

On efficient large margin semisupervised learning: method and theory*

Junhui Wang
Department of Statistics
Columbia University
New York, NY 10027

Xiaotong Shen
School of Statistics
University of Minnesota
Minneapolis, MN 55455

Wei Pan
Department of Biostatistics
University of Minnesota
Minneapolis, MN 55455

Abstract

In classification, semisupervised learning usually involves a large amount of unlabeled data with only a small number of labeled data. This imposes a great challenge in that it is difficult to achieve good classification performance through labeled data alone. To leverage unlabeled data for enhancing classification, this article introduces a large margin semisupervised learning method within the framework of regularization, based on an efficient margin loss for unlabeled data, which seeks efficient extraction of the information from unlabeled data for estimating the Bayes decision boundary for classification. For implementation, an iterative scheme is derived through conditional expectations. Finally, theoretical and numerical analyses are conducted, in addition to an application to gene function prediction. They suggest that the proposed method enables to recover the performance of its supervised counterpart based on complete data in the sense of rate of convergence, as if the label values of unlabeled data were available in advance.

Key words: Difference convex programming, classification, nonconvex minimization, regularization, support vectors.

* Research supported in part by NSF grants IIS-0328802 and DMS-0604394, by NIH grants HL65462 and 1R01GM081535-01, and a UM AHC FRD grant.

1 Introduction

Semisupervised learning occurs in classification, where only a small number of labeled data is available with a large amount of unlabeled data, because of the difficulty of labeling. In artificial intelligence, one central issue is how to integrate human's intelligence with machine's processing capacity. This occurs, for instance, in webpage classification and spam email detection, where webpages and emails are automatically collected, yet require labeling manually or classification by experts. The reader may refer to Blum and Mitchell (1998), Amini and Gallinari (2003), and Balcan et al. (2005) for more details. In genomics applications, functions of many genes in sequenced genomes remain unknown, and are predicted using available biological information, c.f., Xiao and Pan (2005). In situations as such, the primary goal is to leverage unlabeled data to enhance predictive performance of classification (Zhu, 2005).

In semisupervised learning, labeled data $\{(X_i, Y_i)_{i=1}^{n_l}\}$ are sampled from an unknown distribution $P(x, y)$, together with an independent unlabeled sample $\{X_j\}_{j=n_l+1}^n$ from its marginal distribution $q(x)$. Here label $Y_i \in \{-1, 1\}$, $X_i = (X_{i1}, \dots, X_{id})$ is an d -dimensional input, $n_l \ll n_u$ and $n = n_l + n_u$ is the combined size of labeled and unlabeled samples.

Two types of approaches—distributional and margin-based, have been proposed in the literature. The distributional approach includes, among others, co-training (Blum and Mitchell, 1998), the EM method (Nigam et al., 1998), the bootstrap method (Collins and Singer, 1999), Gaussian random fields (Zhu, Ghahramani and Lafferty, 2003), and structure learning models (Ando and Zhang, 2004). The distributional approach relies on an assumption relating the class probability given input $p(x) = P(Y = 1|X = x)$ to $q(x)$ for an improvement to occur. However, the assumption of this sort is often not verifiable or met in practice.

A margin approach uses the concept of regularized separation. It includes Transductive

SVM (TSVM, Vapnik, 1998; Chapelle and Zien, 2005; Wang, Shen and Pan, 2007), and a large margin method of Wang and Shen (2007). These methods utilize the notation of separation to borrow information from unlabeled data to enhance classification, which relies on the clustering assumption (Chapelle and Zien, 2005) that the clustering boundary can precisely approximate the Bayes decision boundary which is the focus of classification.

This article develops a large margin semisupervised learning method, which aims to extract the information from unlabeled data for estimating the Bayes decision boundary. This is achieved by constructing an efficient loss for unlabeled data with regard to reconstruction of the Bayes decision boundary and by incorporating some knowledge from an estimate of p . This permits efficient use of unlabeled data for accurate estimation of the Bayes decision boundary thus enhancing the classification performance based on labeled data alone. The proposed method, utilizing both the grouping (clustering) structure of unlabeled data and the smoothness structure of p , is designed to recover the classification performance based on complete data without missing labels, when possible.

The proposed method has been implemented through an iterative scheme with ψ -loss (Shen et al., 2003), which can be thought of as an analogy of Fisher's efficient scoring method (Fisher, 1946). That is, given a consistent initial classifier, an iterative improvement can be obtained through the constructed loss function. Numerical analysis indicates that the proposed method performs well against several state-of-the-art semisupervised methods including TSVM, and Wang and Shen (2007), where Wang and Shen (2007) compares favorably against several smooth and clustering based semisupervised methods.

A novel statistical learning theory for ψ -loss is developed to provide an insight into the proposed method. The theory reveals that the ψ -learning classifier's generalization performance based on complete data can be recovered by its semisupervised counterpart based on incomplete data in rates of convergence, without knowing the label values of unlabeled data in advance. The theory also says that the least favorable situation for a

semisupervised problem occurs at points near $p(x) = 0$ or 1 because little information can be provided by these points for reconstructing the classification boundary as discussed in Section 2.3. This is in contrast to the fact that the least favorable situation for a supervised problem occurs near $p(x) = 0.5$. In conclusion, this semisupervised method achieves the desired objective of delivering higher generalization performance.

This article also examines one novel application in gene function discovery, which has been a primary focus of biomedical research. In gene function discovery, microarray gene expression profiles can be used to predict gene functions, because genes sharing the same function tend to co-express, c.f., Zhou, Kao and Wong (2002). Unfortunately, biological functions of many discovered genes remain unknown at present. For example, about $1/3$ to $1/2$ of the genes in the genome of bacterium *E. coli* have unknown functions. Therefore, gene function discovery is an ideal application for semisupervised methods and also employed in this article as a real numerical example.

This article is organized in six sections. Section 2 introduces the proposed method. Section 3 develops an iterative algorithm for implementation. Section 4 presents some numerical examples, together with an application to gene function prediction. Section 5 develops a learning theory. Section 6 contains a discussion, and the appendix is devoted to technical proofs.

2 Methodology

2.1 Large margin classification

Consider large margin classification with labeled data $(X_i, Y_i)_{i=1}^{n_i}$. In linear classification, given a class of candidate decision functions \mathcal{F} , a cost function

$$C \sum_{i=1}^{n_i} L(y_i f(x_i)) + J(f) \tag{1}$$

is minimized over $f \in \mathcal{F} = \{f(x) = \tilde{w}_f^T x + w_{f,0} \equiv (1, x^T)w_f\}$ to yield the minimizer \hat{f} leading to classifier $\text{sign}(\hat{f})$. Here $J(f)$ is the reciprocal of the geometric margin of various form with the usual L_2 margin $J(f) = \|\tilde{w}_f\|^2/2$ to be discussed in further detail, and $L(\cdot)$ is a margin loss defined by functional margin $z = yf(x)$, and $C > 0$ is a regularization parameter. In nonlinear classification, a kernel $K(\cdot, \cdot)$ is introduced for flexible representations: $f(x) = \sum_{i=1}^{n_l} \alpha_i K(x, x_i) + b$. For this reason, it is referred to as kernel-based learning, where the reproducing kernel Hilbert spaces (RKHS) are useful, c.f., Gu (2000) and Wahba (1990).

Different margin losses correspond to different learning methodologies. Margin losses include, among others, the hinge loss $L(z) = (1 - z)_+$ for SVM with its variant $L(z) = (1 - z)_+^q$ for $q > 1$; c.f., Lin (2002); the ψ -losses $L(z) = \psi(z)$, with $\psi(z) = 1 - \text{sign}(z)$ if $z \geq 1$ or $z < 0$, and $2(1 - z)$ otherwise, c.f., Shen, et al. (2003), the logistic loss $V(z) = \log(1 + e^{-z})$, c.f., Zhu and Hastie (2005); the η -hinge loss $L(z) = (\eta - z)_+$ for nu-SVM (Schölkopf et al., 2000) with $\eta > 0$ being optimized; the sigmoid loss $L(z) = 1 - \tanh(cz)$; c.f., Mason, et al. (2000). A margin loss $L(z)$ is said to be large margin if $L(z)$ is non-increasing in z , penalizing small margin values. In this paper, we fix $L(z) = \psi(z)$.

2.2 Loss construction for unlabeled data

In classification, the optimal Bayes rule is defined by $\bar{f}_{.5} = \text{sign}(f_{.5})$ with $f_{.5}(x) = P(Y = 1|X = x) - 0.5$ being a global minimizer of the generalization error $GE(f) = EI(Y \neq \text{sign}(f(X)))$, which is usually estimated by labeled data through $L(\cdot)$ in (1). In absence of sufficient labeled data, the focus is on how to improve (1) by utilizing additional unlabeled data. For this, we construct a margin loss U to measure the performance of estimating $\bar{f}_{.5}$ for classification through unlabeled data. Specifically, we seek the best loss U from a class of candidate losses of form $T(f)$, which minimizes the L_2 -distance between the target classification loss $L(yf)$ and $T(f)$. The expression of this loss U is given in Lemma 1.

Lemma 1 (*Optimal loss*) For any margin loss $L(z)$,

$$\operatorname{argmin}_T E(L(Yf(X)) - T(f(X)))^2 = E(L(Yf(X))|X = x) = U(f(x)),$$

where $U(f(x)) = p(x)L(f(x)) + (1 - p(x))L(-f(x))$ and $p(x) = P(Y = 1|X = x)$.

Moreover, $\operatorname{argmin}_{f \in \mathcal{F}} EU(f(X)) = \operatorname{argmin}_{f \in \mathcal{F}} EL(Yf(X))$.

Based on Lemma 1, we define $\hat{U}(f)$ to be $\hat{p}(x)L(f(x)) + (1 - \hat{p}(x))L(-f(x))$ by replacing p in $U(f)$ by \hat{p} . Clearly, $\hat{U}(f)$ approximates the ideal optimal loss $U(f)$ for reconstructing the Bayes decision function $f_{.5}$ when \hat{p} is a good estimate of p , as suggested by Corollary 2. This is analogous to construction of the efficient scores for Fisher's scoring method: an optimal estimate can be obtained iteratively through an efficient score function, provided that a consistent initial estimate is supplied, c.f., McCullagh and Nelder (1983) for more details. Through (approximately) optimal loss $\hat{U}(f)$, an iterative improvement of estimation accuracy is achieved by starting with a consistent estimate \hat{p} of p , which, for instance, can be obtained through SVM or TSVM. For $\hat{U}(f)$, its optimality is established through its closeness to $U(f)$ in Corollary 2, where our iterative method based on \hat{U} is shown to yield an iterative improvement in terms of the classification accuracy, recovering the generalization error rate of its supervised counterpart based on complete data ultimately.

As a technical remark, we note that the explicit relationship between p and f is usually unavailable in practice. As a result, several large margin classifiers such as SVM and ψ -learning do not directly yield an estimate of p given \hat{f} . Therefore p needs to be either assumed or estimated. For instance, the methods of Wahba (1999) and Platt (1999) assume a parametric form of p so that an estimated \hat{f} yields an estimated p , whereas Wang, Shen and Liu (2007) estimates p given \hat{f} nonparametrically.

The preceding discussion leads to our proposed cost function:

$$s(f) = C \left(n_l^{-1} \sum_{i=1}^{n_l} L(y_i f(x_i)) + n_u^{-1} \sum_{j=n_l+1}^n \hat{U}(f(x_j)) \right) + J(f). \quad (2)$$

Minimization of (2) with respect to $f \in \mathcal{F}$ gives our estimated decision function \hat{f} for classification.

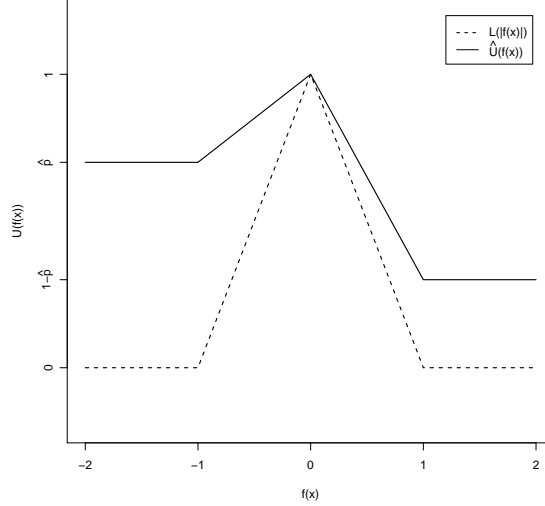
2.3 Connection with clustering assumption

We now intuitively explain advantages of $\hat{U}(f)$ over a popular large margin loss $L(|f|) = (1 - |f(x)|)_+$ (Vapnik, 1998; Wang and Shen, 2007), and its connection with the clustering assumption (Chapelle and Zien, 2005) that assumes closeness between the classification and grouping (clustering) boundaries.

First, $\hat{U}(f)$ has an optimality property, as discussed in Section 2.2, which leads to better performance as suggested by Theorem 2. Second, it has a higher discriminative power over its counterpart $L(|f|)$. To see this aspect, note that $L(|f|) = \inf_p U(f)$ by Lemma 1 of Wang and Shen (2007). This says that $L(|f|)$ is a version of $\hat{U}(f)$ in the least favorable situation where unknown p is estimated by $\text{sign}(f)$, completely ignoring the magnitude of p . As displayed in Figure 1, $\hat{U}(f)$ corresponds to an “uneven” hat function or the solid line, whereas $L(|f|)$ corresponds to an “even” one or the dashed line. By comparison, $\hat{U}(f)$ enables not only to identify the clustering boundary through the hat function as $L(|f|)$ does but also to discriminate $f(x)$ from $-f(x)$ through an estimated $\hat{p}(x)$. That is, $\hat{U}(f)$ has a smaller value for $f(x) > 0$ than for $-f(x) < 0$ when $\hat{p} > 0.5$, and vice versa, whereas $L(|f|)$ is in-discriminative with regard to the sign of $f(x)$.

To reinforce the second point in the foregoing discussion, we examine one specific example with two possible clustering boundaries as described in Figure 3 of Zhu (2005). There $\hat{U}(f)$ favors one clustering boundary for classification if a consistent \hat{p} is provided, whereas $L(|f|)$ fails to discriminate these two. More details are deferred to Section 4.1,

Figure 1: Plots of $L(|f(x)|)$ and $\hat{U}(f(x))$.



where the simulated example 2 of this nature is studied.

In conclusion, $\hat{U}(f)$ yields a more efficient loss for a semisupervised problem as it utilizes the clustering information from the unlabeled data as $L(|f|)$ does, in addition to guidance about labeling through \hat{p} to gain a higher discriminative power.

3 Computation

3.1 Iterative scheme

The effectiveness of \hat{U} depends largely on the accuracy of \hat{p} in estimating p . Given an estimate $\hat{p}^{(0)}$, (2) yields an estimate $\hat{f}^{(1)}$, which leads to a new estimate $\hat{p}^{(1)}$ through **Algorithm 0** below. The $\hat{p}^{(1)}$ is expected to be more accurate than $\hat{p}^{(0)}$ for p because additional information from unlabeled data has been utilized in estimation of $\hat{f}^{(1)}$ through $\hat{p}^{(0)}$ and additional smoothness structure has been used in **Algorithm 0** in estimation of $\hat{p}^{(1)}$ given $\hat{f}^{(1)}$. Specifically, an improvement in the process from $\hat{p}^{(0)}$ to $\hat{f}^{(1)}$ and that from $\hat{f}^{(1)}$ to $\hat{p}^{(1)}$ are assured by Assumptions B and D in Section 5.1, respectively, which are a more

general version of the clustering assumption and a smoothness assumption of p . In other words, the marginal information from unlabeled data has been effectively incorporated in each iteration of **Algorithm 1** for improving estimation accuracy of \hat{f} and \hat{p} .

Detailed implementation of the preceding scheme as well as the conditional probability estimation are summarized as follows.

Algorithm 0: (Conditional probability estimation; Wang, Shen and Liu, 2007)

Step 1. Initialize $\pi_k = (k - 1)/m$, for $k = 1, \dots, m + 1$.

Step 2. Train a weighted margin classifier \hat{f}_{π_k} with $1 - \pi_k$ associated with positive instances and π_k associated with negative instances, for $k = 1, \dots, m + 1$.

Step 3. Estimate labels of x by $\text{sign}\{\hat{f}_{\pi_k}(x)\}$.

Step 4. Sort $\text{sign}\{\hat{f}_{\pi_k}(x)\}$, $k = 1, \dots, m + 1$, to compute $\pi^* = \max[\pi_k : \text{sign}\{\hat{f}_{\pi_k}(x)\} = 1]$, $\pi_* = \min[\pi_k : \text{sign}\{\hat{f}_{\pi_k}(x)\} = -1]$. The estimated class probability is $\hat{p}(x) = \frac{1}{2}(\pi^* + \pi_*)$.

Algorithm 1: (Efficient semisupervised learning)

Step 1. (Initialization) Given any initial classifier $\text{sign}(\hat{f}^{(0)})$, compute $\hat{p}^{(0)}$ through **Algorithm 0**. Specify precision tolerance level $\epsilon > 0$.

Step 2. (Iteration) At iteration $k + 1$; $k = 0, 1, \dots$, minimize $s(f)$ in (2) for $\hat{f}^{(k+1)}$ with $\hat{U} = \hat{U}^{(k)}$ defined by $\hat{p} = \hat{p}^{(k)}$ there. This is achieved through sequential QP for the ψ -loss. Details for sequential QP are deferred to Section 3.2. Compute $\hat{p}^{(k+1)}$ through **Algorithm 0**, based on complete data with unknown labels imputed by $\text{sign}(\hat{f}^{(k+1)})$. Define $\hat{p}^{(k+1)} = \max(\hat{p}^{(k)}, \hat{p}^{(k+1)})$ when $\hat{f}^{(k+1)} \geq 0$ and $\min(\hat{p}^{(k)}, \hat{p}^{(k+1)})$ otherwise.

Step 3. (Stopping rule) Terminate when $|s(\hat{f}^{(k+1)}) - s(\hat{f}^{(k)})| \leq \epsilon$. The final solution \hat{f}_C is $\hat{f}^{(K)}$, with K the number of iterations to termination in **Algorithm 1**.

Theorem 1 (*Monotonicity*) **Algorithm 1** has a monotone property such that $s(\hat{f}^{(k)})$ is non-increasing in k . As a consequence, **Algorithm 1** converges to a stationary point $s(\hat{f}^{(\infty)})$ in that $s(\hat{f}^{(k)}) \geq s(\hat{f}^{(\infty)})$. Moreover, **Algorithm 1** terminates finitely.

Algorithm 1 differs from the EM algorithm and its variant MM algorithm (Hunter and Lange, 2000) in that little marginal information has been used in these algorithms as argued in Zhang and Oles (2000). **Algorithm 1** also differs from Yarowsky’s algorithm (Yarowsky, 1995; Abney, 2004) in that Yarowsky’s algorithm solely relies on the strength of the estimated \hat{p} , ignoring the potential information from the clustering assumption.

There are several important aspects of **Algorithm 1**. First, loss $L(\cdot)$ in (2) may not be a likelihood regardless of if labeling missingness occurs at random. Secondly, the monotonicity property, as established in Theorem 1, is assured by constructing $\hat{p}^{(k+1)}$ to satisfy $(\hat{p}^{(k+1)} - \hat{p}^{(k)})\hat{f}^{(k+1)} \geq 0$, as opposed to the property of likelihood in the EM algorithm. Most importantly, the smoothness and clustering assumptions have been utilized in estimating p , and thus semisupervised learning. This is in contrast to the EM, where only likelihood is used in estimating p in a supervised manner.

Finally, we note that in **Step 2** of **Algorithm 1**, given $\hat{p}^{(k)}$, minimization in (2) involves nonconvex minimization when $L(\cdot)$ is ψ -loss. Next we shall discuss how to solve (2) for $\hat{f}^{(k+1)}$ through difference convex (DC) programming for nonconvex minimization.

3.2 Nonconvex minimization

This section develops a nonconvex minimization method based on DC programming (An and Tao, 1997) for (2) with the ψ -loss, which was previously employed in Liu, Shen and Wong (2005) for supervised ψ -learning. As a technical remark, we note that DC programming has a high chance to locate an ε -global minimizer (An and Tao, 1997), although it can not guarantee globality. In fact, when combined with the method of branch-and-bound, it yields a global minimizer, c.f., Liu, Shen and Wong (2005). For a computational consideration, we shall use the DC programming algorithm without seeking an exact global minimizer.

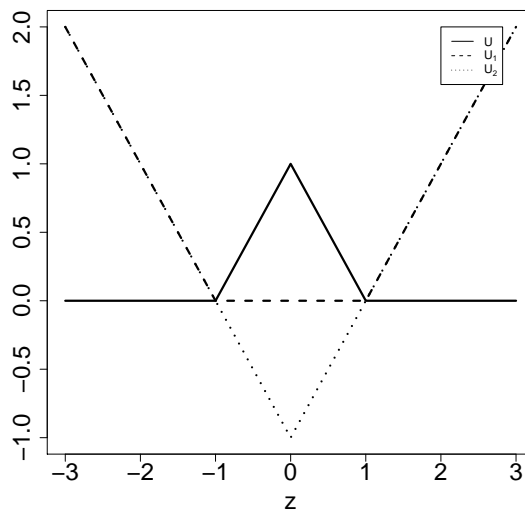
Key to DC programming is decomposing the cost function $s(f)$ in (2) with $L(z) = \psi(z)$

into a difference of two convex functions as follows:

$$\begin{aligned}
s(f) &= s_1(f) - s_2(f); \\
s_1(f) &= C\left(n_l^{-1} \sum_{i=1}^{n_l} \psi_1(y_i f(x_i)) + n_u^{-1} \sum_{j=n_l+1}^n \hat{U}_{\psi_1}^{(k)}(f(x_j))\right) + J(f); \\
s_2(f) &= C\left(n_l^{-1} \sum_{i=1}^{n_l} \psi_2(y_i f(x_i)) + n_u^{-1} \sum_{j=n_l+1}^n \hat{U}_{\psi_2}^{(k)}(f(x_j))\right),
\end{aligned} \tag{3}$$

where $\hat{U}_{\psi_t}^{(k)}(f(x_j)) = \hat{p}^{(k)}(x_j)\psi_t(f(x_j)) + (1 - \hat{p}^{(k)}(x_j))\psi_t(-f(x_j))$; $t = 1, 2$, $\psi_1 = 2(1 - z)_+$ and $\psi_2 = 2(-z)_+$. Here ψ_1 and ψ_2 are obtained through a convex decomposition of $\psi = \psi_1 - \psi_2$ as displayed in Figure 2.

Figure 2: Plot of ψ , ψ_1 and ψ_2 for the DC decomposition of $\psi = \psi_1 - \psi_2$. Solid, dotted and dashed lines represent ψ , ψ_1 and ψ_2 , respectively.



With these decompositions, we treat (2) with the ψ -loss and $\hat{p} = \hat{p}^{(k)}$ by solving a sequence of quadratic problems described in **Algorithm 2**.

Algorithm 2: (Sequential QP)

Step 1. (Initialization) Set initial $\hat{f}^{(k+1,0)}$ to be the solution of $\min_f s_1(f)$. Specify precision tolerance level $\epsilon > 0$.

Step 2. (Iteration) At iteration $l + 1$, compute $\hat{f}^{(k+1,l+1)}$ by solving

$$\min_f (s_1(f) - \langle w_f, \nabla s_2(\hat{f}^{(k+1,l)}) \rangle), \quad (4)$$

where $\nabla s_2(f^{(k+1,l)})$ is a gradient vector of $s_2(f)$ at $w_{\hat{f}^{(k+1,l)}}$.

Step 3. (Stopping rule) Terminate when $|s(\hat{f}^{(k+1,l+1)}) - s(\hat{f}^{(k+1,l)})| \leq \epsilon$.

Then the estimate $\hat{f}^{(k+1)}$ is the best solution among $\hat{f}^{(k+1,l)}$; $l = 0, 1, \dots$.

In (4), gradient $\nabla s_2(f^{(k+1,l)})$ is defined as the sum of partial derivatives of s_2 over each observation, with $\nabla \psi_2(z) = 0$ if $z > 0$ and $\nabla \psi_2(z) = -2$ otherwise. By the definition of $\nabla s_2(f^{(k+1,l)})$ and convexity of $s_2(f^{(k+1,l)})$, (4) gives a sequence of non-increasing upper envelopes of (3), which can be solved via their dual forms.

The speed of convergence of **Algorithm 2** is super-linear, following the proof of Theorem 3 in Liu, Shen and Wong (2005). This means that the number of iterations required for **Algorithm 2** to achieve the precision ϵ is $o(\log(1/\epsilon))$.

4 Numerical comparison

This section examines effectiveness of the proposed method through numerical examples.

A test error, averaged over 100 independent simulation replications, is used to measure a classifier's generalization performance. For simulation comparison, the amount of improvement of our method over $\text{sign}(\hat{f}^{(0)})$ is defined as the percent of improvement in terms of the Bayesian regret

$$\frac{(T(\text{Before}) - \text{Bayes}) - (T(\text{After}) - \text{Bayes})}{T(\text{Before}) - \text{Bayes}}, \quad (5)$$

where $T(\text{Before})$, $T(\text{After})$, and Bayes denote the test errors of $\text{sign}(\hat{f}^{(0)})$, the proposed method based on initial classifier $\text{sign}(\hat{f}^{(0)})$, and the Bayes error. The Bayes error is the

ideal optimal performance and serves as a benchmark for comparison, which can be computed when the distribution is known. For benchmark examples, the amount of improvement over $\text{sign}(\hat{f}^{(0)})$ is defined as

$$\frac{T(\text{Before}) - T(\text{After})}{T(\text{Before})}, \quad (6)$$

which actually underestimates the amount of improvement in absence of the Bayes rule.

Numerical analyses are conducted in R2.1.1. In linear learning, $K(x, y) = \langle x, y \rangle$; in Gaussian kernel learning, $K(x, y) = \exp(-\frac{\|x-y\|^2}{\sigma^2})$, where σ is set to be the median distance between positive and negative classes to reduce computational cost for tuning σ^2 , c.f., Jaakkola, Diekhans and Haussler (1999).

4.1 Simulations and benchmarks

Two simulated and five benchmark data sets are examined, based on four state-of-the-art classifiers $\text{sign}(\hat{f}^{(0)})$'s. They are SVM (with labeled data alone), TSVM (SVM^{light}; Joachims, 1999), and the methods of Wang and Shen (2007) with the hinge loss (SSVM) and with the ψ -loss (SPSI), where SSVM and SPSI compare favorably against their competitors. Corresponding to these methods, our method yields four semisupervised classifiers denoted as ESVM, ETSVM, ESSVM and ESPSI.

Simulated examples: Examples 1 and 2 are taken from Wang and Shen (2007), where 200 and 800 labeled instances are randomly selected for training and testing. For training, 190 out 200 instances are randomly chosen for removing their labels. Here the Bayes errors are 0.162 and 0.089, respectively.

Benchmarks: Six benchmark examples include Wisconsin breast cancer (WBC), Pima Indians diabetes (PIMA), HEART, MUSHROOM, Spam email (SPAM) and Brain computer interface (BCI). The first five datasets are available in the UCI Machine Learning Repository (Blake and Merz, 1998) and the last one can be found in Chapelle, Schölkopf

and Zien (2006). WBC discriminates a benign breast tissue from a malignant one through 9 diagnostic characteristics; PIMA differentiates between positive and negative cases for female diabetic patients of Pima Indian heritage based on 8 biological or diagnostic attributes; HEART concerns diagnosis status of the heart disease based on 13 clinic attributes; MUSHROOM separates an edible mushroom from a poisonous one through 22 biological records; SPAM identifies spam emails using 57 frequency attributes of a text, such as frequencies of particular words and characters; BCI concerns the difference of brain images when imagining left-hand and right-hand movements, based on 117 autoregressive model parameters fitted over human's electroencephalography.

Instances in WBC, PIMA, HEART, and MUSHROOM are randomly divided into two halves with 10 labeled and 190 unlabeled instances for training, and the remaining 400 for testing. Instances in SPAM are randomly divided into halves with 20 labeled and 380 unlabeled instances for training, and the remaining instances for testing. Twelve splits for BCI have already given at <http://www.kyb.tuebingen.mpg.de/ssl-book/benchmarks.html>, with 10 labeled and 390 unlabeled instances while no instance for testing. An averaged error rate over the unlabeled set is used in BCI example to approximate the test error.

In each example, the smallest test errors of all methods in comparison are computed over 61 grid points $\{10^{-3+k/10}; k = 0, \dots, 60\}$ for tuning C in (2) through a grid search. The results are summarized in Tables 1-2.

As suggested in Tables 1-2, ESVM, ETSVM, ESSVM and ESPSI perform no worse than their counterparts in almost all examples, except ESVM in SPAM where the performance is slightly worse but indistinguishable from its counterpart. The amount of improvement, however, varies over examples and different types of classifiers. In linear learning excluding ESVM in SPAM, the improvements of the proposed method are from 1.9% to 67.8% over its counterparts, except in SPAM where ESVM performs slightly worse than SVM; in kernel learning, the improvements range from 0.0% to 23.2% over its counter-

Table 1: **Linear learning.** Averaged test errors as well as estimated standard errors (in parenthesis) of ESVM, ETSVM, ESSVM, ESPSI, and their initial counterparts and testing samples, in the simulated and benchmark examples. SVM_c denotes the performance of SVM with complete labeled data. Here the amount of improvement is defined in (5) or (6).

Data	Example 1	Example 2	WBC	PIMA	HEART	MUSHROOM	SPAM	BCI ^a
Size	1000 × 2	1000 × 2	682 × 9	768 × 8	303 × 13	8124 × 22	4601 × 57	400 × 117
SVM	.344(.0104)	.333(.0129)	.053(.0071)	.351(.0070)	.284(.0085)	.232(.0135)	.216(.0097)	.479(.0059)
ESVM	.281(.0143)	.297(.0177)	.031(.0007)	.320(.0059)	.214(.0066)	.172(.0084)	.217(.0178)	.474(.0052)
Improv.	53.8%	19.8%	41.5%	8.8%	24.6%	25.9%	-0.5%	1.0%
TSVM	.249(.0121)	.222(.0128)	.077(.0043)	.315(.0067)	.270(.0082)	.204(.113)	.227(.0120)	.500(.0054)
ETSVM	.190(.0074)	.147(.0131)	.029(.0009)	.309(.0063)	.211(.0062)	.153(.0054)	.179(.0101)	.474(.0076)
Improv.	67.8%	56.4%	62.3%	1.9%	21.9%	25.0%	21.1%	5.2%
SSVM	.188(.0084)	.129(.0031)	.032(.0025)	.307(.0054)	.240(.0074)	.186(.0095)	.191(.0114)	.479(.0071)
ESSVM	.182(.0065)	.124(.0034)	.028(.0006)	.293(.0029)	.205(.0059)	.162(.0054)	.169(.0107)	.474(.0041)
Improv.	23.1%	28.6%	12.5%	4.6%	14.6%	12.9%	11.5%	1.0%
SPSI	.184(.0084)	.128(.0084)	.029(.0022)	.291(.0032)	.232(.0067)	.184(.0095)	.189(.0107)	.476(.0068)
ESPSI	.182(.0065)	.123(.0029)	.027(.0006)	.284(.0026)	.181(.0052)	.137(.0067)	.167(.0107)	.471(.0046)
Improv.	9.1%	12.8%	6.9%	4.5%	22.0%	25.5%	10.1%	1.1%
SVM _c	.164(.0084)	.115(.0032)	.027(.0020)	.238(.0011)	.176(.0031)	.041(.0018)	.095(.0022)	0

^aThe error rate is computed on the unlabeled data and averaged over 12 splits.

parts. Overall, large improvement occurs for less accurate initial classifiers when they are sufficiently accurate. However, if the initial classifier is too accurate, then the potential for an improvement becomes small or null, as in the cases of SSVM and SPSI for Gaussian kernel learning. If the initial classifier is too poor, then no improvement may occur. This is the case for ESVM, where it performs worse than SVM with $n_l = 10$ labeled data alone. This suggests that a better initial estimate should be used together with unlabeled data.

Furthermore, it is interesting to note that TSVM obtained from SVM^{light} performs worse than SVM with labeled data alone in the WBC example for linear learning, the PIMA example for Gaussian learning and the SPAM example for both linear and Gaussian kernel learning. One possible explanation is that SVM^{light} may not yield a good minimizer for TSVM. This is evident from Wang, Shen and Pan (2007), where DC programming yields a better solution than SVM^{light} in generalization.

In summary, we recommend SPSI to be an initial classifier for $\hat{f}^{(0)}$ based on its overall performance across all the examples. Moreover, ESPSI nearly recovers the classification performance of its counterpart SVM with complete labeled data in the two simulated examples, WBC and HEART.

Table 2: **Gaussian kernel learning.** Averaged test errors as well as estimated standard errors (in parenthesis) of ESVM, ETSVM, ESSVM, ESPSI, and their initial counterparts in the simulated and benchmark examples. Here the amount of improvement is defined in (5) or (6).

Data Size	Example 1 1000 × 2	Example 2 1000 × 2	WBC 682 × 9	PIMA 768 × 8	HEART 303 × 13	MUSHROOM 8124 × 22	SPAM 4601 × 57	BCI ^a 400 × 117
SVM	.385(.0099)	.347(.0119)	.047(.0038)	.342(.0044)	.331(.0094)	.217(.0135)	.226(.0108)	.488(.0073)
ESVM	.368(.0077)	.322(.0109)	.039(.0067)	.335(.0035)	.308(.0107)	.187(.0118)	.212(.0104)	.482(.0076)
Improv.	7.6%	9.7%	17.0%	2.0%	6.9%	13.8%	6.2%	1.2%
TSSVM	.267(.0132)	.258(.0157)	.037(.0015)	.353(.0073)	.331(.0087)	.217(.0117)	.275(.0158)	.492(.0087)
ETSVM	.236(.0090)	.235(.0084)	.030(.0005)	.323(.0028)	.303(.0094)	.201(.0093)	.198(.0106)	.485(.0086)
Improv.	11.6%	13.6%	18.9%	8.5%	8.5%	7.4%	28.0%	1.4%
SSVM	.201(.0072)	.175(.0092)	.030(.0005)	.304(.0044)	.226(.0063)	.173(.0126)	.189(.0120)	.479(.0080)
ESSVM	.201(.0072)	.170(.0083)	.030(.0005)	.304(.0042)	.223(.0054)	.147(.0105)	.170(.0103)	.476(.0085)
Improv.	0.0%	5.8%	0.0%	0.0%	1.3%	15.0%	10.1%	0.6%
SPSI	.200(.0069)	.175(.0092)	.030(.0005)	.295(.0037)	.215(.0057)	.164(.0123)	.189(.0112)	.475(.0072)
ESPSI	.198(.0072)	.169(.0082)	.030(.0005)	.294(.0033)	.215(.0054)	.126(.0083)	.169(.0091)	.475(.0081)
Improv.	1.0%	7.0%	0.0%	0.3%	0.0%	23.2%	10.6%	0.0%
SVM _c	.196(.0015)	.151(.0021)	.030(.0004)	.254(.0013)	.196(.0031)	.021(.0014)	.099(.0018)	0

^aThe error rate is computed on the unlabeled data and averaged over 12 splits.

4.2 Gene function prediction through expression profiles

This section applies the proposed method to predict gene functions through gene data in Hughes et al. (2000), consisting of expression profiles of a total of 6316 genes for yeast *S. cerevisiae* from 300 microarray experiments. In this case almost half of the genes have unknown functions although gene expression profiles are available for almost the entire yeast genome.

Our specific focus is predicting functional categories defined by the MIPS, a multifunctional classification scheme (Mewes et al., 2002). For simplicity, we examine two functional categories, namely “transcriptional control” and “mitochondrion”, with 334 and 346 annotated genes, respectively. The goal is to predict gene functional categories for genes annotated within these two categories by training our semisupervised classifier on expression profiles of genes, where some genes are treated as if their functions are unknown to mimic the semisupervised scenario in complete dataset. At present, detection of novel class is not permitted in our formulation, which remains to be an open research question.

For the purpose of evaluation, we divide the entire dataset into two sets of training and testing. The training set involves a random sample of $n_l = 20$ labeled and $n_u = 380$

unlabeled gene profiles, while the testing set contains 280 remaining profiles.

Table 3: Averaged test errors as well as estimated standard errors (in parenthesis) of ESVM, ETSVM, ESSVM, ESPSI, and their initial counterparts, over 100 pairs of training and testing samples, in gene function prediction. Here I stands for an initial classifier, E stands for our proposed method with the initial method, and the amount of improvement is defined in (5) or (6).

		SVM	TSVM	SSVM	SPSI
Linear	I	.298(.0066)	.321(.0087)	.270(.0075)	.272(.0063)
	E	.278(.0069)	.276(.0080)	.261(.0052)	.252(.0112)
	Improv.	6.7%	14.0%	3.3 %	7.3%
Gaussian	I	.280(.0081)	.323(.0027)	.284(.0111)	.283(.0063)
	E	.279(.0085)	.282(.0076)	.275(.0086)	.256(.0082)
	Improv.	0.4%	12.7%	3.2%	9.5%

As indicated in Table 3, ESVM, ETSVM, ESSVM and ESPSI all improve predictive accuracy of their initial counterparts in linear learning and Gaussian kernel learning. It appears that ESPSI performs best. Most importantly, it demonstrates predictive power of the proposed method for predicting which of the two categories a gene belongs to.

5 Statistical learning theory

This section develops a novel theory for the generalization accuracy of ESPSI \hat{f}_C defined by the ψ -loss in **Algorithm 1**. The accuracy is measured by the Bayesian regret $e(\hat{f}_C, \bar{f}_{.5}) = GE(\hat{f}_C) - GE(\bar{f}_{.5}) \geq 0$ with $GE(f)$ defined in Section 2.2.

5.1 Statistical learning theory

Here error bounds of $e(\hat{f}_C, \bar{f}_{.5})$ are derived in terms of the complexity of the candidate class \mathcal{F} , the sample size n , tuning parameter $\lambda = (nC)^{-1}$, the error rate of the initial classifier $\delta_n^{(0)}$ and the maximum number K of iterations in **Algorithm 1**, which implies that ESPSI, without knowing labels of the unlabeled data, enables to recover the classification accuracy of ψ -learning based on complete data under certain conditions.

We first introduce some notations. Let $L(z) = \psi(z)$ be the ψ -loss. Define the margin loss $V_\pi(f, Z)$ for unequal cost classification to be $S_\pi(y)L(yf(x))$, with cost $0 < \pi < 1$ for

the positive class and $S_\pi(y) = 1 - \pi$ if $y = 1$, and π otherwise. Let $e_{V_\pi}(f, \bar{f}_\pi)E(V_\pi(f, Z) - V_\pi(\bar{f}_\pi, Z)) \geq 0$ for $f \in \mathcal{F}$ with respect to unequal cost π , where $\bar{f}_\pi(x) \text{sign}(f_\pi(x)) = \arg \min_f EV_\pi(f, Z)$ is the Bayes rule, with $f_\pi(x) = p(x) - \pi$.

Assumption A: (Approximation) For any $\pi \in (0, 1)$, there exist some positive sequence $s_n \rightarrow 0$ as $n \rightarrow \infty$ and $f_\pi^* \in \mathcal{F}$ such that $e_{V_\pi}(f_\pi^*, \bar{f}_\pi) \leq s_n$.

Assumption A is an analog of that of Shen et al. (2003), which ensures that the Bayes rule \bar{f}_π can be well approximated by elements in \mathcal{F} .

Assumption B. (Conversion) For any $\pi \in (0, 1)$, there exist constants $0 < \alpha, \beta_\pi < \infty$, $0 \leq \zeta < \infty$, $a_i > 0$; $i = 0, 1, 2$, such that for any sufficiently small $\delta > 0$,

$$\sup_{\{f \in \mathcal{F}: e_{V_{.5}}(f, \bar{f}_{.5}) \leq \delta\}} e(f, \bar{f}_{.5}) \leq a_0 \delta^\alpha, \quad (7)$$

$$\sup_{\{f \in \mathcal{F}: e_{V_\pi}(f, \bar{f}_\pi) \leq \delta\}} \|\text{sign}(f) - \text{sign}(\bar{f}_\pi)\|_1 \leq a_1 \delta^{\beta_\pi}, \quad (8)$$

$$\sup_{\{f \in \mathcal{F}: e_{V_\pi}(f, \bar{f}_\pi) \leq \delta\}} \text{Var}(V_\pi(f, Z) - V_\pi(\bar{f}_\pi, Z)) \leq a_2 \delta^\zeta. \quad (9)$$

Assumption B describes local smoothness of the Bayesian regret $e(f, \bar{f}_{.5})$ in terms of a first-moment function $\|\text{sign}(f) - \text{sign}(\bar{f}_\pi)\|_1$ and a second-moment function $\text{Var}(V_\pi(f, Z) - V_\pi(\bar{f}_\pi, Z))$ relative to $e_{V_\pi}(f, \bar{f}_\pi)$ with respect to unequal cost π . Here the degrees of smoothness are defined by exponents α , β_π and ζ . Note that (7) and (9) have been previously used in Wang and Shen (2007) for quantifying the error rate of a large margin classifier and are necessary conditions of the ‘‘no noise assumption’’ (Tsybakov, 2001); and (8) has been used in Wang, Shen and Liu (2007) for quantifying the error rate of probability estimation, which plays a key role in controlling the error rate of ESPSI. For simplicity, denote $\beta_{.5}$ and $\inf_{\pi \neq 0.5} \{\beta_\pi\}$ as β and γ respectively, where β quantifies the degree to which the positive and negative clusters are distinguishable and γ measures the conversion rate between the classification and probability estimation accuracies.

For Assumption C, we define a complexity measure—the L_2 -metric entropy with brack-

eting, describing the cardinality of \mathcal{F} . Given any $\epsilon > 0$, denote $\{(f_r^l, f_r^u)\}_{r=1}^R$ as an ϵ -bracketing function set of \mathcal{F} if for any $f \in \mathcal{F}$, there exists an r such that $f_r^l \leq f \leq f_r^u$ and $\|f_r^l - f_r^u\|_2 \leq \epsilon; r = 1, \dots, R$. Then the L_2 -metric entropy with bracketing $H_B(\epsilon, \mathcal{F})$ is defined as the logarithm of the cardinality of the smallest ϵ -bracketing function set of \mathcal{F} . See Kolmogorov and Tihomirov (1959) for more details.

Define $\mathcal{F}(k) = \{L(f, z) - L(f_\pi^*, z) : f \in \mathcal{F}, J(f) \leq k\}$ to be a space defined by candidate decision functions, with $J(f) = \frac{1}{2}\|f\|_K^2$. Let $J_\pi^* = \max(J(f_\pi^*), 1)$. In (11), we specify an entropy integral to establish a relationship between the complexity of $\mathcal{F}(k)$ and convergence speed ϵ_n for the Bayesian regret.

Assumption C. (Complexity) For some constants $a_i > 0; i = 3, \dots, 5$ and $\epsilon_n > 0$,

$$\sup_{k \geq 2} \phi(\epsilon_n, k) \leq a_5 n^{1/2}, \quad (10)$$

where $\phi(\epsilon, k) = \int_{a_4 M}^{a_3^{1/2} M^{\min(1, \zeta)/2}} H_B^{1/2}(w, \mathcal{F}(k)) dw / M$, and $M = M(\epsilon, \lambda, k) = \min(\epsilon^2 + \lambda(k/2 - 1)J_\pi^*, 1)$.

Assumption D. (Smoothness of $p(x)$) There exist some constants $0 \leq \eta \leq 1$, $d \geq 0$ and $a_6 > 0$ such that $\|\Delta^j(p)\|_\infty \leq a_6$ for $j = 0, 1, \dots, d$, and $|\Delta^d(p(x_1)) - \Delta^d(p(x_2))| \leq a_6 \|x_1 - x_2\|_1^\eta + d/m$ for any $\|x_1 - x_2\|_1 \leq \delta$ with some sufficiently small $\delta > 0$, where Δ^j is the j -th order difference operator.

Assumption D specifies the degree of smoothness of the conditional density $p(x)$.

Assumption E. (Degree of least favorable situation) There exist some constants $0 \leq \theta \leq \infty$ and $a_7 > 0$ such that $P(X : \min(p(X), 1 - p(X)) \leq \delta) \leq a_7 \delta^\theta$ for any sufficiently small $\delta > 0$.

Assumption E describes the behavior of $p(x)$ near 0 and 1, corresponding to the least favorable situation, as described in Section 2.3.

Theorem 2 *In addition to Assumptions A-E, let the precision parameter m be $[\delta_n^{-\beta\gamma}]$ and*

$\delta_n^2 \min(\max(\epsilon_n^2, 16s_n), 1)$. Then for ESPSI \hat{f}_C , there exist some positive constants a_8 - a_{10} such that

$$\begin{aligned} P\left(e(\hat{f}_C, \bar{f}_{.5}) \geq a_{10} \max(\delta_n^{2\alpha}, (a_{11}\rho_n\delta_n^{(0)})^{2\alpha \max(1, B^K)})\right) &\leq \\ P\left(e_L(\hat{f}_{.5}^{(0)}, \bar{f}_{.5}) \geq 2a_{11}\rho_n(\delta_n^{(0)})^2\right) + 3.5K \exp(-a_8 n_l (\lambda J_{.5}^*)^{\max(1, 2-\zeta)}) + &(11) \\ 3.5K \exp(-a_9 n (\lambda J_{\pi}^*)^{\max(1, 2-\zeta)}) + 2K \rho_n^{-\min(1, \beta)}. & \end{aligned}$$

Here $B = \frac{(\theta+1)(d+\eta)\beta\gamma}{2(1+\max(0, 1-\beta)\theta)(d+\eta+1)}$, $a_{11} = \max\left(1, 2^{\frac{3\gamma(d+\eta)}{d+\eta+1}+2} a_1^{\frac{(2\gamma+1)(d+\eta)}{d+\eta+1}}\right)$ and $\rho_n > 0$ is any real number satisfying $a_{11}\rho_n\delta_n^2 \leq 4\lambda J_{\pi}^*$.

Theorem 2 provides a finite-sample probability bound for the Bayesian regret $e(\hat{f}_C, \bar{f}_{.5})$, where the parameter B measures the level of difficulty of a semisupervised problem, with small value of B indicating more difficulty. Note that the value of B is proportional to those of α , β , γ , d , η and θ , as defined in Assumptions A-E. In fact, α , β and γ quantify the local smoothness of the Bayesian regret $e(f, \bar{f}_{.5})$, and d , η and θ describe the smoothness of $p(x)$ as well as its behavior near 0 and 1.

Next, by letting n_l, n_u tending infinity, we obtain the rates of convergence of ESPSI in terms of the error rate $\delta_n^{2\alpha}$ of its supervised counterpart ψ -learning based on complete data, and the initial error rate $\delta_n^{(0)}$, B , and the maximum number K of iteration.

Corollary 1 Under the assumptions of Theorem 2, as $n_u, n_l \rightarrow \infty$,

$$\begin{aligned} |e(\hat{f}_C, \bar{f}_{.5})| &= O_p\left(\max(\delta_n^{2\alpha}, (\rho_n\delta_n^{(0)})^{2\alpha \max(1, B^K)})\right), \\ E|e(\hat{f}_C, \bar{f}_{.5})| &= O\left(\max(\delta_n^{2\alpha}, (\rho_n\delta_n^{(0)})^{2\alpha \max(1, B^K)})\right), \end{aligned}$$

provided that the initial classifier converges in that $P\left(e_L(\hat{f}_{.5}^{(0)}, \bar{f}_{.5}) \geq 2a_{11}\rho_n(\delta_n^{(0)})^2\right) \rightarrow 0$, with any slow varying sequence $\rho_n \rightarrow \infty$ and $\rho_n\delta_n^{(0)} \rightarrow 0$, and the tuning parameter λ is chosen such that $n(\lambda J_{\pi}^*)^{\max(1, 2-\zeta)}$ and $n_l(\lambda J_{.5}^*)^{\max(1, 2-\zeta)}$ are bounded away from 0.

There are two important cases defined by the value of B . When $B > 1$, ESPSI achieves the convergence rate $\delta_n^{2\alpha}$ of its supervised counterpart ψ -learning based on complete data, c.f., Theorem 1 of Shen and Wang (2003). When $B \leq 1$, ESPSI performs no worse than its initial classifier because $(\delta_n^{(0)})^{2\alpha \max(1, B^K)} (\delta_n^{(0)})^{2\alpha}$. Therefore, it is critical to compute the value of B . For instance, if the two classes are perfectly separated and located very densely within respective regions, then $B = \infty$ and our method recovers the rate $\delta_n^{2\alpha}$; if the two classes are totally indistinguishable, then $B = 0$ and our method yields the rate $(\delta_n^{(0)})^2$.

For the optimality claimed in Section 2.2, we show that $\hat{U}(f)$ is sufficiently close to $U(f)$ so that the ideal optimality of $U(f)$ can be translated into $\hat{U}(f)$. As a result, minimization of $\hat{U}(f)$ over f mimics that of $U(f)$.

Corollary 2 (Optimality) *Under the assumptions of Corollary 1, as $n_u, n_l \rightarrow \infty$,*

$$\sup_{f \in \mathcal{F}} \|\hat{U}(f) - U(f)\|_1 = O_p\left(\max(\delta_n^{\beta\gamma}, (\rho_n \delta_n^{(0)})^{\beta\gamma \max(1, B^K)})\right),$$

where $\hat{U}(f)$ is estimated $U(f)$ loss with p estimated based on \hat{f}_C .

To argue that the approximation error rate of $\hat{U}(f)$ to $U(f)$ is sufficiently small, note that \hat{f}_C obtained from minimizing (2) recovers the classification error rate of its supervised counterpart based on complete data, as suggested by Corollary 1. Otherwise, a poor approximation precision could impede the error rate of ESPSI.

In conclusion, ESPSI, without knowing label values of unlabeled instances, enables to reconstruct the classification and estimation performance of ψ -learning based on complete data in rates of convergence.

5.2 Theoretical example

We now apply Corollary 1 to linear and kernel learning examples to derive generalization errors rates for ESPSI in terms of the Bayesian regret. In all cases, ESPSI (nearly) achieves the generalization error rates of ψ -learning for complete data when unlabeled data provides

useful information, and yields no worse performance than its initial classifier otherwise. Technical details are deferred to Appendix B.

Consider a learning example in which $X = (X_{.1}, X_{.2})$ are independent, following marginal distribution $q(x) = \frac{1}{2}(\kappa_1 + 1)|x|^{\kappa_1}$ for $x \in [-1, 1]$ for $\kappa_1 > 0$. Given $X = 1$, $P(Y = 1|X = x) = p(x) = \frac{2}{5} \text{sign}(x_{.1})|x_{.1}|^{\kappa_2} + \frac{1}{2}$ with $\kappa_2 > 0$. Note that $f_\pi(x)$ is $x_{.1} - \text{sign}(\pi - \frac{1}{2})(\frac{5}{4}|2\pi - 1|)^{\frac{1}{\kappa_2}}$, which in turn yields the vertical line as the decision boundary for classification with unequal cost π . The value of κ_i ; $i = 1, 2$ describe the behavior of the marginal distribution around the origin, and that of the conditional distribution $p(x)$ in the neighborhood of $1/2$, respectively.

Linear learning: Here it is natural to consider linear learning in which candidate decision functions are linear in $\mathcal{F} = \{f(x) = (1, x^T)w : w \in \mathcal{R}^3, x = (x_{.1}, x_{.2}) \in \mathcal{R}^2\}$.

For ESPSI \hat{f}_C , we choose $\delta_n^{(0)} = n_l^{-1} \log n_l$ to be the convergence rate of supervised linear ψ -learning, $\rho_n \rightarrow \infty$ to be an arbitrarily slow sequence and tuning parameter $C = O((\log n)^{-1})$. An application of Corollary 1 yields that $E|e(\hat{f}_C, \bar{f}_{.5})| = O(\max(n^{-1} \log n, (n_l^{-1} (\log n_l)^2)^{\max(1, 2B^K)}))$, with $B = \frac{(1+\kappa_1)^2}{2\kappa_2(1+\kappa_1+\kappa_2)}$. When $B > 1$, equivalently, $\kappa_1 + 1 > (1 + \sqrt{3})\kappa_2$, this rate reduces to $O(n^{-1} \log n)$ when K is sufficiently large. Otherwise, the rate is $O(n_l^{-1} \log n_l)$.

The fast rate $n^{-1} \log n$ is achieved when κ_1 is large but κ_2 is relatively small. Interestingly, large κ_1 value implies that $q(x)$ has a low density around $x = 0$, corresponding to the low density separation assumption in Chapelle and Zien (2005) for a semisupervised problem, whereas large κ_1 value and small κ_2 value indicates that $p(x)$ has a small probability to be close to the decision boundary $p(x) = 1/2$ for a supervised problem.

Kernel learning: Consider a flexible representation defined by a Gaussian kernel, where $\mathcal{F} = \{x \in \mathcal{R}^2 : f(x)w_{f,0} + \sum_{k=1}^n w_{f,k}K(x, x_k) : w_f = (w_{f,1}, \dots, w_{f,n})^T \in \mathcal{R}^n\}$ by the representation theorem of RKHS, c.f., Wahba (1990). Here $K(x, z) = \exp(-\frac{\|x-z\|^2}{2\sigma^2})$ is the Gaussian kernel.

Similarly, we choose $\delta_n^{(0)} = n_l^{-1}(\log n_l)^3$ to be the convergence rate of supervised ψ -learning with Gaussian kernel, $\rho_n \rightarrow \infty$ to be an arbitrarily slow sequence and $C = O((\log n)^{-3})$. According to Corollary 1, $E|e(\hat{f}_C, \bar{f}_5)| = O(\max((n_l^{-1}(\log n_l)^3)^{\max(1, 2B^K)}, n^{-1}(\log n)^3,)) = O(n^{-1}(\log n)^3)$ when $\kappa_1 + 1 > 2\kappa_2(1 + \kappa_2)$ and K is sufficiently large, and $O(n_l^{-1}(\log n_l)^3)$ otherwise. Again, large κ_1 and small κ_2 lead to the fast rate.

6 Summary

This article introduces a large margin semisupervised learning method through an iterative scheme based on an efficient loss for unlabeled data. In contrast to most methods assuming a relationship between the conditional and the marginal distributions, the proposed method integrates labeled and unlabeled data through utilizing the clustering structure of unlabeled data as well as the smoothness structure of the estimated p . The theoretical and numerical results suggest that the method compares favorably against top competitors, and achieves the desired goal of reconstructing the classification performance of its supervised counterpart on complete labeled data.

With regard to tuning parameter C , further investigation is necessary. One critical issue is how to utilize unlabeled data to enhance the accuracy of estimating the generalization error so that adaptive tuning is possible.

Appendix A: Technical Proofs

Proof of Lemma 1: Let $U(f(x)) = E(L(Yf(X))|X = x)$. By orthogonality, we have $E(L(Yf(X)) - T(f(X)))^2 = E(L(Yf(X)) - U(f(X)))^2 + E(U(f(X)) - T(f(X)))^2$, implying that $U(f(x))$ minimizes $E(L(Yf(X)) - T(f(X)))^2$ over any T . Then the proof follows from the fact that $EL(Yf(X)) = E(E(L(Yf(X))|X))$.

Proof of Theorem 1: For clarity, we write $s(\hat{f})$ as $s(\hat{f}, \hat{p})$ in this proof. Then it suffices

to show that $s(\hat{f}^{(k)}, \hat{p}^{(k)}) \geq s(\hat{f}^{(k+1)}, \hat{p}^{(k+1)})$. First, $s(\hat{f}^{(k)}, \hat{p}^{(k)}) \geq s(\hat{f}^{(k+1)}, \hat{p}^{(k)})$ since $\hat{f}^{(k+1)}$ minimizes $s(f, \hat{p}^{(k)})$. Then $s(\hat{f}^{(k+1)}, \hat{p}^{(k)}) - s(\hat{f}^{(k+1)}, \hat{p}^{(k+1)}) = \sum_{j=n_l+1}^n (\hat{p}^{(k)} - \hat{p}^{(k+1)})(L(\hat{f}^{(k+1)}(x_j)) - L(-\hat{f}^{(k+1)}(x_j)))$, which is nonnegative by the definition of $\hat{p}^{(k+1)}$.

Proof of Theorem 2: The proof involves two steps. In **Step 1**, given $f^{(k)}$, we derive a probability upper bound for $\|\hat{p}^{(k)} - p\|_1$, where $\hat{p}^{(k)}$ is obtained from **Algorithm 0**. In **Step 2**, based on the result of **Step 1**, the difference between the tail probability of $e_\pi(\hat{f}_\pi^{(k+1)}, \bar{f}_\pi)$ and that of $e_\pi(\hat{f}_\pi^{(k)}, \bar{f}_\pi)$ is bounded through a large deviation inequality of Wang and Shen (2007); $k = 0, 1, \dots$. This in turn results in a faster rate for $e(\hat{f}_{.5}^{(k+1)}, \bar{f}_{.5})$, thus $e(\hat{f}_C, \bar{f}_{.5})$.

Step 1: First we bound the probability of the percentage of wrongly labeled unlabeled instances by $\text{sign}(\hat{f}^{(k)})$ by the tail probability of $e_{V_{.5}}(\hat{f}^{(k)}, \bar{f}_{.5})$. For this purpose, let $D_f = \{\text{sign}(\hat{f}^{(k)}(X_j)) \neq \text{sign}(\bar{f}_{.5}(X_j)); n_l + 1 \leq j \leq n\}$ be the set of unlabeled data that are wrongly labeled by $\text{sign}(\hat{f}^{(k)})$, with $n_f = \#\{D_f\}$ being its cardinality. By Markov's inequality, the fact that $E(\frac{n_f}{n}) = \frac{n_u}{n} E\|\text{sign}(\hat{f}^{(k+1)}) - \text{sign}(\bar{f}_{.5})\|_1$, and (8),

$$\begin{aligned} P\left(\frac{n_f}{n} \geq a_1(a_{11}\rho_n^2(\delta_n^{(k)})^2)^\beta\right) &\leq P\left(\|\text{sign}(\hat{f}^{(k)}) - \text{sign}(\bar{f}_{.5})\|_1 \geq a_1(a_{11}\rho_n(\delta_n^{(k)})^2)^\beta\right) \\ &\quad + P\left(\frac{n_f}{n} \geq \rho_n^\beta \|\text{sign}(\hat{f}^{(k+1)}) - \text{sign}(\bar{f}_{.5})\|_1\right) \\ &\leq P\left(e_{V_{.5}}(\hat{f}^{(k)}, \bar{f}_{.5}) \geq a_{11}\rho_n(\delta_n^{(k)})^2\right) + \rho_n^{-\beta}. \end{aligned} \quad (12)$$

Next we bound the tail probability of $\|\hat{p}^{(k)} - p\|_1$ based on ‘‘complete’’ data consisting of unlabeled data assigned by $\text{sign}(\hat{f}^{(k)})$. An application of a treatment similar to that in the proof of Theorem 3 of Wang, Shen and Liu (2007) leads to

$$\begin{aligned} P(\|\hat{p}^{(k)} - p\|_1 \geq 8^\gamma a_1^{2\gamma+1} (a_{11}\rho_n \delta_n^{(k)})^{\beta\gamma}) &\leq \\ P(\exists j : \|\text{sign}(\hat{f}_{\pi_j}^{(k)}) - \text{sign}(\bar{f}_{\pi_j})\|_1 \geq 8^\gamma a_1^{2\gamma+1} (a_{11}\rho_n \delta_n^{(k)})^{2\beta\gamma}), \end{aligned} \quad (13)$$

with $\pi_j = j / \lceil (a_{11}\rho_n \delta_n^{(k)})^{-\beta\gamma} \rceil$. According to (8), it suffices to bound $P(e_{V_\pi}(\hat{f}_{\pi_j}^{(k)}, \bar{f}_{\pi_j}) \geq 8^\gamma a_1^{2\gamma+1} (a_{11}\rho_n \delta_n^{(k)})^{2\beta})$ for all π_j in what follows.

We now introduce some notations to be used. Let $\tilde{V}_\pi(f, z) = V_\pi(f, z) + \lambda J(f)$, and $Z_j = (X_j, Y_j)$ with $Y_j = \text{sign}(\hat{f}^{(k)}(X_j))$; $n_l + 1 \leq j \leq n$. Define a scaled empirical process $E_n(\tilde{V}_\pi(f_\pi^*, Z) - \tilde{V}_\pi(f, Z)) = n^{-1} \left(\sum_{i \in D_f} + \sum_{i \notin D_f} \right) \left(\tilde{V}_\pi(f_\pi^*, Z_i) - \tilde{V}_\pi(f, Z_i) - E(\tilde{V}_\pi(f_\pi^*, Z_i) - \tilde{V}_\pi(f, Z_i)) \right) \equiv E_n(V_\pi(f_\pi^*, Z) - V_\pi(f, Z))$.

By the definition of $\hat{f}_\pi^{(k)}$ and (12),

$$\begin{aligned} P\left(e_{V_\pi}(\hat{f}_\pi^{(k)}, \bar{f}_\pi) \geq \delta_k^2\right) &\leq P\left(\frac{n_f}{n} \geq a_1(a_{11}\rho_n^2(\delta_n^{(k)})^2)^\beta\right) + P^*\left(\sup_{N_k} \frac{1}{n} \sum_{i=1}^n (\tilde{V}_\pi(f_\pi^*, Z_i) - \tilde{V}_\pi(f, Z_i)) \geq 0, \frac{n_f}{n} \leq a_1(a_{11}\rho_n^2(\delta_n^{(k)})^2)^\beta\right) \\ &\leq P\left(e_{V_{.5}}(\hat{f}_\pi^{(k)}, \bar{f}_{.5}) \geq a_{11}\rho_n(\delta_n^{(k)})^2\right) + \rho_n^{-\beta} + I_1, \end{aligned} \quad (14)$$

where $N_k = \{f \in \mathcal{F} : e_{V_\pi}(f, \bar{f}_\pi) \geq \delta_k^2\}$, $I_1 = P^*\left(\sup_{N_k} E_n(V_\pi(f_\pi^*, Z) - V_\pi(f, Z)) \geq \inf_{N_k} \nabla(f, f_\pi^*), \frac{n_f}{n} \leq a_1(a_{11}\rho_n^2(\delta_n^{(k)})^2)^\beta\right)$, $\nabla(f, f_\pi^*) = \frac{n_f}{n} E_{i \in D_f}(\tilde{V}_\pi(f, Z_i) - \tilde{V}_\pi(f_\pi^*, Z_i)) + \frac{n-n_f}{n} E_{i \notin D_f}(\tilde{V}_\pi(f, Z_i) - \tilde{V}_\pi(f_\pi^*, Z_i))$, and $\delta_k^2 = 8a_1^2(a_{11}\rho_n\delta_n^{(k)})^{2\beta}$.

To bound I_1 , we partition N_k into a union of $A_{s,t}$ with $A_{s,t} = \{f \in \mathcal{F} : 2^{s-1}\delta_k^2 \leq e_{V_\pi}(f, \bar{f}_\pi) < 2^s\delta_k^2, 2^{t-1}J_\pi^* \leq J(f) < 2^tJ_\pi^*\}$ and $A_{s,0} = \{f \in \mathcal{F} : 2^{s-1}\delta_k^2 \leq e_{V_\pi}(f, \bar{f}_\pi) < 2^s\delta_k^2, J(f) < J_\pi^*\}$; $s, t = 1, 2, \dots$. Then it suffices to bound the corresponding probability over each $A_{s,t}$. Toward this end, we need to bound the first and second moments of $\tilde{V}_\pi(f, Z) - \tilde{V}_\pi(f_\pi^*, Z)$ over $f \in A_{s,t}$. Without loss of generality, assume that $4s_n < \delta_k^2 < 1$, $J(f_\pi^*) \geq 1$, and thus $J_\pi^* = \max(J(f_\pi^*), 1) = J(f_\pi^*)$.

For the first moment, note that $\nabla(f, f_\pi^*) \geq e_{V_\pi}(f, f_\pi^*) - \frac{n_f}{n} E|V_\pi(f, Z) - V_\pi(f_\pi^*, Z) + \bar{V}_\pi(f, Z) - \bar{V}_\pi(f_\pi^*, Z)| \geq e_{V_\pi}(f, f_\pi^*) - 4\frac{n_f}{n}$ with $\bar{V}_\pi(f, z) = S_\pi(-y)L(-yf(x))$. Using the assumption that $4\lambda J(f_\pi^*) \leq \delta_k^2$, and Assumptions A and B, we obtain

$$\begin{aligned} \inf_{A_{s,t}} \nabla(f, f_\pi^*) &\geq M(s, t) = (2^{s-1} - 1/2)\delta_k^2 + \lambda(2^{t-1} - 1)J(f_\pi^*), \\ \inf_{A_{s,0}} \nabla(f, f_\pi^*) &\geq (2^{s-1} - 3/4)\delta_k^2 \geq M(s, 0) = 2^{s-3}\delta_k^2. \end{aligned}$$

For the second moment, by Assumptions A and B and $|\bar{V}_\pi(f, Z) - \bar{V}_\pi(\bar{f}_\pi, Z)| \leq 2$ for

any $0 < \pi < 1$, we have, for any $s, t = 1, 2, \dots$ and some constant $a_3 > 0$,

$$\begin{aligned}
& \sup_{A_{s,t}} \text{Var}(V_\pi(f, Z) - V_\pi(f_\pi^*, Z)) \\
& \leq \sup_{A_{s,t}} \frac{2(n - n_f)}{n} \left(\text{Var}(V_\pi(f, Z) - V_\pi(\bar{f}_\pi, Z)) + \text{Var}(V_\pi(f_\pi^*, Z) - V_\pi(\bar{f}_\pi, Z)) \right) + \\
& \quad \frac{2n_f}{n} \left(\text{Var}(\bar{V}_\pi(f, Z) - \bar{V}_\pi(\bar{f}_\pi, Z)) + \text{Var}(\bar{V}_\pi(f_\pi^*, Z) - \bar{V}_\pi(\bar{f}_\pi, Z)) \right) \\
& \leq 2a_2 M(s, t)^\zeta + 8a_1 (a_{11} \rho_n^2 (\delta_n^{(k)})^2)^\beta + 4s_n \leq a_3 M(s, t)^{\min(1, \zeta)} = v^2(s, t),
\end{aligned}$$

Note that $I_1 \leq I_2 + I_3$ with $I_2 = \sum_{s,t=1}^\infty P^*(\sup_{A_{s,t}} E_n(V_\pi(f_\pi^*, Z) - V_\pi(f, Z))) \geq M(s, t)$, $\frac{n_f}{n} \leq a_1 (a_{11} \rho_n^2 (\delta_n^{(k)})^2)^\beta$, and $I_3 = \sum_{s=1}^\infty P^*(\sup_{A_{s,0}} E_n(V_\pi(f_\pi^*, Z) - V_\pi(f, Z))) \geq M(s, 0)$, $\frac{n_f}{n} \leq a_1 (a_{11} \rho_n^2 (\delta_n^{(k)})^2)^\beta$. Then we bound I_2 and I_3 separately using Lemma 1 of Wang, Shen and Pan (2007). For I_2 , we verify the conditions (8)-(10) there. Note that $\int_{aM(s,t)}^{v(s,t)} H_B^{1/2}(w, \mathcal{F}(2^w)) dw / M(s, t)$ is non-increasing in s and $M(s, t)$, we have

$$\int_{aM(s,t)}^{v(s,t)} H_B^{1/2}(w, \mathcal{F}(2^t)) dw / M(s, t) \leq \int_{a_3 M(1,t)}^{aM(1,t)^{\min(1, \zeta)/2}} H_B^{1/2}(w, \mathcal{F}(2^t)) dw / M(1, t),$$

which is bounded by $\phi(\epsilon_n^2, 2^t)$ with $a = 2a_4 \epsilon$ and $\epsilon_n^2 \leq \delta_k^2$. Then Assumption C implies (8)-(10) there with $\epsilon = 1/2$, the choices of $M(s, t)$ and $v(s, t)$ and some constants $a_i > 0$; $i = 3, 4$. It then follows that for some constant $0 < \xi < 1$,

$$\begin{aligned}
I_2 & \leq \sum_{s,t=1}^\infty 3 \exp\left(-\frac{(1 - \xi)n(M(s, t))^2}{2(4(v(s, t))^2 + 2M(s, t)/3)}\right) \\
& \leq \sum_{s,t=1}^\infty 3 \exp(-a_8 n (M(s, t))^{\max(1, 2-\zeta)}) \\
& \leq \sum_{s,t=1}^\infty 3 \exp(-a_8 n (2^{s-1} \delta_k^2 + \lambda(2^{t-1} - 1) J_\pi^*)^{\max(1, 2-\zeta)}) \\
& \leq 3 \exp(-a_8 n (\lambda J_\pi^*)^{\max(1, 2-\zeta)}) / (1 - \exp(-a_8 n (\lambda J_\pi^*)^{\max(1, 2-\zeta)}))^2.
\end{aligned}$$

Similarly $I_3 \leq 3 \exp(-a_8 n (\lambda J_\pi^*)^{\max(1, 2-\zeta)}) / (1 - \exp(-a_8 n (\lambda J_\pi^*)^{\max(1, 2-\zeta)}))^2$. Combining

the bounds for I_i ; $i = 2, 3$, we have $I_1 \leq 3.5 \exp(-a_8 n (\lambda J_\pi^*)^{\max(1, 2-\zeta)})$. Consequently, by (8), (13) and (14)

$$\begin{aligned} P\left(\|\hat{p}^{(k)} - p\|_1 \geq 8^\gamma a_1^{2\gamma+1} (a_{11} \rho_n \delta_n^{(k)})^{\beta\gamma}\right) \leq \\ P\left(e_{V_{.5}}(\hat{f}_{.5}^{(k)}, \bar{f}_{.5}) \geq a_{11} \rho_n (\delta_n^{(k)})^2\right) + \rho_n^{-\beta} + 3.5 \exp(-a_8 n (\lambda J_\pi^*)^{\max(1, 2-\zeta)}). \end{aligned} \quad (15)$$

Step 2: To begin, note that $P\left(e_{V_{.5}}(\hat{f}_{.5}^{(k+1)}, \bar{f}_{.5}) \geq a_{11} \rho_n (\delta_n^{(k+1)})^2\right) \leq I_4 + I_5$ with $I_4 = P\left(e_{V_{.5}}(\hat{f}_{.5}^{(k+1)}, \bar{f}_{.5}) \geq a_{11} \rho_n (\delta_n^{(k+1)})^2 \|\hat{p}^{(k)} - p\|_1 < 8^\gamma a_1^{2\gamma+1} (a_{11} \rho_n \delta_n^{(k)})^{\beta\gamma}\right)$, and $I_5 = P\left(\|\hat{p}^{(k)} - p\|_1 \geq 8^\gamma a_1^{2\gamma+1} (a_{11} \rho_n \delta_n^{(k)})^{\beta\gamma}\right)$, where $a_{12} = 2a_6^{1/(d+\eta+1)} (4a_7)^{-\frac{1}{\theta}}$ and $a_{11} \rho_n (\delta_n^{(k+1)})^2 = (a_{12} (a_{11} \rho_n \delta_n^{(k)})^{\frac{(d+\eta)\beta\gamma}{d+\eta+1}})^{\frac{\theta+1}{1+\max(0, 1-\beta)\theta}}$. By (15), it suffices to bound I_4 .

For I_4 , we need some notations. Let the ideal cost function be $V_{.5}(f, z) + U_{.5}(f(x))$, the ideal version of (2), where $V_{.5}(f, z) = \frac{1}{2}L(yf(x))$, and $U_{.5}(f(x)) = \frac{1}{2}(p(x)L(f(x)) + (1-p(x))L(-f(x)))$ is the ideal optimal loss for unlabeled data. Denote by $\hat{U}_{.5}^{(k)}(f(x)) = \frac{1}{2}(\hat{p}^{(k)}(x)L(f(x)) + (1-\hat{p}^{(k)}(x))L(-f(x)))$ an estimate of $U_{.5}(f(x))$ at **Step** k . Therefore the regularized cost function in (2) can be written as $\tilde{W}(f, z) = W(f, z) + \lambda J(f)$ with $W(f, z) = V_{.5}(f, z) + U_{.5}(f(x))$. For simplicity, denote the weighted empirical process as $E_n(W(f_{.5}^*, z) - W(f, z)) = n_l^{-1} \sum_{i=1}^{n_l} (V_{.5}(f_{.5}^*, Z_i) - V_{.5}(f, Z_i) - E(V_{.5}(f_{.5}^*, Z) - V_{.5}(f, Z))) + n_u^{-1} \sum_{j=n_l+1}^n (U_{.5}(f_{.5}^*(X_j)) - U_{.5}(f(X_j)) - E(U_{.5}(f_{.5}^*(X)) - U_{.5}(f(X))))$.

By definition of $\hat{f}_{.5}^{(k+1)}$, we have $I_4 \leq P\left(\sup_{N'_k} n_l^{-1} \sum_{i=1}^{n_l} (V_{.5}(f_{.5}^*, Z_i) - V_{.5}(f, Z_i)) + n_u^{-1} \sum_{j=n_l+1}^n (\hat{U}_{.5}^{(k)}(f_{.5}^*(X_j)) - \hat{U}_{.5}^{(k)}(f(X_j))) + \lambda(J(f_{.5}^*) - J(f)) \geq 0, \|\hat{p}^{(k)} - p\|_1 < 8^\gamma a_1^{2\gamma+1} (a_{11} \rho_n \delta_n^{(k)})^{\beta\gamma}\right)$, where $N'_k = \{f \in \mathcal{F} : e_{V_{.5}}(f, \bar{f}_{.5}) \geq a_{11} \rho_n (\delta_n^{(k+1)})^2\}$. Then $I_4 \leq I_6 + I_7$ with

$$\begin{aligned} I_6 = P\left(\sup_{N'_k} n_u^{-1} \sum_{j=n_l+1}^n D(f, X_j) \geq \right. \\ \left. 8^{\frac{\gamma(d+\eta)}{d+\eta+1}} a_1^{\frac{(2\gamma+1)(d+\eta)}{d+\eta+1}} \rho_n (e_{V_{.5}}(f, f_{.5}^*))^{\frac{\min(1, \beta)\theta}{\theta+1}} (a_{11} \rho_n (\delta_n^{(k+1)})^2)^{\frac{1+\max(0, 1-\beta)\theta}{\theta+1}}\right), \end{aligned}$$

$$I_7 = P\left(\sup_{N'_k} E_n(W(f_{.5}^*, Z) - W(f, Z)) \geq \inf_{N'_k} E(\tilde{W}(f, Z) - \tilde{W}(f_{.5}^*, Z)) - 8 \frac{\gamma(d+\eta)}{d+\eta+1} a_1 \frac{(2\gamma+1)(d+\eta)}{d+\eta+1} \rho_n(e_{V_{.5}}(f, f_{.5}^*)) \frac{\min(1, \beta)\theta}{\theta+1} (a_{11}\rho_n(\delta_n^{(k+1)})^2)^{\frac{1+\max(0, 1-\beta)\theta}{\theta+1}}\right),$$

where $D(f, X_j) = \hat{U}_{.5}^{(k)}(f_{.5}^*(X_j)) - \hat{U}_{.5}^{(k)}(f(X_j)) - U_{.5}(f_{.5}^*(X_j)) + U_{.5}(f(X_j))$.

For I_6 , we note that $E|D(f, X)| = \frac{1}{2}E|\hat{p}^{(k)}(X) - p(X)| |L(f_{.5}^*(X)) - L(f(X)) - L(-f_{.5}^*(X)) + L(-f(X))| \leq \frac{1}{2}\|\hat{p}^{(k)} - p\|_\infty E(|L(f_{.5}^*(X)) - L(f(X))| + |L(-f_{.5}^*(X)) - L(-f(X))|)$. We thus bound $\|\hat{p}^{(k)} - p\|_\infty$ and $E(|L(f_{.5}^*(X)) - L(f(X))| + |L(-f_{.5}^*(X)) - L(-f(X))|)$ separately.

To bound $\|\hat{p}^{(k)} - p\|_\infty$, note that $EJ(\hat{f}_{\pi_j}^{(k)})$ is bounded for all π_j following the same argument as in Lemma 5 of Wang and Shen (2007). By Sobolev's interpolation theorem (Adams, 1975), $\|\hat{p}^{(k)}\|_\infty \leq 1$ and the fact that $|\hat{p}^{(k,d)}(x_1) - \hat{p}^{(k,d)}(x_2)| \leq \sup_j |\hat{f}_{\pi_j}^{(k,d)}(x_1) - \hat{f}_{\pi_j}^{(k,d)}(x_2)| + d(m^{(k)})^{-1}$ for any x_1 and x_2 based on the construction of $\hat{p}^{(k)}$ with $\hat{f}_{\pi_j}^{(k,d)} = \Delta^d(\hat{f}_{\pi_j}^{(k)})$, there exists a constant a_{13} such that $\|\hat{p}^{(k,d)}\|_\infty \leq a_{13}$ and $|\hat{p}^{(k,d)}(x_1) - \hat{p}^{(k,d)}(x_2)| \leq a_{13}|x_1 - x_2|^\eta + d(m^{(k)})^{-1}$ when $|x_1 - x_2|$ is sufficiently small. Without loss of generality, we assume $a_{13} \leq a_6$. By Assumption D and $m^{(k)} = \lceil (a_{11}\rho_n\delta_n^{(k)})^{-\beta\gamma} \rceil$, we have $|\hat{p}^{(k,d)}(x_1) - p^{(d)}(x_1)| - |\hat{p}^{(k,d)}(x_2) - p^{(d)}(x_2)| \leq |\hat{p}^{(k,d)}(x_1) - \hat{p}^{(k,d)}(x_2)| + |p^{(d)}(x_1) - p^{(d)}(x_2)| \leq 2a_6|x_1 - x_2|^\eta + 2d(a_{11}\rho_n\delta_n^{(k)})^{\frac{\beta\gamma(d+\eta)}{d+\eta+1}}$. It then follows from Proposition 6 of Shen (1997) that $\|\hat{p}^{(k)} - p\|_\infty \leq 2a_6 \frac{1}{d+\eta+1} \|\hat{p}^{(k)} - p\|_1^{\frac{d+\eta}{d+\eta+1}} + 2d(a_{11}\rho_n\delta_n^{(k)})^{\frac{\beta\gamma(d+\eta)}{d+\eta+1}} \leq 2a_6 \frac{1}{d+\eta+1} (8^\gamma a_1^{2\gamma+1})^{\frac{d+\eta}{d+\eta+1}} (a_{11}\rho_n\delta_n^{(k)})^{\frac{\beta\gamma(d+\eta)}{d+\eta+1}}$.

To bound $E(|L(f_{.5}^*(X)) - L(f(X))| + |L(-f_{.5}^*(X)) - L(-f(X))|)$, we note that $E|V_{.5}(f, Z) - V_{.5}(f_{.5}^*, Z)| = \frac{1}{2}E(p(X)|L(f_{.5}^*) - L(f)| + (1-p(X))|L(-f) - L(-f_{.5}^*)|) \geq E\frac{\delta}{2}(|L(f_{.5}^*) - L(f)| + |L(-f) - L(-f_{.5}^*)|)I(\min(p(X), 1-p(X)) \geq \delta) \geq \delta(E\frac{1}{2}(|L(f_{.5}^*) - L(f)| + |L(-f) - L(-f_{.5}^*)|) - 2a_7\delta^\theta)$ by Assumption E. With $\delta = (E(|L(f_{.5}^*) - L(f)| + |L(-f) - L(-f_{.5}^*)|)/8a_7)^{1/\theta}$, it yields that $E(|L(f_{.5}^*) - L(f)| + |L(-f) - L(-f_{.5}^*)|) \leq (4a_7)^{-1/\theta} (E|V_{.5}(f, Z) - V_{.5}(f_{.5}^*, Z)|)^{\theta/\theta+1}$, where $E|V_{.5}(f, Z) - V_{.5}(f_{.5}^*, Z)| \leq E|V_{.5}(f, Z) - V_{.5}(\bar{f}_{.5}, Z)| + E|V_{.5}(f_{.5}^*, Z) - V_{.5}(\bar{f}_{.5}, Z)| \leq P(\text{sign}(f) \neq \text{sign}(\bar{f}_{.5})) + E(V_{.5}(f, Z) -$

$V_{.5}(\bar{f}_{.5}, Z)) + P(\text{sign}(f_{.5}^*) \neq \text{sign}(\bar{f}_{.5})) + E(V_{.5}(f_{.5}^*, Z) - V_{.5}(\bar{f}_{.5}, Z)) \leq (e_{V_{.5}}(f, f_{.5}^*))^\beta + e_{V_{.5}}(f, f_{.5}^*) + s_n^\beta + s_n \leq 4(e_{V_{.5}}(f, f_{.5}^*))^{\min(1, \beta)}$ by Assumptions A and B. Therefore, we have $E|D(f, X)| \leq 8 \frac{\beta(d+\eta)}{d+\eta+1} a_1 \frac{(2\beta+1)(d+\eta)}{d+\eta+1} (e_{V_{.5}}(f, f_{.5}^*))^{\frac{\min(1, \beta)\theta}{\theta+1}} (a_{11}\rho_n(\delta_n^{(k+1)})^2)^{\frac{1+\max(0, 1-\beta)\theta}{\theta+1}}$ and $I_6 \leq \rho_n^{-1}$ by Markov's inequality.

To bound I_7 , we apply a similar treatment as in bounding I_1 in **Step 1** to yield that $I_7 \leq 3.5 \exp(-a_8 n_l (\lambda J_{.5}^*)^{\max(1, 2-\zeta)})$. Adding the upper bounds of I_6 and I_7 , $P(e_{V_{.5}}(\hat{f}_{.5}^{(k+1)}, \bar{f}_{.5}) \geq a_{11}\rho_n(\delta_n^{(k+1)})^2) \leq P(e_{V_{.5}}(\hat{f}_{.5}^{(k)}, \bar{f}_{.5}) \geq a_{11}\rho_n(\delta_n^{(k)})^2) + 3.5 \exp(-a_9 n (\lambda J_{.5}^*)^{\max(1, 2-\zeta)}) + 3.5 \exp(-a_8 n_l (\lambda J_{.5}^*)^{\max(1, 2-\zeta)}) + \rho_n^{-\beta} + \rho_n^{-1}$. Iterating this inequality yields that

$$\begin{aligned} & P\left(e_{V_{.5}}(\hat{f}_{.5}^{(K)}, \bar{f}_{.5}) \geq (a_{12}^{\frac{2B(d+\eta+1)}{\beta\gamma(d+\eta)}} (a_{11}\rho_n)^{2B-1})^{\frac{B^{K+1}-1}{B-1}} (\delta_n^{(0)})^{2B^K}\right) \\ & \leq P\left(e_{V_{.5}}(\hat{f}_{.5}^{(0)}, \bar{f}_{.5}) \geq a_{11}\rho_n(\delta_n^{(0)})^2\right) + 3.5K \exp(-a_8 n_l (\lambda J_{.5}^*)^{\max(1, 2-\zeta)}) + \\ & \quad 3.5K \exp(-a_9 n (\lambda J_{.5}^*)^{\max(1, 2-\zeta)}) + K\rho_n^{-\beta} + K\rho_n^{-1}, \end{aligned}$$

where $B = \frac{(\theta+1)(d+\eta)\beta\gamma}{2(1+\max(0, 1-\beta)\theta)(d+\eta+1)}$. Then the desired result follows from Assumption B and the fact that $\delta_k^2 = 8a_1^2 (a_{11}\rho_n \delta_n^{(k)})^{2\beta} \geq \max(\epsilon_n^2, 16s_n) = \delta_n^2$ for any k .

Proof of Corollary 1: It follows from Theorem 1 immediately and the proof is omitted.

Proof of Corollary 2: It follows from (15) and Corollary 1 that $\|\hat{p}_C - p\|_1 = O_p\left(\max(\delta_n^{\beta\gamma}, (\rho_n \delta_n^{(0)})^{\max(1, 1\beta\gamma B^K)})\right)$, where \hat{p}_C is the estimated probability through \hat{f}_C . The desired results follows from the fact that $\|\hat{U}_C(f) - U(f)\|_1 \leq 4\|\hat{p}_C - p\|_1$.

Appendix B: Verification of Assumptions in Section 5.2

Linear learning: Note that $(X_{.1}, Y)$ is independent of $X_{.2}$, which implies $ES(f; C) = E(E(S(f; C)|X_{.2})) \geq ES(\tilde{f}_C^*; C)$ for any $f \in \mathcal{F}$, where $\tilde{f}_C^* = \arg \min_{\tilde{f} \in \mathcal{F}_1} ES(\tilde{f}; C)$ with $\mathcal{F}_1 = \{x_{.1} \in \mathcal{R} : \tilde{f}(x) = (1, x_{.1})^T w : w \in \mathcal{R}^2\} \subset \mathcal{F}$ and $S(f; C) = C(L(Yf(X)) + U(f(X))) + J(f)$.

Assumption A follows from $e_{V_\pi}(f_\pi^*, \bar{f}_\pi) \leq 2P(|nf_\pi(X)| \leq 1) \leq (\kappa_1 + 1)n^{-1} = s_n$

with $f_\pi^* = nf_\pi$. Easily, (7) in Assumption B holds for $\alpha = 1$. To verify (8), direct calculation yields that there exist some constants $b_1 > 0$ and $b_2 > 0$ such that for any $f \in \mathcal{F}_1$, we have $e_{V_\pi}(f, \bar{f}_\pi) \geq e_\pi(f, \bar{f}_\pi) = b_1((\frac{5}{4}(2\pi - 1) + e)^{\kappa_1 + \kappa_2 + 1} - (\frac{5}{4}(2\pi - 1))^{\kappa_1 + \kappa_2 + 1})$ and $E|\text{sign}(f_\pi) - \text{sign}(f)| = b_2((\frac{5}{4}(2\pi - 1) + e)^{\kappa_1 + 1} - (\frac{5}{4}(2\pi - 1))^{\kappa_1 + 1})$ with $w_f = w_{f_\pi} + (e_0, e_1)^T$ and $e = -\frac{50e_1(\frac{5}{4}(2\pi - 1) + 10e_0)}{4 + 10e_1} > 0$. This implies (8) with $\beta = \gamma = \frac{1 + \kappa_1}{1 + \kappa_1 + \kappa_2}$. For (9) in Assumption B, by the triangle inequality, $\text{Var}(V_\pi(f, Z) - V_\pi(f_\pi^*, Z)) \leq 2E|V_\pi(f, Z) - V_\pi(\bar{f}_\pi, Z)| \leq 2(\Lambda_1 + \Lambda_2)$, where $\Lambda_1 = E|l_\pi(f, Z) - V_\pi(\bar{f}_\pi, Z)|E|S_\pi(Y)||\text{sign}(f) - \text{sign}(\bar{f}_\pi)| \leq (2^{1 + \kappa_2}(\kappa_1 + 1)^{\kappa_2})^{\frac{1 + \kappa_1}{1 + \kappa_1 + \kappa_2}} e_{V_\pi}(f, \bar{f}_\pi)^{\frac{1 + \kappa_1}{1 + \kappa_1 + \kappa_2}}$, and $\Lambda_2 = E(V_\pi(f, Z) - l_\pi(f, Z)) = E(V_\pi(f, Z) - V_\pi(\bar{f}_\pi, Z)) + E(l_\pi(\bar{f}_\pi, Z) - l_\pi(f, Z)) \leq 2e_{V_\pi}(f, \bar{f}_\pi)$. Therefore (9) is met with $\zeta = \frac{1 + \kappa_1}{1 + \kappa_1 + \kappa_2}$. For Assumption C, we define $\phi_1(\epsilon, k) = a_3(\log(1/M^{1/2}))^{1/2}/M^{1/2}$ with $M = M(\epsilon, \lambda, k)$. By Lemma 6 of Wang and Shen (2007), solving (10) yields $\epsilon_n = (\frac{\log n}{n})^{1/2}$ when $C/J_\pi^* \sim \delta_n^{-2}n^{-1} \sim (\log n)^{-1}$. Assumption D is satisfied with $d = \infty$ and $\eta = 0$, and Assumption E is met with $\theta = \infty$ by noting that $\min(p(x), 1 - p(x)) \geq 1/10$. In this case, $B = \frac{(1 + \kappa_1)^2}{2\kappa_2(1 + \kappa_1 + \kappa_2)}$, and the desired result follows from Corollary 1.

Kernel learning: Similary to the linear case, we restrict our attention to $\mathcal{F}_1 = \{x_{\cdot 1} \in \mathcal{R} : f(x_{\cdot 1}) = w_{f,0} + \sum_{k=1}^n w_{f,k}K(x_{\cdot 1}, x_{k1}) : w_f = (w_{f,1}, \dots, w_{f,n})^T \in \mathcal{R}^n\}$.

Note that \mathcal{F}_1 is rich for sufficiently large n in that for function f_π^* as defined in the linear example, there exists a $\tilde{f}_\pi^* \in \mathcal{F}_1$ such that $\|\tilde{f}_\pi^* - f_\pi^*\|_\infty \leq s_n$, and hence $e_{V_\pi}(\tilde{f}_\pi^*, \bar{f}_\pi) \leq 2s_n$. Assumption A is then met. Easily, (7) is satisfied for $\alpha = 1$. To verify (8), note that there exists constant $b_3 > 0$ such that for small $\delta > 0$, $P(|p(x) - 1/2| \geq \delta) = 2P(0 \leq p(x) - 1/2 \leq \delta) = 2P(0 \leq x_{(1)} \leq \frac{5}{2}\delta^{\frac{1}{\kappa_2}}) \leq b_3\delta^{\frac{1 + \kappa_1}{\kappa_2}}$. Therefore, $e_{V_{.5}}(f, \bar{f}_{.5}) \geq e_{.5}(f, \bar{f}_{.5}) \geq \delta E|\text{sign}(f) - \text{sign}(\bar{f}_{.5})|I(|p(x)| \geq \delta) \geq 2^{-1}(4b_3)^{-\frac{\kappa_2}{1 + \kappa_1}} \|\text{sign}(f) - \text{sign}(\bar{f}_{.5})\|_1^{\frac{1 + \kappa_1 + \kappa_2}{1 + \kappa_1}}$ with $\delta = (\|\text{sign}(f) - \text{sign}(\bar{f}_{.5})\|_1 / 4b_3)^{\frac{\kappa_2}{1 + \kappa_1}}$. This implies $\beta = \frac{1 + \kappa_1}{1 + \kappa_1 + \kappa_2}$ in (8). Similarly, we can verify that there exists a constant $b_4 > 0$ such that $P(|p(x) - \pi| \geq \delta) = 2P(\frac{5}{2}(\pi - \frac{1}{2}) \leq x_{(1)} \leq \frac{5}{2}(\pi - \frac{1}{2} + \delta^{\frac{1}{\kappa_2}})) \leq b_4\delta^{\frac{1}{\kappa_2}}$ when $\pi > \frac{1}{2}$, which implies (8) with $\gamma = \frac{1}{1 + \kappa_2}$. For (9),

an application of the similar argument leads to $\zeta = \frac{1}{1+\kappa_2}$. For Assumption C, we define $\phi_1(\epsilon, k) = a_3(\log(1/M^{1/2}))^{3/2}/M^{1/2}$ with $M = M(\epsilon, \lambda, k)$. By Lemma 7 of Wang and Shen (2007), solving (10) yields $\epsilon_n = ((\log n)^3 n^{-1})^{1/2}$ when $C/J_\pi^* \sim \delta_n^{-2} n^{-1} \sim (\log n)^{-3}$. Assumption D is satisfied with $d = \infty$ and $\eta = 0$, and Assumption E is met with $\theta = \infty$. Finally, $B = \frac{(1+\kappa_1)}{2\kappa_2(1+\kappa_2)}$, and the desired result follows from Corollary 1.

References

- [1] ABNEY, S. (2004). Understanding the Yarowsky Algorithm. *Computational Linguistics*, **30**, 365-395.
- [2] ADAMS, R.A. (1975). Sobolev spaces. Academic press, New York.
- [3] AMINI, M. AND GALLINARI, P. (2003). Semi-supervised learning with an explicit label-error model for misclassified data. In *Proc. Int. Joint Conf. on Artificial Intelligence*, 555-561.
- [4] AN, L. AND TAO, P. (1997). Solving a class of linearly constrained indefinite quadratic problems by D.C. algorithms. *J. Global Optimization*, **11**, 253-285.
- [5] ANDO, R. AND ZHANG, T. (2004). A framework for learning predictive structures from multiple tasks and unlabeled data. *J. Mach. Learn. Res.*, **6**, 1817-1853.
- [6] BALCAN, M., BLUM, A., CHOI, P., LAFFERTY, J., PANTANO, B., RWEBANGIRA, M., AND ZHU, X. (2005). Person identification in webcam images: an application of semi-supervised learning. *ICML 2005 Workshop on Learning with Partially Classified Training Data*.
- [7] BLAKE, C.L. AND MERZ, C.J. (1998). UCI repository of machine learning databases. UC-Irvine, Dept. Information and Computer Science.

- [8] BLUM, A. AND MITCHELL, T. (1998). Combining labeled and unlabeled data with co-training. In *Proc. Annual Conf. on Comput. Learn. Theory*, 92-100.
- [9] CHAPELLE, O., SCHÖLKOPF, B. AND ZIEN, A. (2006). *Semi-supervised learning*. MIT press.
- [10] CHAPELLE, O. AND ZIEN, A. (2005). Semi-supervised classification by low density separation. In *Proc. Int. Workshop on Artificial Intelligence and Statist.*, 57-64.
- [11] COZMAN, F.G., COHEN, I. AND CIRELO, M.C. (2003) Semi-supervised learning of mixture models and Bayesian networks. *Proc. Int. Conf. on Mach. Learn.*.
- [12] COLLINS, M. AND SINGER. Y. (1999). Unsupervised models for named entity classification. In *Empirical Methods in Natural Language Processing and Very Large Corpora*.
- [13] FISHER, R.A. (1946). A system of scoring linkage data, with special reference to the pied factors in mice. *Amer. Nat.*, **80** , 568-578.
- [14] GU, C. (2000). Multidimension smoothing with splines. *Smoothing and Regression: Approaches, Computation and Application*, Schimek.
- [15] HUGHES, T., MARTON, M., JONES, A., ROBERTS, C., STOUGHTON, R., ARMOUR, C., BENNETT, H., COFFEY, E., DAI, H., HE, Y., KIDD, M., KING, A., MEYER, M., SLADE, D., LUM, P., STEPANIANTS, S., SHOEMAKER, D., GACHOTTE, D., CHAKRABURTTY, K., SIMON, J., BARD, M. AND FRIEND, S. (2000). Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109-126.
- [16] HUNTER, D. AND LANGE, K. (2000). Quantile regression via an MM algorithm. *J. Computa. & Graph. Statist.*, **9**, 60-77.

- [17] JAAKKOLA, T., DIEKHANS, M. AND HAUSSLER, D. (1999). Using the Fisher kernel method to detect remote protein homologies. In *Proc. Int. Conf. on Intelligent Systems for Molecular Biology*, 149-158.
- [18] KOLMOGOROV, A. N. AND TIHOMIROV, V. M. (1959). ε -entropy and ε -capacity of sets in function spaces. *Uspekhi Mat. Nauk.*, **14**, 3-86. [In Russian. English translation, *Ameri. Math. Soc. Transl.*, **14**, 277-364. (1961)]
- [19] LIN, Y. (2002). Support vector machines and the Bayes rule in classification. *Data Mining and Knowledge Discovery*, **6**, 259-275.
- [20] LIU, Y. AND SHEN, X. (2006). Multicategory ψ -learning. *J. Ameri. Statist. Assoc.*, **101**, 500-509.
- [21] LIU, S., SHEN, X. AND WONG, W. (2005). Computational development of ψ -learning. *Proc. SIAM 2005 Int. Data Mining Conf.*, P1-12.
- [22] MCCULLAGH, P. AND NELDER, J. (1983). *Generalized Linear Models*, 2nd edition. Chapman and Hall.
- [23] MASON, P., BAXTER, L., BARTLETT, J. AND FREAN, M. (2000) Boosting algorithms as gradient descent. In *Advances in Neural Information Processing Systems*, **12**, 512-518.
- [24] MEWES, H., ALBERMANN, K., HEUMANN, K., LIEBL, S. AND PFEIFFER, F. (2002) MIPS: a database for protein sequences, homology data and yeast genome information. *Nucleic Acids Res.*, **25**, 28-30.
- [25] NIGAM, K., MCCALLUM, A., THRUN, S. AND MITCHELL T. (1998). Text classification from labeled and unlabeled documents using EM. *Mach. Learn.*, **39**, 103–134.

- [26] PLATT, J. C. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in large margin classifiers*, MIT press.
- [27] SCHÖLKOPF, B., SMOLA, A., WILLIAMSON, R. AND BARTLETT, P. (2000). New support vector algorithms. *Neural Computation*, **12** 1207-1245.
- [28] SHEN, X. (1997). On method of sieves and penalization. *Ann. Statist.*, **25**, 2555-2591.
- [29] SHEN, X., TSENG, G., ZHANG, X. AND WONG, W. H. (2003). On ψ -learning. *J. Ameri. Statist. Assoc.*, **98**, 724-734.
- [30] SHEN, X. AND WANG, L. (2007). Generalization error for multi-class margin classification. *Electronic J. of Statist.*, **1**, 307-330.
- [31] SHEN, X. AND WONG, W.H. (1994). Convergence rate of sieve estimates. *Ann. Statist.*, **22**, 580-615.
- [32] TSYBAKOV, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *Ann. Statist.*, **32**, 135-166.
- [33] VAPNIK, V. (1998). *Statistical Learning Theory*. Wiley, New York.
- [34] WAHBA, G. (1990). *Spline models for observational data*. CBMS-NSF Regional Conference Series, Philadelphia.
- [35] WAHBA, G. (1999). Support vector machines, reproducing kernel Hilbert spaces and the randomized GACV. Technical report 984, Department of Statistics, University of Wisconsin.
- [36] WANG, J. AND SHEN, X. (2007). Large margin semi-supervised learning. *J. Mach. Learn. Res.* **8**, 1867-1891.

- [37] WANG, J., SHEN, X. AND LIU, Y. (2007). Probability estimation for large margin classifiers. *Biometrika*. To appear.
- [38] WANG, J., SHEN, X. AND PAN, W. (2007). On transductive support vector machines. *Proc. of the Snowbird Mach. Learn. Workshop*. In press.
- [39] YAROWSKY, D (1995). Unsupervised word sense disambiguation rivaling supervised methods. In *Proc. of the 33rd Annual Meeting of the Association for Computational Linguistics*, 189-196.
- [40] ZHANG, T. AND OLES, F. (2000). A probability analysis on the value of unlabeled data for classification problems. In *Proc. Int. Conf. on Mach. Learn. (ICML2000)*, 1191-1198.
- [41] ZHOU, X., KAO, M. AND WONG, W. (2002). Transitive functional annotation by shortest-path analysis of gene expression data. *Proc. Nat. Acad. Sci.*, **99**, 12783-12788.
- [42] ZHU, J. AND HASTIE, T. (2005). Kernel logistic regression and the import vector machine. *J. Comp. and Graph. Statist.*, **14**, 185-205.
- [43] ZHU, X., GHAMRANI, Z., AND LAFFERTY, J. (2003). Semi-supervised learning using Gaussian fields and harmonic functions. *Int. Conf. on Mach. Learn.*.
- [44] ZHU, X. (2005). Semi-supervised learning literature survey. *Technical Report 1530*, University of Wisconsin, Madison.