing (4.2) subject to (4.3), the solution $\hat{\zeta}$ of the dual problem is obtained. Then the solution of α_i's are determined by $\hat{\alpha}_i = y_i\hat{\zeta}_i/C$;

$i = 1, \cdots, n$. The solution of b is determined by the instances with $0 < \hat{\alpha}_i < 1$, called the *support vectors*. By Karush-Kuhn-Tucker (KKT) conditions,

$$\hat{b} = \frac{\sum_{i=1}^n \hat{\zeta}_i(1-\hat{\zeta}_i)(y_i - \sum_{j=1}^n \hat{\zeta}_j K(x_i, x_j))}{\sum_{i=1}^n \hat{\zeta}_i(1-\hat{\zeta}_i)}.$$

For linear classification, $K(x_i, x_j)$ is replaced by $\langle x_i, x_j \rangle$. For reference, see Vapnik (1995, 1998), and Cristianini and Shawe-Taylor (2000).

4.1.2 Kernel logistic regression

KLR can be solved using the Newton-Raphson method. However, one drawback of the Newton-Raphson method is that it involves inversion of $n \times n$ matrix at each iteration. Zhu and Hastie (2005) suggested an algorithm, called import vector machine (IVM), that finds a submodel $f(x) = \sum_{x_i \in S} \alpha_i K(x, x_i)$ approximating the full model $f(x) = \sum_{i=1}^n \alpha_i K(x, x_i)$ where S is a subset of the training data $\{x_i\}_{i=1}^n$ and the instances in S are called the *import vectors*.

The import vector machine (IVM) algorithm of Zhu and Hastie (2005) can be described as follows:

1. For $k = 1$, start with $S = \phi$ and $L = \{x_i\}_{i=1}^n$.

2. For $x_l \in L$, set $f_l(x) = \sum_{x_i \in S \cup \{x_l\}} \alpha_i K(x, x_i)$. Find α minimizing the cost function $s(x_l) = \frac{1}{2}\|f_l\|_K^2 + C\sum_{i=1}^n \ln(1 + \exp(-y_i f_l(x_i)))$ and set $k = |S| + 1$.

3. Let $x_{l^*} = \arg\min_{x_l \in L} s(x_l)$. Set $S = S \cup \{x_{l^*}\}$, $L = L - \{x_{l^*}\}$, $s^{(k)} = s(x_{l^*})$, and $k = k + 1$.

4. Iterate steps 2 and 3 until $s^{(k)}$ converges.

They also proposed the revised algorithm. Basically, it is a further approximation to reduce the calculation in step 2 of the basic algorithm, where the Newton-Raphson method is adopted to find α. For the discussion of other issues in KLR, such as stopping rule for adding points to S and the choice of tuning parameter, see Zhu and Hastie (2005).

4.1.3 ψ-learning

The theory in Chapter 3 indicates that ψ losses achieve faster rates of convergence than convex losses in terms of generalization error. Although the optimization of ψ-learning is nonconvex, we can deal with the optimization problem via a global optimization technique, called difference convex (DC) programming. DC programming can solve the optimization problem using a sequential quadratic programming (SQP) if a cost function has a DC representation (An and Tao, 1997).

For the computation of ψ-learning, two computational strategies, SQP and SQP with the method of Branch-and-Bound (SQP-BB), were proposed in Liu, Shen, and Wong (2004). In this chapter, ψ learning is implemented by using SQP. The advantages of SQP are (i) it is simple to implement (ii) it yields reasonably good result in a few iterations of quadratic program (usually 4 or 5 iterations). The other method, SQP-BB, is more computationally intensive. However, this method seems to be able to produce global optima whereas SQP usually yields suboptimal local minima. Here we describe only the first method based on the simplified Differenced Convex algorithm (DCA).

For presentational convenience, consider a linear classification problem where a decision function f is a hyperplane defined by $f(\tilde{x}) = \langle \tilde{w}, \tilde{x} \rangle$ where $\tilde{w} = (w_1, \cdots, w_d, b)$ $\in \mathbb{R}^{d+1}$ and $\tilde{x} = (x_1, \cdots, x_d, 1) \in \mathcal{X} \times \{1\}$.

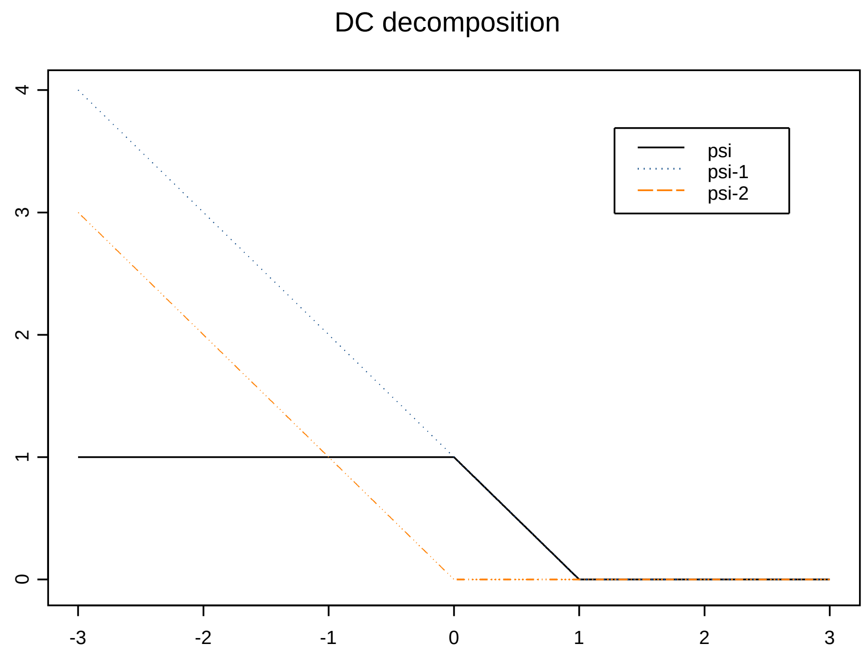

Figure 4.1: DC decomposition of ψ loss into ψ_1 and ψ_2.

A ψ-loss $\psi(z)$, defined by

$$\psi(z) = \begin{cases} 1, & \text{if } z < 0 \\ 1 - z, & \text{if } 0 \le z \le 1 \\ 0, & \text{otherwise,} \end{cases}$$

has a DC decomposition of $\psi(z) = \psi_1(z) - \psi_2(z)$ where $\psi_1(z) = [1 - z]_+$ and $\psi_2(z) = [z]_+$. Figure 4.1 shows the DC decomposition of this ψ-loss.

Let s be the cost function (2.3) with ψ as the surrogate loss. Then it can be decomposed into

$$s(\tilde{w}) = s_1(\tilde{w}) - s_2(\tilde{w}) \tag{4.4}$$

where $s_1(\tilde{w}) = \frac{1}{2}\|\tilde{w}\|^2 + C\sum_{i=1}^{n} \psi_1(y_i f(\tilde{x}_i))$ and $s_2(\tilde{w}) = C\sum_{i=1}^{n} \psi_2(y_i f(\tilde{x}_i))$.

Since a DC decomposition is available, we can construct nonincreasing upper envelopes of s. DCA solves this problem by a series of primal and dual subproblems by constructing $(\tilde{w}^{(k)}, y^{(k)})$ iteratively; Given $(\tilde{w}^{(k)}, y^{(k)})$, the k-th primal subproblem, $s_1(\tilde{w}) - s_2(\tilde{w}^{(k)}) - \langle \tilde{w} - \tilde{w}^{(k)}, y^{(k)} \rangle$, is obtained by replacing $s_2(\tilde{w})$ in (4.4) with its affine minimizer $s_2(\tilde{w}^{(k)}) + \langle \tilde{w} - \tilde{w}^{(k)}, y^{(k)} \rangle$. Then $\tilde{w}^{(k+1)}$ is obtained by minimizing the k-th primal subproblem with respect to $\tilde{w}$. In a similar fashion, we may obtain $y^{(k+1)}$ through the minimization of the k-th dual subproblem after obtaining $\tilde{w}^{(k+1)}$. This corresponds to select an appropriate subgradient of s_2 at $\tilde{w}^{(k)}$. The subgradient, $\nabla s_2(\tilde{w}^k)$, is given by $(v_1^{(k)}, v_2^{(k)})$ where $v_1^{(k)} = C \sum_{i=1}^n \nabla \psi_2(y_i f^{(k)}(\tilde{x}_i)) y_i x_i$ and $v_2^{(k)} = C \sum_{i=1}^n \nabla \psi_2(y_i f^{(k)}(\tilde{x}_i)) y_i$.

Hence we need to solve the primal subproblem

$$\min_{\tilde{w}} s_1(\tilde{w}) - \langle \tilde{w}, \nabla s_2(\tilde{w}^{(k)}) \rangle$$

at each iteration k via quadratic programming. By KKT's condition, it is equivalent to the dual subproblem

$$\max_{\zeta} W(\zeta) = \sum_{i=1}^{n} \zeta_i (1 - y_i \langle v_1^{(k)}, x_i \rangle) - \frac{1}{2} \sum_{i=1}^{j} \sum_{j=1}^{j} \zeta_i \xi_j y_i y_j \langle x_i, x_j \rangle, \qquad (4.5)$$

subject to $\sum_{i=1}^n \zeta_i y_i = -v_2^{(k)}$, $0 \leq \zeta_i \leq 2C$ for $i = 1, \cdots, n$ where $\zeta = (\zeta_1, \cdots, \zeta_n)$. Then the solution of the primal subproblem is given by $\tilde{w}^{(k+1)} = v_1^{(k)} + \sum_{i=1}^n \zeta_i^{(k)} y_i x_i$ where $\zeta^{(k)}$ is the solution of the dual subproblem. Here $\tilde{w}^{(k+1)}$ satisfies KKT's condition : $y_i \langle \tilde{w}^{(k+1)}, \tilde{x}_i \rangle = 1$ for any i such that $0 < \zeta_i^{(k)} < 2C$.

The following is the algorithm for SQP; Given initial value $w^{(0)}$ and tolerance $\epsilon_{tol} > 0$, we compute $w^{(k+1)}$ for each k by solving (4.5). If $|s(\tilde{w}^{(k+1)}) - s(\tilde{w}^{(k)})| \leq \epsilon_{tol}$, then stop the iteration. The final solution $\hat{w} = (w_1^{(k+1)}, \cdots, w_d^{(k+1)})$ and $\hat{b} = w_{d+1}^{(k+1)}$ yield $\hat{f}(x) = \langle \hat{w}, \tilde{x} \rangle + \hat{b}$.

For a nonlinear classification, $\langle x_i, x_j \rangle$ and $\langle v_1^{(k)}, x_i \rangle$ are replaced by $K(x_i, x_j)$ and $C \sum_{j=1}^n \nabla\psi_2(y_j f^{(k)}(\tilde{x}_j)) y_j K(x_i, x_j)$ in (4.5). The solution $\zeta^{(k)}$ yields the solution for the primal subproblem given by $\tilde{w}_j^{(k+1)} = y_j(\zeta_j^{(k)} + C\nabla\psi_2(y_j f^{(k)}(\tilde{x}_j)))$ for $j = 1, \cdots, n$. Here $\tilde{w}^{(k+1)}$ satisfies KKT's condition : $y_i \langle \tilde{w}^{(k+1)}, \tilde{x}_i \rangle = 1$ for any i such that $0 < \zeta_i^{(k)} < 2C$.

4.2 Simulated data

In this section, we compared the performance of KLR, SVM, and ψ-learning in terms of the excess risk over $\mathcal{F}$ for the linear example in 3.5.1. Linear classification in $\mathbb{R}^2$ uses decision functions $f(x) = w_1x_1 + w_2x_2 + b$. Data generation scheme for training sample $(X_{1,i}, X_{2,i}, Y_i)_{i=1}^n$ is as follows: First, $(X_{1,i}, X_{2,i})_{i=1}^n$ are generated according to the uniform distribution on $\mathcal{X} = \{(x_1, x_2) \in \mathbb{R}^2 : x_1^2 + x_2^2 \leq 1\}$. Since $f_t(x) = x_1$, the class label is assigned by $Y_i = Sign(X_{1,i})$ for $i = 1, \cdots, n$. Each Y_i's are randomly flipped with probability r. This results in a nonseparable case with the Bayes risk r.

KLR, SVM, and ψ-learning were compared for three levels of r: .05, .10, and .20. Grid search on $[10^{-5}, 10^5]$ was used to determine the optimal value of the tuning parameter C for each classification method. The following is the expression of the excess risk from Shen et al. (2003). Let $\theta_1 = \Psi - \cos^{-1}(|b|/\|w\|)$, $\theta_2 = \Psi + \cos^{-1}(|b|/\|w\|)$, and $\Psi = |w_1|/\|w\|$. Because $e(f_0, \bar{f}) = 0$ in this example, $|e(f, f_0)| = e(f, \bar{f}) = (1-r)A$ where $2\pi A$ is the area between $f_t(x)$ and $f(x)$ on $\mathcal{X}$ and

$$A = \begin{cases} (|\theta_1 - \pi/2|/2 + |\theta_2 - \pi/2|/2 + |b/w_2|(|\cos(\theta_1)| - |\cos(\theta_2)|)/(2\pi) & \text{if } w_2 \neq 0, \\ (\pi/2 - (\theta_2 - \theta_1 - \sin(\theta_2 - \theta_1))/2)/\pi & \text{otherwise.} \end{cases}$$

The average of the excess risk with its standard error in parenthesis for KLR, SVM, and ψ-learning are summarized in Table 4.1 for sample sizes $n = 50, 100$, and

		r		
n	classifier	.05	.10	.20
50	KLR	.0283 (.0262)	.0324 (.0240)	.0352 (.0259)
	SVM	.0241 (.0192)	.0264 (.0209)	.0301 (.0251)
	ψ	.0195 (.0162)	.0206 (.0205)	.0232 (.0224)
100	KLR	.0175 (.0119)	.0238 (.0174)	.0259 (.0211)
	SVM	.0163 (.0127)	.0224 (.0166)	.0232 (.0183)
	ψ	.0118 (.0103)	.0145 (.0137)	.0196 (.0185)
200	KLR	.0128 (.0095)	.0169 (.0118)	.0161 (.0124)
	SVM	.0123 (.0095)	.0157 (.0112)	.0145 (.0103)
	ψ	.0074 (.0070)	.0081 (.0081)	.0098 (.0090)

Table 4.1: Average of the excess risk and the standard error in parenthesis for KLR, SVM, and ψ-learning over 100 simulation replications for $n = 50, 100$, and 200 with r = .05, .10, and .20.

200. The same seed was used for each n in generating random numbers. It seems that ψ-learning outperforms KLR and SVM in terms of the excess risk for each level of r as n increases. This is consistent with the linear example in section 3.5.1. The error rate of SVM appears to be slightly better than that of KLR. This simulation study confirms that ψ-learning has better generalization accuracy than classifiers with convex losses if the sample size is reasonably large.

4.3 Benchmark data

We analyzed the following data sets from UC Irvine Machine Learning Repository to compare the performance of KLR, SVM, and ψ-learning. Data sets and their information are available at `http://www.ics.uci.edu/~mlearn/MLSummary.html`.

- Shuttle landing control database

  The database was used to generate rules for determining when an auto landing is better than a manual landing of a space shuttle. This database consists of 7 attributes including the class attribute denoting auto and manual landing. The other attributes are stability, error, sign, wind, magnitude, and visibility. Among 253 observations, 125 and 128 cases are automatic and manual landing, respectively.

- Statlog heart disease database

  Statlog heart disease database has 13 attributes: age, sex, chest pain type (4 values), resting blood pressure, serum cholestoral in mg/dl, fasting blood sugar $>$ 120 mg/dl, resting electrocardiographic results (values 0,1,2), maximum heart rate achieved, exercise induced angina, oldpeak (ST depression induced by exercise relative to rest), the slope of the peak exercise ST segment, number of major vessels (0-3) colored by flourosopy, thal (3, 6, 7 denote normal, fixed defect, and reversible defect, respectively.) These attributes are used to predict absence or presence of heart disease. In this database, there are 270 observations; 150 patients do not have heart disease and 120 patients have heart disease.

- Promoter gene sequences database

  The number of instances is 106 with 53 positive and 53 negative instances. The database has 59 numerical attributes: class attribute (positive or negative), instance name (non-promoters named by position in the 1500-long nucleotide sequence), and 57 sequential nucleotide ("base-pair") positions.

The performance of KLR, SVM, and ψ learning was compared using a linear kernel in terms of testing error. To determine the optimal values of tuning parameter for each classification method, grid search on $[10^{-5}, 10^{5}]$ was adopted. For each

data set	classifier	training error	testing error	(n,d)
shuttle landing	KLR	.0687 (.0152)	.0658 (.0151)	(253,6)
	SVM	.0686 (.0150)	.0658 (.0151)	
	ψ	.0686 (.0150)	.0658 (.0151)	
statlog heart disease	KLR	.1410 (.0230)	.1630 (.0232)	(270,13)
	SVM	.1472 (.0241)	.1614 (.0229)	
	ψ	.1479 (.0266)	.1587 (.0259)	
promoter gene	KLR	.0545 (.0242)	.2208 (.0470)	(106,57)
	SVM	.0577 (.0276)	.2194 (.0428)	
	ψ	.0049 (.0190)	.0121 (.0444)	

Table 4.2: Average of training and testing errors, and the standard error in parenthesis for KLR, SVM, and ψ-learning over 100 random partitions.

databases, about half and half of data were used as training and testing data. For each data set, the same partitions of training vs testing samples were used for these classification methods. Table 4.2 shows the average of training and testing errors with their standard errors in parentheses for KLR, SVM, and ψ-learning over 100 random partitions of training and testing data. For shuttle landing database, KLR, SVM, and ψ-learning show almost same performance in terms of testing error. The situation is somewhat different in statlog heart disease example. In terms of testing error, SVM seems to perform slightly better than KLR, and ψ-learning performs slightly better than SVM. However, relative gain of ψ-learing is not clear for these two examples. For promoter gene sequence database, the relative gain of ψ-learning appears to be strikingly higher than the other classifiers. However, it should be noted that our analysis was conducted using a linear kernel. For nonlinear kernels, situation may be different as our Gaussian kernel example in section 3.5.3 indicates.

Our result suggests that the ratio of n and d may have some effect in the performance of classification methods. The gain in the performance for ψ and SVM may increase as the dimension becomes large, implying that these classification methods may be effective for microarray data. Theory in Chapter 3 can deal with the situation of microarray data where $d \gg n$ by computing metric entropy more accurately as a function of n as well as d. Another situation where $d \to \infty$ and $n \to \infty$ is also interesting in both practice and theory.

CHAPTER 5

DISCUSSION

We have compared the generalization accuracy of margin-based classification methods with general losses, convex and nonconvex, in terms of the excess risk (over $\mathcal{F}$). Nonconvex losses such as ψ are shown to yield faster rates of convergence than convex losses such as SVM and KLR, in both theory and numerical examples. Our results suggest that ψ is recommendable when one is interested only in classification while KLR seems to be more natural for obtaining probability estimates.

To make our contribution clear, let us compare our result with other relevant recent work. Bartlett, Jordan, and McAuliffe (2003) obtained the rates of convergence to the Bayes risk for convex losses. They adopted the low noise assumption as well as the one that the margin is uniformly bounded, or equivalently, the function class is uniformly bounded. As noted in Chapter 3, their formulation using the global minimizer is appropriate only when the approximation error is (or tends to) zero. Also, their result may not be applicable to SVM. Our formulation is somewhat different in that we use the class minimizer instead of the global minimizer. Because f_0 is the feasible minimizer in practice, our formulation seems to be more appropriate. Another advantage is that we may avoid the approximation error under our formulation.

In Bartlett, Bousquet, and Mendelson (2004), they obtained data dependent bounds using local Rademacher averages. Data dependent bounds have the advantage that no assumption on underlying distribution, for example, the low noise assumption, is necessary. However, they assumed that the second moment of the empirical process is bounded by some constant times its first moment. Because the rates of convergence are determined by the first and second moments, this assumption may be a restrictive one. Furthermore, the low noise assumption seems to be generally accepted in statistical learning theory.

Scovel and Steinwart (2004) and Blanchard, Bousquet, and Massart (2005) studied the rates for SVM using Gaussian kernels. The difference of these two studies is the penalty term in the cost function. Blanchard et al. (2005) adopted L_1 penalty in their penalized cost function and obtained fast rates under the low noise assumption with $\alpha = \infty$. Because of the difference in penalty term, it may not be relevant to compare their result with ours. Moreover, their result for SVM using Gaussian kernels can be applicable only for the special when the noise level $\alpha = \infty$. In Scovel and Steinwart (2004), they adopted the low noise assumption as well as the geometric noise assumption on the underlying distribution. As we pointed out, to make the approximation error negligible, the class of functions should be sufficiently large and the underlying distribution should be restricted in some fashion. The geometric noise assumption may be regarded as a condition on the underlying distribution such that an upper bound of the approximation error tends to zero.

In this thesis, there are some issues yet to be resolved. First, the rate of convergence for SVM is not available when the approximation error does not tend to zero.

It is expected that the approximation error may not be negligible when finite dimensional kernels are used. Even when infinite dimensional kernels are adopted, it may not be negligible depending on the size of $\mathcal{F}$ and the underlying distribution. To deal with general situation, the connection between the excess and the excess surrogate risk using the class minimizer should be established under mild conditions on the underlying distribution. For this task, the low noise assumption, which enables us to obtain fast rates by restricting the class of underlying distributions, may not be sufficient. It may be necessary to restrict the class of underlying distributions further by imposing some mild conditions.

Second, the choice of penalty term is limited to L_2 penalties here. However, our theory may be extended to cover other types of penalties such as L_1 penalties. For L_1 type penalties, the metric entropy should be the L_1 norm, which yields rougher entropy bounds than L_2 norm. This may result in slower rates of convergence.

Third, our lower bound needs to be sharpened. Unless $\alpha = \infty$, the lower bound may not be sharp. It appears that there is some loss of power in lower rates during the conversion of the pseudo-metric for sets into that for functions. This requires further investigation.

Aside from the issues mentioned above, there ia a gap in statistical learning theory for multicategory classification. Zhang (2004b) extended his study on consistency for margin-based classifiers to multicategory classification. Liu and Shen (2004) obtained rates of convergence for multicategory ψ-learning. To our knowledge, these are the only available theoretical results for multicategory classification. We expect that our study in binary classification can give some insight into multicategory classification.

The key is to understand the characteristics of multicategory classification that differ from the binary situation.

APPENDIX A

PROOFS

Define f_m, the truncated f at m_1 and m_2, as

$$f_m = \begin{cases} m_1, & \text{if } f \le m_1 \\ f, & \text{if } m_1 < f < m_2 \\ m_2, & \text{otherwise.} \end{cases}$$

where $m = (m_1, m_2)$ and $m_1 < 0 < m_2$. Let $\mathcal{F}_m = \{f_m : f \in \mathcal{F}\}$ be the class of truncated functions. The truncation constants can be chosen so that V^*-risk with respect to $\mathcal{F}$ is equivalent to V-risk with respect to $\mathcal{F}_m$, i.e., $EV^*(Yf(X)) = EV(Yf_m(X))$ for $f \in \mathcal{F}$ and $f_m \in \mathcal{F}_m$. Using the equivalence, we may work on derivatives on V over $\mathcal{F}_m$ instead of V over $\mathcal{F}$.

Proof of Lemma 1: We apply a truncation argument together with Taylor's expansion: First, we show that there is an appropriate truncation such that f_0 remains the risk minimizer defined by a truncated loss over $\mathcal{F}$. Second, the term with first derivative of the truncated loss is shown to be zero by perturbing f around f_0. Finally, the inequality (3.1) is obtained by applying Assumption A.

Let us prove that there exists a truncation such that f_0 is the minimizer of $EV^*(Yf(X))$ over $\mathcal{F}$. To make f_0 invariant with respect to the truncation, let us take the truncation constant $m = (m_1, m_2)$ so that $m_2 > a$ and $m_1 < -a$. Suppose that f_0 is not the minimizer of $EV^*(Yf(X))$ for any truncation constant $m = (m_1, m_2)$

with $m_2 > a$ and $m_1 < -a$. For each $k > a$, there exists $f^{(k)} \in \mathcal{F}$ such that $f^{(k)} \neq f_0$ with positive probability and

$$EV_k^*(Yf^{(k)}(X)) < EV_k^*(Yf_0(X)) = EV(Yf_0(X)) \qquad \text{(A.1)}$$

where V_k^* is the truncated loss with $m = (-k, k)$. Let $f = \limsup_{k\to\infty} f^{(k)}$. Because $\mathcal{F}$ is a linear space, $f \in \mathcal{F}$. We have $EV(Yf(X)) \leq EV(Yf_0(X))$ by taking $\limsup_{k\to\infty}$ in (A.1) because $\limsup_{k\to\infty} EV_k^*(Yf^{(k)}(X)) \leq EV(Yf(X))$ and $EV(Yf_0(X))$ does not depend on k. Moreover, $f \in \mathcal{F}$ implies $EV(Yf(X)) \geq EV(Yf_0(X))$. Hence $EV(Yf(X)) = EV(Yf_0(X))$. Because strictly convex risk function has a unique minimizer, $f = \limsup_{k\to\infty} f^{(k)} = f_0$ a.s., which implies that $f^{(l)} = f_0$ a.s. for all $l \geq k$ and sufficiently large $k > a$. This is a contradiction to our assumption that $f^{(k)} \neq f_0$ with positive probability for any $k > a$.

It follows from (C-2) that V'' exists. Taylor's expansion of $V(Yf_m)$ at Yf_0 yields

$$0 \leq e_{V^*}(f, f_0) = E[V'(Yf_0)(Yf_m - Yf_0) + \frac{1}{2}V''(Yg_m)(Yf_m - Yf_0)^2], \qquad \text{(A.2)}$$

where g_m, an intermediate value in Taylor's expansion, is between f_m and f_0. It then follows from (A.2) that $E[V'(Yf_0)(Yf_m - Yf_0)] = 0$, by setting $f_m = f_0 + h_1 \in \mathcal{F}$ for sufficiently small $h_1 > 0$ and $f_m = f_0 - h_2 \in \mathcal{F}$ for sufficiently small $h_2 > 0$.

Consequently, by Assumption A,

$$\begin{aligned} & e_{V*}(f, f_0) \\ = \; & \frac{1}{2}E\left(V''(Yg_m(X))(f_m(X) - f_0(X))^2\right) \\ \geq \; & \frac{1}{2}\inf_{m_1 \leq z \leq m_2} V''(z)E(f_m(X) - f_0(X))^2 \\ \geq \; & cE\{(f_m(X) - f_0(X))^2 I(Sign(f) \neq Sign(f_0), |f_0| \geq \delta)\} \end{aligned}$$

$$
\begin{aligned}
&\geq c\delta^2(P(Sign(f)\neq Sign(f_0)) - P(|f_0|\leq\delta))\\
&\geq c\delta^2(P(Sign(f)\neq Sign(f_0)) - c_1^*\delta^\beta) \geq c^* P(Sign(f)\neq Sign(f_0))^{\frac{\beta+2}{\beta}}, \quad \text{(A.3)}
\end{aligned}
$$

with the choice of $\delta = (P(Sign(f)\neq Sign(f_0))/2c_1^*)^{\frac{1}{\beta}}$, where $c = \frac{1}{2}\inf_{m_1\leq z\leq m_2} V''(z) > 0$ and $c^* = 2^{-1}c(2c_1^*)^{-1/\beta}$. By triangle inequality,

$$
\begin{aligned}
|e(f,f_0)| &= \frac{1}{2}\,|E|Y - Sign(f)| - E|Y - Sign(f_0)||\\
&\leq \frac{1}{2}E|Sign(f) - Sign(f_0)| = P(Sign(f)\neq Sign(f_0)),
\end{aligned}
$$

which yields the desired result together with (A.3). ■

Proof of Lemma 2: For losses satisfying (C-4), let $V^* = V$. By Proposition 1 of Shen et al. (2003), $e(f,\bar{f}) \leq e_V(f,\bar{f})$. Hence (3.3) follows.

Now, suppose V satisfies (C-3) with $q = 1$. Let $T_1 = 2$ and $T_2 = 0$. Since $|\bar{f}|\leq 1$, $\bar{f}$ is the global minimizer of $EV^*(Yf(X))$. For any given x, let $A_{V^*}(f(x))$ be $E(V^*(Yf(X))|X = x)$, where $A_{V^*}(z) = p^*(x)V^*(z) + (1-p^*(x))V^*(-z)$. Because $V^*(f) + V^*(-f) = 2$, we have $A_{V^*}(f) - A_{V^*}(\bar{f}) = (2p^*(x)-1)(V^*(f) - V^*(\bar{f}))$. Consequently, $A_{V^*}(f) - A_{V^*}(\bar{f}) \geq |2p^*(x) - 1|$ when $Sign(f)\neq Sign(f^*)$, implying

$$
\begin{aligned}
e_{V^*}(f,\bar{f}) &\geq E\{(A_{V^*}(f) - A_{V^*}(\bar{f}))I(Sign(f)\neq Sign(f^*))\}\\
&\geq E\{|2p^*(X) - 1|I(Sign(f)\neq Sign(f^*))\}\\
&= e(f,\bar{f}).
\end{aligned}
$$

The last equality is from Theorem 2.2 of Devryoe, Györfi, and Lugosi (1996). This implies $e(f,\bar{f}) \leq e_{V^*}(f,\bar{f})$. The desired result follows. ■

Proof of Theorem 1: Let V^* be the truncated version of V defined in Lemma 1, where $0\leq T_2\leq V^*(z)\leq T_1 < \infty$ for all z and T_1 and T_2 are the truncation constants.

Let $\tilde{l}_{V^*}(f,Z_i) = l_{V^*}(f,Z_i) + \lambda J[f]$ be the cost function to be minimized, with $l_{V^*}(f,Z_i) = V^*(Y_i f(X_i))$ and $\lambda = 1/(Cn)$ where $Z_i = (X_i,Y_i)$ is a training example.

Let $\tilde{l}(f, Z_i) = l(f, Z_i) + \lambda J[f]$ be the corresponding cost function defined by the misclassification loss function where $l(f, Z_i) = L(Y_i f(X_i))$. Define the scaled empirical process $E_n(\tilde{l}_{V^*}(f, Z) - \tilde{l}_{V^*}(f_0, Z))$ as

$$n^{-1} \sum_{i=1}^{n} (\tilde{l}_{V^*}(f, Z_i) - \tilde{l}_{V^*}(f_0, Z_i) - E(\tilde{l}_{V^*}(f, Z_i) - \tilde{l}_{V^*}(f_0, Z_i)))$$

where $Z = (X, Y)$. Let $A_{i,j} = \{f \in \mathcal{F} : 2^{i-1}\delta_n^2 \leq e_{V^*}(f, f_0) < 2^i \delta_n^2,\ 2^{j-1} \max \{J[f_0], 1\} \leq J[f] < 2^j \max\{J[f_0], 1\}\}$ and $A_{i,0} = \{f \in \mathcal{F} : 2^{i-1}\delta_n^2 \leq e_{V^*}(f, f_0) < 2^i \delta_n^2,$ $J[f] < \max\{J[f_0], 1\}\}$; $j = 1, 2, \cdots$, $i = 1, 2, \cdots$. Without loss of generality, assume that $J[f_0] \geq 1$.

To bound $P(|e(\hat{f}, f_0)| \geq \delta_n^{\frac{2\beta}{\beta+2}})$, we apply Theorem 3 of Shen and Wong (1994), a large deviation inequality for empirical processes, to $P(A_{i,j})$ by controlling the mean and variance defined by $l_{V^*}(f, Z_i)$ and λ. First, let us establish the connection between $e(\hat{f}, \bar{f})$ and $E_n(l_{V^*}(f, Z) - l_{V^*}(f_0, Z))$. By Lemma 1, we obtain that

$$\begin{aligned} & \{|e(\hat{f}, f_0)| \geq \delta_n^{2\beta/(\beta+2)}\} \\ \subset\ & \{e_{V^*}(\hat{f}, f_0) \geq c_{V^*}^{-(\beta+2)/\beta} \delta_n^2\} \\ \subset\ & \{ \sup_{\{f \in \mathcal{F} : e_{V^*}(f, f_0) \geq c_{V^*}^{-(\beta+2)/\beta} \delta_n^2\}} n^{-1} \sum_{i=1}^{n} (\tilde{l}_{V^*}(f_0, Z_i) - \tilde{l}_{V^*}(f, Z_i)) \geq 0\}. \end{aligned}$$

Hence

$$\begin{aligned} & P(|e(\hat{f}, f_0)| \geq \delta_n^{2\beta/(\beta+2)}) \\ \leq\ & P^* \left(\sup_{\{f \in \mathcal{F} : e_{V^*}(f, f_0) \geq c_{V^*}^{-(\beta+2)/\beta} \delta_n^2\}} n^{-1} \sum_{i=1}^{n} (\tilde{l}_{V^*}(f_0, Z_i) - \tilde{l}_{V^*}(f, Z_i)) \geq 0 \right) = I \end{aligned}$$

where P^* denotes the outer probability measure.

To bound I, it suffices to bound $P(A_{i,j})$ for each $i, j = 1, \cdots$. To this end, we need some inequalities regarding the first and second moments of $\tilde{l}_{V^*}(f, Z) - \tilde{l}_{V^*}(f_0, Z)$ for

$f \in A_{i,j}$. Using the assumption that $\max\{J[f_0], 1\}\lambda \le \delta_n^2$, we have

$$\inf_{f \in A_{i,j}} E(\tilde{l}_{V^*}(f, Z) - \tilde{l}_{V^*}(f_0, Z)) \ge M(i, j) = 2^{i-1}\delta_n^2 + \lambda(2^{j-1} - 1)J[f_0] \qquad \text{(A.4)}$$

for any integer $i, j \ge 1$ and

$$\inf_{f \in A_{i,0}} E(\tilde{l}_{V^*}(f, Z) - \tilde{l}_{V^*}(f_0, Z)) \ge M(i, 0) = 2^{i-2}\delta_n^2. \qquad \text{(A.5)}$$

We now compute the second moment. By the truncation argument in Lemma 1 and Taylor's expansion to f_0, we have

$$\begin{aligned} E(l_{V^*}(f, Z) - l_{V^*}(f_0, Z))^2 &= E(V'(Yfg_m(X))(Yf_m(X) - Yf_0(X)))^2 \\ &\le \sup_{m_1 \le z \le m_2} [V'(z)]^2 E(f_m(X) - f_0(X))^2 \\ &\le c^* e_{V^*}(f, f_0) \end{aligned}$$

where g_m is an intermediate value between f_m and f_0 and $c^* = 2\sup_{m_1 \le z \le m_2} [V'(z)]^2 / \inf_{m_1 \le z \le m_2} V''(z)$. The last inequality follows from Taylor's expansion in (A.3). Consequently,

$$\sup_{A_{i,j}} E(l_{V^*}(f, Z) - l_{V^*}(f_0, Z))^2 \le v(i, j)^2 = c_3 M(i, j);$$

$i = 1, 2, \cdots$, $j = 0, 1, \cdots$, where $c_3 = c^*$.

Using the assumption that $\max\{J[f_0], 1\}\lambda \le \delta_n^2/2$, (A.4) and (A.5), we have

$$\begin{aligned} I \quad \le \quad & \sum_{i,j} P^* \left(\sup_{A_{i,j}} E_n(l_{V^*}(f_0, Z) - l_{V^*}(f, Z)) \ge M(i, j) \right) + \\ & \sum_{i} P^* \left(\sup_{A_{i,0}} E_n(l_{V^*}(f_0, Z) - l_{V^*}(f, Z)) \ge M(i, 0) \right) = I_1 + I_2. \end{aligned}$$

Next, we bound I_i separately. For I_1, we verify the conditions (4.5)-(4.7) in Theorem 3 of Shen and Wong (1994). To compute the metric entropy in (4.7), define a bracketing function of $l_{V^*}(f_0, Z) - l_{V^*}(f, Z)$. Denote $\{(f_k^l, f_k^u)\}_{k=1,\cdots,n_c}$ as an ε-bracketing function of $\mathcal{F}$. Let $z = (y, x)$. Because V is strictly convex, V^* is either nonincreasing or not monotone with a minimum. If V^* is nonincreasing, let $l_k^l = \min\{l_{V^*}(f_k^l, z), l_{V^*}(f_k^u, z)\} - l_{V^*}(f_0, z)$ and $l_k^u = \max\{l_{V^*}(f_k^l, z), l_{V^*}(f_k^u, z)\} - l_{V^*}(f_0, z)$. If V^* is not monotone with a minimum at z_{min}, then let

$$\begin{aligned} l_k^l &= \min\{l_{V^*}(f_k^l, z), l_{V^*}(f_k^u, z)\} I((yf_k^l - z_{min})(yf_k^u - z_{min}) \geq 0) \\ &\quad + V^*(z_{min}) I((yf_k^l - z_{min})(yf_k^u - z_{min}) < 0) - l_{V^*}(f_0, z) \end{aligned}$$

and $l_k^u = \max\{l_{V^*}(f_k^l, z), l_{V^*}(f_k^u, z)\} - l_{V^*}(f_0, z)$. For any $f \in \mathcal{F}$ with $J[f] \leq 2^j$, there is $k = 1, \cdots, m$ such that $f_k^l \leq f \leq f_k^u$, which implies that $l_k^l \leq l_{V^*}(f, z) - l_{V^*}(f_0, z) \leq l_k^u$. Hence $\{(l_k^l, l_k^u)\}_{k=1,\cdots,n_c}$ is an ϵ-bracketing function of $l_{V^*}(f, z) - l_{V^*}(f_0, z)$. By Taylor's expansion, $\|l_k^u - l_k^l\|_2 = \|V'(Yg_m(X))(Yf_m^u - Yf_m^l)\|_2 \leq c\|f_m^u - f_m^l\|_2 \leq c\|f_k^u - f_k^l\|_2$ for $c = \sup_{m_1 \leq z \leq m_2}[V'(z)]^2 > 0$ where f_m^u and f_m^l denotes the truncated f_k^u and f_k^l, respectively, and g_m denotes an intermediate value between f_m^u and f_m^l. Hence $H_B(u, \mathcal{F}_{V^*}(2^j)) \leq H_B(cu, \mathcal{F}(2^j))$.

Using the fact that $\int_{aM(i,j)}^{v(i,j)} H_B^{1/2}(u, \mathcal{F}_{V^*}(2^j))du/M(i,j)$ is nonincreasing in i and $M(1,j)$ for $j = 1, 2, \cdots$, we have

$$\begin{aligned} \int_{aM(i,j)}^{v(i,j)} H_B^{1/2}(u, \mathcal{F}_{V^*}(2^j))du/M(i,j) &\leq \int_{aM(1,j)}^{\sqrt{c_3}M(1,j)^{1/2}} H_B^{1/2}(u, \mathcal{F}_{V^*}(2^j))du/M(1,j) \\ &\leq \phi(\bar{\varepsilon}_n, 2^j), \end{aligned}$$

where $a = \varepsilon/32$. The metric entropy condition implies (4.7) with a choice of $\varepsilon = 1/2$ and c_i for $i = 3, 4$. Moreover, (4.5)-(4.6) are satisfied with $\varepsilon = 1/2$ with the choice of $M(i,j)$ and $v(i,j)$ in that $M(i,j)/v^2(i,j) \leq 1/4c^*$.

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

$$\begin{aligned} I_1 &\leq \sum_{j=1}^{\infty}\sum_{i=1}^{\infty} 3\exp\left(-\frac{(1-\varepsilon)nM(i,j)^2}{2(4v^2(i,j)+M(i,j)T/3)}\right) \\ &\leq \sum_{j=1}^{\infty}\sum_{i=1}^{\infty} 3\exp\left(-c_5 nM(i,j)\right) \\ &\leq \sum_{j=1}^{\infty}\sum_{i=1}^{\infty} 3\exp\left(-c_5 n[(2^{i-1}\delta_n^2) + ((2^{j-1}-1)\lambda J[f_0])]\right) \\ &\leq 3\exp\left(-c_5 n\lambda J[f_0]\right)/[1-\exp\left(-c_5 n\lambda J[f_0]\right)]^2, \end{aligned}$$

where $c_5 > 0$ is a generic constant. Similarly, I_2 can be bounded. Finally, $I \leq 6\ \exp(-c_5 n\lambda\ J[f_0])/[1-\exp(-c_5 n\lambda J[f_0])]^2$. This implies that $I^{1/2} \leq (5/2 + I^{1/2})$ $\exp(-c_5 n\ \lambda J[f_0]))$. Since $I \leq I^{1/2} \leq 1$, $I \leq 3.5\exp\left(-c_5 n\lambda J[f_0]\right)$. ■

Proof of Theorem 2: Because the proof for V satisfying (C-4) is essentially the same for V satisfying (C-3) with $q = 1$, let us sketch the proof for V satisfying (C-3) with $q = 1$. We may follow the proof of Theorem 1 except for orders of exponents and constants.

Let V^* be the truncated version of V defined in the proof of Lemma 2, so that $T_2 \leq V^*(z) \leq T_1$ for all z where $T_2 = 0$ and $T_1 = 2$. Let $m = (-1, 1)$. By Assumption B, we have $e_V(f_0, \bar{f}) \leq \delta_n^2$.

Consider the empirical process $E_n(\tilde{l}_{V^*}(f, Z) - \tilde{l}(f_0, Z))$. Let $A_{i,j} = \{f \in \mathcal{F} : 2^{i-1}\delta_n^2 \leq e_{V^*}(f, f_0) < 2^i\delta_n^2,\ 2^{j-1}\ \max\ \{J[f_0], 1\} \leq J[f] < 2^j \max\{J[f_0], 1\}\}$ and $A_{i,0} = \{f \in \mathcal{F} : 2^{i-1}\delta_n^2 \leq e_{V^*}(f, f_0) < 2^i\delta_n^2,\ J[f] < \max\{J[f_0], 1\}\}$; $j = 1, 2, \cdots$, $i = 1, 2, \cdots$. By an analogous argument, we have

$$\begin{aligned} &P(e(\hat{f}, \bar{f}) \geq \delta_n^2) \\ \leq\ &P^*\left(\sup_{\{f\in\mathcal{F}: e_{V^*}(f,f_0)\geq c_{V^*}^{-1}\delta_n^2\}} n^{-1}\sum_{i=1}^{n}(\tilde{l}_{V^*}(f_0, Z_i) - \tilde{l}_{V^*}(f, Z_i)) \geq 0\right) = I. \end{aligned}$$

To bound I, it suffices to bound $P(A_{i,j})$ for each $i, j = 1, \cdots$. For the first moment of the empirical process, it can be shown that

$$\inf_{f \in A_{i,j}} E(\tilde{l}_{V^*}(f, Z) - \tilde{l}_{V^*}(f_0, Z)) \geq M(i,j) = 2^{i-1}\delta_n^2 + \lambda(2^{j-1} - 1)J[f_0] \qquad \text{(A.6)}$$

for any integer $i, j \geq 1$ and

$$\inf_{f \in A_{i,0}} E(\tilde{l}_{V^*}(f, Z) - \tilde{l}_{V^*}(f_0, Z)) \geq M(i,0) = 2^{i-2}\delta_n^2. \qquad \text{(A.7)}$$

We now compute the second moment. It follows from Assumption A that, for any $f \in \mathcal{F}$,

$$\begin{aligned}
e_{V^*}(f, f_0) + \delta_n^2 &\geq e_{V^*}(f, \bar{f}) \\
&= 2E|f^*(X)||f_m(X) - \bar{f}(X)| \\
&\geq \delta E|f_m(X) - \bar{f}(X)|I(|f^*(X)| \geq \delta) \\
&\geq \delta(E|f_m(X) - \bar{f}(X)| - 4c_1\delta^\alpha) \\
&\geq 2^{-1}(8c_1)^{-1/\alpha}(E|f_m(X) - \bar{f}(X)|)^{(\alpha+1)/\alpha}
\end{aligned}$$

with a choice of $\delta = (E|f_m(X) - \bar{f}(X)|)^{1/\alpha}$. Hence,

$$\begin{aligned}
& E(V^*(Yf(X)) - V^*(Yf_0(X)))^2 \\
= \; & E\left(V(Yf_m(X)) - V(Yf_m^0(X))\right)^2 \\
= \; & E(f_m(X) - f_m^0(X))^2 \\
\leq \; & 2E|f_m(X) - f_m^0(X)| \\
\leq \; & 2(E|f_m(X) - \bar{f}(X)| + E|f_m^0(X) - \bar{f}(X)|) \\
\leq \; & 4(4c_1)^{1/(\alpha+1)}((e_{V^*}(f, f_0) + \delta_n^2)^{\alpha/(\alpha+1)} + e_{V^*}(f_0, \bar{f})^{\alpha/(\alpha+1)}) \\
\leq \; & c^*(e_{V^*}(f, f_0)^{\alpha/(\alpha+1)} + 2\delta_n^{2\alpha/(\alpha+1)}) \leq c^* e_{V^*}(f, f_0)^{\alpha/(\alpha+1)}
\end{aligned}$$

for a positive generic constant c^* because δ_n^2 can be absorbed into $e_{V^*}(f, f_0)$. We have proved that $E(l_{V^*}(f,Z) - l_{V^*}(f_0,Z))^2 \leq c^* e_{V^*}(f, f_0)^{\alpha/(\alpha+1)}$ for some positive constant c^*.

Denote $\{(f_k^l, f_k^u)\}_{k=1,\cdots,n_c}$ an ϵ-bracketing function of $\mathcal{F}$. For any $z = (x, y)$ where $f \in A_{i,j}$, let $l_k^l = \min\{l_{V^*}(f_k^l, z), l_{V^*}(f_k^u, z)\} - l_{V^*}(f_0, z)$ and $l_k^u = \max\{l_{V^*}(f_k^l, z)$ $, l_{V^*}(f_k^u, z)\} - l_{V^*}(f_0, z)$ because V^* is nonincreasing. Using the analogous argument in Theorem 1, $\{(l_k^l, l_k^u)\}_{k=1,\cdots,n_c}$ is an ϵ-bracketing function of $l_{V^*}(f,z) - l_{V^*}(f_0, z)$ because V^* is decreasing. Easily, we can show that $\|l_k^u - l_k^l\|_2 = \|f_m^u - f_m^l\|_2 \leq \|f_k^u - f_k^l\|_2$ because V^* satisfies Lipschitz condition. Hence $H_B(u, \mathcal{F}_{V^*}(2^j)) \leq H_B(cu, \mathcal{F}(2^j))$.

Thus

$$\sup_{A_{i,j}} E(l_{V^*}(f,Z) - l_{V^*}(f_0,Z))^2 \leq v(i,j)^2 = c_3 M(i,j)^{\alpha/(\alpha+1)}$$

for $i = 1, 2, \cdots$ and $j = 0, 1, \cdots$ where $c_3 = c^*$. Following the proof of Theorem 1, we obtain the desired result. ■

Proof of Theorem 4: The main idea of the proof is to construct a least favorable subfamily of $\mathcal{F}$ that is as difficult as the original problem. This yields the lower bound by an application of Fano's lemma; c.f., Ibragimov and Has'minskii (1981). In our classification framework, only decision functions whose signs define classifiers are specified, whereas the existence of conditional probabilities is not required.

Let p^* be the true conditional density of $Y = 1$ given x. For any $f \in \mathcal{F}$, let $\tilde{p}(f)$ represents an equivalence class of probability densities that yield the same decision boundary in the sense that $Sign(f) = Sign(\tilde{p}(f) - 1/2)$. Note that the mapping from f to $\tilde{p}(f)$ is not unique but we can choose a representative without loss of generality. This implies that classification based on $f \in \mathcal{F}$ is as difficult as that based on $\tilde{p}(f)$ for $f \in \mathcal{F}$.

We now construct a least favorable family of $\{\tilde{p}(f) : f \in \mathcal{F}\}$. Our construction uses truncation as well as the behavior of $\tilde{p}(f)$ near the decision boundary, which characterizes the most difficult situation for classification. For any $f \in \mathcal{F}$ and any $0 < \delta < 1/4$, define

$$\bar{p}(f) = \begin{cases} \tilde{p}(f) & \text{if } |(\tilde{p}(f)/p^*)^{1/2} - 1| \leq \delta^{1/2} \text{ and } |p^* - 1/2| \leq \delta, \\ (1 + Sign(f)\delta^{1/2})^2 p^* & \text{otherwise.} \end{cases}$$

This in turn defines a density $p(f) = \bar{p}(f)/c(f)$ after normalization, where $c(f)$ is a normalizing constant. By construction, $\bar{p}(f)$ yields the same the decision boundary as $\tilde{p}(f)$, in addition that for any $f_i \in \mathcal{F}$; $i = 1, \cdots, r$, the likelihood ratio $\frac{(1-\delta^{1/2})^2}{(1+\delta^{1/2})^2} \leq p(f_i)/p(f_j) \leq \frac{(1+\delta^{1/2})^2}{(1-\delta^{1/2})^2}$, but each $p(f_i)$ may not be bounded away from zero, depending on p^*.

Next, we introduce a maximal $\underline{\varepsilon}_n^2$-separated subset of $\mathcal{G}(\underline{\varepsilon}_n, \mathcal{F})$ $N = \{G_{f_1}, \cdots, G_{f_r}\}$, whose existence is implied by assumption. Denote C_n as an $\underline{\varepsilon}_n$-covering of $\mathcal{G}(\underline{\varepsilon}_n, \mathcal{F})$. Because an $\underline{\varepsilon}_n^2$-separated set for $\mathcal{G}(\underline{\varepsilon}_n, \mathcal{F})$ can also serve as an $\underline{\varepsilon}_n^2$-cover of $\mathcal{G}(\underline{\varepsilon}_n, \mathcal{F})$, we may take C_n so that $C(\underline{\varepsilon}_n^2, \mathcal{G}(\underline{\varepsilon}_n, \mathcal{F})) = H(2\underline{\varepsilon}_n^2, \mathcal{G}(\underline{\varepsilon}_n, \mathcal{F}))$ in view of the relation in (3.6). Define $G_{\tilde{f}} = \arg\min_{1 \leq i \leq r} d_\Delta(G_{f_i}, G_{f^*})$.

The likelihood of $(X_i, Y_i)_{i=1}^n$ for any $f \in \mathcal{F}$ is

$$\left(\prod_{i=1}^n h(x_i)\right) \prod_{i=1}^n \left(p(f(x_i))^{y_i} (1 - p(f(x_i)))^{1-y_i}\right)$$

where h is the marginal density of X_i.

For any f_i and $\delta < 1/4$; $i = 1, \cdots, r$, define p_i as above, so that the likelihood ratio between p_i and p_j is uniformly bounded away from zero and infinity. Let $A_i = (|(\tilde{p}(f_i)/p^*)^{1/2} - 1| \leq \delta^{1/2}) \cap (|p^* - 1/2| \leq \delta)$ for $i = 1, \cdots, r$. By assumption, it can be verified that p_i satisfies the low noise condition on A_i. Furthermore, the Kullback-Leibler (K-L) divergence between p_i and p_j is bounded above by the Hellinger distance between p_i and p_j:

$$\begin{aligned}
&K(p_i,p_j) \leq \|p_i^{1/2} - p_j^{1/2}\|^2 \\
\leq\ & \frac{1}{c(f_i)}\|\bar{p}(f_i)^{1/2} - \bar{p}(f_j)^{1/2}\|^2 + \left(\frac{1}{c(f_i)^{1/2}} - \frac{1}{c(f_j)^{1/2}}\right)^2 \|\bar{p}(f_j)^{1/2}\|^2 \\
\leq\ & c\|\bar{p}(f_i)^{1/2} - \bar{p}(f_j)^{1/2}\|^2 \\
\leq\ & c(\|(\bar{p}(f_i)^{1/2} - \bar{p}(f_j)^{1/2})I(A_i \cup A_j)\|^2 + \|(\bar{p}(f_i)^{1/2} - \bar{p}(f_j)^{1/2})I(A_i^C \cap A_j^C)\|^2) \\
\leq\ & c\left(\|2(p^*)^{1/2}\delta^{1/2}I_{A_i \cup A_j}\|^2 + \|(p^*)^{1/2}\delta^{1/2}(Sign(f_i) - Sign(f_j))I(A_i^C \cap A_j^C)\|^2\right) \\
\leq\ & c\delta\left(P(A_i \cup A_j) + d_\Delta(G_{f_i}, G_{f_j})\right) \\
\leq\ & c\delta\left(P\left(|p^* - 1/2| \leq \delta\right) + d_\Delta(G_{f_i}, G_{f_j})\right) \\
\leq\ & c\delta^{1+\alpha} + \delta d_\Delta(G_{f_i}, G_{f_j}) \leq c d_\Delta^{\frac{\alpha+1}{\alpha}}(G_{f_i}, G_{f_j}) \leq c\underline{\varepsilon}_n^{\frac{2(\alpha+1)}{\alpha}},
\end{aligned}$$

by minimizing with respect to δ, with a choice of $\delta = d_\Delta^{\frac{1}{\alpha}}(G_{f_i}, G_{f_j})/4 \geq \underline{\varepsilon}_n^{\frac{2}{\alpha}}/4$, where $c > 0$ is a generic constant.

To apply Fano's Lemma, note that N is an $\underline{\varepsilon}_n$-separated set, $d_\Delta(G_{f_i}, G_{f^*}) + d_\Delta(G_{f_j}, G_{f^*}) \geq d_\Delta(G_{f_i}, G_{f_j}) \geq \underline{\varepsilon}_n^2$. Thus $d_\Delta(G_f, G_{f^*}) \geq \underline{\varepsilon}_n^2/2$ if $G_f \neq G_{\tilde{f}}$. Note that (3.7) implies that $\max_{1 \leq i,j \leq r} nK(p_i,p_j) \leq \frac{1}{2}\log(r-1) - \log 2$ for sufficiently small $\underline{\varepsilon}_n > 0$. It then follows from Fano's Lemma that

$$\max_{1\leq i\leq r} P(d_\Delta(G_{f_i}, G_{f^*}) \geq \underline{\varepsilon}_n^2/2) \geq \max_{1\leq i\leq r} P(G_{f_i} \neq G_{\tilde{f}}) \geq \frac{1}{r}\sum_{i=1}^{r} P(G_{f_i} \neq G_{\tilde{f}}) \geq \frac{1}{2}.$$

To obtain the lower bound for $\mathcal{G}(\underline{\varepsilon}_n, \mathcal{F})$, note that for $G_{f_1}, \cdots, G_{f_r}$, there exists $G_{\tilde{f}_1}, \cdots, G_{\tilde{f}_r} \in \mathcal{G}(\underline{\varepsilon}_n, \mathcal{F})$ such that $d_\Delta(G_{\tilde{f}_i}, G_{f_i}) \leq \underline{\varepsilon}_n^d$ by the definition of the metric entropy for some sufficiently large $d > 0$ such that $n\underline{\varepsilon}_n^{2d} \leq 1/4$. This implies that

$$\max_{1\leq i\leq r} P(d_\Delta(G_{\tilde{f}_i}, G_{f^*}) \geq \underline{\varepsilon}_n^2/2) \geq \max_{1\leq i\leq r} P(d_\Delta(G_{f_i}, G_{f^*}) \geq \underline{\varepsilon}_n^2/2) - n\underline{\varepsilon}_n^2 \geq 1/4.$$

Hence

$$\sup_{G_f \in \mathcal{G}(\underline{\varepsilon}_n, \mathcal{F})} P(d_\Delta(G_f, G_{f^*}) \geq \underline{\varepsilon}_n^2/2) \geq \max_{1\leq i\leq r} P(d_\Delta(G_{\tilde{f}_i}, G_{f^*}) \geq \underline{\varepsilon}_n^2/2) \geq 1/4. \quad \text{(A.8)}$$

Finally, by Assumption A

$$e(f,\bar{f}) \geq 2^{-1}(4c_1)^{-\frac{1}{\alpha}} d_{\Delta}^{\frac{\alpha+1}{\alpha}}(G_f, G_{f^*});$$

c.f., the proof of Theorem 1 in Shen et al. (2003) or Lemma 2 in Mammen and Tsybakov (1999). Then

$$\sup_{f\in\mathcal{F}} P(e(f,\bar{f}) \geq \underline{\varepsilon}_n^2/2) \geq \sup_{G_f\in\mathcal{G}(\underline{\varepsilon}_n,\mathcal{F})} P(d_{\Delta}(G_f, G_{f^*}) \geq (4c_1)^{-\frac{1}{\alpha}} \underline{\varepsilon}_n^{\frac{2\alpha}{\alpha+1}}). \tag{A.9}$$

Hence, if

$$C(\underline{\varepsilon}_n^{\frac{2\alpha}{\alpha+1}}, \mathcal{G}(\underline{\varepsilon}_n, \mathcal{F})) \geq c^{'} n \left(\underline{\varepsilon}_n^{\frac{2\alpha}{\alpha+1}}\right)^{\frac{\alpha+1}{\alpha}} = n\underline{\varepsilon}_n^2,$$

then, from (A.8) and (A.9),

$$\sup_{f\in\mathcal{F}} P(e(f,\bar{f}) \geq \underline{\varepsilon}_n^2/2) \geq 1/2.$$

■

The following proposition provides the rate of convergence for SVM in the linear example through direct calculations.

Proposition 2. *Under the assumptions of section 3.5.1, if V satisfies (C-3) with $q = 1$, then $E|e(\hat{f}, f_0)| = O(n^{-1/2})$.*

Proof of Proposition 2: Because the proof is lengthy, we divide the proof into three steps.

Step 1. We establish connection between the excess surrogate risk over $\mathcal{F}$ and the excess risk over $\mathcal{F}$:

$$|e(f, f_0)| \leq c_{V^*} e_{V^*}(f, f_0)^{1/2} \tag{A.10}$$

for some truncated loss V^* of V with a positive constant c_{V^*} depending on V^*.

The idea is to use a truncation argument together with Taylor's expansion: First, we show that f_0 remains as the risk minimizer defined by a truncated loss. Second, the risk of a decision rule defined by the truncated loss is shown to be strictly convex as a function of a vector of coefficients corresponding to the decision rule. Then we apply Taylor's expansion argument to bound the excess risk over $\mathcal{F}$ defined by the truncated loss by l_2-norm of their vectors of coefficients. Finally, (A.10) is obtained.

We can show that there is a truncated loss V^* of V such that f_0 is the minimizer of the risk defined by the truncated loss. The truncations constants can be chosen so that f_0 is invariant with respect to the truncation.

Before proceeding, let us introduce a few notations. Define

$$\mathcal{W} = \{w \in \mathbb{R}^{d+1} : w \text{ is the vector of coefficients corresponding to } f \in \mathcal{F}\}.$$

Because there is 1-1 correspondence between $\mathcal{F}$ and $\mathcal{W}$, we may define $v(w) = EV^*(Yf(X))$ for any $f \in \mathcal{F}$ with the representation $f(x) = a_1x_1 + \cdots a_dx_d + b$ where $(a_1, \cdots, a_d, b) \in \mathcal{W}$ is the vector of coefficients. Denote the vector of coefficients for f_0 by w_0.

Now, let us show that w_0 is the minimizer of v and $v(w)$ is strictly convex around w_0. Note that w_0 is a minimizer of v because $EV^*(Yf_0(X)) \leq EV^*(Yf(X))$ for all $f \in \mathcal{F}$ implies that $v(w_0) \leq v(w)$ for any $w \in \mathcal{W}$. Easily, $v(w)$ is a convex function in w around w_0. For f around f_0, $v(w) = 1 - b - \sum_{i=1}^d a_i E(X_i) + E(Yf(X) - 1)_+$. If $\|f\|_\infty \leq 1$, then $v(w)$ is a linear function in w because $E(Yf(X) - 1)_+$ vanishes. Otherwise, $v(w)$ is nonlinear in w. Through tedious calculations, $E(Yf(X) - 1)_+$ can be shown to be smooth in w. Because v is globally convex, it is strictly convex in some neighborhood of w_0. This, together with $\|f_0\|_\infty = \sqrt{\frac{1-r}{r}} > 1$ for $0 < r < 1/2$, imply that the risk is strictly convex around w_0. Hence w_0 is unique minimizer of $v(w)$.

Take a threshold $0 < t_c < \sqrt{\frac{1-r}{r}} - 1$ of $e_{V^*}(f, f_0)$ so that f is invariant with respect to the truncation for f satisfying $e_{V^*}(f, f_0) \leq t_c$. If $e_{V^*}(f, f_0) > t_c$, then $\frac{1}{t_c} e_{V^*}(f, f_0) > 1 \geq e(f, f_0)^2$, which holds trivially because $e(f, f_0) \leq 1$. The solution w_0 does not belong to the region of f where $e_{V^*}(f, f_0)$ is bounded away from t_c. Hence $e_{V^*}(f, f_0) \leq t_c$ corresponds to the least favorable situation. The set of $f \in \mathcal{F}$ satisfying $e_{V^*}(f, f_0) \leq t_c$ can be translated into some neighborhood of coefficients w_0 defined by $N(w_0, t_w) = \{w \in \mathbb{R}^{d+1} : \|w - w_0\| \leq t_w\}$ for some positive constant t_w.

Now restrict our attention to this neighborhood. On $N(w_0, t_w)$, $v(w)$ is a strictly convex function in w. By Taylor's expansion,

$$0 \leq e_{V^*}(f, f_0) = v(w) - v(w_0) = \nabla v(w_0)'(w - w_0) + \frac{1}{2}(w - w_0)' H_v(w^*)(w - w_0)$$

where ∇v and H_v denote the gradient and the Jacobian of v, respectively, and w^* is an intermediate value between w and w_0. Note that w_0 is the unique minimizer of v in the expansion and v is a strictly convex function of w on the neighborhood of w_0. Thus $H_v(w^*)$ is positive definite for w in the neighborhood of w_0. Using the perturbation argument around w_0 as before, the first term must be zero. Hence,

$$0 \leq e_{V^*}(f, f_0) = \frac{1}{2}(w - w_0)' H_v(w^*)(w - w_0).$$

Let $\lambda_i(w)$; $i = 1, \cdots, d+1$ be the eigenvalues of $H_v(w^*)$. Because $H_v(w^*)$ is positive definite and v is smooth on $N(w_0, t_w)$, we have $\inf_{w \in N(w_0, t_w)} \min_{1 \leq i \leq d+1} \lambda_i(w) \geq c \min_{1 \leq i \leq d+1} \lambda_i(w_0) > 0$ for constant $c > 0$. Then

$$e_{V^*}(f, f_0) = \frac{1}{2} \sum_{i=1}^{d+1} \lambda_i(w)(w_i - w_i^0)^2 \geq \frac{c}{2} \min_{1 \leq i \leq d+1} \lambda_i(w_0) \|w - w_0\|^2. \tag{A.11}$$

To bound $|e(f, f_0)|$ by $\|w - w_0\|$, note that

$$\begin{aligned} |e(f, f_0)| &\leq P(Sign(f) \neq Sign(f_0)) \\ &= P(|f + f_0| < |f - f_0|) \\ &\leq P\left(|f + f_0| \leq |b - b_0 + \sum_{i=1}^{d}(a_i - a_i^0)X_i|\right) \\ &\leq P\left(|f + f_0| \leq c'\|w - w_0\|_1\right) \\ &\leq \min\{c'\|w - w_0\|, 1\} \end{aligned}$$

for a generic constant $c' > 0$ where $\| \cdot \|_1$ denotes the l_1 norm. This, together with (A.11), yields $e_{V^*}(f, f_0) \geq c^* e(f, f_0)^2$ for some positive constant c^* depending on w_0. Because $|e(f, f_0)| \leq \frac{1}{\sqrt{c^*}} e_{V^*}(f, f_0)^{1/2}$ for the least favorable situation, $|e(f, f_0)| \leq c_{V^*} e_{V^*}(f, f_0)^{1/2}$.

Step 2. Let us show that

$$E(l_{V^*}(f, Z) - l_{V^*}(f_0, Z))^2 \leq c^* e_{V^*}(f, f_0) \tag{A.12}$$

for some constant $c^* > 0$.

Define $v_{w_0}(w) = E(V^*(Yf(X)) - V^*(Yf_0(X)))^2 \geq v_{w_0}(w_0) = 0$. We can apply the argument of Step 1 for $v_{w_0}(w)$. Since $v_{w_0}(w) \leq T^2$, the threshold t_c and the neighborhood $N(w_0, t_w)$ can be determined in a similar fashion. Obviously, $v_{w_0}(w)$ is a strictly convex function in w and it has the minimizer at w_0. By Taylor's expansion and the perturbation argument around w_0, $v_{w_0}(w) = \frac{1}{2}(w - w_0)' H_{v_{w_0}}(w^*)(w - w_0)$ where $H_{v_{w_0}}$ denotes the Jacobian of v_{w_0} and w^* is an intermediate value between w and w_0. Following the argument of Step 1, we have $v(w, w_0) \leq \frac{c^*}{2} \max_{1 \leq l \leq d+1} \lambda_l(w_0) \|w - w_0\|^2$ for some positive constant c^* where $\lambda_l(w_0); l = 1, \cdots, d+1$ are the eigenvalues of $H_{v_{w_0}}(w_0)$. By (A.11), $E(l_{V^*}(f, Z) - l_{V^*}(f_0, Z))^2 \leq c^* e_{V^*}(f, f_0)$ for some positive constant c^* depending on $\max_{1 \leq l \leq d+1} \lambda_l(w_0) \ / \min_{1 \leq l \leq d+1} \lambda_l^*(w_0)$ where $\lambda_l^*(w_0)$; $l = 1, \cdots, d+1$ are the eigenvalues of the Jacobian matrix at w_0 for $v^*(w) = e_{V^*}(f, f_0)$.

Step 3. Following the proof of Theorem 1, we obtain the probability bound, which implies the risk bound. By solving entropy equations, $\bar{\varepsilon}_n = n^{-1/2}$. Therefore, $E|e(\hat{f}, f_0)| = O(n^{-1/2})$. ■

Lemma 3. *Let f and g be Lipschitz functions from $\mathbb{R}^d$ to $\mathbb{R}$. For any $t \in \mathbb{R}$, if $\mathcal{L}^d(f \geq t) < \infty$ and $\mathcal{L}^d(g \geq t) < \infty$, then*

$$\mathcal{L}^d((f \geq t)\Delta(g \geq t)) = \int_t^\infty \mathcal{L}^{d-1}((f = s)\Delta(g = s))ds$$

where $\mathcal{L}^d$ denotes the Lebesgue measure on $\mathbb{R}^d$.

Proof of Lemma 3: By Theorem 1 (Coarea Formula) in Evans and Gariepy (1992), we can prove $\mathcal{L}^d(f \geq t) = \int_t^\infty \mathcal{L}^{d-1}(f = s)ds$. By applying this result to $\mathcal{L}^d((f \geq t)\Delta(g \geq t)) = \mathcal{L}^d(f \geq t) + \mathcal{L}^d(g \geq t) - \mathcal{L}^d((f \geq t) \cap (g \geq t)$, the result follows. ■

In Theorem 1 of Belyakov (1986), the metric entropy for 0-level sets of polynomials is provided. The following lemma bounds the metric entropy for classification sets using that for level sets.

Lemma 4. *Under the assumptions of section 3.5.2,*

$$H(\epsilon, \mathcal{G}(k)) \leq O(\log(k/\epsilon)).$$

Proof of Lemma 4: Let $f, g \in \mathcal{F}(k)$, polynomials of degree m_p in $\mathcal{X}$, such that f/k and g/k satisfies the conditions of Lemma 2 in Belyakov (1986). Denote $\Gamma_f(t)$ and $\Gamma_g(t)$ as t-level sets induced by polynomials f and g, respectively, and the sup-norm distance in $\mathcal{X}$ as $\omega(\cdot, \cdot)$.

By Lemma 3, the volume of a symmetric difference of two classification sets can be expressed as

$$P(G_f \Delta G_g) = \int_0^k \mathcal{L}^{d-1}(\Gamma_f(t)\Delta\Gamma_g(t))dt/\mathcal{L}^d(\mathcal{X}),$$

for some $k > 0$. Since levels sets are compact in a bounded space $\mathcal{X}$, $\sup_{0\le t\le k}$ $\mathcal{L}^{d-1}(\Gamma_f(t)\Delta\Gamma_g(t))$ is attained at some $t_0 \in [0, k]$. Then the volume can be bounded by a rectangle containing both the t_0-level sets $\Gamma_f(t_0)$ and $\Gamma_g(t_0)$ with length ω $(\Gamma_f(t_0)$ $, \Gamma_g(t_0))$ multiplied by some constant. That is,

$$P(G_f \Delta G_g) \le C_{m_p,d,k}\omega(\Gamma_f(t_0), \Gamma_g(t_0))^d,$$

where $C_{m_p,d,k}$ is a positive generic constant depending on m_p, d, and k. Applying Lemma 2 of Belyakov (1986), we have $\omega(\Gamma_f(t_0), \Gamma_g(t_0)) \le C_{m_p,d}(\epsilon/k)^{1/m_p}$, implying that

$$P(G_f \Delta G_g) \le C_{m_p,d,k}(\epsilon/k)^{d/m_p}.$$

Plugging this in the proof of Theorem 1 in Belyakov (1986), the result follows. ■

The following lemma is useful in local entropy calculations for polynomial kernel.

Lemma 5. *Let f and g be two polynomials of degree m on a compact and connected set $\mathcal{X}$ in $\mathbb{R}^d$ defined by $f(x) = \sum_{l_0+\cdots+l_d=m_p} a_{l_0,\cdots,l_d} x_1^{l_1}\cdots x_d^{l_d}$ and $g(x) = \sum_{l_0+\cdots+l_d=m_p}$ $b_{l_0,\cdots,l_d}$ $x_1^{l_1}\cdots x_d^{l_d}$ such that $Sign(a_{l_0,\cdots,l_d}) = Sign(b_{l_0,\cdots,l_d})$ for each $l_0, \cdots, l_d$ with $l_1 + \cdots + l_d = m_p$. If $\mathcal{L}^d(G_f \Delta G_g) \le \epsilon$, then $|a_{l_0,\cdots,l_d} - b_{l_0,\cdots,l_d}| < c\epsilon$ for any $l_0, \cdots, l_d$ with $l_1 + \cdots + l_d = m_p$ where c is a positive constant.*

Proof of Lemma 5: First, consider the case polynomials in one variable. Without loss of generality, we may put $f(x) = x^{m_p} + a_{m_p-1}x^{m_p-1}\cdots + a_0$ and $g(x) = x^{m_p} + b_{l-1}x^{m_p-1}\cdots + b_0$. Denote roots of f and g as τ_i and τ_i^*; $i = 1, \cdots, m_p$, respectively.

Assume that $\tau_1 \leq \cdots \leq \tau_{m_p}$ and $\tau_1^* \leq \cdots \leq \tau_{m_p}^*$. Since $G_f \Delta G_g$ is a finite union of disjoint intervals in $\mathcal{X} = [u, v]$ $(-\infty < u < v < \infty)$ with end points τ_i or τ_i^*; $i = 1, \cdots, l$, we can easily see that $|\tau_i - \tau_i^*| \leq c\epsilon$; $i = 1, \cdots, l$ for some positive generic constant c, which implies $|a_i^* - a_i| \leq c\epsilon$ for $i = 0, \cdots, l-1$ using the relation between roots and coefficients.

Suppose that the assertion holds for polynomials in $(d-1)$-variables. We may rearrange the polynomials f and g as a polynomial of degree m_p in $x_1, \cdots, x_{d-1}$ for any fixed x_d as follows: $f(x_1, \cdots, x_d) = \sum_{l_0+\cdots+l_{d-1} \leq m_p} a'_{l_0,\cdots,l_{d-1}}\ x_1^{l_1} \cdots x_{d-1}^{l_{d-1}}$ and $g(x_1, \cdots, x_d) = \sum_{l_0+\cdots+l_{d-1} \leq m_p} b'_{l_0,\cdots,l_{d-1}}\ x_1^{l_1} \cdots x_{d-1}^{l_{d-1}}$ where $a'_{l_0,\cdots,l_{d-1}} = a_{l_0,\cdots,l_d} x_d^{l_d}$ and $b'_{l_0,\cdots,l_{d-1}} = b_{l_0,\cdots,l_d} x_d^{l_d}$ for $l_d = m_p - l_0 - \cdots - l_{d-1}$. Take x_d^* so that $x_d - x_d^*$ is bounded away from zero. Denote the translation of set G_f with respect to d-th coordinate by x_d^* as $T_{x_d^*}(G_f)$. The volume of $G_f \Delta G_g$ can be expressed as $\mathcal{L}^d(G_f \Delta G_g)$ $= \int \mathcal{L}^{d-1}(G_f \Delta G_g)(x_d) dx_d$ where $\mathcal{L}^{d-1}$ $(G_f \Delta G_g)(x_d) = \int \cdots \int I_{G_f \Delta G_g}$ $(x_1, \cdots, x_d)$ $dx_1 \cdots dx_{d-1}$. The volume condition on $G_f \Delta G_g$ implies $\mathcal{L}^d(G_f \Delta G_g) \leq \epsilon$. Since the volume is invariant with respect to translations, $\mathcal{L}^d(G_f \Delta G_g) = \mathcal{L}^d(T_{x_d^*}(G_f \Delta G_g)) =$ $\int_{H_d(T_{x_d^*}(\mathcal{X}))} \mathcal{L}^{d-1}(G_f \Delta G_g)(x_d - x_d^*) dx_d$ where $H_d : \mathcal{X} \to \mathbb{R}$ is a projection map defined by $H_d(x_1, \cdots, x_d) = x_d$. By Mean Value Theorem in calculus, there is x_d^0 such that $\mathcal{L}^{d-1}(G_f \Delta G_g)$ $(x_d^0 - x_d^*)$ $\mathcal{L}^1(H_d(T_{x_d^*}(\mathcal{X}))) = \mathcal{L}^{d-1}(G_f \Delta G_g)(x_d^0 - x_d^*) \mathcal{L}^1(H_d(\mathcal{X})) =$ $\mathcal{L}^d(G_f \Delta G_g)$. Hence $\mathcal{L}^{d-1}$ $(G_f \Delta G_g)$ $(x_d^0 - x_d^*) < c\epsilon$ for some positive constant c. By inductive assumption, $|a'_{l_0,\cdots,l_{d-1}} - b'_{l_0,\cdots,l_{d-1}}| \leq c\epsilon$, which implies $|a_{l_0,\cdots,l_d} - b_{l_0,\cdots,l_d}| < c\epsilon$.

■

BIBLIOGRAPHY

[1] An, L. T. H., and Tao, P. D. (1997). Solving a class of linearly constrained indefinite quadratic problems by D.C. algorithms. *J. Global Opt.* **11**, 253-285.

[2] Bartlett, P. L., Bousquet, O., and Mendelson, S. (2004). Local Rademacher complexities. *Ann. Statist.* To appear.

[3] Bartlett, P. L., Jordan, M. I. and McAuliffe, J. D. (2003). Convexity, Classification, and Risk Bounds. *J. Amer. Statist. Assoc.* To appear.

[4] Belyakov, A. V. (1986). Estimates for the ϵ-entropy of a set of 0-levels of polynomials of degree at most n in p complex variables. In *Geometric questions in the theory of functions and sets.* 19-25, Kalinin. Gos. Univ., Kalinin. [In Russian.]

[5] Blanchard, G., Bousquet, O., and Massart, P. (2004). Statistical Performance of Support Vector Machines. Manuscript.

[6] Bousquet, O., Boucheron, S., and Lugosi, G. (2004a). Introduction to statistical learning theory. In *Advanced Lectures in Machine Learning.* 169-207. Springer, New York.

[7] Bousquet, O., Boucheron, S., and Lugosi, G. (2004b). Thoery of Classification : A Survey of Recent Advances. *ESIAM : Probability and Statistics.*

[8] Breiman, L. (1996). Bagging Predictors. *Mach. Learn.* **36**, 105-139.

[9] Breiman, L., Friedman, J. H., Olshen, R., and Stone, C. J. (1984). *Classification and Regression Trees.* Belmont, California.

[10] Cristianini, N. and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines.* Cambridge University Press, Cambridge.

[11] Corduneanu, A. and Jaakola, T. (2003). On Information Regularization. In *Proceedings of the Nineteenth Annual Conference on Uncertainty in Artificial Intelligence.* 151-158. Morgan Kaufmann Publishers, San Fracisco, California.

[12] Cortes, C. and Vapnik, V. (1995). Support vector networks. *Machine Learning* **20**, 273-297.

[13] Devroye, L., Györfi, L. and Lugosi, G. (1996). *A Probability Theory of Pattern Recognition.* Springer, New York.

[14] Dudley, R. M. (1974). Metric Entropy of Some Classes of Sets with Differentiable Boundaries. *J. of Approx. Theory* **10**, 227-236.

[15] Evans, L. C. and Gariepy, R. F. (1992). *Measure theory and fine properperties of functions.* CRC Press, Florida.

[16] Freund, Y., and Schapire, R. (1997). A decision-theoretic generalization of online learning and an application to boosting. *J. Comput. System Sci.* **55**, 119-139.

[17] Hastie, T., Tibshirani, R. and Friedman, J. H. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, New York.

[18] Ibragimov, I. A. and Has'minskii, R. Z. (1981). *Statistical Estimation.* Springer, New York.

[19] Jiang, W. (2004). Process consistency for Adaboost. *Ann. Statist.* **32**, 12-29.

[20] Kolmogorov, A. N. and Tikhomirov, V. M. (1959). ε-entropy and ε-capacity of sets in a functional space. *Uspekhi Mat. Nauk* **14**, 3-86. [In Russian. English Translations in American Society Translations **17**, 277-364 (1961).]

[21] Lin, Y. (2000). Some asymptotic properties of the support vector machine. Technical report 1029, Department of Statistics, University of Wisconsin-Madison.

[22] Lin, Y. (2002a). A note on margin-based loss functions in classification. *Statist. Probab. Lett.* **68**, 73-82.

[23] Lin, Y. (2002b). Support Vector Machines and the Bayes Rule in Classification. *Data Mining and Knowledge Discovery* **6**, 259-275.

[24] Liu, S., Shen, X., and Wong, W. H. (2005). Computational developments of ψ-learning. Manuscript.

[25] Liu, Y. and Shen, X. (2004). On multicategory ψ-learning and support vector machine. *J. Amer. Statist. Assoc.* To appear.

[26] Lugosi, G. and Vayatis, N. (2004). On the Bayes-risk consistency of regularized boosting methods. *Ann. Statist.* **32**, 30-55.

[27] Mammen, E. and Tsybakov, A. B. (1999). Smooth discriminant analysis. *Ann. Statist.* **27**, 1808-1829.

[28] Marron, J. S. and Todd, M. (2002). Distance weighted discrimination. Technical report 1339, Department of Statistical Science, Cornell University.

[29] Mason, L., Baxter, J., Bartlett, P. and Frean, M. R. (2000). Boosting algorithms as gradient descent in function space. *Adv. Neural Inf. Process. Syst.* **12**, 512-518.

[30] Mercer, J. (1909). Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society, London A* **209**, 415-446.

[31] Scovel, J. C. and Steinwart, I. (2004). Fast rates for support vector machines. Technical report LA-UR03-9117, Los Alamos Nationl Laboratory.

[32] Scott, C. and Nowak, R. (2004). Minimax-Optimal Classification with Dyadic Decision Trees. Technical report TREE 0403, Department of Electrical and Computer Engineering, University of Wisconsin-Madison.

[33] Shen, X. and Wong, W. H. (1994). Convergence rate of sieve estimates. *Ann. Statist.* **22**, 580-615.

[34] Shen, X., Zhang, X., Tseng, G. C. and Wong, W. H. (2003). On ψ-learning. *J. Amer. Statist. Assoc.* **98**, 724-734.

[35] Tsybakov, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *Ann. Statist.* **32**, 135-166.

[36] Van de Geer, S. (1993). Hellinger-consistency of certain nonparametric maximum likelihood estimators. *Ann. Statist.* **21**, 14-44.

[37] Vapnik, V. (1995). *The Nature of Statistical Learning Theory.* Springer, New York.

[38] Vapnik, V. (1998). *Statistical Learning Theory.* Wiley, New York.

[39] Wahba, G. (1990). *Spline Methods for Observational Data.* CBMS-NSF Regional Conference Series, Philadelphia.

[40] Yang, Y. (1999). Minimax nonparametric classification - Part I : Rates of convergence, Part II : Model selection for adaptation. *IEEE Trans. Inform. Theory* **45**, 2271-2292.

[41] Zhang, T. (2004a). Statistical behavior and consistency of classification methods based on convex risk minimization. *Ann. Statist.* **32**, 56-84.

[42] Zhang, T. (2004b). Statistical Analysis of Some Multi-Category Large Margin Classification Mehtods. *J. Mach. Learn. Res.* **5**, 1225-1251.

[43] Zhang, T. and Yu, B. (2005). Boosting with Early Stopping: Convergence and Consistency. *Ann. Statist.* To appear.

[44] Zhou, D. X. (2002). The covering number in learning theory. *J. Complexity* **18**, 739-767.

[45] Zhu, J. and Hastie, T. (2005). Kernel Logistic Regression and the Import Vector Machines. *J. Comput. Graph. Statist.* **14**, 185-205.