Journal of Machine Learning Research 8 (2007) 1867-1891 Submitted 2/06; Revised 1/07; Published 8/07

Large Margin Semi-supervised Learning

Junhui Wang WANGJH@STAT.UMN.EDU
Xiaotong Shen XSHEN@STAT.UMN.EDU
School of Statistics
University of Minnesota
Minneapolis, MN 55455, USA

Editor: Tommi Jaakkola

Abstract

In classification, semi-supervised learning occurs when a large amount of unlabeled data is available with only a small number of labeled data. In such a situation, how to enhance predictability of classification through unlabeled data is the focus. In this article, we introduce a novel large margin semi-supervised learning methodology, using grouping information from unlabeled data, together with the concept of margins, in a form of regularization controlling the interplay between labeled and unlabeled data. Based on this methodology, we develop two specific machines involving support vector machines and ψ-learning, denoted as SSVM and SPSI, through difference convex programming. In addition, we estimate the generalization error using both labeled and unlabeled data, for tuning regularizers. Finally, our theoretical and numerical analyses indicate that the proposed methodology achieves the desired objective of delivering high performance in generalization, particularly against some strong performers.

Keywords: generalization, grouping, sequential quadratic programming, support vectors

1. Introduction

In many classification problems, a large amount of unlabeled data is available, while it is costly to obtain labeled data. In text categorization, particularly web-page classification, a machine is trained with a small number of manually labeled texts (web-pages), as well as a huge amount of unlabeled texts (web-pages), because manually labeling is impractical; compare with Joachims (1999). In spam detection, a small group of identified e-mails, spam or non-spam, is used, in conjunction with a large number of unidentified e-mails, to train a filter to flag incoming spam e-mails, compare with Amini and Gallinari (2003). In face recognition, a classifier is trained to recognize faces with scarce identified and enormous unidentified faces, compare with Balcan et al. (2005). In a situation as such, one research problem is how to enhance accuracy of prediction in classification by using both unlabeled and labeled data. The problem of this sort is referred to as semi-supervised learning, which differs from a conventional "missing data" problem in that the size of unlabeled data greatly exceeds that of labeled data, and missing occurs only in response. The central issue that this article addresses is how to use information from unlabeled data to enhance predictability of classification.

In semi-supervised learning, a sample $\{Z_i = (X_i, Y_i)\}_{i=1}^{n_l}$ is observed with labeling $Y_i \in \{-1, 1\}$, in addition to an independent unlabeled sample $\{X_j\}_{j=n_l+1}^{n}$ with $n = n_l + n_u$, where $X_k = (X_{k1}, \cdots, X_{kp})$; $k = 1, \cdots, n$ is an p-dimensional input. Here the labeled sample is independently and identically distributed according to an unknown joint distribution $P(x, y)$, and the unlabeled sample is

independently and identically distributed from distribution $P(x)$ that may not be the marginal distribution of $P(x,y)$.

A number of semi-supervised learning methods have been proposed through some assumptions relating $P(x)$ to the conditional distribution $P(Y=1|X=x)$. These methods include, among others, co-training (Blum and Mitchell, 1998), the EM method (Nigam, McCallum, Thrun and Mitchell, 1998), the bootstrap method (Collins and Singer, 1999), information-based regularization (Szummer and Jaakkola, 2002), Bayesian network (Cozman, Cohen and Cirelo, 2003), Gaussian random fields (Zhu, Ghahramani and Lafferty, 2003), manifold regularization (Belkin, Niyogi and Sindhwani, 2004), and discriminative-generative models (Ando and Zhang, 2004). Transductive SVM (TSVM; Vapnik, 1998) uses the concept of margins.

Despite progress, many open problems remain. Essentially all existing methods make various assumptions about the relationship between $P(Y=1|X=x)$ and $P(x)$ in a way for an improvement to occur when unlabeled data is used. Note that an improvement of classification may not be expected when simply imputing labels of X through an estimated $P(Y=1|X=x)$ from labeled data, compare with Zhang and Oles (2000). In other words, the potential gain in classification stems from an assumption, which is usually not verifiable or satisfiable in practice. As a consequence, any departure from such an assumption is likely to degrade the "alleged" improvement, and may yield worse performance than classification with labeled data alone.

The primary objective of this article is to develop a large margin semi-supervised learning methodology to deliver high performance of classification by using unlabeled data. The methodology is designed to adapt to a variety of situations by identifying as opposed to specifying a relationship between labeled and unlabeled data from data. It yields an improvement when unlabeled data can reconstruct the optimal classification boundary, and yields a no worse performance than its supervised counterpart otherwise. This is in contrast to the existing methods.

Through three key ingredients, our objective is achieved, including (1) comparing all possible grouping boundaries from unlabeled data for classification, (2) using labeled data to determine label assignment for classification as well as a modification of the grouping boundary, and (3) interplay between (1) and (2) through tuning to connect grouping to classification for seeking the best classification boundary. These ingredients are integrated in a form of regularization involving three regularizers, each controlling classification with labeled data, grouping with unlabeled data, and interplay between them. Moreover, we introduce a tuning method using unlabeled data for tuning the regularizers.

Through the proposed methodology and difference convex programming, we develop two specific machines based on support vector machines (SVM; Cortes and Vapnik, 1995) and ψ-learning (Shen, Tseng, Zhang and Wong, 2003), denoted as SSVM and SPSI. Numerical analysis indicates that SSVM and SPSI achieve the desired objective, particularly against TSVM and a graphical method in simulated and benchmark examples. Moreover, a novel learning theory is developed to quantify SPSI's generalization error as a function of complexity of the class of candidate decision functions, the sample sizes (n_l, n_u), and the regularizers. To our knowledge, this is the first attempt to relate a classifier's generalization error to (n_l, n_u) and regularizers in semisupervised learning. This theory not only explains SPSI's performance, but also supports our aforementioned discussion concerning the interplay between grouping and classification, as evident from Section 5 that SPSI can recover the optimal classification performance at a speed in n_u because of grouping from unlabeled data.

This article is organized in eight sections. Section 2 introduces the proposed semi-supervised learning methodology. Section 3 treats non-convex minimization through difference convex programming. Section 4 proposes a tuning methodology that uses both labeled and unlabeled data to enhance of accuracy of estimation of the generalization error. Section 5 presents some numerical examples, followed by a novel statistical learning theory in Section 6. Section 7 contains a discussion, and the appendix is devoted to technical proofs.

2. Methodology

In this section, we present our proposed margin-based semi-supervised learning method as well its connection to other existing popular methodologies.

2.1 Proposed Methodology

We begin with our discussion in linear margin classification with labeled data $(X_i,Y_i)_{i=1}^{n_l}$ alone. Given a class of linear decision functions of the form $f(x)=\tilde{w}_f^T x+w_{f,0}\equiv(1,x^T)w_f$, a cost function $C\sum_{i=1}^{n_l}L(y_i f(x_i))+J(f)$ is minimized with respect to $f\in\mathcal{F}$, a class of candidate decision functions, to obtain the minimizer $\hat{f}$ yielding a classifier $\mathrm{Sign}(\hat{f})$, where $J(f)=\|\tilde{w}_f\|^2/2$ is the reciprocal of the L_2 geometric margin, and $L(\cdot)$ is a margin loss defined by functional margins $z_i=y_i f(x_i)$; $i=1,\cdots,n_l$.

Different learning methodologies are defined by different margin losses. Margin losses include, among others, the hinge loss $L(z)=(1-z)_+$ for SVM with its variants $L(z)=(1-z)_+^q$ for $q>1$; compare with Lin (2002); the ρ-hinge loss $L(z)=(\rho-z)_+$ for nu-SVM (Schölkopf, Smola, Williamson and Bartlett, 2000) with $\rho>0$ to be optimized; the ψ-loss $L(z)=\psi(z)$, with $\psi(z)=1-\mathrm{Sign}(z)$ if $z\geq 1$ or $z<0$, and $2(1-z)$ otherwise, compare with Shen et al. (2003), the logistic loss $L(z)=\log(1+e^{-z})$, compare with Zhu and Hastie (2005); the sigmoid loss $L(z)=1-\tanh(cz)$; compare with Mason, Baxter, Bartlett and Frean (2000). A margin loss $L(z)$ is said to be a large margin if $L(z)$ is nonincreasing in z, which penalizes small margin values.

In order to extract useful information about classification from unlabeled data, we construct a loss $U(\cdot)$ for a grouping decision function $g(x)=(1,x^T)w_g\equiv\tilde{w}_g^T x+w_{g,0}$, with $\mathrm{Sign}(g(x))$ indicating grouping. Towards this end, we let $U(z)=\min_{\{y=\pm1\}}L(yz)$ by minimizing y in $L(\cdot)$ to remove its dependency of y. As shown in Lemma 1, $U(z)=L(|z|)$, which is symmetric in z and indicates that it can only determine the grouping boundary that occurs near in an area with low value of $U(z)$ but provide no information regarding labeling.

While U can be used to extract the grouping boundary, it needs to yield the Bayes decision function $f^*=\arg\min_{f\in\mathcal{F}}EL(Yf(X))$ in order for it to be useful for classification, where E is the expectation with respect to (X,Y). More specifically, it needs $f^*=\arg\min_{g\in\mathcal{F}}EU(g(X))$. However, it does not hold generally since $\arg\min_{g\in\mathcal{F}}EU(g(X))$ can be any $g\in\mathcal{F}$ satisfying $|g(x)|\geq 1$. Generally speaking, U gives no information about labeling Y. To overcome this difficulty, we regularize U and introduce our regularized loss for semi-supervised learning to induce a relationship between classification f and grouping g:

$$S(f,g;C)=C_1L(yf(x))+C_2U(g(x))+\frac{C_3}{2}\|w_f-w_g\|^2+\frac{1}{2}\|\tilde{w}_g\|^2, \tag{1}$$

where $C=(C_1,C_2,C_3)$ are non-negative regularizers, and $\|w_f-w_g\|^2=\|\tilde{w}_f-\tilde{w}_g\|^2+(w_{f,0}-w_{g,0})^2$ is the usual L_2-Euclidean norm in R^{p+1}. Whereas $L(yf(x))$ regularizes the contribution from labeled

data, $U(g(x))$ controls the information extracted from unlabeled data, and $\|w_f - w_g\|^2$ penalizes the disagreement between f and g, specifying a loose relationship between f and g. The interrelation between f and g is illustrated in Figure 3. Note that in (1) the geometric margin $\frac{2}{\|\tilde{w}_f\|^2}$ does not enter as it is regularized implicitly through $\frac{2}{\|w_f - w_g\|^2}$ and $\frac{2}{\|\tilde{w}_g\|^2}$.

In nonlinear learning, a kernel $K(\cdot,\cdot)$ that maps from $S \times S$ to $\mathcal{R}^1$ is usually introduced for flexible representations: $f(x) = (1, K(x,x_1), \cdots, K(x,x_n))w_f$ and $g(x) = (1, K(x,x_1), \cdots, K(x,x_n))w_g$ with $w_f = (\tilde{w}_f, w_{f,0})$ and $w_g = \tilde{w}_g + w_{g,0}$. Then nonlinear surfaces separate instances of two classes, implicitly defined by $K(\cdot,\cdot)$, where the reproducing kernel Hilbert spaces (RKHS) plays an important role; compare with Wahba (1990) and Gu (2000). The forgoing treatment for the linear case is applicable when the Euclidean inner product $\langle x_i, x_j \rangle$ is replaced by $K(x_i, x_j)$. In this sense, the linear case may be regarded as a special case of nonlinear learning.

Lemma 1 says that the regularized loss (1) allows U to yield precise information about the Bayes decision function f^* when after tuning. Specifically, U targets at the Bayes decision function in classification when C_1 and C_3 are large, and grouping can differ from classification at other C values.

Lemma 1 *For any large margin loss $L(z)$, $U(z) = \min_{y \in \{-1,1\}} L(yz) = L(|z|)$, where $y = \text{Sign}(z) = \arg\min_{y \in \{-1,1\}} L(yz)$ for any given z. Additionally,*

$$(f_C^*, g_C^*) = \arg \inf_{f,g \in \mathcal{F}} ES(f,g;C) \to (f^*, f^*) \text{ as } C_1, C_3 \to \infty.$$

In the case that (f_C^*, g_C^*) is not unique, we choose it as any minimizer of $ES(f,g;C)$.

Through (1), we propose our cost function for semi-supervised learning:

$$s(f,g) = C_1 \sum_{i=1}^{n_l} L(y_i f(x_i)) + C_2 \sum_{j=n_l+1}^{n} U(g(x_j)) + \frac{C_3}{2}\|f-g\|^2 + \frac{1}{2}\|g\|_-^2, \tag{2}$$

where in the linear case, $\|g\|_- = \|\tilde{w}_g\|$ and $\|f-g\| = \|w_f - w_g\|$; in the nonlinear case $\|g\|_-^2 = \tilde{w}_g^T K \tilde{w}_g$, $\|f-g\|^2 = (\tilde{w}_f - \tilde{w}_g)^T K (\tilde{w}_f - \tilde{w}_g) + (\tilde{w}_{f,0} - \tilde{w}_{g,0})^2$ is the RKHS norm, with an $n \times n$ matrix $\mathbf{K}$ whose ijth element is $K(x_i, x_j)$. Minimization of (2) with respect to (f,g) yields an estimated decision function $\hat{f}$ thus classifier $\text{Sign}(\hat{f})$. The constrained version of (2), after introducing slack variables $\{\xi_k \geq 0; k = 1, \cdots, n\}$, becomes

$$C_1 \sum_{i=1}^{n_l} \xi_i + C_2 \sum_{j=n_l+1}^{n} \xi_j + \frac{C_3}{2}\|f-g\|^2 + \frac{1}{2}\|g\|_-^2, \tag{3}$$

subject to $\xi_i - L(y_i f(x_i)) \geq 0$; $i = 1, \cdots, n_l$; $\xi_j - U(g(x_j)) \geq 0$; $j = n_l+1, \cdots, n$. Minimization of (2) with respect to (f,g), equivalently, minimization of (3) with respect to $(f, g, \xi_k; k = 1, \cdots, n)$ subject to the constraints gives our estimated decision function $(\hat{f}, \hat{g})$, where $\hat{f}$ is for classification.

Two specific machines SSVM and SPSI will be further developed in what follows. In (2), SSVM uses $L(z) = (1-z)_+$ and $U(z) = (1-|z|)_+$, and SPSI uses $L(z) = \psi(z)$ and $U(z) = 2(1-|z|)_+$.

2.2 Connection Between SSVM and TSVM

To better understand the proposed methodology, we now explore the connection between SSVM and TSVM. In specific, TSVM uses a cost function in the form of

$$C_1 \sum_{i=1}^{n_l} (1 - y_i f(x_i))_+ + C_2 \sum_{j=n_l+1}^{n} (1 - y_j f(x_j))_+ + \frac{1}{2}\|f\|_-^2,$$

where minimization with respect to $(y_j : j = n_l+1, \cdots, n; f)$ yields the estimated decision function $\hat{f}$. It can be thought of as the limiting case of SSVM as $C_3 \to \infty$ forcing $f = g$ in (2).

SSVM in (3) stems from grouping and interplay between grouping and classification, whereas TSVM focuses on classification. Placing TSVM in the framework of SSVM, we see that SSVM relaxes TSVM in that it allows grouping (g) and classification (f) to differ, whereas $f \equiv g$ for TSVM. Such a relaxation yields that $|e(\hat{f}, f^*)| = |GE(\hat{f}) - GE(f^*)|$ is bounded by $|e(\hat{f}, \hat{g})| + |e(\hat{g}, g_C^*)| + |e(g_C^*, f^*)|$, with $|e(\hat{f}, \hat{g})|$ controlled by C_3, the estimation error $|e(\hat{g}, g_C^*)|$ controlled by $C_2 n_u^{-1}$ and the approximation error $|e(g_C^*, f^*)|$ controlled by C_1 and C_3. As a result, all these error terms can be reduced simultaneously with a suitable choice of (C_1, C_2, C_3), thus delivering better generalization. This aspect will be demonstrated by our theory in Section 6 and numerical analysis in Section 5. In contrast, TSVM is unable to do so, and needs to increase the size of one error in order to reduce the other error, and vice versa, compare with Wang, Shen and Pan (2007). This aspect will be also confirmed by our numerical results.

The forgoing discussion concerning SSVM is applicable to (2) with a different large margin loss L as well.

3. Non-convex Minimization Through Difference Convex Programming

Optimization in (2) involves non-convex minimization, because of non-convex $U(z)$ and/or possibly $L(z)$ in z. On the basis of recent advances in global optimization, particularly difference convex (DC) programming, we develop our minimization technique. Key to DC programming is decomposition of our cost function into a difference of two convex functions, based on which iterative upper approximations can be constructed to yield a sequence of solutions converging to a stationary point, possibly an ε-global minimizer. This technique is called DC algorithms (DCA; An and Tao, 1997), permitting a treatment of large-scale non-convex minimization.

To use DCA for SVM and ψ-learning in (2), we construct DC decompositions of the cost functions of SPSI and SSVM s^{Ψ} and s^{SVM} in (2):

$$s^{\Psi} = s_1^{\Psi} - s_2^{\Psi}; \quad s^{SVM} = s_1^{SVM} - s_2^{SVM},$$

where $L(z) = \psi(z)$ and $U(z) = 2(1-|z|)_+$ for SPSI,

$$\begin{aligned} s_1^{\Psi} &= C_1 \textstyle\sum_{i=1}^{n_l} \psi_1(y_i f(x_i)) + C_2 \sum_{j=n_l+1}^{n} 2U_1(g(x_j)) + \frac{C_3}{2}\|f-g\|^2 + \frac{1}{2}\|g\|_-^2, \\ s_2^{\Psi} &= C_1 \textstyle\sum_{i=1}^{n_l} \psi_2(y_i f(x_i)) + C_2 \sum_{j=n_l+1}^{n} 2U_2(g(x_j)); \end{aligned}$$

and $L(z) = (1-z)_+$ and $U(z) = (1-|z|)_+$ for SSVM,

$$\begin{aligned} s_1^{SVM} &= C_1 \textstyle\sum_{i=1}^{n_l} (1-y_i f(x_i))_+ + C_2 \sum_{j=n_l+1}^{n} U_1(g(x_j)) + \frac{C_3}{2}\|f-g\|^2 + \frac{1}{2}\|g\|_-^2, \\ s_2^{SVM} &= C_2 \textstyle\sum_{j=n_l+1}^{n} U_2(g(x_j)). \end{aligned}$$

These DC decompositions are obtained through DC decompositions of $(1-|z|)_+ = U_1(z) - U_2(z)$ and $\psi(z) = \psi_1(z) - \psi_2(z)$, where $U_1 = (|z|-1)_+$, $U_2 = |z|-1$, $\psi_1 = 2(1-z)_+$, and $\psi_2 = 2(-z)_+$. The decompositions are displayed in Figure 1.

With these decompositions, we treat the nonconvex minimization in (2) by solving a sequence of quadratic programming (QP) problems. Algorithm 1 solves (2) for SPSI and SSVM.

Algorithm 1: (Sequential QP)

Step 1. (Initialization) Set initial values $f^{(0)} = g^{(0)}$ as the solution of SVM with labeled data alone,

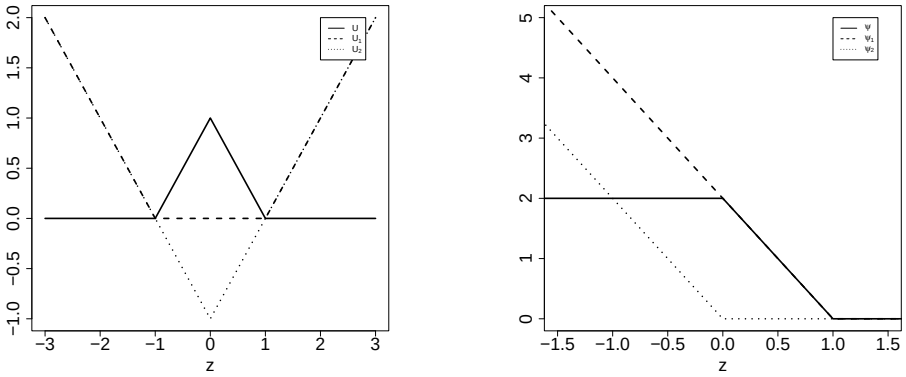

Figure 1: The left panel is a plot of U, U_1 and U_2, for the DC decomposition of $U = U_1 - U_2$. Solid, dotted and dashed lines represent U, U_1 and U_2, respectively. The right panel is a plot of ψ, ψ_1 and ψ_2, for the DC decomposition of $\psi = \psi_1 - \psi_2$. Solid, dotted and dashed lines represent ψ, ψ_1 and ψ_2, respectively.

and an precision tolerance level $\varepsilon > 0$.

Step 2. (Iteration) At iteration $k+1$, compute $(f^{(k+1)}, g^{(k+1)})$ by solving the corresponding dual problems given in (4).

Step 3. (Stopping rule) Terminate when $|s(f^{(k+1)}, g^{(k+1)}) - s(f^{(k)}, g^{(k)})| \le \varepsilon$.

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

At iteration $k+1$, after omitting constants that are independent of (4), the primal problems are required to solve

$$\begin{aligned} &\min_{w_f, w_g} s_1^{\psi}(f,g) - \langle (f,g), \nabla s_2^{\psi}(f^{(k)}, g^{(k)}) \rangle, \\ &\min_{w_f, w_g} s_1^{SVM}(f,g) - \langle (f,g), \nabla s_2^{SVM}(f^{(k)}, g^{(k)}) \rangle. \end{aligned} \tag{4}$$

Here $\nabla s_2^{SVM} = (\nabla_{1f}^{SVM}, \nabla_{2f}^{SVM}, \nabla_{1g}^{SVM}, \nabla_{2g}^{SVM})$ is the gradient vector of s_2^{SVM} with respect to (f,g), with $\nabla_{1g}^{SVM} = C_2 \sum_{j=n_l+1}^{n} \nabla U_2(g(x_j)) x_j$, $\nabla_{2g}^{SVM} = C_2 \sum_{j=n_l+1}^{n} \nabla U_2(g(x_j))$, $\nabla_{1f}^{SVM} = \mathbf{0}_p$, and $\nabla_{2f}^{SVM} = 0$, where $\nabla U_2(z) = 1$ if $z > 0$, and $\nabla U_2(z) = -1$ otherwise. Similarly, $\nabla s_2^{\psi} = (\nabla_{1f}^{\psi}, \nabla_{2f}^{\psi}, \nabla_{1g}^{\psi}, \nabla_{2g}^{\psi})$ is the gradient vector of s_2^{ψ} with respect to (w_f, w_g), with $\nabla_{1f}^{\psi} = C_1 \sum_{i=1}^{n_l} \nabla \psi_2(y_i f(x_i)) y_i x_i$, $\nabla_{2f}^{\psi} = C_1 \sum_{i=1}^{n_l} \nabla \psi_2(y_i f(x_i)) y_i$, $\nabla_{1g}^{\psi} = 2\nabla_{1g}^{SVM}$, and $\nabla_{2g}^{\psi} = 2\nabla_{2g}^{SVM}$, where $\nabla \psi_2(z) = 0$ if $z > 0$ and $\nabla \psi_2(z) = -2$ otherwise. By Karush-Kuhn-Tucker(KKT)'s condition, the primal problems in (4) are equivalent to their dual forms, which are generally easier to work with and given in the Appendix C.

By Theorem 3 of Liu, Shen and Wong (2005), $\lim_{k\to\infty} \| f^{(k+1)} - f^{(\infty)} \| = 0$ for some $f^{(\infty)}$, and convergence of Algorithm 1 is superlinear in that $\lim_{k\to\infty} \| f^{(k+1)} - f^{(\infty)} \| / \| f^{(k)} - f^{(\infty)} \| = 0$ and $\lim_{k\to\infty} \| g^{(k+1)} - g^{(\infty)} \| / \| g^{(k)} - g^{(\infty)} \| = 0$, if there does not exist an instance $\tilde{x}$ such that $f^{(\infty)}(\tilde{x}) = g^{(\infty)}(\tilde{x}) = 0$ with $f^{(\infty)}(x) = (1, K(x,x_1), \cdots, K(x,x_n)) w_f^{(\infty)}$ and $g^{(\infty)}(x) = (1, K(x,x_1), \cdots, K(x,x_n)) w_g^{(\infty)}$. Therefore, the number of iterations required for Algorithm 1 is $o(\log(1/\varepsilon))$ to achieve the precision $\varepsilon > 0$.

4. Tuning Involving Unlabeled Data

This section proposes a novel tuning method based on the concept of generalized degrees of freedom (GDF) and the technique of data perturbation (Shen and Huang, 2006; Wang and Shen, 2006), through both labeled and unlabeled data. This permits tuning of three regularizers $C = (C_1, C_2, C_3)$ in (2) to achieve the optimal performance.

The generalization error (GE) of a classification function f is defined as $GE(f) = P(Yf(X) < 0) = EI(Y \neq \text{Sign}(f(X)))$, where $I(\cdot)$ is the indicator function. The $GE(f)$ usually depends on the unknown truth, and needs to be estimated. Minimization of the estimated $GE(f)$ with respect to the range of the regularizers gives the optimal regularization parameters.

For tuning, write $\hat{f}$ as $\hat{f}_C$, and write $(X^l, Y^l) = (X_i, Y_i)_{i=1}^{n_l}$ and $X^u = \{X_j\}_{j=n_l+1}^{n}$. By Theorem 1 of Wang and Shen (2006), the optimal estimated $GE(\hat{f}_C)$, after ignoring the terms independent of $\hat{f}_C$, has the form of

$$EGE(\hat{f}_C) + \frac{1}{2n_l}\sum_{i=1}^{n_l} \text{Cov}(Y_i, \text{Sign}(\hat{f}_C(X_i))|X^l) + \frac{1}{4}D_1(X^l, \hat{f}_C). \tag{5}$$

Here, $EGE(\hat{f}_C) = \frac{1}{2n_l}\sum_{i=1}^{n_l}(1 - Y_i\,\text{Sign}(\hat{f}_C(X_i)))$ is the training error, and $D_1(X^l, \hat{f}_C) = E\left(E(\triangle(X)) - \frac{1}{n_l}\sum_{i=1}^{n_l}\triangle(X_i)|X^l\right)$ with $\triangle(X) = (E(Y|X) - \text{Sign}(\hat{f}_C(X)))^2$, where $E(\cdot|X)$ and $E(\cdot|X^l)$ are conditional expectations with respect to Y and Y^l respectively. As illustrated in Wang and Shen (2006), the estimated (5) based on GDF is optimal in the sense that it performs no worse than the method of cross-validation and other tuning methods; see Efron (2004).

In (5), $\text{Cov}(Y_i, \text{Sign}(\hat{f}_C(X_i))|X^l); i = 1\cdots, n_l$ and $D_1(X^l, \hat{f}_C)$ need to be estimated. It appears that $\text{Cov}(Y_i, \text{Sign}(\hat{f}_C(X_i))|X^l)$ is estimated only through labeled data, for which we apply the data perturbation technique of Wang and Shen (2006). On the other hand, $D_1(X^l, \hat{f}_C)$ is estimated directly through (X^l, Y^l) and X^u jointly.

Our method proceeds as follows. First generate pseudo data Y_i^* by perturbing Y_i:

$$Y_i^* = \begin{cases} Y_i & \text{with probability } 1-\tau, \\ \tilde{Y}_i & \text{with probability } \tau, \end{cases} \tag{6}$$

where $0 < \tau < 1$ is the size of perturbation, and $(\tilde{Y}_i + 1)/2$ is sampled from a Bernoulli distribution with $\hat{p}(x_i)$, an rough probability estimate of $p(x_i) = P(Y = 1|X = x_i)$, which may be obtained through the same classification method that defines $\hat{f}_C$ or through logistic regression when it doesn't yield an estimated $p(x)$, such as SVM and ψ-learning. The estimated covariance is proposed to be

$$\widehat{\text{Cov}}(Y_i, \text{Sign}(\hat{f}_C(X_i))|X^l) = \frac{1}{k(Y_i, \hat{p}(X_i))}\text{Cov}^*(Y_i^*, \text{Sign}(\hat{f}_C^*(X_i))|X^l);\ i = 1, \cdots, n_l, \tag{7}$$

where $k(Y_i, \hat{p}(X_i)) = \tau + \tau(1-\tau)\frac{((Y_i+1)/2 - \hat{p}(X_i))^2}{\hat{p}(X_i)(1-\hat{p}(X_i))}$, and f_C^* is an estimated decision function through the same classification method trained through $(X_i, Y_i^*)_{i=1}^{n_l}$.

To estimate D_1, we express it as a difference between the true model error $E(E(Y|X) - Sign(\hat{f}_C(X)))^2$ and its empirical version $n_l^{-1}\sum_{i=1}^{n_l}(E(Y_i|X_i) - \text{Sign}(\hat{f}_C(X_i)))^2$, where the former can

be estimated through (X^l, Y^l) and X^u. The estimated D_1 becomes

$$\widehat{D}_1(X^l, \hat{f}_C) = E^* \left(\frac{1}{n_u} \sum_{j=n_l+1}^{n} ((2\hat{p}(X_j) - 1) - \text{Sign}(\hat{f}_C^*(X_j)))^2 - \frac{1}{n_l} \sum_{i=1}^{n_l} ((2\hat{p}(X_i) - 1) - \text{Sign}(\hat{f}_C^*(X_i)))^2 \middle| X^l \right), \tag{8}$$

Generally, $\widehat{\text{Cov}}$ in (7) and $\widehat{D}_1$ in (8) can be always computed using a Monte Carlo (MC) approximation of Cov^*, E^*, when it is difficult to obtain their analytic forms. Specifically, when Y^l is perturbed D times, a MC approximation of $\widehat{\text{Cov}}$ and $\widehat{D}_1$ can be derived:

$$\widehat{\text{Cov}}(Y_i, \text{Sign}(\hat{f}_C(X_i)) | X^l) \approx \frac{1}{D-1} \sum_{d=1}^{D} \frac{1}{k(Y_i, \hat{p}(X_i))} \text{Sign}(\hat{f}_C^{*d}(X_i))(Y_i^{*d} - \overline{Y}_i^*), \tag{9}$$

$$\widehat{D}_1(X^l, \hat{f}_C) \approx \frac{1}{D-1} \sum_{d=1}^{D} \left(\frac{1}{n_u} \sum_{j=n_l+1}^{n} ((2\hat{p}(X_j) - 1) - \text{Sign}(\hat{f}_C^{*d}(X_j)))^2 - \frac{1}{n_l} \sum_{i=1}^{n_l} ((2\hat{p}(X_i) - 1) - \text{Sign}(\hat{f}_C^{*d}(X_i)))^2 \right),$$

where $Y_i^{*d}; d = 1, \cdots, D$ are perturbed samples according to (6), $\overline{Y}_i^* = \frac{1}{D} \sum_d Y_i^{*d}$, and $\hat{f}_C^{*d}$ is trained through $(X_i, Y_i^{*d})_{i=1}^{n_l}$. Our proposed estimate $\widehat{GE}$ becomes

$$\widehat{GE}(\hat{f}_C) = EGE(\hat{f}_C) + \frac{1}{2n_l} \sum_{i=1}^{n_l} \widehat{\text{Cov}}(Y_i, \text{Sign}(\hat{f}_C(X_i)) | X^l) + \frac{1}{4} \widehat{D}_1(X^l, \hat{f}_C), \tag{10}$$

By the law of large numbers, $\widehat{GE}$ converges to (5) as $D \to \infty$. In practice, we recommend D to be at least n_l to ensure the precision of MC approximation and τ to be 0.5. In contrast to the estimated GE with labeled data alone, the $\widehat{GE}(\hat{f}_C)$ in (10) requires no perturbation of X when X^u is available. This permits more robust and computationally efficient estimation.

Minimization of (10) with respect to C yields the minimizer $\hat{C}$, which is optimal in terms of GE as suggested by Theorem 2, under similar technical assumptions as in Wang and Shen (2006).

(C.1): (Loss and risk) $\lim_{n_l \to \infty} \sup_C |GE(\hat{f}_C)/E(GE(\hat{f}_C)) - 1| = 0$ in probability.

(C.2): (Consistency of initial estimates) For almost all x, $\hat{p}_i(x) \to p_i(x)$, as $n_l \to \infty$; $i = 1, \cdots, n_l$.

(C.3): (Positivity) Assume that $\inf_C E(GE(\hat{f}_C)) > 0$.

Theorem 2 *Under Conditions C.1-C.3,* $\lim_{n_l, n_u \to \infty} \left(\lim_{\tau \to 0^+} GE(\hat{f}_{\hat{C}}) / \inf_C GE(\hat{f}_C) \right) = 1.$

Theorem 2 says the ideal optimal performance $\inf_C GE(\hat{f}_C)$ can be realized by $GE(\hat{f}_{\hat{C}})$ when $\tau \to 0^+$ and $n_l, n_u \to \infty$ against any other tuning method.

5. Numerical Examples

This section examines effectiveness of SSVM and SPSI and compare them against SVM with labeled data alone, TSVM and a graphical method of Zhu, Ghahramani and Lafferty (2003), in both simulated and benchmark examples. A test error, averaged over 100 independent replications, is used to measure their performances.

For simulation comparison, we define the amount of improvement of a method over SVM with labeled data alone as the percent of improvement in terms of the Bayesian regret,

$$\frac{(T(SVM)-T(Bayes))-(T(\cdot)-T(Bayes))}{T(SVM)-T(Bayes)}, \tag{11}$$

where $T(\cdot)$ and $T(Bayes)$ are the test error of any method and the Bayes error. This metric seems to be sensible, which is against the baseline—the Bayes error $T(Bayes)$, which is approximated by the test error over a test sample of large size, say 10^5.

For benchmark comparison, we define the amount of improvement over SVM as

$$\frac{T(SVM)-T(\cdot)}{T(SVM)}, \tag{12}$$

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

Numerical analyses are performed in R2.1.1. For TSVM, SVM^{light} (Joachims, 1999) is used. For the graphical method, a MATLAB code provided in Zhu, Ghahramani and Lafferty (2003) is employed. In the linear case, $K(s,t)=\langle s,t\rangle$; in the Gaussian kernel case, $K(s,t)=\exp\left(-\frac{\|s-t\|^2}{\sigma^2}\right)$, where σ^2 is set to be p, a default value in the "svm" routine of R, to reduce computational cost for tuning σ^2.

5.1 Simulations and Benchmarks

Two simulated and three benchmark examples are examined. In each example, we perform a grid search to minimize the test error of each classifier with respect to tuning parameters, in order to eliminate the dependency of the classifier on these parameters. Specifically, one regularizer for SVM and one tuning parameter σ in the Gaussian weight matrix for the graphical method, two regularization regularizers for TSVM, and three regularizers for SSVM and SPSI are optimized over $[10^{-2},10^3]$. For SSVM and SPSI, C is searched through a set of unbalanced grid points, based on our small study of the relative importance among (C_1,C_2,C_3). As suggested by Figure 2, C_3 appears to be most crucial to $\widehat{GE}(\hat{f}_C)$, whereas C_2 is less important than (C_1,C_3), and C_1 is only useful when its value is not too large. This leads to our unbalanced search over C, that is, $C_1\in\{10^{-2},10^{-1},1,10,10^2\}$, $C_2\in\{10^{-2},1,10^2\}$, and $C_3\in\{10^{m/4};m=-8,-7,\cdots,12\}$. This strategy seems reasonable as suggested by our simulation. Clearly, a more refined search is expected to yield better performance for SSVM and SPSI.

Example 1: A random sample $\{(X_{i1},X_{i2},Y_i);i=1,\cdots,1000\}$ is generated as follows. First, 1000 independent instances (Y_i,X_{i1},X_{i2}) are sampled according to $(Y_i+1)/2\sim Bernoulli(0.5)$, $X_{i1}\sim Normal(Y_i,1)$, and $X_{i2}\sim Normal(0,1)$. Second, 200 instances are randomly selected for training, and the remaining 800 instances are retained for testing. Next, 190 unlabeled instances (X_{i1},X_{i2}) are obtained by removing labels from a randomly chosen subset of the training sample, whereas the remaining 10 instances are treated as labeled data. The Bayes error is 0.162.

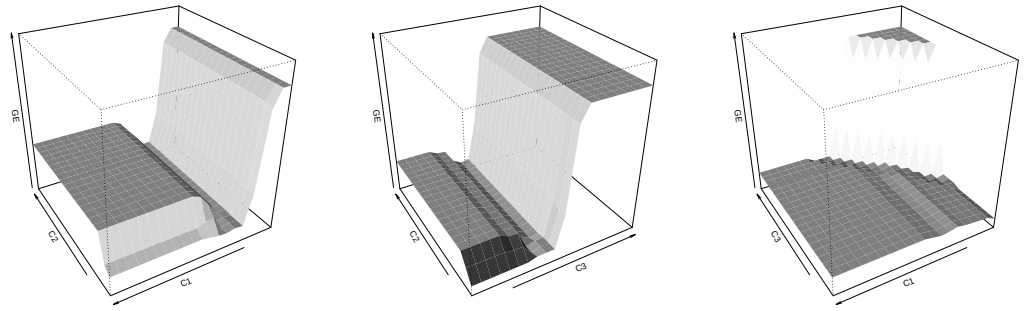

Figure 2: Plot of $\widehat{GE}(\hat{f}_C)$ as a function of (C_1, C_2, C_3) for one random selected sample of the WBC example. The top left, the top right and the bottom left are plots of $\widehat{GE}(\hat{f}_C)$ versus (C_1, C_2), (C_2, C_3) and (C_3, C_1), respectively. Here (C_1, C_2, C_3) take values in set $\{10^{-2+m/4}; m = 0, 1, \cdots, 20\}$.

Example 2: A random sample $\{(X_{i1}, X_{i2}, Y_i); i = 1, \cdots, 1000\}$ is generated. First, a random sample (X_{i1}, X_{i2}) of size 1000 is generated: $X_{i1} \sim Normal(3\cos(k_i\pi/2 + \pi/8), 1)$, $X_{i2} \sim Normal(3\sin(k_i\pi/2 + \pi/8), 4)$, with k_i sampled uniformly from $\{1, \cdots, 4\}$. Second, their labels Y_i; $i = 1, \cdots, 1000$ are assigned: $Y_i = 1$ if $k_i \in \{1, 4\}$, and -1 if $k_i \in \{2, 3\}$. As in Example 1, we obtain 200 (10 labeled and 190 unlabeled) instances for training as well as 800 instances for testing. The Bayes error is 0.089.

Benchmarks: Three benchmark examples are examined, including Wisconsin Breast Cancer (WBC), Mushroom and Spam email, each available in the UCI Machine Learning Repository (Blake and Merz, 1998). The WBC example concerns discrimination of a benign breast tissue from a malignant tissue through 9 clinic diagnostic characteristics; the Mushroom example separates an edible mushroom from a poisonous one through 22 biological records; the Spam email example discriminates texts to identify spam emails through 57 frequency attributes such as frequencies of particular words and characters. All these benchmarks are suited for linear and Gaussian kernel semi-supervised learning (Blake and Merz, 1998).

Instances in the WBC and Mushroom examples are randomly divided into halves with 10 labeled and 190 unlabeled instances for training, and the remaining instances for testing. Instances in the Spam email example are randomly divided into halves with 20 labeled and 580 unlabeled instances for training, and the remaining instances for testing.

In each example, the smallest averaged test errors of SVM with labeled data alone, TSVM, the graphical method and our proposed methods are reported in Tables 1 and 2.

As indicated in Tables 1-2, SPSI and SSVM outperform both SVM and TSVM in all cases, and the graphical method in all examples except the Mushroom example. The amount of improvement, however, varies over examples and types of classifiers. Specifically, we make the following observations.

Data $n \times dim$	Method	SVM_l	TSVM Improv.	Graph Improv.	SSVM Improv.	SPSI Improv.	SVM_c
Example 1	Linear	.344(.0104)	.249(.0134)		.188(.0084)	.184(.0084)	.164(.0084)
1000×2			52.2%	.232(.0108)	85.7%	87.9%	
	Gaussian	.385(.0099)	.267(.0132)	61.5%	.201(.0072)	.200(.0069)	.196(.0015)
			52.9%		82.5%	83.0%	
Example 2	Linear	.333(.0129)	.222(.0128)		.129(.0031)	.128(.0031)	.115(.0032)
1000×2			45.5%	.213(.0114)	83.6%	84.0%	
	Gaussian	.347(.0119)	.258(.0157)	49.2%	.175(.0092)	.175(.0098)	.151(.0021)
			34.5%		66.7%	66.7%	

Table 1: Averaged test errors as well as the estimated standard errors (in parenthesis) of SVM with labeled data alone, TSVM, the graphical method, SSVM and SPSI, over 100 pairs of training and testing samples, in the simulated examples. Here Graph, SVM_l and SVM_c denote performances of the graphical method, SVM with labeled data alone, and SVM with complete data without missing. The amount of improvement is defined in (11), where the Bayes error serves as a baseline for comparison.

Data $n \times dim$	Method	SVM_l	TSVM Improv.	Graph Improv.	SSVM Improv.	SPSI	SVM_c
WBC	Linear	.053(.0071)	.077(.0113)		.032(.0025)	.029(.0022)	.027(.0020)
682×9			-45.3%	.080(.0235)	39.6%	45.3%	
	Gaussian	.047(.0038)	.037(.0015)	-70.2%	.030(.0005)	.030(.0005)	.030(.0004)
			21.3%		36.2%	36.2%	
Mushroom	Linear	.232(.0135)	.204(.0113)		.186(.0095)	.184(.0095)	.041(.0018)
8124×22			12.1%	.126(.0090)	19.8%	20.7%	
	Gaussian	.217(.0135)	.217(.0117)	41.9%	.173(.0126)	.164(.0123)	.021(.0014)
			0.0%		20.3%	24.4%	
Email	Linear	.216(.0097)	.227(.0120)		.191(.0114)	.189(.0107)	.095(.0022)
4601×57			-5.09%	.232(.0101)	11.6%	12.5%	
	Gaussian	.226(.0108)	.275(.0158)	-7.41%	.189(.0120)	.189(.112)	.099(.0018)
			-21.7%		16.4%	16.4%	

Table 2: Averaged test errors as well as the estimated standard errors (in parenthesis) of SVM with labeled data alone, TSVM, the graphical method, SSVM and SPSI, over 100 pairs of training and testing samples, in the benchmark examples. The amount of improvement is defined in (12), where the performance of SVM with labeled data alone serves as a baseline for comparison in absence of the Bayes error.

- In the simulated examples, the improvements of SPSI and SSVM are from 66.9% to 87.9% over SVM, while the improvements of TSVM and the graphical method are from 34.5% to 52.9% and 49.2% to 61.5%, over SVM.
- In the benchmark examples, the improvements of SPSI, SSVM, TSVM, and the graphical method, over SVM, range from 19.8% to 45.3%, from -45.3% to 21.3%, and from -70.2% to 41.9%.
- It appears that the ψ-loss performs slightly better than the SVM hinge loss in almost all examples.

- SPSI and SSVM nearly reconstruct all relevant information about labeling in the two simulated examples and the WBC example, when they are compared with SVM with full label data. This suggests that room for further improvement in these cases is small.

To understand how SPSI and SSVM perform, we examine one randomly chosen realization in Example 1 for SPSI. As displayed in Figure 3, SVM fails to provide an accurate estimate of the true decision boundaries, because of the small size of labeled data. In contrast, the grouping boundaries estimated by unlabeled covariates, almost recover the true decision boundaries for classification. This, together with the information obtained from the labeled data regarding the sign of labeling, results in much better estimated classification boundaries.

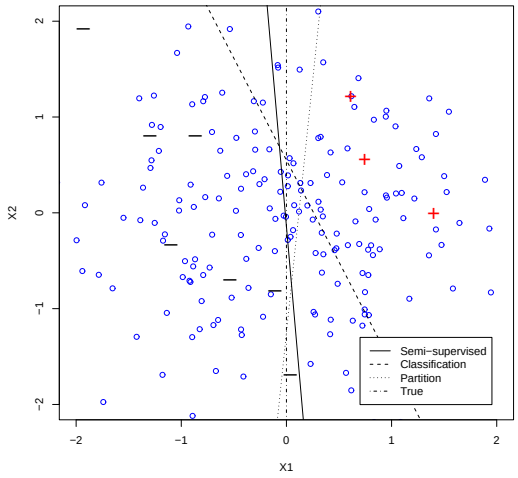

Figure 3: Illustration of SPSI in one randomly selected replication of Example 1. The solid, dashed, dotted and dotted-dashed (vertical) lines represent our ψ-learning-based decision function, the SVM decision function with labeled data alone, the partition decision function defined by unlabeled data, and the true decision boundary for classification. Here $C_1 = 0.1$, $C_2 = 0.01$ and $C_3 = 0.5$.

5.2 Performance After Tuning

This section compares the performances of the six methods in Section 5.1 when tuning is done using our proposed method in Section 4 and the training sample only. Particularly, SVM is tuned using the method of Wang and Shen (2006) with labeled data alone, and SPSI, SSVM , TSVM and the graphical method are tuned by minimizing the $\widehat{GE}(\hat{f}_C)$ in (10) involving both labeled and unlabeled data over a set of grid points in the same fashion as in Section 5.1. Performances of all the methods are evaluated by a test error on an independent test sample. The averaged test errors of these methods are summarized in Table 3.

As expected, SPSI and SSVM outperform both SVM with labeled data alone and TSVM in all cases, and the graphical method in all examples except Mushroom, with improvements ranging from 2.15% to 77.5% over SVM.

Data	Method	SVM_l	TSVM Improv.	Graph Improve.	SSVM Improv.	SPSI Improv.	SVM_c
Example 1	Linear	.350(.0107)	.281(.0153) 36.7%	.244(.0112)	.234(.0106) 61.7%	.233(.0106) 62.2%	.167(.0085)
	Gaussian	.395(.0101)	.331(.0211) 27.5%	56.4%	.280(.0176) 49.4%	.273(.0177) 52.4%	.258(.0102)
Example 2	Linear	.338(.0146)	.252(.0144) 34.5%	.227(.0129)	.148(.0104) 76.3%	.145(.0111) 77.5%	.118(.0084)
	Gaussian	.375(.0153)	.303(.0196) 25.2%	44.6%	.248(.0167) 44.4%	.233(.175) 49.7%	.201(.0123)
WBC	Linear	.060(.0081)	.094(.0131) -56.7%	.087(.0247)	.045(.0044) 25.0%	.042(.0035) 30.0%	.037(.0027)
	Gaussian	.051(.0039)	.044(.0047) 13.7%	-70.6%	.039(.0016) 21.6%	.039(.0018) 21.6%	.038(.0005)
Mushroom	Linear	.241(.0141)	.211(.0120) 12.4%	.137(.0101)	.209(.0108) 13.3%	.209(.0111) 13.3%	.053(.0037)
	Gaussian	.230(.0148)	.232(.0140) -0.87%	40.4%	.219(.0156) 4.78%	.210(.0131) 8.69%	.036(.0045)
Email	Linear	.236(.0109)	.241(.0128) -2.12%	.240(.0117)	.228(.0130) 3.39%	.224(.0125) 5.08%	.099(.0024)
	Gaussian	.233(.0107)	.296(.0136) -27.0%	-1.69%	.227(.0130) 2.58%	.228(.0131) 2.15%	.123(.0056)

Table 3: Averaged test errors as well as the estimated standard errors (in parenthesis) of SVM with labeled data alone, TSVM, the graphical method, SSVM and SPSI after tuning, over 100 pairs of training and testing samples, for the simulated and benchmark examples.

In conclusion, our proposed methodology achieves the desired objective of delivering high performance and is highly competitive against the top performers in the literature, where the loss $U(\cdot)$ plays a critical role in estimating decision boundaries for classification. It is also interesting to note that TSVM obtained from SVM^{light} performs even worse than SVM with labeled data alone in the WBC example for linear learning, and the Spam email example for both linear and Gaussian kernel learning. One possible explanation is that SVM^{light} may not have some difficulty in reaching good minimizers for TSVM. Moreover, the graphical method compares favorably against SVM and TSVM, but its performance does not seem to be robust in different examples. This may be due to the required Gaussian assumption.

6. Statistical Learning Theory

This section derives a finite-sample probability upper bound measuring the performance of SPSI in terms of complexity of the class of candidate decision functions $\mathcal{F}$, sample sizes (n_l, n_u) and tuning parameter C. Specifically, the generalization performance of the SPSI decision function $\hat{f}_C$ is measured by the Bayesian regret $e(f, f^*) = GE(f) - GE(f^*) \geq 0$ that is the difference between the actual performance of f and the ideal performance defined by the Bayes rule f^*. This yields SPSI's performance $\inf_C |e(\hat{f}_C, f^*)|$ after tuning.

6.1 Assumptions and Theorems

Our statistical learning theory involves risk minimization and the empirical process theory. The reader may consult Shen and Wang (2006) for a discussion about a learning theory of this kind.

First we introduce some notations. Let $(f_C^*, g_C^*) = \arg\inf_{f,g\in\mathcal{F}} ES(f,g;C)$ is a minimizer for surrogate risk $ES(f,g;C)$, as defined in Lemma 1. Let $e_f = e(f,f^*)$ be the Bayesian regret for f and $e_g = e(g,g_C^*)$ be the corresponding version for g relative to g_C^*. Denote by $V_f(X) = L(Yf(X)) - L(Yf^*(X))$ and $V_g(X) = \tilde{U}(g(X)) - \tilde{U}(g_C^*(X))$ be the differences between f and f^*, and g and g_C^* with respect to surrogate loss L and regularized surrogate loss $\tilde{U}(g) = U(g) + \frac{C_3}{2n_uC_2}\|g - f_C^*\|^2$.

To quantify complexity of $\mathcal{F}$, we define the L_2-metric entropy with bracketing. Given any $\varepsilon > 0$, denote $\{(f_m^l, f_m^u)\}_{m=1}^M$ as an ε-bracketing function set of $\mathcal{F}$ if for any $f \in \mathcal{F}$, there exists an m such that $f_m^l \le f \le f_m^u$ and $\|f_m^l - f_m^u\|_2 \le \varepsilon; m = 1,\cdots,M$, where $\|\cdot\|_2$ is the usual L_2 norm. Then the L_2-metric entropy with bracketing $H(\varepsilon,\mathcal{F})$ is defined as the logarithm of the cardinality of smallest ε-bracketing function set of $\mathcal{F}$.

Three technical assumptions are formulated based upon local smoothness of L, complexity of $\mathcal{F}$ as measured by the metric entropy, and a norm relationship.

Assumption A. (Local smoothness: Mean and variance relationship) For some some constants $0 < \alpha_h < \infty$, $0 \le \beta_h < 2$, $a_j > 0; j = 1,2$,

$$\sup_{\{h\in\mathcal{F}:\, E(V_h(X))\le\delta\}} |e_h| \le a_1\delta^{\alpha_h}, \tag{13}$$

$$\sup_{\{h\in\mathcal{F}:\, E(V_h(X))\le\delta\}} \mathrm{Var}(V_h(X)) \le a_2\delta^{\beta_h}, \tag{14}$$

for any small $\delta > 0$ and $h = f, g$.

Assumption A describes the local behavior of mean (e_h)-and-variance $(\mathrm{Var}(V_h(X)))$ relationship. In (13), Taylor's expansion usually leads to $\alpha_h = 1$ when f and g can be parameterized. In (14), the worst case is $\beta_h = 0$ because $\max(|L(yf)|, |U(g)|) \le 2$. In practice, values for α_h and β_h depend on the distribution of (X,Y).

Let $J_0 = \max(J(g_C^*), 1)$ with $J(g) = \frac{1}{2}\|g\|_-^2$ the regularizer. Let $\mathcal{F}_l(k) = \{L(yf) - L(yf^*) : f \in \mathcal{F}, J(f) \le k\}$ and $\mathcal{F}_u(k) = \{U(g) - U(g_C^*) : g \in \mathcal{F}, J(g) \le kJ_0\}$ be the regularized decision function spaces for f's and g's.

Assumption B. (Complexity) For some constants $a_i > 0; i = 3,\cdots,5$ and ε_{n_v} with $v = l$ or u,

$$\sup_{k\ge 2} \phi_v(\varepsilon_{n_v}, k) \le a_5 n_v^{1/2}, \tag{15}$$

where $\phi_u(\varepsilon,k) = \int_{a_4T_u}^{a_3^{1/2}T_u^{\beta_g/2}} H^{1/2}(w,\mathcal{F}_u(k))dw/T_u$ with $T_u = T_u(\varepsilon,C,k) = \min(1, \varepsilon^{2/\beta_g}/2 + (n_uC_2)^{-1}(k/2-1)J_0)$, and $\phi_l(\varepsilon,k) = \int_{a_4T_l}^{a_3^{1/2}T_l^{\beta_f/2}} H^{1/2}(w,\mathcal{F}_l(k))dw/T_l$ with $T_l = T_l(\varepsilon,C,k) = \min(1, \varepsilon^{2/\beta_f}/2 + (n_lC_1)^{-1}(k/2-1)\max(J(f^*),1))$.

Although Assumption B is always satisfied by some ε_{n_v}, the smallest possible ε_{n_v} from (15) yields the best possible error rate, for given $\mathcal{F}_v$ and sample size n_v. This is to say that the rate is indeed governed by the complexity of $\mathcal{F}_v(k)$. An equation of this type, originated from the empirical process theory, has been widely used in quantifying the error rates in function estimation, see, for example, Shen and Wong (1994).

Assumption C. (Norm relationship) For some constant $a_6 > 0$, $\|f\|_1 \le a_6\|f\|$ for any $f \in \mathcal{F}$, where $\|\cdot\|_1$ is the usual L_1-norm.

Assumption C specifies a norm relationship between norm $\|\cdot\|$ defined by a RKHS and $\|\cdot\|_1$. This is usually met when $\mathcal{F}$ is a RKHS, defined, for instance, by Gaussian and Sigmoid kernels, compare with Adams (1975).

Theorem 3 *(Finite-sample probability bound for SPSI) In addition to Assumptions A-C, assume that $n_l \le n_u$. For the SPSI classifier* $\mathrm{Sign}(\hat{f}_C)$*, there exist constants $a_j > 0$; $j = 1,6,7,10,11$, and $J_l > 0$, $J_u > 0$ and $B \ge 1$ defined as in Lemma 5, such that*

$$\begin{aligned} P\big(\inf_C |e(\hat{f}_C, f^*)| \ge a_1 s_n\big) \le\ & 3.5\exp(-a_7 n_u((n_u C_2^*)^{-1} J_0)^{\max(1,2-\beta_g)}) + \\ & 6.5\exp(-a_{10} n_l((n_l C_1^*)^{-1}\min(J_l, J(f^*)))^{\max(1,2-\beta_f)}) + \\ & 6.5\exp(-a_{11} n_u((n_u C_2^*)^{-1} J_u)^{\max(1,2-\beta_g)}), \end{aligned}$$

where $s_n = \min\big(\delta_{n_l}^{2\alpha_f}, \max(\delta_{n_u}^{2\alpha_g}, \inf_{C\in\mathcal{C}}|e(g_C^*, f^*)|)\big)$, $\delta_{n_\nu} = \min(\varepsilon_{n_\nu}, 1)$ *with* $\nu = l, u$, $C^* = (C_1^*, C_2^*, C_3^*) = \arg\inf_{C\in\mathcal{C}}|e(g_C^*, f^*)|)$, *and* $\mathcal{C} = \{C : n_l C_1 \ge 2\delta_{n_l}^{-2}\max(J_l, J(f^*), 1), n_u C_2 \ge 2\delta_{n_u}^{-2}\max(J_0, 2C_3(2B + J(f_C^*) + J(g_C^*))), C_3 \ge a_6^2 B\delta_{n_u}^{-4}\}$.

Corollary 4 *Under the assumptions of Theorem 3, as $n_u \ge n_l \to \infty$,*

$$\inf_C |e(\hat{f}_C, f^*)| = O_p(s_n),\ \ s_n = \min\big(\delta_{n_l}^{2\alpha_f}, \max(\delta_{n_u}^{2\alpha_g}, \inf_{C\in\mathcal{C}}|e(g_C^*, f^*)|)\big).$$

Theorem 3 provides a probability bound for the upper tail of $|e(\hat{f}_C, f^*)|$ for any finite (n_l, n_u). Furthermore, Corollary 4 says that the Bayesian regret $\inf_{C\in\mathcal{C}}|e(g_C^*, f^*)|$ for the SPSI classifier $\mathrm{Sign}(\hat{f}_C)$ after tuning is of order of no larger than s_n, when $n_u \ge n_l \to \infty$. Asymptotically, SPSI performs no worse than its supervised counterpart in that $\inf_C |e(\hat{f}_C, f^*)| = O_p(\delta_{n_l}^{2\alpha_f})$. Moreover, SPSI can outperform its supervised counterpart in the sense that $\inf_C |e(\hat{f}_C, f^*)| = O_p(\min(\delta_{n_u}^{2\alpha_g}, \delta_{n_l}^{2\alpha_f})) = O_p(\delta_{n_u}^{2\alpha_g})$, when $\{g_C^* : C \in \mathcal{C}\}$ provides a good approximation to the Bayes rule f^*.

Remark: Theorem 3 and Corollary 4 continue to hold when the "global" entropy in (15) is replaced by a "local" entropy, compare with Van De Geer (1993). Let $\mathcal{F}_{l,\xi}(k) = \{L(yf) - L(yf^*) : f \in \mathcal{F}, J(f) \le k, |e(f, f^*)| \le \xi\}$ and $\mathcal{F}_{u,\xi}(k) = \{U(g) - U(g_C^*) : g \in \mathcal{F}, J(g) \le k, |e(g, g_C^*)| \le \xi\}$ be the "local" entropy of $\mathcal{F}_l(k)$ and $\mathcal{F}_u(k)$. The proof requires only a slight modification. The local entropy avoids a loss of $\log n_u$ factor in the linear case, although it may not be useful in the nonlinear case.

6.2 Theoretical Examples

We now apply the learning theory to one linear and one kernel learning examples to obtain the generalization error rates for SPSI, as measured by the Bayesian regret. We will demonstrate that the error in the linear case can be arbitrarily fast while that in the nonlinear case is fast. In either case, SPSI's performance is better than that of its supervised counterpart.

Linear learning: Consider linear classification where $X = (X_{(1)}, X_{(2)})$ is sampled independently according to the same probability density $q(z) = \frac{1}{2}(\theta+1)|z|^\theta$ for $z \in [-1,1]$ with $\theta \ge 1$. Given X, assign label Y to 1 if $X_{(1)} > 0$ and -1 otherwise; then Y is chosen randomly to flip with constant

probability τ for $0 < \tau < \frac{1}{2}$. Here the true decision function $f_t(x) = x_{(1)}$ yielding the vertical line as the classification boundary.

In this case, the degree of smoothness of this problem is characterized by exponent $\theta > 0$ in the density $q(z)$, which describes the level of difficulty of linear classification but may not be so in the nonlinear case.

For classification, we minimize (2) over $\mathcal{F}$, consisting of linear decision functions of form $f(x) = (1,x)^T w$ for $w \in \mathcal{R}^3$ and $x = (x_{(1)}, x_{(2)}) \in \mathcal{R}^2$. To apply Corollary 4, we verify Assumptions A-C with detailed verification given in Appendix B. In fact, Assumption A follows from the smoothness of $E(V_h(X))$ and $\mathrm{Var}(V_h(X))$ with respect to h, where a local Taylor expansion yields the degree of smoothness exponents α and β. Assumption B is automatically met, and the entropy Equation (15) is solved for the smallest possible ε_{n_v} satisfying it. Assumption C is always true for RKHS. It then follows from Corollary 4 that $\inf_C |e(\hat{f}_C, f^*)| = O_p(n_u^{-(\theta+1)/2} (\log n_u)^{(\theta+1)/2})$ as $n_u \geq n_l \to \infty$. This says that the optimal ideal performance of the Bayes rule is recovered by SPSI at speed of $n_u^{-(\theta+1)/2} (\log n_u)^{(\theta+1)/2}$ as $n_u \geq n_l \to \infty$. This rate is arbitrarily fast as $\theta \to \infty$.

Kernel learning: Consider, in the preceding case, kernel learning with a different candidate decision function class defined by the Gaussian kernel. To specify $\mathcal{F}$, we may embed a finite-dimensional Gaussian kernel representation into an infinite-dimensional space $\mathcal{F} = \{x \in \mathcal{R}^2 : f(x) = w_f^T \phi(x) = \sum_{k=0}^{\infty} w_{f,k} \phi_k(x) : w_f = (w_{f,0}, \cdots)^T \in \mathcal{R}^\infty\}$ by the representation theorem of RKHS, compare with Wahba (1990). Here $\langle \phi(x), \phi(z) \rangle = K(x,z) = \exp(-\frac{\|x-z\|^2}{2\sigma^2})$.

To apply Corollary 4, we verify Assumptions A-C as before, with detailed verification given in Appendix B. The function space $\mathcal{F}$ generated by the Gaussian kernel is rich enough to well approximate the ideal performer $\mathrm{Sign}(E(Y|X))$ (Steinwart, 2001), and yields the exponents α and β in Assumption A with smoothness and Soblev's inequality (Adams, 1975). Similarly, it follows from Corollary 4 that $\inf_C |e(\hat{f}_C, f^*)| = O_p(\min(n_l^{-1} (\log n_l J_l)^3, n_u^{-1/2} (\log n_u J_u)^{3/2}))$ as $n_u \geq n_l \to \infty$. Therefore, the optimal ideal performance of the Bayes rule is recovered by SPSI at fast speed of $\min(n_l^{-1} (\log n_l J_l)^3, n_u^{-1/2} (\log n_u J_u)^{3/2})$ as $n_u \geq n_l \to \infty$.

7. Discussion

This article proposed a novel large margin semi-supervised learning methodology that is applicable to a class of large margin classifiers. In contrast to most semi-supervised learning methods assuming various dependencies between the marginal and conditional distributions, the proposed methodology integrates labeled and unlabeled data through regularization to identify such dependencies for enhancing classification. The theoretical and numerical results show that our methodology outperforms SVM and TSVM in situations when unlabeled data provides useful information, and performs no worse when unlabeled data does not so. For tuning, further investigation of regularization paths of our proposed methodology is useful as in Hastie, Rosset, Tibshirani and Zhu (2004), to reduce computational cost.

Acknowledgments

This research is supported by NSF grants IIS-0328802 and DMS-0604394. We thank Wei Pan for many constructive comments. We also thank three referees and the editor for helpful comments and suggestions.

Appendix A. Technical Proofs

Proof of Theorem 2: The proof is similar to that of Theorem 2 of Wang and Shen (2006), and thus is omitted.

Proof of Theorem 3: The proof uses a large deviation empirical technique for risk minimization. Such a technique has been previously developed in function estimation as in Shen and Wong (1994). The proof proceeds in three steps. In **Step 1**, the tail probability of $\{e_{\widetilde{U}}(\hat{g}_C, g_C^*) \geq \delta_{n_u}^2\}$ is bounded through a large deviation probability inequality of Shen and Wong (1994). In **Step 2**, a tail probability bound of $\{|e(\hat{f}_C, f^*)| \geq \delta_{n_u}^2\}$ is induced from **Step 1** using a conversion formula between $e_{\widetilde{U}}(\hat{g}_C, g_C^*)$ and $|e(\hat{f}_C, f^*)|$. In **Step 3**, a probability upper bound for $\{|e(\hat{f}_C, f^*)| \geq \delta_{n_l}^2\}$ is obtained using the same treatment as above. The desired bound is obtained based on the bounds in **Step 2** and **Step 3**.

Step 1: It follows from Lemma 5 that $\max(\|\hat{f}_C\|^2, \|\hat{g}_C\|^2) \leq B$ for a constant $B \geq 1$, where $(\hat{f}_C, \hat{g}_C)$ is the minimizer of (2). Furthermore, $\hat{g}_C$ defined in (2) can be written as $\hat{g}_C = \arg\min_{g \in \mathcal{F}} \{C_2 \sum_{j=n_l+1}^{n} \widetilde{U}(g(x_j)) + J(g) + \frac{C_3}{2}(\|\hat{f}_C - g\|^2 - \|f_C^* - g\|^2)\}$.

By the definition of $\hat{g}_C$, $P(e_{\widetilde{U}}(\hat{g}_C, g_C^*) \geq \delta_{n_u}^2)$ is upper bounded by

$$
\begin{aligned}
& P(J(\hat{g}_C) \geq B) + P^*\Big(\sup_{g \in N} n_u^{-1} \sum_{j=n_l+1}^{n} (\widetilde{U}(g_C^*(x_j)) - \widetilde{U}(g(x_j))) + \lambda(J(g_C^*) - J(g)) \\
& \qquad + \frac{\lambda C_3}{2}(\|\hat{f}_C - g_C^*\|^2 - \|f_C^* - g_C^*\|^2 - \|\hat{f}_C - g\|^2 + \|f_C^* - g\|^2) \geq 0\Big) \\
\leq \quad & P(J(\hat{g}_C) \geq B) + P^*\Big(\sup_{g \in N} n_u^{-1} \sum_{j=n_l+1}^{n} (\widetilde{U}(g_C^*(x_j)) - \widetilde{U}(g(x_j))) + \lambda(J(g_C^*) - J(g)) \\
& \qquad + \lambda C_3(2B + J(f_C^*) + J(g_C^*)) \geq 0\Big) \equiv P(J(\hat{g}_C) \geq B) + I,
\end{aligned}
$$

where $\lambda = (n_u C_2)^{-1}$, $N = \{g \in \mathcal{F}, J(g) \leq B, e_{\widetilde{U}}(g, g_C^*) \geq \delta_{n_u}^2\}$, and P^* denotes the outer probability. By Lemma , there exists constants $a_{10}, a_{11} > 0$ such that $P(J(\hat{g}_C) \geq B) \leq 6.5\exp(-a_{10} n_l (n_l C_1)^{-1} J_l)$ $+6.5\exp(-a_{11} n_u (n_u C_2)^{-1} J_u)$, where J_l and J_u are defined in Lemma 5.

To bound I, we introduce some notations. Define the scaled empirical process as $E_u(\widetilde{U}(g_C^*) - \widetilde{U}(g)) = n_u^{-1} \sum_{j=n_l+1}^{n} \big(\widetilde{U}(g_C^*(x_j)) - \widetilde{U}(g(x_j)) + \lambda(J(g_C^*) - J(g))\big) - E(\widetilde{U}(g_C^*(X_j)) - \widetilde{U}(g(X_j)) + \lambda(J(g_C^*) - J(g))) = E_u(U(g_C^*) - U(g))$. Thus

$$
\begin{aligned}
I = P^* \Bigg(& \sup_{g \in N} E_u(U(g_C^*) - U(g)) \geq \\
& \inf_{g \in N} E(\widetilde{U}(g(X)) - \widetilde{U}(g_C^*(X))) + \lambda(J(g_C^*) - J(g)) - \lambda C_3(2B + J(f_C^*) + J(g_C^*))\Bigg).
\end{aligned}
$$

Let $A_{s,t} = \{g \in \mathcal{F} : 2^{s-1}\delta_{n_u}^2 \leq e_{\widetilde{U}}(g, g_C^*) < 2^s \delta_{n_u}^2, 2^{t-1} J_0 \leq J(g) < 2^t J_0\}$, and let $A_{s,0} = \{g \in \mathcal{F} : 2^{s-1}\delta_{n_u}^2 \leq e_{\widetilde{U}}(g, g_C^*) < 2^s \delta_{n_u}^2, J(g) < J_0\}$; $s, t = 1, 2, \cdots$. Without loss of generality, we assume that $\varepsilon_{n_u} < 1$. Then it suffices to bound the corresponding probability over $A_{s,t}$; $s, t = 1, 2, \cdots$. Toward this end, we control the first and second moment of $\widetilde{U}(g_C^*(X)) - \widetilde{U}(g(X))$ over $f \in A_{s,t}$.

For the first moment, by assumption $\delta_{n_u}^2 \geq 2\lambda \max(J_0, 2C_3(2B + J(f_C^*) + J(g_C^*)))$,

$$
\inf_{A_{s,t}} E(\widetilde{U}(g(X)) - \widetilde{U}(g_C^*(X))) + \lambda(J(g_C^*) - J(g)) \geq 2^{s-1}\delta_{n_u}^2 + \lambda(2^{t-1} - 1)J_0; s, t = 1, 2, \cdots,
$$

$$\inf_{A_{s,0}} E(\widetilde{U}(g(X)) - \widetilde{U}(g_C^*(X))) + \lambda(J(g_C^*) - J(g)) \geq (2^{s-1} - 1/2)\delta_{n_u}^2 \geq 2^{s-2}\delta_{n_u}^2; s = 1, 2, \cdots.$$

Therefore, $\inf_{A_{s,t}} E(\widetilde{U}(g(X)) - \widetilde{U}(g_C^*(X))) + \lambda(J(g_C^*) - J(g)) - \lambda C_3(2B + J(f_C^*) + J(g_C^*)) \geq M(s,t) = 2^{s-2}\delta_{n_u}^2 + \lambda(2^{t-1} - 1)J_0$, and $\inf_{A_{s,0}} E(\widetilde{U}(g(X)) - \widetilde{U}(g_C^*(X))) + \lambda(J(g_C^*) - J(g)) - \lambda C_3(2B + J(f_C^*) + J(g_C^*)) \geq M(s,0) = 2^{s-3}\delta_{n_u}^2$, for all $s,t = 1, 2, \cdots$.

For the second moment, by Assumptions A,

$$\begin{aligned} & \sup_{A_{s,t}} \mathrm{Var}(\widetilde{U}(g(X)) - \widetilde{U}(g_C^*(X))) \leq \sup_{A_{s,t}} a_2 (e_{\widetilde{U}}(g, g_C^*))^{\beta_g} \leq a_2 (2^s \delta_{n_u}^2 + (2^t - 1)\lambda J_0)^{\beta_g} \\ \leq \; & a_2 2^{3\beta_g} (2^{s-2}\delta_{n_u}^2 + (2^{t-1} - 1)\lambda J_0)^{\beta_g} \leq a_3 M(s,t)^{\beta_g} = v^2(s,t), \end{aligned}$$

for and $s,t = 1, 2, \cdots$ and some constant $a_3 > 0$.

Now $I \leq I_1 + I_2$ with $I_1 = \sum_{s,t=1}^{\infty} P^*(\sup_{A_{s,t}} E_u(U(g_C^*) - U(g)) \geq M(s,t))$; $I_2 = \sum_{s=1}^{\infty} P^*(\sup_{A_{s,0}} E_u(U(g_C^*) - U(g)) \geq M(s,0))$. Next we bound I_1 and I_2 separately using Theorem 3 of Shen and Wong (1994). We now verify conditions (4.5)-(4.7) there. To compute the metric entropy of $\{U(g) - U(g_C^*) : g \in A_{s,t}\}$ in (4.7) there, we note that $\int_{aM(s,t)}^{v(s,t)} H^{1/2}(w, \mathcal{F}_u(2^t))dw/M(s,t)$ is nonincreasing in s and $M(s,t)$ and hence that

$$\begin{aligned} \int_{aM(s,t)}^{v(s,t)} H^{1/2}(w, \mathcal{F}_u(2^t))dw/M(s,t) &\leq \int_{aM(1,t)}^{a_3^{1/2} M(1,t)^{\beta_g/2}} H^{1/2}(w, \mathcal{F}_u(2^t))dw/M(1,t) \\ &\leq \phi(\varepsilon_{n_u}, 2^t), \end{aligned}$$

with $a = 2a_4\varepsilon$. Assumption B implies (4.7) there with $\varepsilon = 1/2$ and some $a_i > 0$; $i = 3, 4$. Furthermore, $M(s,t)/v^2(s,t) \leq 1/8$ and $T = 1$ imply (4.6), and (4.7) implies (4.5). By Theorem 3 of Shen and Wong (1994), for some constant $0 < \zeta < 1$,

$$\begin{aligned} I_1 &\leq \sum_{s,t=1}^{\infty} 3\exp\left(-\frac{(1-\zeta)n_u M^2(s,t)}{2(4v^2(s,t) + M(s,t)/3)}\right) \leq \sum_{s,t=1}^{\infty} 3\exp(-a_7 n_u (M(s,t))^{\max(1,2-\beta_g)}) \\ &\leq \sum_{s,t=1}^{\infty} 3\exp(-a_7 n_u (2^{s-1}\delta_{n_u}^2 + \lambda(2^{t-1} - 1)J_0)^{\max(1,2-\beta_g)}) \\ &\leq 3\exp(-a_7 n_u (\lambda J_0)^{\max(1,2-\beta_g)})/(1 - \exp(-a_7 n_u (\lambda J_0)^{\max(1,2-\beta_g)}))^2. \end{aligned}$$

Similarly, $I_2 \leq 3\exp(-a_7 n_u (\lambda J_0)^{\max(1,2-\beta_g)})/(1 - \exp(-a_7 n_u (\lambda J_0)^{\max(1,2-\beta_g)}))^2$. Thus $I \leq I_1 + I_2 \leq 6\exp(-a_7 n_u ((n_u C_2)^{-1} J_0)^{\max(1,2-\beta_g)})/(1 - \exp(-a_7 n_u ((n_u C_2)^{-1} J_0)^{\max(1,2-\beta_g)}))^2$, and $I^{1/2} \leq (2.5 + I^{1/2}) \exp(-a_7 n_u ((n_u C_2)^{-1} J_0)^{\max(1,2-\beta_g)})$. Thus $P(e_{\widetilde{U}}(\hat{g}_C, g_C^*) \geq \delta_{n_u}^2) \leq 3.5\exp(-a_7 n_u ((n_u C_2)^{-1} J_0)^{\max(1,2-\beta_g)}) + 6.5\exp(-a_{10} n_l ((n_l C_1)^{-1} J_l)^{\max(1,2-\beta_f)}) + 6.5\exp(-a_{11} n_u ((n_u C_2)^{-1} J_u)^{\max(1,2-\beta_f)})$.

Step 2: By Lemma 5 and Assumption C, $|e_{\widetilde{U}}(\hat{f}_C, \hat{g}_C)| \leq E|\hat{f}_C(X) - \hat{g}_C(X)| \leq a_6 \|\hat{f}_C - \hat{g}_C\| \leq a_6\sqrt{B/C_3} \leq \delta_{n_u}^2$ when $C_3 \geq a_6^2 B \delta_{n_u}^{-4}$. By Assumption A and the triangle inequality, $|e(\hat{f}_C, g_C^*)| \leq a_1(e_{\widetilde{U}}(\hat{f}_C, g_C^*))^{\alpha_g} \leq a_1(e_{\widetilde{U}}(\hat{g}_C, g_C^*) + |e_{\widetilde{U}}(\hat{f}_C, \hat{g}_C)|)^{\alpha_g} \leq a_1(e_{\widetilde{U}}(\hat{g}_C, g_C^*) + \delta_{n_u}^2$, implying that $P\big(|e(\hat{f}_C, g_C^*)| \geq a_1(2\delta_{n_u}^2)^{\alpha_g}\big) \leq P(e_{\widetilde{U}}(\hat{g}_C, g_C^*) \geq \delta_{n_u}^2)$, $\forall C \in \mathcal{C}$. Then $P\big(\inf_C |e(\hat{f}_C, f^*)| \geq a_1(2\delta_{n_u}^2)^{\alpha_g} + \inf_{C \in \mathcal{C}} |e(g_C^*, f^*)|\big) \leq P(e_{\widetilde{U}}(\hat{g}_{C^*}, g_{C^*}^*) \geq \delta_{n_u}^2) \leq 3.5\exp(- a_7 n_u ((n_u C_2^*)^{-1} J_0)^{\max(1,2-\beta_g)}) +$

$6.5\exp(-a_{10}n_l((n_lC_1^*)^{-1}J_l)^{\max(1,2-\beta_f)}) \;+\; 6.5\exp(-a_{11}n_u\;((n_uC_2^*)^{-1}J_u)^{\max(1,2-\beta_g)})$, where $C^* = \arg\inf_{C\in\mathcal{C}}|e(g_C^*,f^*)|$.

Step 3: Note that $\hat{f}_C = \text{argmax}_{f\in\mathcal{F}}\{C_1\sum_{i=1}^{n_l}L(y_if(x_i)) + \frac{1}{2}\|f\|_-^2\}$ when $C_2 = 0$ and $C_3 = \infty$. An application of the same treatment yields that $P(\inf_C e(\hat{f}_C,f^*) \geq a_1\delta_{n_l}^2) \leq P(\inf_C e_L(\hat{f}_C,f^*) \geq a_1\delta_{n_l}^2) \leq 3.5\exp(-a_{10}n_l((n_lC_1^*)^{-1}J(f^*))^{\max(1,2-\beta_f)})$ when $n_lC_1^* \geq 2\delta_{n_l}^{-2}\max(J(f^*),1)$. The desired result follows.

Lemma 5 *Under the assumptions of Theorem 3, for $(\hat{f}_C,\hat{g}_C)$ as the minimizer of (2), there exists constants $B>0$, depending only on C_1, such that*

$$\max(E(C_3\|\hat{f}_C-\hat{g}_C\|^2+\|\hat{g}_C\|^2), E\|\hat{f}_C\|^2, 2C_1) \quad\leq\quad B.$$

Proof: It suffices to show $E(C_3\|\hat{f}_C-\hat{g}_C\|^2+\|\hat{g}_C\|^2)\leq B$. Let $\widetilde{W}(f,g) = \frac{1}{C_1}s(f,g) = \sum_{i=1}^{n_l}\widetilde{W}_l(y_if(x_i)) + \frac{C_2}{C_1}\sum_{j=n_l+1}^{n}\widetilde{W}_u(g(x_j))$, where $\widetilde{W}_l(f(x_i)) = L(y_if(x_i)) + \frac{C_3}{4n_lC_1}\|f-g\|^2$, and $\widetilde{W}_u(g(x_j)) = U(g(x_j)) + \frac{1}{2n_uC_2}\|g\|^2 + \frac{C_3}{4n_uC_2}\|f-g\|^2$. For convenience, write $J_l(f,g) = \frac{C_3}{4}\|f-g\|^2$, $J_u(f,g) = \frac{C_3}{4}\|f-g\|^2 + \frac{1}{2}\|g\|_-^2$, $\lambda_l = (C_1n_l)^{-1}$, and $\lambda_u = (C_2n_u)^{-1}$. We then define a new empirical process $E_{l,u}(\widetilde{W}(f,g) - \widetilde{W}(f_C^*,g_C^*)) = E_l(\widetilde{W}_l(f) - \widetilde{W}_l(f_C^*)) + \frac{C_2n_u}{C_1n_l}E_u(\widetilde{W}_u(g) - \widetilde{W}_u(g_C^*))$ as

$$\frac{1}{n_l}\sum_{i=1}^{n_l}\Big(\widetilde{W}_l(f(x_i)) - \widetilde{W}_l(f_C^*(x_i)) - E(\widetilde{W}_l(f(X_i)) - \widetilde{W}_l(f_C^*(X_i)))\Big) + \frac{C_2n_u}{C_1n_l}\frac{1}{n_u}\sum_{i=n_l+1}^{n}\Big(\widetilde{W}_u(g(x_j)) - \widetilde{W}_u(g_C^*(x_i)) - E(\widetilde{W}_u(g(X_j)) - \widetilde{W}_u(g_C^*(X_i)))\Big).$$

An application of the same argument as in the proof of Theorem 3 yields that for constants $a_8,a_9>0$, $P(e_W(\hat{f}_C,\hat{g}_C;f_C^*,g_C^*)\geq\tilde{\delta}_w^2)$ is upper bounded by

$$3.5\exp(-a_8n_l((n_lC_1)^{-1}J_l)^{\max(1,2-\beta_f)}) + 3.5\exp(-a_9n_u((n_uC_2)^{-1}J_u)^{\max(1,2-\beta_g)}),$$

provided that $2J_l\leq n_lC_1\tilde{\delta}_{n_l}^2$ and $2J_u\leq n_uC_2\tilde{\delta}_{n_u}^2$, where $e_W(f,g;f_C^*,g_C^*) = e_L(f,f_C^*) + \frac{C_2}{C_1}e_U(g,g_C^*)$, $\tilde{\delta}_w^2 = \tilde{\delta}_{n_l}^2 + \frac{C_2n_u}{C_1n_l}\tilde{\delta}_{n_u}^2$, $J_l = \max(J_l(f_C^*,g_C^*),1)$ and $J_u = \max(J_u(f_C^*,g_C^*),1)$.

Without loss of generality, assume $\min(J_l(f_C^*,g_C^*),J_u(f_C^*,g_C^*))\geq 1$. Let $J(f,g) = J_l(f,g) + J_u(f,g)$ and $A_t = \{f,g\in\mathcal{F} : e_W(f,g;f_C^*,g_C^*)\leq\tilde{\delta}_w^2, 2^{t-1}J(f_C^*,g_C^*)\leq J(f,g) < 2^tJ(f_C^*,g_C^*)\}; t=1,\cdots$. Then, $P\left(J(\hat{f}_C,\hat{g}_C)\geq J(f_C^*,g_C^*)\right)$ is upper bounded by

$$\begin{aligned}
& P(e_W(\hat{f}_C,\hat{g}_C;f_C^*,g_C^*)\geq\tilde{\delta}_w^2) + \\
& \sum_{t=1}^{\infty}P^*\left(\sup_{A_t}E_{l,u}(\widetilde{W}(f_C^*,g_C^*) - \widetilde{W}(f,g)) \geq E(\widetilde{W}(f,g) - \widetilde{W}(f_C^*,g_C^*))\right) \\
\leq\; & P(e_W(\hat{f}_C,\hat{g}_C;f_C^*,g_C^*)\geq\tilde{\delta}_w^2) + \\
& \sum_{t=1}^{\infty}P^*\left(\sup_{A_t}E_{l,u}(\widetilde{W}(f_C^*,g_C^*) - \widetilde{W}(f,g)) \geq (2^{t-1}-1)\lambda_lJ(f_C^*,g_C^*) + \tilde{\delta}_w^2\right) \\
\leq\; & P(e_W(\hat{f}_C,\hat{g}_C;f_C^*,g_C^*)\geq\tilde{\delta}_w^2) + \\
& \sum_{t=1}^{\infty}P^*\left(\sup_{A_t}E_l(\widetilde{W}_l(f_C^*) - \widetilde{W}_l(f)) \geq (2^{t-1}-1)\lambda_lJ_l + \tilde{\delta}_{n_l}^2\right) + \\
& \sum_{t=1}^{\infty}P^*\left(\sup_{A_t}E_u(\widetilde{W}_u(g_C^*) - \widetilde{W}_u(g)) \geq (2^{t-1}-1)\lambda_uJ_u + \tilde{\delta}_{n_u}^2\right).
\end{aligned}$$

An application of the same argument in the proof of Theorem 3 yields that for some constants $0 < a_{10} \le a_8$ and $0 < a_{11} \le a_9$ that $P\left(J(\hat{f}_C, \hat{g}_C) \ge J(f_C^*, g_C^*)\right)$ is upper bounded by

$$\begin{aligned} & P(e_W(f,g; f_C^*, g_C^*) \ge \tilde{\delta}_w^2) + \sum_{t=1}^{\infty} \left(3\exp(-a_{11} n_l ((n_l C_1)^{-1} J_l(f_C^*, g_C^*) 2^{t-1})^{\max(1,2-\beta_f)}) + \right. \\ & \left. 3\exp(-a_{12} n_u ((n_u C_2)^{-1} J_u(f_C^*, g_C^*) 2^{t-1})^{\max(1,2-\beta_g)})\right) \\ \le \quad & 6.5\exp(-a_{10} n_l ((n_l C_1)^{-1} J_l)^{\max(1,2-\beta_f)}) + 6.5\exp(-a_{11} n_u ((n_u C_2)^{-1} J_u)^{\max(1,2-\beta_g)}). \end{aligned}$$

Note that $J(\hat{f}_C, \hat{g}_C) \le s(\hat{f}, \hat{g}) \le s(1,1) \le 2C_1 n_l$. There exists a constant $B_1 > 0$ such that

$$E(C_3 \| \hat{f}_C - \hat{g}_C \|^2 + \| \hat{g}_C \|_-^2) \le J(f_C^*, g_C^*) + B_1 \le 2C_1 + B_1, \tag{16}$$

since $J(f_C^*, g_C^*) \le ES(f_C^*, g_C^*) \le ES(1,1) \le 2C_1$. It follows from the KKT condition and (16) that $E|w_{\hat{g}_C,0}|$ is bounded by a constant B_2, depending only on C_1. The desired result follows with a choice of $B = 2C_1 + B_1 + B_2^2$.

Lemma 6 *(Metric entropy in Example 6.2.1) Under the assumptions there, for $v = l$ or u,*

$$H(\varepsilon, \mathcal{F}_{v,\xi}(k)) \le O(\log(\xi^{1/(\theta+1)}/\varepsilon)).$$

Proof: We first show the inequality for $\mathcal{F}_{u,\xi}(k)$. Suppose lines $g(x) = 0$ and $g_C^*(x) = 0$ intersect lines $x_{(2)} = \pm 1$ with two points $(u_g, 1), (v_g, -1)$ and $(u_{g_C^*}, 1), (v_{g_C^*}, -1)$, respectively. Note that $e(g, g_C^*) \le \xi$ implies $P(\Delta(g, g_C^*)) \le \frac{\xi}{1-2\tau}$ with $\Delta(g, g_C^*) = \{\text{Sign}(g(x)) \ne \text{Sign}(g_C^*(x))\}$. Direct calculation yields that $P(\Delta(g, g_C^*)) \ge \frac{1}{2}\max(|u_g - u_{g_C^*}|, |v_g - v_{g_C^*}|)^{\theta+1}$, $\max(|u_g - u_{g_C^*}|, |v_g - v_{g_C^*}|) \le a' \xi^{1/(\theta+1)}$ for a constant $a' > 0$. We then cover all possible $(u_g, 1)$ and $(v_g, -1)$ with intervals of length ε^*. The covering number for these possible points is no more than $(2a'\xi^{1/(\theta+1)}/\varepsilon^*)^2$. After these points are covered, we then connect the endpoints of the covering intervals to form bracket planes $l(x) = 0$ and $u(x) = 0$ such that $l \le g \le u$, and $\|u - l\|_2 \le \|u - l\|_\infty \le \varepsilon^*$. Let $U^l(g) = 2 - 2\max(|l_{\pm 1}|, |u_{\pm 1}|)$ and $U^u(g) = 2 - 2I(l(x)u(x) > 0)\min(|l_{\pm 1}|, |u_{\pm 1}|)$, then $U^l(g) \le U(g) \le U^u(g)$ and $\|U^u(g) - U^l(g)\|_\infty \le 2\|u - l\|_\infty \le 2\varepsilon^*$. With $\varepsilon = 2\varepsilon^*$, $\{(U^l(g), U^u(g))\}$ forms an ε-bracketing set of $U(g)$. Therefore, the ε-covering number for $\mathcal{F}_{u,\xi}(k)$ is at most $(4a'\xi^{1/(\theta+1)}/\varepsilon)^2$, implying $H(\varepsilon, \mathcal{F}_{u,\xi}(k))$ is upper bounded by $O(\log(\xi^{\frac{1}{\theta+1}}/\varepsilon))$. Furthermore, it is similar to show the inequality for $\mathcal{F}_{l,\xi}(k)$ since $\left(2\min(1, 1 - \max(yl(x), yu(x))_+), 2\min(1, 1 - \min(yl(x), yu(x))_+)\right)$ forms a bracket for $L(yf(x))$ when $l \le f \le u$.

Lemma 7 *(Metric entropy in Example 6.2.2) Under the assumptions there, for $v = l$ or u,*

$$H(\varepsilon, \mathcal{F}_v(k)) \le O((\log(k/\varepsilon))^3).$$

Proof: We first show the inequality for $\mathcal{F}_u(k)$. Suppose there exist ε-brackets $(g_m^l, g_m^u)_{m=1}^M$ for some M such that for any $g \in \mathcal{F}(k) = \{g \in \mathcal{F} : J(g) \le k\}$, $g_m^l \le g \le g_m^u$ and $\|g_m^u - g_m^l\|_\infty \le \varepsilon$ for some $1 \le m \le M$. Let $U^l(g) = 2 - 2\max(|g_{m,\pm 1}^l|, |g_{m,\pm 1}^u|)$ and $U^u(g) = 2 - 2I(g_m^l g_m^u > 0)\min(|g_{m,\pm 1}^l|, |g_{m,\pm 1}^u|)$, then $U^l(g) \le U(g) \le U^u(g)$ and $\|U^u(g) - U^l(g)\|_\infty \le 2\|g_m^u - g_m^l\|_\infty \le 2\varepsilon$. Therefore, $(U^l(g) - U(g_C^*), U^u(g) - U(g_C^*))$ forms a bracket of length 2ε for $U(g) - U(g_C^*)$. The desired inequality then follows from the Example 4 in Zhou (2002) that $H_\infty(\varepsilon, \mathcal{F}(k)) \le O(\log(k/\varepsilon)^3)$ under the $L_\infty-$metric: $\|g\|_\infty = \sup_{x \in \mathcal{R}^2} |g(x)|$. Furthermore, it is similar to show the inequality for $\mathcal{F}_l(k)$ as in Lemma 6.

Lemma 8 *For any functions f, g and any constant $\rho > 0$,*

$$E|\operatorname{Sign}(f(X)) - \operatorname{Sign}(g(X))|I(|f(X)| \geq \rho) \leq 2\rho^{-1}E|f(X) - g(X)|.$$

Proof: The left hand side is $2P(|f(X)| \geq \rho, \operatorname{Sign}(f(X)) \neq \operatorname{Sign}(g(X))) \leq 2P(|f(x) - g(x)| \geq \rho) \leq 2\rho^{-1}E|f(X) - g(X)|$ by Chebyshev's inequality.

Appendix B. Verification of Assumptions A-C in the Theoretical Examples

Linear learning: Since $(X_{(1)}, Y)$ is independent of $X_{(2)}$, $ES(f,g;C) = E(E(S(f,g;C)|X_{(2)})) \geq ES(\tilde{f}_C^*, \tilde{g}_C^*; C)$ for any $f, g \in \mathcal{F}$, and $(\tilde{f}_C^*, \tilde{g}_C^*) = \arg\min_{\tilde{f},\tilde{g}\in\mathcal{F}_1} ES(\tilde{f},\tilde{g};C)$ with $\mathcal{F}_1 = \{x_{(1)} \in \mathcal{R} : \tilde{f}(x) = (1, x_{(1)})^T w : w \in \mathcal{R}^2\} \subset \mathcal{F}$. It then suffices to verify Assumptions A-C over $\mathcal{F}_1$ rather than $\mathcal{F}$. By Lemma 1, the approximation error $\inf_{C\in\mathcal{C}} e(\tilde{g}_C^*, f^*) = 0$. For (13), note that f^* minimizes $EL(Yf(X))$ and $\tilde{g}_C^*$ minimizes $E\widetilde{U}(\tilde{g})$ given $\tilde{f}_C^*$. Direct computation, together with Taylor's expansion yields that $E(V_{\tilde{h}}(X)) = (e_0, e_1)\Gamma_{\tilde{h}}(e_0, e_1)^T$ for any function $\tilde{h} = (1, x_{(1)})^T w_{\tilde{h}} \in \mathcal{F}_1$ with $w_{\tilde{h}} = w_{\tilde{h}^*} + (e_0, e_1)^T$, where $\tilde{h}^* = f^*$ or $\tilde{g}_C^*$ and $\Gamma_{\tilde{h}}$ is a positive definite matrix. Thus $E(V_{\tilde{h}}(X)) \geq \lambda_1(e_0^2 + e_1^2)$ for constant $\lambda_1 > 0$. Moreover, straightforward calculation yields that $|e_{\tilde{h}}| \leq \frac{1}{2}(1 - 2\tau)\min(|w_{\tilde{h}^*,1}|, |w_{\tilde{h}^*,1} + e_1|)^{-(\theta+1)}|e_0|^{\theta+1} \leq \lambda_2(e_0^2 + e_1^2)^{(\theta+1)/2}$ for some constant $\lambda_2 > 0$, where $w_{\tilde{h}^*} = (w_{\tilde{h}^*,0}, w_{\tilde{h}^*,1})$. A combination of these two inequalities leads to (13) with $\alpha_{\tilde{h}} = (\theta+1)/2$. For (14), note that $\operatorname{Var}(V_{\tilde{h}}(X)) \leq \|\tilde{h} - \tilde{h}^*\|_2^2 = e_0^2 + e_1^2 EX_{(1)}^2 \leq \max(1, EX_{(1)}^2)(e_0^2 + e_1^2)$. This implies (14) with $\beta_{\tilde{h}} = 1$. For Assumption B, by Lemma 6, $H(\varepsilon, \mathcal{F}_{v,\varepsilon}(k)) \leq O(\log(\varepsilon^{1/(\theta+1)})/\varepsilon)$ for any given k, thus $\phi_v(\varepsilon, k) = a_3(\log(T_v^{-\theta/2(\theta+1)}))^{1/2}/T_v^{1/2}$ with $T_v = T_v(\varepsilon, C, k)$. Hence $\sup_{k\geq 2}\phi_v(\varepsilon,k) \leq O\big((\log(\varepsilon^{-\theta/(\theta+1)}))^{1/2}/\varepsilon\big)$ in (15). Solving (15), we obtain $\varepsilon_{n_l} = (\frac{\log n_l}{n_l})^{1/2}$ when $C_1 \sim J(f^*)\delta_{n_l}^{-2} n_l^{-1} \sim \log n_l$ and $\varepsilon_{n_u} = (\frac{\log n_u}{n_u})^{1/2}$ when $C_2 \sim J_0\delta_{n_u}^{-2} n_u^{-1} \sim \log n_u$. Assumption C is fulfilled because $E(X^2) < \infty$. In conclusion, we obtain, by Corollary 4, that $\inf_C |e(\hat{f}_C, f^*)| = O_p((n_u^{-1}\log n_u)^{(\theta+1)/2})$. Surprisingly, this rate is arbitrarily fast as $\theta \to \infty$.

Kernel learning: Similarly, we restrict our attention to $\mathcal{F}_1 = \{x \in \mathcal{R} : f(x) = w_f^T\tilde{\phi}(x) = \sum_{k=0}^{\infty} w_{f,k}\tilde{\phi}_k(x) : w_f \in \mathcal{R}^\infty\}$, where $\langle\tilde{\phi}(x), \tilde{\phi}(z)\rangle = \exp(-\frac{(x-z)^2}{2\sigma^2})$.

For (13), note that $\mathcal{F}$ is rich for sufficiently large n_l in that for any continuous function f, there exists a $\tilde{f} \in \mathcal{F}$ such that $\|f - \tilde{f}\|_\infty \leq \varepsilon_{n_l}^2$, compare with Steinwart (2001). Then $f^* = \arg\min_{f\in\mathcal{F}} EL(Yf)$ implies $\|f^* - \operatorname{Sign}(E(Y|X))\|_\infty \leq \varepsilon_{n_l}^2$ and $|EL(Yf^*) - GE(f^*)| \leq 2\varepsilon_{n_l}^2$. Consequently, $|e(f, f^*)| \leq E(V_f(X)) + 2\varepsilon_{n_l}^2$ and $\alpha_f = 1$. On the other hand, $E(V_g(X)) \geq -E|g - f_C^*| - E|g_C^* - f_C^*| + \frac{C_3}{2n_uC_2}\|g - f_C^*\|^2 - \frac{C_3}{2}\|g_C^* - f_C^*\|^2$. Using the fact that (f_C^*, g_C^*) is the minimizer of $ES(f,g;C)$, we have $\frac{C_3}{2}\|g_C^* - f_C^*\|^2 \leq ES(f_C^*, g_C^*) \leq ES(1,1) \leq 2C_1$. By Sobolev's inequality (Adams, 1975), $E|g_C^* - f_C^*| \leq \lambda_3\|g_C^* - f_C^*\| \leq \lambda_3(4C_1/C_3)^{1/2}$ and $E|g - f_C^*| \leq \lambda_3\|g - f_C^*\|$, for some constant $\lambda_3 > 0$. Plugging these into the previous inequality, we have $e_{\tilde{U}}(g, g_C^*) \geq \frac{C_3}{2n_uC_2}\|g - f_C^*\|^2 - \lambda_3\|g - f_C^*\| - \frac{2C_1}{n_uC_2} - \lambda_3(4C_1/C_3)^{1/2}$. By choosing suitable C, we obtain $\frac{1}{2}\|g - f_C^*\|^2 - e_{\tilde{U}}(g, g_C^*)^{1/2}\|g - f_C^*\| - e_{\tilde{U}}(g, g_C^*) \leq 0$. Solving this inequality yields $\|g - f_C^*\| \leq (1 + \sqrt{5})e_{\tilde{U}}(g, g_C^*)^{1/2}$. Furthermore, by Lemma 8 and Sobolev's inequality, for sufficient small $\lambda_4 > 0$, $e(g, g_C^*) \leq E2\lambda_4^{-1}|f_C^*(X) - g(X)| + 2P(|f_C^*(X)| \leq \lambda_4) + e(f_C^*, g_C^*) \leq 2\lambda_4^{-1}(1 + \sqrt{5})E(V_g(X))^{1/2} + 2P(|f_C^*(X)| \leq \lambda_4) + e(f_C^*, g_C^*)$. However, by Lemma 1, $e(f_C^*, g_C^*) \to 0$, and $P(|f_C^*(X)| \leq \lambda_4) \leq P(|f^*(X)| - |f^*(X) - f_C^*(X)| \leq \lambda_4) = P(|f^*(X)| \leq |f^*(X) - f_C^*(X)| + \lambda_4) \to 0$, as $C_1, C_2, C_3 \to \infty$, because of linearity of f^*. This yields (13) with $\alpha_g = 1/2$. For (14), $\operatorname{Var}(L(Yf(X)) - L(Yf^*(X))) \leq 2E(L(Yf(X)) - L(Yf^*(X)))^2 =$

$(w_f - w_{f^*})^T \Gamma_2 (w_f - w_{f^*})$ where Γ_2 is a positive definite matrix, and similar to Example 6.2.1, $E(V_f(X)) = (w_f - w_{f^*})\Gamma_{\tilde{h}}(w_f - w_{f^*})^T$ since f^* minimizes $E(L(Yf(X)))$. Therefore, there exists a constant $\lambda_5 > 0$ such that $\mathrm{Var}(L(Yf(X)) - L(Yf^*(X))) \le \lambda_5 E(V_f(X))$. Also, $\mathrm{Var}(U(g(X)) - U(g_C^*(X))) \le \|g - g_C^*\|_2^2 \le 2(\|g - f_C^*\|^2 + \|g_C^* - f_C^*\|^2) \le 2((1+\sqrt{3})^2 e_{\tilde{U}}(g, g_C^*) + \frac{4C_1}{C_3}) \le (8 + 2(1+\sqrt{3})^2) E(V_g(X)))$, implying (14) with $\beta_f = \beta_g = 1$. For Assumption B, by Lemma 7, $H(\varepsilon, \mathcal{F}_v(k)) \le O((\log(kJ_v/\varepsilon))^3)$ for any given k. Similarly, we have $\varepsilon_{n_l} = (n_l^{-1}(\log n_l J_l)^3)^{1/2}$ when $C_1 \sim J(f^*)\delta_{n_l}^{-2} n_l^{-1} \sim (\log n_l J_l)^{-3}$ and $\varepsilon_{n_u} = (n_u^{-1}(\log n_u J_u)^3)^{1/2}$ when $C_2 \sim J_0 \delta_{n_u}^{-2} n_u^{-1} \sim (\log n_u J_u)^{-3}$. Assumption C is fulfilled with the Gaussian kernel.

Appendix C. The Dual Form of (5)

Let $\nabla^{\psi(k)} = (\nabla_1^{\psi(k)}, \nabla_2^{\psi(k)})^T$, $\nabla_1^{\psi(k)} = C_1(\nabla\psi_2(y_1 f^{(k)}(x_1))y_1, \cdots, \nabla\psi_2(y_{n_l} f^{(k)}(x_{n_l}))y_{n_l})$ and $\nabla_2^{\psi(k)} = 2C_2(\nabla U_2(g^{(k)}(x_{n_l+1})), \cdots, \nabla U_2(g^{(k)}(x_n)))$. Further, let $\alpha = (\alpha_1, \cdots, \alpha_{n_l})^T$, $\beta = (\beta_{n_l+1}, \cdots, \beta_n)^T$, $\gamma = (\gamma_{n_l+1}, \cdots, \gamma_n)^T$, $\mathbf{y}_\alpha = (y_1\alpha_1, \cdots, y_{n_l}\alpha_{n_l})^T$, and

Theorem 9 *(ψ-learning) The dual problem of (4) with respect to (α, β, γ) is*

$$\max_{\alpha,\beta,\gamma} \left\{ -\begin{pmatrix} \mathbf{y}_\alpha \\ \beta - \gamma \end{pmatrix}^T \begin{pmatrix} (1+\frac{1}{C_3})\mathbf{K}_{ll} + \frac{1}{C_3}\mathbf{I}_l & \mathbf{K}_{lu} \\ \mathbf{K}_{ul} & \mathbf{K}_{uu} \end{pmatrix} \begin{pmatrix} \mathbf{y}_\alpha \\ \beta - \gamma \end{pmatrix} + (\alpha - (\beta+\gamma))^T \mathbf{1}_n - (\mathbf{y}_\alpha - (\beta - \gamma))^T \left(\mathbf{K}\nabla^{\psi(k)} + \begin{pmatrix} \nabla_1^{\psi(k)} \\ \mathbf{0}_{n_u} \end{pmatrix} \right) \right\}, \tag{17}$$

subject to $\left(2\begin{pmatrix} \mathbf{y}_\alpha \\ \gamma - \beta \end{pmatrix} + \nabla^{\psi(k)}\right)^T \mathbf{1}_n = 0$, $\mathbf{0}_n \le \alpha \le C_1\mathbf{1}_n$, $\mathbf{0}_n \le \beta$, $\mathbf{0}_n \le \gamma$, *and* $\mathbf{0}_n \le \beta + \gamma \le C_2\mathbf{1}_n$.

Proof of Theorem 9: For simplicity, we only prove the linear case as the nonlinear case is essentially the same. The kth primal problem in (4), after introducing slack variable ξ, is equivalent to $\min_{(w_f, w_g, \xi_i, \xi_j)} C_1 \sum_{i=1}^{n_l} \xi_i + C_2 \sum_{j=n_l+1}^{n} \xi_j + \frac{C_3}{2}\|w_f - w_g\|^2 + \frac{1}{2}\|\tilde{w}_g\|^2 - \langle w, \nabla s_2^{\psi}(f^{(k)}, g^{(k)})\rangle$ subject to constraints $2(1 - y_i(\langle w_f, x_i\rangle)) \le \xi_i$, $x_i \ge 0$; $i = 1, \cdots, n_l$, and $2(|\langle w_g, x_j\rangle| - 1) \le \xi_j$, $\xi_j \ge 0$; $j = n_l + 1, \cdots, n$.

To solve this minimization problem, the Lagrangian multipliers are employed to yield

$$\begin{aligned} & L(w_f, w_g, \xi_i, \xi_j) \\ = \; & C_1 \sum_{i=1}^{n_l} \xi_i + C_2 \sum_{j=n_l+1}^{n} \xi_j + \frac{C_3}{2}\|w_f - w_g\|^2 + \frac{1}{2}\|\tilde{w}_g\|^2 - \langle w, \nabla s_2^{\psi}(w_f^{(k)}, w_g^{(k)})\rangle + \\ & 2\sum_{i=1}^{n_l} \alpha_i (1 - y_i(\langle w_f, x_i\rangle) - \frac{\xi_i}{2}) + 2\sum_{j=n_l+1}^{n} \beta_j(\langle w_g, x_j\rangle - 1 - \frac{\xi_j}{2}) - \\ & 2\sum_{j=n_l+1}^{n} \gamma_j(\langle w_g, x_j\rangle + 1 + \frac{\xi_j}{2}) - \sum_{i=1}^{n_l} \gamma_i \xi_i - \sum_{j=n_l+1}^{n} \eta_j \xi_j, \end{aligned} \tag{18}$$

where $\alpha_i \ge 0$; $i = 1, \cdots, n_l$, $\beta_j \ge 0$, $\gamma_j \ge 0$, $j = n_l + 1, \cdots, n$. Differentiate L with respect to (w_f, w_g, ξ_i, ξ_j) and let the partial derivatives be zero, we obtain that $\frac{\partial L}{\partial \tilde{w}_f} = C_3(\tilde{w}_f - \tilde{w}_g) - 2\sum_{i=1}^{n_l} \alpha_i y_i x_i - \nabla_{1f}^{\psi(k)} = 0$, $\frac{\partial L}{\partial \tilde{w}_g} = \tilde{w}_g - C_3(\tilde{w}_f - \tilde{w}_g) - 2\sum_{j=n_l+1}^{n}(\gamma_j - \beta_j)x_j - \nabla_{1g}^{\psi(k)} = 0$, $\frac{\partial L}{\partial w_{f,0}} = C_3(w_{f,0} - w_{g,0}) -$

$2\sum_{i=1}^{n_l}\alpha_i y_i = 0$, $\frac{\partial L}{\partial w_{g,0}} = -C_3(w_{f,0}-w_{g,0}) - 2\sum_{j=n_l+1}^{n}(\gamma_j-\beta_j) - \nabla_{2g}^{\psi(k)} = 0$, $\frac{\partial L}{\partial \xi_i} = C_1-\alpha_i-\gamma_i = 0$, and $\frac{\partial L}{\partial \xi_j} = C_2-\beta_j-\gamma_j-\eta_j = 0$. Solving these equations yields that $\tilde{w}_f^* = 2(1+C_3^{-1})\sum_{i=1}^{n_l}\alpha_i y_i x_i + \sum_{j=n_l+1}^{n}(\gamma_j-\beta_j)x_j + (1+C_3^{-1})\nabla_{1f}^{\psi(k)} + \nabla_{1g}^{\psi(k)}$, $\tilde{w}_g^* = 2\sum_{i=1}^{n_l}\alpha_i y_i x_i + \sum_{j=n_l+1}^{n}(\gamma_j-\beta_j)x_j + \nabla_{1f}^{\psi(k)} + \nabla_{1g}^{\psi(k)}$, $2\sum_{i=1}^{n_l}\alpha_i y_i + 2\sum_{j=n_l+1}^{n}(\gamma_j-\beta_j) + \nabla_{2f}^{\psi(k)} + \nabla_{2g}^{\psi(k)} = 0$, $\alpha_i+\gamma_i = C_1$; $i = 1,\cdots,n_l$, and $\beta_j+\gamma_j+\eta_j = 0$; $j = n_l+1,\cdots,n$. Substituting $\tilde{w}_f^*$, $\tilde{w}_g^*$ and these identities into (18), we obtain (17) after ignoring all constant terms. To derive the corresponding constraints, note that $C_1-\alpha_i-\gamma_i = 0$, $\gamma_i \geq 0$ and $\alpha_i \geq 0$ implies $0 \leq \alpha_i \leq C_1$, $\eta_j \geq 0$ and $C_2-\beta_j-\gamma_j-\eta_j = 0$ implies $\beta_j+\gamma_j \leq C_2$. Furthermore, KKT's condition requires that $\alpha_i(1-y_i(\langle w_f, x_i\rangle)-\xi_i) = 0$, $\beta_j(\langle w_g, x_j\rangle - 1 - \xi_j)$, $\gamma_j(\langle w_g, x_j\rangle + 1 + \xi_j) = 0$, $\gamma_i\xi_i = 0$, and $\eta_j\xi_j = 0$. That is, $\xi_i \neq 0$ implies $\gamma_i = 0$ and $\alpha_i = C_1$, and $\xi_j \neq 0$ implies $\eta_j = 0$ and $\beta_j+\gamma_j = C_2$. Therefore, if $0 < \alpha_i < C_1$, then $\xi_i = 0$ and $1-y_i(\langle w_f, x_i\rangle) = 0$, if $0 < \beta_j+\gamma_j < C_2$, then $\xi_j = 0$ and $\langle w_g, x_j\rangle + 1 = 0$ or $\langle w_g, x_j\rangle - 1 = 0$.

Write the solution of (17) as $(\alpha^{(k+1)}, \beta^{(k+1)}, \gamma^{(k+1)})$, which yields the solution of (4): $\tilde{w}_f^{(k+1)} = 2\mathbf{X}^T\begin{pmatrix}(1+\frac{1}{C_3})\mathbf{y}_\alpha \\ \beta-\gamma\end{pmatrix} + \nabla^{\psi(k)}\begin{pmatrix}(1+\frac{1}{C_3})\mathbf{1}_{n_l} \\ \mathbf{1}_{n_u}\end{pmatrix}$, and $\tilde{w}_g^{(k+1)} = 2\mathbf{X}^T\begin{pmatrix}\mathbf{y}_\alpha \\ \beta-\gamma\end{pmatrix} + \nabla^{\psi(k)}\mathbf{1}_n$, and $(w_{f,0}^{(k+1)}, w_{g,0}^{(k+1)})$ satisfies KKT's condition in that $y_{i_0}(K(\tilde{w}_f^{(k+1)}, x_{i_0}) + w_{f,0}^{(k+1)}) = 1$ for any i_0 with $0 < \alpha_{i_0} < C_1$, and for any j_0 with $0 < \beta_{j_0}+\gamma_{j_0} < C_2$, $K(\tilde{w}_g^{(k+1)}, x_{j_0}) + w_{g,0}^{(k+1)} = 1$ if $\beta_{j_0} > 0$ or $K(\tilde{w}_g^{(k+1)}, x_{j_0}) + w_{g,0}^{(k+1)} = -1$ if $\gamma_{j_0} > 0$. Here $K(\tilde{w}_f^{(k+1)}, x_{i_0}) = (1+\frac{1}{C_3})\sum_{i=1}^{n_l}(2\alpha_i^{(k+1)}y_i + C_1\nabla\psi_2(f^{(k)}(x_i)))K(x_i, x_{i_0}) + 2\sum_{j=n_l+1}^{n}(\gamma_j^{(k+1)} - \beta_j^{(k+1)})K(x_j, x_{i_0}) + 2C_2\sum_{j=n_l+1}^{n}\nabla U_2(g^{(k)}(x_j))K(x_j, x_{i_0})$, and $K(\tilde{w}_g^{(k+1)}, x_{j_0}) = \sum_{i=1}^{n_l}2\alpha_i^{(k+1)}y_i K(x_i, x_{j_0}) + \sum_{i=1}^{n_l}C_1\nabla\psi_2(f^{(k)}(x_i))K(x_i, x_{j_0}) + 2\sum_{j=n_l+1}^{n}(\gamma_j^{(k+1)} - \beta_j^{(k+1)} + C_2\nabla U_2(g^{(k)}(x_j)))K(x_j, x_{j_0})$. When KKT's condition is not applicable to determine $(w_{f,0}^{(k+1)}, w_{g,0}^{(k+1)})$, that is, there does not exist an i such that $0 < \alpha_i < C_1$ or an j such that $0 < \beta_j+\gamma_j < C_2$, we may compute $(w_{f,0}^{(k+1)}, w_{g,0}^{(k+1)})$ through quadratic programming by substituting $(\tilde{w}_f^{(k)}, \tilde{w}_g^{(k)})$ into (4).

Theorem 10 *(SVM) The dual problem of (4) for SVM with respect to* (α, β, γ) *is the same as (17) with* $(\alpha, \beta, \gamma, \mathbf{y}_\alpha)$ *replaced by* $\frac{1}{2}(\alpha, \beta, \gamma, \mathbf{y}_\alpha)$, *and* $\nabla^{\psi(k)}$ *replaced by* $\nabla^{S(k)} = (0,\cdots,0, C_2\nabla U_2(g^{(k)}(x_{n_l+1})),\cdots, C_2\nabla U_2(g^{(k)}(x_n)))^T$. *Here KKT's condition remains the same.*

Proof of Theorem 10: The proof is similar to that of Theorem 9, and thus is omitted.

References

R. A. Adams. *Sobolev Spaces*. Academic Press, New York, 1975.

M. Amini, and P. Gallinari. Semi-supervised learning with an explicit label-error model for misclassified data. In *IJCAI* 2003.

L. An and P. Tao. Solving a class of linearly constrained indefinite quadratic problems by D.C. algorithms. *J. of Global Optimization*, 11:253-285, 1997.

R. Ando and T. Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *Technical Report RC23462*, IBM T.J. Watson Research Center, 2004.

M. Balcan, A. Blum, P. Choi, J. Lafferty, B. Pantano, M. Rwebangira and X. Zhu. Person identification in webcam images: an application of semi-supervised learning. In *ICML* 2005.

P. L. Bartlett, M. I. Jordan and J. D. McAuliffe. Convexity, classification, and risk bounds. *J. Amer. Statist. Assoc.*, 19:138-156, 2006.

M. Belkin, P. Niyogi and V. Sindhwani. Manifold Regularization : A Geometric Framework for Learning From Examples. Technical Report, Univ. of Chicago, Department of Computer Science, TR-2004-06, 2004.

C. L. Blake and C. J. Merz. UCI repository of machine learning databases [http://www.ics.ci.edu/∼mlearn/MLRepository.html]. University of California, Irvine, Department of Information and Computer Science, 1998.

A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, 1998.

M. Collins and Y. Singer. Unsupervised models for named entity classification. In *Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 100-110, 1999.

C. Cortes and V. Vapnik. Support vector networks. *Machine Learning*, 20:273-297, 1995.

F. G. Cozman, I. Cohen and M. C. Cirelo. Semi-supervised learning of mixture models and Bayesian networks. In *ICML* 2003.

B. Efron. The estimation of prediction error: Covariance penalties and cross-validation. *J. Amer. Statist. Assoc.*, 99:619-632, 2004.

C. Gu. Multidimension smoothing with splines. In M.G. Schimek, editor, *Smoothing and Regression: Approaches, Computation and Application*, 2000.

T. Hastie, S. Rosset, R. Tibshirani and J. Zhu. The entire regularization path for the support vector machine. *J. of Machine Learning Research*, 5: 1391-1415, 2004.

T. Joachims. Transductive inference for text classification using support vector machines. In *ICML* 1999.

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

Y. Lin and L. D. Brown. Statistical properties of the method of regularization with periodic Gaussian reproducing kernel . *Ann. Statist.*, 32:1723-1743, 2004.

S. Liu, X. Shen and W. Wong. Computational development of ψ-learning. In *SIAM 2005 International Data Mining Conference*, pages 1-12, 2005.

Y. Liu and X. Shen. Multicategory ψ-learning. *J. Amer. Statist. Assoc.*, 101:500-509, 2006.

P. Mason, L. Baxter, J. Bartlett and M. Frean. Boosting algorithms as gradient descent. In *Advances in Neural Information Processing Systems 12*, pages 512-518. The MIT Press, 2000.

K. Nigam, A. McCallum, S. Thrun and T. Mitchell . Text classification from labeled and unlabeled documents using EM. In *AAAI* 1998.

B. Schölkopf, A. Smola, R. Williamson and P. Bartlett. New support vector algorithms. *Neural Computation*, 12:1207-1245, 2000.

X. Shen and W. Wong. Convergence rate of sieve estimates. *Ann. Statist.*, 22:580-615, 1994.

X. Shen. On the method of penalization. *Statist. Sinica*, 8:337-357, 1998.

X. Shen and H. C. Huang. Optimal model assessment, selection and combination. *J. Amer. Statist. Assoc.*, 101:554-568, 2006.

X. Shen, G. C. Tseng, X. Zhang and W. Wong. On psi-learning. *J. Amer. Statist. Assoc.*, 98:724-734, 2003.

X. Shen and L. Wang. Discussion of 2004 IMS Medallion Lecture: "Local Rademacher complexities and oracle inequalities in risk minimization". *Ann. Statist.*, in press.

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

M. Szummer and T. Jaakkola. Information regularization with partially labeled data. In *NIPS* 2003.

S. Van De Geer. Hellinger-consistency of certain nonparametric maximum likelihood estimators. *Ann. Statist.*, 21:14-44, 1993.

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

G. Wahba. Spline models for observational data. *Series in Applied Mathematics*, Vol. 59, SIAM, Philadelphia, 1990.

J. Wang and X. Shen. Estimation of generalization error: random and fixed inputs. *Statist. Sinica*, 16:569-588, 2006.

J. Wang, X. Shen and W. Pan. On transductive support vector machines. In *Proc. of the Snowbird Machine Learning Conference*, in press.

T. Zhang and F. Oles. A probability analysis on the value of unlabeled data for classification problems. In *ICML* 2000.

D. Zhou. The covering number in learning theory. *J. of Complexity*, 18:739-767, 2002.

J. Zhu and T. Hastie. Kernel logistic regression and the import vector machine. *J. Comp. Graph. Statist.*, 14:185-205, 2005.

X. Zhu, Z. Ghahramani and J. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *ICML* 2003.

X. Zhu and J. Lafferty. Harmonic mixtures: combining mixture models and graph-based methods for inductive and scalable semi-supervised learning. In *ICML* 2005.