On Transductive Support Vector Machines*

Junhui Wang
School of Statistics
University of Minnesota
Minneapolis, MN 55455

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

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

Abstract

Transductive support vector machines (TSVM) has been widely used as a means of treating partially labeled data in semi-supervised learning. Around it, there has been mystery because of lack of understanding its foundation in generalization. This article aims to clarify several controversial aspects regarding TSVM. Two main results are established. First, TSVM performs no worse than its supervised counterpart SVM when tuning is performed, which is contrary to several studies indicating otherwise. The "alleged" inferior performance of TSVM is mainly because it was not tuned in the process, in addition to the involved minimization routines. Second, we utilize difference convex programming to derive a nonconvex minimization routine for TSVM, which compares favorably against some state-of-the-art methods. This, together with our learning theory lands some support to TSVM.

1 Introduction

In many real-world applications, labeling is often costly, while an enormous amount of unlabeled data is available with little cost. Examples of this type include, but are not limited to, webpage classification, medical diagnosis, spam email detection, text categorization, image processing, c.f., Baluja (1998); Blum and Mitchell (1998); Amini and Gallinari (2003); Balcan, et. al. (2005). In situation as such, how to enhance classification by utilizing additional unlabeled data becomes critical, which is referred to as the problem of semi-supervised learning in what follows.

In the semi-supervised learning literature, methods have been proposed from different perspectives, including margin-based classification (Vapnik, 1998; Wang and Shen, 2006), the EM method (Nigam, McCallum, Thrun and Mitchell, 1998), graph-based method (Blum and Chawla, 2001; Zhu, Ghahramani and Lafferty, 2003), and information regularization (Szummer and Jaakkola, 2002). The central topic this article concerns is the generalization performance of transductive support vector machine (TSVM; Vapnik, 1998), which remains mysterious, particularly its "alleged" unstable performance in empirical studies.

TSVM seeks the largest separation between labeled and unlabeled data through regularization. In empirical studies, it performs well in text classification (Joachims, 1999) but can perform substantially worse than its supervised counterpart SVM (Cortes and Vapnik, 1995) in other applications (Wu, Bennett, Cristianini and Shawe-Taylor, 1999). This unstable performance has been criticized. Zhang and Oles (2000) argued that there is lack of evidence that the notion of separation leads to correct classification. Chapelle and Zien (2005) suggested that the cost function of TSVM is appropriate but implementation of TSVM is inadequate. Astorino and Fuduli (2005) also noted that implementation of TSVM is an issue.

In this article, we address the aforementioned issues. We argue that in principle TSVM performs no worse than its supervised counterpart SVM after tuning. Key to it is tuning, which has been commonly ignored in the literature. Tuning guards against potential unstable performance by tuning regularizers towards labeled data. Furthermore, we develop a statistical learning theory to demonstrate this aspect with regard to TSVM's generalization ability. To treat the implementation issue, we develop a nonconvex minimization routine based on recent advances in global optimization, particularly difference convex (DC) programming. Numerical analysis indicates that the proposed routine delivers a better solution than that of

* 2000 Math Subject Classification Numbers: 68T10

Joachims (1999), and confirms that TSVM performs no worse than SVM.

At the time this article is nearly completed, we noted that Collobert, Sinz, Weston and Bottou (2006) developed a similar implementation of TSVM using a different DC decomposition of the hat function. Nevertheless, some overlapping is inevitable between their implementation and ours.

The rest of the paper is organized as follows. Section 2 introduces TSVM. Section 3 solves TSVM with a DC algorithm. Section 4 presents some numerical examples, followed by a novel statistical learning theory in Section 5. Section 6 contains summary and discussion. Technical details are deferred to the appendix.

2 TSVM

In semi-supervised learning, a sample $(X^l, Y^l) = \{(X_i, Y_i)\}_{i=1}^{n_l}$ is observed with an independent unlabeled sample $X^u = \{X_j\}_{j=n_l+1}^{n}$ and $n = n_l + n_u$. Here $X_i = (X_{i1}, \cdots, X_{ip})$ is an p-dimensional input and $Y_i \in \{-1, 1\}$, independently and identically distributed according to an unknown distribution $P(x, y)$, and X^u is distributed according to distribution $P(x)$.

TSVM uses an idea of maximizing separation between labeled and unlabeled data, c.f., Vapnik (1998). It solves

$$\min_{y_j, f\in\mathcal{F}} C_1 \sum_{i=1}^{n_l} L(y_i f(x_i)) + C_2 \sum_{j=n_l+1}^{n} L(y_j f(x_j)) + J(f), \tag{1}$$

where f is a decision function in $\mathcal{F}$, a candidate function class, $L(z) = (1-z)_+$ is the hinge loss, and $J(f)$ is the inverse of the geometric separation margin. In the linear case, $f(x) = w^T x + b$ and $J(f) = \frac{1}{2}\|w\|^2$. In the nonlinear kernel case, $f(x) = (K(x, x_1), \cdots, K(x, x_n))w^T + b$, $J(f) = \frac{1}{2} w^T \mathbf{K} w$, where K is a kernel satisfying Mercer's condition to assure $w^T \mathbf{K} w$ with $\mathbf{K} = (K(x_i, x_j))_{i,j=1}^n$ being a proper norm; see Wahba (1990) and Gu (2000) for more details.

Minimization of (**??**) with respect to $f \in \mathcal{F}$ is nonconvex, which can be solved through integer programming, and is known to be NP (Bennett and Demiriz, 1998). To solve (**??**), Joachims (1999) proposed an efficient local search algorithm that is the basis of SVM^{Light}. This algorithm may fail to deliver a good local solution, resulting in worse performance of TSVM against SVM. This aspect is confirmed by our numerical results in Section 4.1 as well as empirical studies in the literature. Chapelle and Zien (2005) aimed to correct this problem by approximating (**??**) by a smooth convex problem through gradient descent. Astorino and Fuduli (2005) used an extended bundle method to treat nonconvexity and nonsmoothness of the cost function. In what follows, we shall develop our nonconvex minimization routine effectively utilizing the DC property of the cost function.

3 Difference convex programming

Key to DC programming is a decomposition of a cost function into a difference of two convex functions, based on which a sequence of upper approximations of the cost function yields 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), and has been used in the implementation of ψ-learning (Shen et. al, 2003; Liu, Shen and Wong, 2005) and large margin semi-supervised learning (Wang and Shen, 2006) for large problems.

In (**??**), direct calculation gives an equivalent cost function $s(f)$ as:

$$C_1 \sum_{i=1}^{n_l} L(y_i f(x_i)) + C_2 \sum_{j=n_l+1}^{n} L(|f(x_j)|) + J(f). \tag{2}$$

Minimization of (**??**) yields an estimated decision function $\hat{f}$ thus classifier Sign$(\hat{f})$.

To utilize DCA, we construct a DC decomposition of $s(f)$: $s(f) = s_1(f) - s_2(f)$; where $s_1(f) = C_1 \sum_{i=1}^{n_l} L(y_i f(x_i)) + C_2 \sum_{j=n_l+1}^{n} U_1(f(x_j)) + \frac{1}{2}\|f\|^2$ and $s_2(f) = C_2 \sum_{j=n_l+1}^{n} U_2(f(x_j))$ for TSVM with $U_1(z) = (|z|-1)_+$ and $U_2(z) = |z| - 1$. This DC decomposition is obtained through a DC decomposition of the hat function $L(|z|) = U(z) = U_1(z) - U_2(z)$, as displayed in Figure 1.

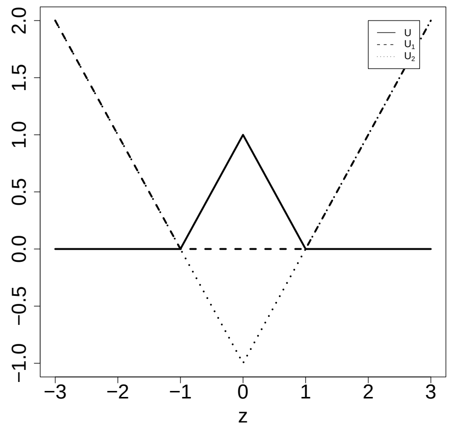
Given the decomposition, DCA solves a sequence of subproblems $\min_f \; s_1(f) - s_2(f^{(k)}) - \langle w - w_{f^{(k)}}, \nabla s_2(f^{(k)})\rangle$ with $\nabla s_2(f^{(k)})$ is a gradient vector of $s_2(f)$ at $f^{(k)}$, or equivalently,

$$\min_f \; (s_1(f) - \langle w, \nabla s_2(f^{(k)})\rangle), \tag{3}$$

after omitting the constant terms that are independent of f. Here s_2 is approximated by its tangent hyperplane at $f^{(k)}$. By convexity, (**??**) is an upper approximation to $s(f)$. Algorithm 1 below solves (**??**) for TSVM based on sequential quadratic programming.

Algorithm 1: (TSVM^{DCA})
Step 1. (Initialization) Set initial value $f^{(0)}$ as the solution of SVM with labeled data alone, and an precision tolerance level $\epsilon > 0$.
Step 2. (Iteration) At iteration $k+1$, solve (**??**) yielding solution $f^{(k+1)}$. The dual problem of (**??**) can be solved yielding the solution of (**??**), as described in

Figure 1: 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.



Appendix B.
Step 3. (Stopping rule) Terminate when $|s(f^{(k+1)}) - s(f^{(k)})| \leq \epsilon$. Then the estimate $\hat{f}$ is the best solution among $f^{(k)}$; $k = 0, 1, \cdots$.

A good initial value nevertheless enhances the chance of Algorithm 1 to locate the global minima. Our numerical experience suggests that SVM is an acceptable choice.

For the convergence property and complexity of **Algorithm 1**, we refer to Theorem 3 of Liu, Shen and Wong (2005) for more details.

4 Numerical examples

This section examines the performance of TSVM^{DCA} and compares it against some state-of-the-art methods–its counterpart TSVM^{Light} (Joachims, 1999) and SVM with labeled data alone in generalization, and ∇TSVM (Chapelle and Zien, 2005) and TSVM^{Bundle} (Astorino and Fuduli, 2005) in transduction.

4.1 Generalization performance

All numerical analyses are performed in R2.1.1, and TSVM^{Light} is trained through SVM^{Light} 6.01. 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.

Two simulated and five benchmark examples are examined for SVM, TSVM^{Light} and TSVM^{DCA}. In each example, an independent test error is used to evaluate a classifier's generalization performance, which approximates the generalization error. Each classifier is optimized with respect to its tuning parameter(s). In particular, a grid search is employed to minimize the test error over the domain $[10^{-3}, 10^3]$ of its tuning parameter(s).

Simulated examples: Simulated examples include Examples 1 and 2 of Wang and Shen (2006), where 800 and 200 instances are randomly selected for testing and training, among which 190 randomly chosen instances are removed their labels to generate unlabeled data whereas the remaining 10 treated as labeled data.

Benchmarks: Five benchmark examples are examined, including Wisconsin Breast Cancer (WBC), Pima Indians Diabetes (Pima), Ionosphere, Mushroom and Spam email, each available in the UCI Machine Learning Repository (Blake and Merz, 1998). Except for Spam email example, instances are randomly divided into halves with 10 labeled and 190 unlabeled instances for training, and the remaining for testing. For the Spam email, instances are randomly divided into halves with 20 labeled and 380 unlabeled instances for training and the remaining for testing.

The smallest averaged testing errors of SVM, TSVM^{Light}, and TSVM^{DCA} are summarized in Tables 1 and 2.

Table 1: **Linear learning:** Averaged test errors as well as the estimated standard errors (in parenthesis) of SVM with labeled data alone, TSVM^{Light}, and TSVM^{DCA}, over 100 pairs of training and testing samples, in the simulated and benchmark examples.

Data	SVM	TSVM^{Light}	TSVM^{DCA}
Example 1	.345(.0081)	.230(.0081)	.220(.0103)
Example 2	.333(.0129)	.222(.0128)	.203(.0088)
WBC	.053(.0071)	.077(.0113)	.037(.0024)
Pima	.328(.0092)	.316(.0121)	.314(.0086)
Ionosphere	.257(.0097)	.295(.0085)	.197(.0071)
Mushroom	.232(.0135)	.204(.0113)	.206(.0113)
Email	.216(.0097)	.227(.0120)	.196(.0132)

Tables 1 and 2 indicate that TSVM^{DCA} performs no worse than its SVM counterpart in all the cases, which agrees with the theoretical result of Corollary 1. Overall TSVM^{DCA} yields better solutions than TSVM^{Light} in all the cases except in the linear Mushroom case where it performs slightly worse. The superiority of TSVM^{DCA} may be due to the DC minimization strategy, where the DC property of the cost function has been effectively used.

Table 2: **Nonlinear learning with Gaussian kernel:** Averaged test errors as well as the estimated standard errors (in parenthesis) of SVM with labeled data alone, TSVM^{Light}, and TSVM^{DCA}, over 100 pairs of training and testing samples, in the simulated and benchmark examples.

Data	SVM	TSVM^{Light}	TSVM^{DCA}
Example 1	.385(.0099)	.267(.0132)	.232(.0122)
Example 2	.347(.0119)	.258(.0157)	.205(.0091)
WBC	.047(.0038)	.037(.0015)	.037(.0045)
Pima	.353(.0089)	.362(.0144)	.330(.0107)
Ionosphere	.232(.0088)	.214(.0097)	.183(.0103)
Mushroom	.217(.0135)	.217(.0117)	.185(.0080)
Email	.226(.0108)	.275(.0158)	.192(.0110)

4.2 Transductive performance

As discussed above, ∇TSVM and TSVM^{Bundle} are two state-of-the art proposals designed to repair the problem of TSVM^{Light}. This section compares TSVM^{DCA} with ∇TSVM and TSVM^{Bundle} in the two simulated examples g50c and g10n in Chapelle and Zien (2005), and Astorino and Fuduli (2005), in addition to two benchmark examples Heart and Ionosphere used in Astorino and Fuduli (2005). To make a fair comparison, we use the average transductive error (Vapnik, 1998) based on their unlabeled sets under the exactly same setting as theirs. Note that the datasets for g50c and g10n were given in Chapelle and Zien (2005) whereas those of Heart and Ionosphere were sampled at random according to Astorino and Fuduli (2005).

Table 3: Averaged transductive errors of TSVM^{Light}, ∇TSVM, TSVM^{Bundle} and TSVM^{DCA}, over 10 pairs of labeled and unlabeled sets, in two simulated examples and two real examples. The **bold** case indicates the best performer in each example.

Data	g50c	g10n	Heart	Ionosphere
TSVM^{Light}	.069	.144	.163	.157
∇TSVM	.058	.098	-	-
TSVM^{Bundle}	**.040**	.086	.120	.114
TSVM^{DCA}	.047	**.081**	**.110**	**.069**

Table 3 indicates that TSVM^{DCA} outperforms ∇TSVM in all the cases, while outperforming TSVM^{Bundle} in all cases except g50c. More importantly, DCA is efficient in that it usually converges in about 5 iterations in here. In summary, DCA compares favorably against ∇TSVM and TSVM^{Bundle}, in addition to its fast convergence speed.

5 Statistical learning theory

This section derives a probability bound for quantifying TSVM's generalization performance after tuning, as measured by $\inf_C |e(\hat{f}_C, f^*)|$, where $f^* = \text{arginf}_{f \in \mathcal{F}} EL(Yf(X))$ denotes the optimal Bayes rule in $\mathcal{F}$, and $e(\hat{f}_C, f^*) = GE(\hat{f}_C) - GE(f^*)$ measures TSVM $\hat{f}_C$'s generalization performance relative to f^*.

5.1 Statistical learning theory

Before proceeding, we introduce some notations. Define the surrogate loss $W(f)$ to be $\frac{n_l C_1}{n_u C_2} L(yf(x)) + U(f(x))$ when $C_2 > 0$, and $L(yf(x))$ when $C_2 = 0$. Define $e_W(f, f_C^*)$ to be $E(W(f(X)) - W(f_C^*(X))) = \frac{n_l C_1}{n_u C_2} e_L(f, f_C^*) + e_U(f, f_C^*) \geq 0$, the surrogate risk measuring the performance of f under L and U. Here $e_L(f, f_C^*) = EL(Yf(X)) - EL(Yf_C^*(X))$, $e_U(f, f_C^*) = EU(f(X)) - EU(f_C^*(X))$, $C = (C_1, C_2)$, and $f_C^* = \text{arginf}_{f \in \mathcal{F}} EW(f(X))$.

Now define $L^T(z)$, truncated version of $L(z)$, to be $L^T(z) = L(yf_C^*) + T$ if $L(z) - L(yf_C^*) \geq T$ and $L^T(z) = L(z)$ otherwise, for any $f \in \mathcal{F}$ and some truncation constant $T \geq 2$. Then $W^T(f) = \frac{n_l C_1}{n_u C_2} L^T(yf) + U(f)$ when $C_2 > 0$ and $L^T(yf)$ when $C_2 = 0$, and $e_{W^T}(f, f_C^*) = E(W^T(f(X)) - W^T(f_C^*(X)))$.

Assumption A. (Conversion formula) There exist constants $0 < \alpha(C) < \infty$ and $a_1(C) > 0$ depending on tuning parameter C such that for any small $\delta > 0$,

$$\sup_{\{e_{W^T}(f, f_C^*) \leq \delta\}} |e(f, f_C^*)| \leq a_1(C) \delta^{\alpha(C)}. \tag{4}$$

Assumption B. (Variance) There exist constants $0 < \beta(C) < 2$ and $a_2(C) > 0$ depending on C such that for any small $\delta > 0$,

$$\sup_{\{e_{W^T}(f, f_C^*) \leq \delta\}} \text{Var}(W^T(f(X)) - W(f_C^*(X))) \leq a_2(C) \delta^{\beta(C)}. \tag{5}$$

Assumptions A and B describe the local behavior of $e(f, f_C^*)$ and $\text{Var}(W^T(f(X)) - W(f_C^*(X)))$ in a neighborhood of f_C^* defined by $e_{W^T}(f, f_C^*)$. In the parametric case, $\alpha(C) = 1$ in (**??**) and $\beta(C) = 1$ in (**??**). In general, $\alpha(C) = \beta(C) = 0$ are always true because $GE(f)$ and $|W^T(f)|$ are bounded. See Shen and Wang (2006) for a discussion of the relation of this type of conditions to the "low noise" assumption.

To quantify complexity of $\mathcal{F}$, we define the L_2-metric entropy with bracketing. For any $\epsilon > 0$, 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 $f_m^l \leq f \leq f_m^u$ and $\|f_m^l - f_m^u\|_2 \leq \epsilon; m = 1, \cdots, M$, where $\|\cdot\|_2$ is the usual L_2 norm. Then the L_2-metric entropy with bracketing $H(\epsilon, \mathcal{F})$ is defined as the logarithm of the cardinality of smallest ϵ-bracketing function set of $\mathcal{F}$.

Let $J_0 = \max(J(f^*_C), 1)$, $\mathcal{F}(k) = \{f \in \mathcal{F} : J(f) \leq kJ_0\}$, and $r(n_l, n_u, C_1, C_2) \geq \sup_{f\in\mathcal{F}} \left\{ \frac{1}{n_l} \frac{A(f)}{A(f)+B(f)} + \frac{1}{n_u} \frac{B(f)}{A(f)+B(f)} \right\}$ with $A(f) = (\frac{n_l C_1}{n_u C_2})^2 \operatorname{Var}(L^T(Yf(X)) - L(Yf^*_C(X)))$ and $B(f) = \operatorname{Var}(U(f(X)) - U(f^*_C(X)))$ when $C_2 > 0$.

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

$$\sup_{k\geq 2} \phi(\epsilon_n, k) \leq a_5 n^{1/2}, \tag{6}$$

where $\phi(\epsilon, k) = \int_{a_4 R}^{(a_3 n r)^{1/2} R^{\beta(C)/2}} H^{1/2}(\frac{v}{2}, \mathcal{F}(k)) dv/R$, $r = r(n_l, n_u, C_1, C_2)$ and $R = R(\epsilon, C, k) = \min(1, \epsilon^{2/\beta(C)} + (n_u C_2)^{-1}(k/2-1)J_0)$.

Assumption C′. (Complexity) For some constants $a_i > 0$; $i = 3, \cdots, 5$ and $\epsilon_{n_l} > 0$,

$$\sup_{k\geq 2} \phi(\epsilon_{n_l}, k) \leq a_5 n_l^{1/2}, \tag{7}$$

where $\phi(\epsilon, k) = \int_{a_4 R}^{a_3^{1/2} R^{\beta(C)/2}} H^{1/2}(\frac{v}{2}, \mathcal{F}(k)) dv/R$, and $R = R(\epsilon, C, k) = \min(1, \epsilon^{2/\beta(C)} + (n_l C_1)^{-1}(k/2 - 1)J_0)$.

The equation (**??**) yields ϵ_n for $\mathcal{F}$. Such an assumption has been used in Shen et al. (2003) in quantifying the rates of convergence of ψ-learning.

Theorem 1 *(TSVM) In addition to Assumptions A-C and C′, $n_l \leq n_u$. For $\hat{f}_C$, the minimizer of (**??**), there exist constants $a_k(C) > 0$; $k = 1, 6$ such that when $C^*_2 > 0$,*

$$P\big(\inf_C |e(\hat{f}_C, f^*)| \geq s_n\big) \leq$$
$$3.5 \exp\Big(- a_6(C^*)(r^*)^{-1}((n_u C^*_2)^{-1} J_0)^{\max(1, 2-\beta(C^*))}\Big);$$

when $C^*_2 = 0$,

$$P\big(\inf_C |e(\hat{f}_C, f^*)| \geq a_1(C^*)\delta_{n_l}^{2\alpha(C^*)}\big) \leq$$
$$3.5 \exp\Big(- a_6(C^*) n_l((n_l C^*_1)^{-1} J_0)^{\max(1, 2-\beta(C^*))}\Big),$$

where $s_n = 2\max(a_1(C^*)\delta_n^{2\alpha(C^*)}, \inf_{C\in\mathcal{C}} |e(f^*_C, f^*)|)$, $\delta_n = \min(\epsilon_n, 1)$, $r^* = r(n_l, n_u, C^*_1, C^*_2)$ and $C^* = (C^*_1, C^*_2) = \operatorname{arginf}_{C\in\mathcal{C}} |e(f^*_C, f^*)|$ with $\mathcal{C} = \{C : n_u C_2 \geq 2\delta_n^{-2} J_0, n_l C_1 \geq 2\delta_{n_l}^{-2} J_0\}$.

Corollary 1 *Under the assumptions of Theorem 2,*

$$\inf_C \ |e(\hat{f}_C, f^*)| = O_p\big(\min(s_n, \delta_{n_l}^{2\alpha(C^*)})\big),$$

provided that $a_6(C^*) n_l((n_l C^*_1)^{-1} J_0)^{\max(1, 2-\beta(C^*))}$ and $a_6(C^*)(r^*)^{-1}((n_u C^*_2)^{-1} J_0)^{\max(1, 2-\beta(C^*))}$ are bounded away from 0.

As suggested by Corollary 1, TSVM outperforms its supervised counterpart when $\{f^*_C : C \in \mathcal{C}\}$ provides an adequate approximation to the Bayes rule f^* in that $\inf_{C\in\mathcal{C}} |e(f^*_C, f^*)| = 0$ for some C^* with $C^*_2 > 0$. In this case, the rate of TSVM $O_p(\delta_n^{2\alpha(C^*)})$ is usually faster than $O_p(\delta_{n_l}^{2\alpha(C^*)})$ of its counterpart. On the other hand, TSVM never performs worse than its supervised counterpart SVM asymptotically in view the fact that $\inf_{C\in\mathcal{C}} |e(f^*_C, f^*)| = 0$ is always true for $C^*_2 = 0$. In this process, tuning is critical to achieve the aforementioned result.

Remark: Note that Theorem 1 and Corollary 1 continue to hold when the "global" entropy in (**??**) is replaced by a "local" entropy, c.f., Van De Geer (1993). Let $\mathcal{F}_v(k) = \{f \in \mathcal{F} : J(f) \leq kJ_0, |e(f, f^*)| \leq 2v\}$ be the "local" entropy of $\mathcal{F}(k)$. The proof requires only a sight 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.

5.2 Illustrative Example

Consider a two-dimensional linear example, where $X = (X_{(1)}, X_{(2)})$ is the input with $X_{(1)}$ and $X_{(2)}$ distributed independently according to probability densities $q_1(z) = \frac{1}{2}(\theta + 1)|z|^\theta$ for $z \in [-1, 1]$ and $q_2(z) = \frac{1}{2}$, and given X, $Y = 1$ if $X_{(1)} > 0$ and $Y = -1$ otherwise, and $\theta \geq 0$. To generate the nonseparable case, Y is randomly flipped with a constant probability τ with $0 < \tau < \frac{1}{2}$. Here the candidate decision function class $\mathcal{F} = \{f(x) = w^T x + b : \ w \in \mathcal{R}^2, \ b \in \mathcal{R}\}$, which contains $f_t(x) = x_{(1)}$ yielding the true classification boundary.

Case I: $\boldsymbol{0 < \theta < \infty}$. Note that $E(W(f(X))) = E(E(W(f(X))|X_{(2)})) \geq E(W(\tilde{f}^*_C(X)))$, where $\tilde{f}^*_C = \operatorname{argmin}_{f\in\mathcal{F}_1} E(W(f(X)))$ and $\mathcal{F}_1 = \{\tilde{f}(x) = wx + b : w, b \in \mathcal{R}\}$, because $(X_{(1)}, Y)$ is independent of $X_{(2)}$. Without loss of generality, we restrict our attention to $\mathcal{F}_1$.

It can be verified that $\inf_{C\in\mathcal{C}} e(f^*_C, f^*) = 0$ because $e(f^*_C, f^*) \to 0$ when $C_1 \to \infty$. Denote by $C^* = \operatorname{arginf}_C e(f^*_C, f^*)$. We verify Assumptions A-C and C′. For Assumption A, note that $f^*_{C^*}$ minimizes $E(W^T(f))$, direct calculation yields that $e_{W^T}(f, f^*_{C^*}) = (e_0, e_1)\Gamma(e_0, e_1)^T$ when $w_f = w_{f^*_{C^*}} + e_1$, $b_f = b_{f^*_{C^*}} + e_0$ and Γ is a positive definite matrix. Thus there exists a constant $\lambda_1 > 0$ such that $e_{W^T}(f, f^*_{C^*}) \geq \lambda_1(e_0^2 + e_1^2)$. Furthermore, $|e(f, f^*_{C^*})| \leq \frac{1}{2}(1 - 2\tau)\min(|w_{f^*_{C^*}}|, |w_{f^*_{C^*}} + e_1|)^{-(\theta+1)}|e_0|^{\theta+1} \leq \lambda_2(e_0^2 + e_1^2)^{(\theta+1)/2}$ for some constant $\lambda_2 > 0$. A combination of the two inequalities leads to (**??**) with $\alpha(C^*) = (\theta+1)/2$. For Assumption B, $\mathrm{Var}(W^T(f(X)) - W(f^*_{C^*}(X))) \leq (1 + (\frac{n_l C_1^*}{n_u C_2^*})^2) E(f(X) - f^*_{C^*}(X))^2 \leq \frac{C_1^{*2}+C_2^{*2}}{C_2^{*2}} \max(1, E(X^2_{(1)}))(e_0^2 + e_1^2)$, which implies (**??**) with $\beta(C^*) = 1$. For Assumption C, by Lemma 3 of Wang and Shen (2006), $H(v, \mathcal{F}_v(k)) \leq O(\log(v^{-\theta/(\theta+1)}))$ for any given k. Note that $\sup_{k\geq 2} \phi(\epsilon, k) \leq O(\log(((nr^*)^{1/2}\epsilon)^{-\theta/(\theta+1)}))^{1/2}/\epsilon$. Solving (**??**), we obtain $\epsilon_n = (\frac{nr^* \log n}{n})^{1/2} = (r^* \log n)^{1/2}$ when $n_u C_2^*/J_0 \sim \delta_n^{-2} \sim (r^* \log n)^{-1}$. For Assumption C′, solving (**??**) yielding $\epsilon_{n_l} = (\frac{\log n_l}{n_l})^{1/2}$ when $n_l C_1^*/J_0 \sim \delta_{n_l}^{-2} \sim n_l (\log n_l)^{-1}$.

When $C_1^* \geq \frac{1}{2n_l}$ and C_1^*/C_2^* is small enough such that $\frac{n_u}{n_l} A(f) \leq \frac{n_l C_1^{*2} T^2}{n_u C_2^{*2}} \leq B(f)$, $r(n_l, n_u, C_1^*, C_2^*) = \frac{2}{n_u}$ and $|e(f^*_{C^*}, f^*)| = 0$. By Corollary 1, $\inf_C \ |e(\hat{f}_C, f^*)| = O_p\big(n_u^{-(\theta+1)/2} (\log n)^{(\theta+1)/2}\big)$, which is arbitrary fast as $\theta \to \infty$.

Case II: $\boldsymbol{\theta = 0}$. Let $q_1(z) = \frac{1}{2}$ and X following the uniform distribution. The assumptions can be verified similarly as in **Case I**. Note that the approximation error $\inf_{C \in \mathcal{C}} |e(f^*_C, f^*)|) = 0$ implies that $\frac{n_l C_1 \tau}{n_u C_2} + \frac{1}{\theta+2} \leq E(W(f^*_C)) \leq E(W(\mathbf{1})) = \frac{n_l C_1}{2 n_u C_2}$ or $C_1/C_2 \geq \frac{n_u}{n_l(1-2\tau)}$, where $\mathbf{1}(x) = 1$ for all x. Using this inequality and the fact that $\mathrm{Var}(L^T(Yf(X)) - L(Yf^*_C(X)))$ and $\mathrm{Var}(U(f(X)) - U(f^*_C(X)))$ are two constants independent of (n_l, n_u), we have $r(n_l, n_u, C_1, C_2) = \frac{1}{n_l}$ and $\inf_C \ |e(\hat{f}_C, f^*)| = O_p(n_l^{-(\theta+1)/2} (\log n_l)^{(\theta+1)/2})$. This says that TSVM is of the same order of speed of SVM and unlabeled data contributes little to classification in this uniform situation.

In conclusion, TSVM yields better performance in some situations and the same performance in other situations than its supervised counterpart SVM. This depends entirely on if unlabeled data is informative with respect to classification. In this process, tuning is critical to assure the no-worse performance.

6 Summary

This article investigates computational and theoretical aspects of TSVM. With regard to implementation of TSVM, we solve the non-convex minimization using DC programming. Our numerical analysis suggests that our implementation compares favorably against the existing ones. Most importantly, TSVM equipped our implementation performs no worse than its supervised counterpart SVM, which is in contrast to the unstable performance of TSVM reported in the literature. With respect to learning theory, we develop a novel theory to quantify TSVM's generalization ability. In conclusion, the results in this article land some support to TSVM. When TSVM is tuned, its regularizers guard against potential unstable performance due to unlabeled data.

Acknowledgment

This research was supported in part by NSF grants IIS-0328802 and DMS-0604394. Pan's research was partially supported by NIH grant HL65462 and a UM AHC FRD grant. We would also like to thank J. Zhu for bringing our attention to the work of Collobert, Sinz, Weston and Bottou (2006).

Appendix A: Proof of Theorem 1

Without loss of generality, we just prove the case of $C_2 > 0$. The proof of $C_2 = 0$ is similar and thus omitted. Let $\tilde{W}(f) = W(f) + \lambda J(f) = \frac{n_l C_1}{n_u C_2} \tilde{L}(yf) + \tilde{U}(f)$ with $\lambda = \frac{1}{n_u C_2}$, $\tilde{L}(yf(x)) = L(yf(x)) + \frac{1}{2 n_l C_1} J(f)$ and $\tilde{U}(f(x)) = U(f(x)) + \frac{1}{2 n_u C_2} J(f)$. By the definition of $\hat{f}_C$,

$$\begin{aligned} &\Big\{ e_{W^T}(f, f^*_C) \geq \delta_n^2 \Big\} \\ \subset&\Big\{ \sup_{e_{W^T}(f, f^*_C) \geq \delta_n^2} \frac{n_l C_1}{n_u C_2} \frac{1}{n_l} \sum_{i=1}^{n_l} (\tilde{L}(y_i f^*_C(x_i)) - \tilde{L}(y_i f(x_i))) \\ &\qquad + \frac{1}{n_u} \sum_{j=n_l+1}^{n} (\tilde{U}(f^*_C(x_j)) - \tilde{U}(f(x_j))) \geq 0 \Big\} \\ \subset&\Big\{ \sup_{e_{W^T}(f, f^*_C) \geq \delta_n^2} \frac{n_l C_1}{n_u C_2} \frac{1}{n_l} \sum_{i=1}^{n_l} (\tilde{L}(y_i f^*_C(x_i)) - \tilde{L}^T(y_i f(x_i))) \\ &\qquad + \frac{1}{n_u} \sum_{j=n_l+1}^{n} (\tilde{U}(f^*_C(x_j)) - \tilde{U}(f(x_j))) \geq 0 \Big\}, \end{aligned}$$

where P^* denotes the outer probability measure.

Before proceeding, we introduce some notations to be used below. Define the scaled empirical process as $E_n(W(f^*_C) - W^T(f)) = \frac{1}{n}\Big(\sum_{i=1}^{n_l} \frac{n_l C_1}{n_u C_2} \frac{n}{n_l} \big(L(y_i f^*_C(x_i)) - L^T(y_i f(x_i)) - E(L(Y f^*_C(X)) - L^T(Y f(X))) \big) + \sum_{j=n_l+1}^{n} \frac{n}{n_u} \big(U(f^*_C(x_j)) - U(f(x_j)) - E(U(f^*_C(X)) - U(f(X))) \big) \Big)$. Then $P(e_{W^T}(\hat{f}_C, f^*_C)$ is upper bounded by

$$\begin{aligned} P^*\Big(&\sup_{e_{W^T}(f, f^*_C) \geq \delta_n^2} E_n(W(f^*_C) - W^T(f)) \geq \\ &\inf_{e_{W^T}(f, f^*_C) \geq \delta_n^2} E(\tilde{W}^T(f(X)) - \tilde{W}(f^*_C(X))) \Big) = \mathcal{I}. \end{aligned}$$

To bound $\mathcal{I}$, we apply a large deviation empirical technique for risk minimization. Such a technique has been previously developed in function estimation as in Shen and Wong (1994). Specifically, we bound $\mathcal{I}$ through a sequence of empirical processes over a partition and by controlling their means and variances.

Let $A_{s,t} = \{f \in \mathcal{F} : 2^{s-1}\delta_n^2 \leq e_{W^T}(f, f_C^*) \leq 2^s\delta_n^2, 2^{t-1}J_0 \leq J(f) \leq 2^t J_0\}$ and $A_{s,0} = \{f \in \mathcal{F} : 2^{s-1}\delta_n^2 \leq e_{W^T}(f, f_C^*) \leq 2^s\delta_n^2, J(f) < J_0\}$; $s, t = 1, 2, \cdots$. Then it suffices to bound the corresponding probability over $A_{s,t}$.

For the first moment, by assumption $\delta_n^2 \geq 2\lambda J_0$,

$$\begin{aligned}
& \inf_{A_{s,t}} E(\tilde{W}^T(f(X)) - \tilde{W}(f_C^*(X))) \\
\geq \quad & 2^{s-1}\delta_n^2 + \lambda(2^{t-1} - 1)J_0 = M(s,t); s, t = 1, 2, \cdots, \\
& \inf_{A_{s,0}} E(\tilde{W}^T(f(X)) - \tilde{W}(f_C^*(X))) \\
\geq \quad & 2^{s-1}\delta_n^2 - \lambda J_0 \geq 2^{s-2}\delta_n^2 = M(s,0); s = 1, 2, \cdots .
\end{aligned}$$

For the second moment,

$$\begin{aligned}
& \sup_{A_{s,t}} \frac{1}{n}\Big(n_l \operatorname{Var}\big(\frac{n_l C_1}{n_u C_2}\frac{n}{n_l}(L^T(Yf(X)) - L(Yf_C^*(X)))\big) \\
& \qquad + n_u \operatorname{Var}\big(\frac{n}{n_u}(U(f(X)) - U(f_C^*(X)))\big)\Big) \\
\leq & \sup_{A_{s,t}} \frac{n}{n_l}\Big(\big(\frac{n_l C_1}{n_u C_2}\big)^2 \operatorname{Var}\Big(L^T(Yf(X)) - L(Yf_C^*(X))\Big) \\
& \qquad + \frac{n}{n_u}\operatorname{Var}\Big(U(f(X)) - U(f_C^*(X))\Big)\Big) \\
\leq & nr(n_l, n_u, C_1, C_2) \sup_{A_{s,t}} \operatorname{Var}\Big(W^T(f(X)) - W(f_C^*(X))\Big) \\
\leq & nr(n_l, n_u, C_1, C_2) \sup_{A_{s,t}} a_2(C)\big(e_{W^T}(f, f_C^*)\big)^{\beta(C)} \\
\leq & a_2(C) nr(n_l, n_u, C_1, C_2)\big(2^s\delta_n^2\big)^{\beta(C)} \\
\leq & 2^{\beta(C)} a_2(C) nr(n_l, n_u, C_1, C_2) M(s,t)^{\beta(C)}.
\end{aligned}$$

Now $\mathcal{I} \leq \mathcal{I}_1 + \mathcal{I}_2$ with $\mathcal{I}_1 = \sum_{s,t=1}^{\infty} P^*(\sup_{A_{s,t}} E_n(W(f_C^*) - W^T(f)) \geq M(s,t))$ and $\mathcal{I}_2 = \sum_{s=1}^{\infty} P^*(\sup_{A_{s,0}} E_n(W(f_C^*) - W^T(f)) \geq M(s,0))$. We bound $\mathcal{I}_1$ and $\mathcal{I}_2$ separately using Lemma 1. First, we verify conditions (**??**)-(**??**) there.

To compute the metric entropy of $\{L(yf) - L(yf_C^*) : f \in A_{s,t}\}$ and $\{U(f) - U(f_C^*) : f \in A_{s,t}\}$, we define bracketing functions for them. Suppose $(f_m^l, f_m^u)_{m=1}^M$ with $M = \exp(H(\epsilon, \mathcal{F}))$ forms a ϵ-bracket for $\mathcal{F}$. That is, for any $f \in \mathcal{F}$, there exists a m such that $f_m^l \leq f \leq f_m^u$ and $\|f_m^u - f_m^l\|_2 \leq \epsilon$. Let $f_{\pm 1}$ be truncated version of f such that $f_{\pm 1} = f$ if $|f| \leq 1$ and $\operatorname{Sign}(f)$ otherwise. Furthermore, let $L^l(yf) = 1 - \max(yf_{m,\pm 1}^l, yf_{m,\pm 1}^u)$, $L^u(yf) = 1 - \min(yf_{m,\pm 1}^l, yf_{m,\pm 1}^u)$, $U^l(f) = 1 - \max(|f_{m,\pm 1}^l|, |f_{m,\pm 1}^u|)$ and $U^u(f) = 1 - I(f_m^l f_m^u > 0)\min(|f_{m,\pm 1}^l|, |f_{m,\pm 1}^u|)$. Then $(L^l(yf) - L(yf_C^*), L^u(yf) - L(yf_C^*))$ and $U^l(f) - U(f_C^*), U^u(f) - U(f_C^*))$ form 2ϵ-brackets for $L(yf) - L(yf_C^*)$ and $U(f) - U(f_C^*)$ respectively. Thus Assumption C implies (**??**) using the fact that $\int_{aM(s,t)}^{v(s,t)} H^{1/2}(w, \mathcal{F}(2^t))dw/M(s,t)$ is nonincreasing in s and $M(s,t)$. In addition, (**??**) and (**??**) are satisfied by setting $M = n^{1/2}M(s,t)$, $v = 2^{\beta(C)}a_2(C)nr(n_l, n_u, C_1, C_2)M(s,t)^{\beta(C)}$, $\epsilon = 1/2$ and $\max(|\frac{n}{n_u}\big(U(f_C^*(x_j)) - U(f(x_j)) - E(U(f_C^*(X)) - U(f(X)))|,\ |\frac{n_l C_1}{n_u C_2}\frac{n}{n_l}\big(L(y_i f_C^*(x_i)) - L^T(y_i f(x_i)) - E(L(Yf_C^*(X)) - L^T(Yf(X)))| \leq \frac{\max(C_1, C_2)}{C_2}T$. An application of Lemma 1 yields that $\mathcal{I}_1$ is upper bounded by

$$\begin{aligned}
& \sum_{s,t=1}^{\infty} 3\exp\Big(-\frac{a_6(C)}{r(n_l, n_u, C_1, C_2)}M(s,t)^{\max(1, 2-\beta(C))}\Big) \\
\leq & \sum_{s,t=1}^{\infty} 3\exp\Big(-\frac{a_6(C)}{r(n_l, n_u, C_1, C_2)}(2^{s-1}\delta_n^2 \\
& \qquad + \lambda(2^{t-1} - 1)J_0)^{\max(1, 2-\beta(C))}\Big) \\
\leq & \frac{3\exp(-\frac{a_6(C)}{r(n_l, n_u, C_1, C_2)}(\lambda J_0)^{\max(1, 2-\beta(C))})}{(1 - 3\exp(-\frac{a_6(C)}{r(n_l, n_u, C_1, C_2)}(\lambda J_0)^{\max(1, 2-\beta(C))}))^2}.
\end{aligned}$$

Similarly $\mathcal{I}_2$ is bounded. Then $\mathcal{I} \leq 6\exp(-\frac{a_6(C)}{r(n_l, n_u, C_1, C_2)}(\lambda J_0)^{\max(1, 2-\beta(C))})/(1 - 3\exp(-\frac{a_6(C)}{r(n_l, n_u, C_1, C_2)}(\lambda J_0)^{\max(1, 2-\beta(C))}))^2$. The desired result follows from Assumption A and the fact that $|e(\hat{f}_C, f^*)| \leq |e(\hat{f}_C, f_C^*)| + |e(f_C^*, f^*)|$.

Lemma 1 *Let $\mathcal{F}$ and $\mathcal{G}$ be classes of functions bounded above by T such that $\max(f, g) < T$ for $(f, g) \in \mathcal{F}\bigcup\mathcal{G}$. Let $v_n(f, g) = n^{-1/2}(\sum_{i=1}^{n_l}(f(Z_i) - E(f(Z_i))) + \sum_{j=n_l+1}^{n}(g(Z_i) - E(g(Z_i)))$ for $f \in \mathcal{F}$ and $g \in \mathcal{G}$, and $v \geq \frac{1}{n}(n_l v_1 + n_u v_2)$ with $v_1 = \sup_{\mathcal{F}} \operatorname{Var}(f)$ and $v_2 = \sup_{\mathcal{G}} \operatorname{Var}(g)$. For $M > 0$ and $\epsilon \in (0,1)$, let $\psi_2(M, v, \mathcal{F}, \mathcal{G}) = \frac{M^2}{2(4v + MT/3n^{1/2})}$ and $s = \epsilon M/8n^{1/2}$. Suppose*

$$\frac{n_l}{n}H(v_1^{1/2}, \mathcal{F}) + \frac{n_u}{n}H(v_2^{1/2}, \mathcal{G}) \leq \frac{\epsilon}{4}\psi_2(M, v, \mathcal{F}, \mathcal{G}), \quad (8)$$

$$M \leq \epsilon n^{1/2}\frac{v}{4T}, \qquad \max(v_1, v_2)^{1/2} \leq T, \qquad (9)$$

and if $s < \min(v_1, v_2)^{1/2}$,

$$\frac{n_l}{n}\int_{\frac{s}{4}}^{v_1^{\frac{1}{2}}} (H(u, \mathcal{F}))^{\frac{1}{2}}du + \frac{n_u}{n}\int_{\frac{s}{4}}^{v_2^{\frac{1}{2}}} (H(u, \mathcal{G}))^{\frac{1}{2}}du \leq \frac{M\epsilon^{3/2}}{2^{10}}. \qquad (10)$$

Then

$$P^*\Big(\sup_{f,g} v_n(f, g) \geq M\Big) \leq 3\exp\big(-(1-\epsilon)\psi_2(M, v, \mathcal{F}, \mathcal{G})\big). \qquad (11)$$

Proof: The proof is similar to that of Theorem 3 in Shen and Wong (1994), and thus is omitted.

Appendix B: Dual form of (??)

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 $\nabla = (\nabla_1, \nabla_2)^T$ with $\nabla_1 = \mathbf{0}_l$ and $\nabla_2 = C_2(\nabla U_2(f^{(k)}(x_{n_l+1})), \cdots, \nabla U_2(f^{(k)}(x_n)))$.

Theorem 2 *The dual problem of (**??**) with respect to (α, β, γ) is*

$$\begin{aligned}\max_{\alpha,\beta,\gamma}\Big\{ &-\frac{1}{2}(\mathbf{y}_\alpha^T, (\gamma-\beta)^T)\mathbf{K}(\mathbf{y}_\alpha^T, (\gamma-\beta)^T)^T \\ &+ (\alpha^T, -(\beta+\gamma)^T)\mathbf{1}_n - ({\mathbf{y}_\alpha}^T, (\gamma-\beta)^T)\mathbf{K}\nabla\Big\}\end{aligned} \tag{12}$$

subject to $({\mathbf{y}_\alpha}^T, (\gamma-\beta)^T + \nabla_2)\mathbf{1}_n = 0$, $\mathbf{0}_{n_l} \le \alpha \le C_1\mathbf{1}_{n_l}$, $\mathbf{0}_{n_u} \le \beta$, $\mathbf{0}_{n_u} \le \gamma$, *and* $\mathbf{0}_{n_u} \le \beta + \gamma \le C_2\mathbf{1}_{n_u}$.

Proof: For simplicity, we only prove the linear case as the nonlinear case is essentially the same. Rewrite (**??**) as $\min_{w,b} C_1\sum_{i=1}^{n_l}\xi_i + C_2\sum_{j=n_l+1}^{n}\xi_j + \frac{1}{2}\|w\|^2 - \langle w, D_1\rangle - \langle b, D_2\rangle$ subject to $1 - y_i(w^Tx_i + b) \le \xi_i$, $\xi_i \ge 0$, $|w^Tx_j + b| - 1 \le \xi_j$ and $\xi_j \ge 0$, where $D_1 = (\nabla_2 X^u)^T$ and $D_2 = \nabla_2\mathbf{1}_u$. To solve this minimization problem, the Lagrangian multipliers are employed to yield $L(w, b, \xi_i, \xi_j)$ as

$$\begin{aligned}&C_1\sum_{i=1}^{n_l}\xi_i + C_2\sum_{j=n_l+1}^{n}\xi_j + \frac{1}{2}\|w\|^2 - \langle w, D_1\rangle - \langle b, D_2\rangle + \\ &\sum_{i=1}^{n_l}\alpha_i(1 - y_i(w^Tx_i + b) - \xi_i) - \sum_{i=1}^{n_l}\zeta_i\xi_i + \\ &\sum_{j=n_l+1}^{n}\beta_j(w^Tx_j + b - 1 - \xi_j) - \\ &\sum_{j=n_l+1}^{n}\gamma_j(w^Tx_j + b + 1 + \xi_j) - \sum_{j=n_l+1}^{n}\eta_j\xi_j,\end{aligned}$$

where α_i, β_j, γ_j, ζ_i, η_j are nonnegative. Differentiate L with respect to (w, b, ξ_i, ξ_j) and set the partial derivatives to be zero, we obtain that $\frac{\partial L}{\partial w} = w - D_1 - \sum_{i=1}^{n_l}\alpha_i y_i x_i + \sum_{j=n_l+1}^{n}(\beta_j - \gamma_j)x_j = 0$, $\frac{\partial L}{\partial b} = -D_2 - \sum_{i=1}^{n_l}\alpha_i y_i + \sum_{j=n_l+1}^{n}(\beta_j - \gamma_j) = 0$, $\frac{\partial L}{\partial \xi_i} = C_1 - \alpha_i - \zeta_i = 0$ and $\frac{\partial L}{\partial \xi_j} = C_2 - (\beta_j + \gamma_j) - \eta_j = 0$. Solving these equations yields that $w^* = D_1 + \sum_{i=1}^{n_l}\alpha_i y_i x_i - \sum_{j=n_l+1}^{n}(\beta_j - \gamma_j)x_j$, $D_2 + \sum_{i=1}^{n_l}\alpha_i y_i - \sum_{j=n_l+1}^{n}(\beta_j - \gamma_j) = 0$, $0 \le \alpha_i \le C_1$ and $0 \le \beta_j + \gamma_j \le C_2$. Substituting w^* and these identities into (**??**), we obtain (**??**) after ignoring all constant terms. The solution of (**??**) is $w = D_1 + \sum_{i=1}^{n_l}\alpha_i y_i x_i - \sum_{j=n_l+1}^{n}(\beta_j - \gamma_j)x_j$ and b satisfying KKT's condition: $y_i(w^Tx_i + b) = 1$ for any i with $0 < \alpha_i < C_1$.

Remark: When there is no i such that $y_i(w^Tx_i + b) = 1$, KKT's condition is not applicable. Linear programming (LP) is applied to determine the solution b by substituting w in (**??**).

and thus is omitted.

References

[1] Amini, M., and Gallinari, P. (2003). Semi-supervised learning with an explicit label-error model for misclassified data. *IJCAI2003.*

[2] 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.

[3] Astorino, A., and Fuduli, A. (2005). Nonsmooth optimization techniques for semi-supervised classification. Preprint.

[4] 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 Workshop on Learning with Partially Classified Training Data.*

[5] Baluja, S. (1998). Probabilistic modeling for face orientation discrimination: Learning from labeled and unlabeled data. *Neural Information Processing Systems (NIPS).*

[6] Bennett, K. and Demiriz, A. (1998). Semi-Supervised Support Vector Machines. In *Advances in Neural Information Processing Systems*, **12**, 368-374.

[7] Blum, A., and Chawla, S. (2001). Learning from labeled and unlabeled data using graph mincuts. In *Proceedings of ICML 2001*, 19-26.

[8] Blum, A., and Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. *Proceedings of the* 11*th Annual Conference on Computational Learning Theory.*

[9] Chapelle, O., and Zien, A. (2005). Semi-Supervised Classification by Low Density Separation. In *Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics*, 57-64.

[10] Collobert, R., Sinz, F., Weston, J., and Bottou, L. large scale transductive SVMs. (2006). Submitted.

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

[12] Joachims, T. (1999). Transductive inference for text classification using support vector machines. *ICML1999.*

[13] Liu, S., Shen, X., and Wong, W. (2005). Computational development of ψ-learning. In *The SIAM 2005 International Data Mining Conference*, P1-12.

[14] Nigam, K., McCallum, A., Thrun, S., and Mitchell T. (1998). Text classification from labeled and unlabeled documents using EM. *AAAI1998*.

[15] Schölkopf, B., Smola, A., Williamson, R., and Bartlett, P. (2000). New support vector algorithms. *Neural Computation*, **12** 12071245.

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

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

[18] Shen, X., and Wang, L. (2006). Discussion of 2004 IMS Medallion Lecture: "Local Rademacher complexities and oracle inequalities in risk minimization". *Ann. Statist.*, to appear.

[19] Szummer, M. and Jaakkola, T. (2002) Information regularization with partially labeled data. In *Advances in Neural Information Processing Systems* **15**.

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

[21] Wang, J., and Shen, X. (2006). Large margin semi-supervised learning. Submitted.

[22] Wu, D., Bennett, K., Cristianini, N. and Shawe-Taylor, J. (1999). Large Margin Decision Trees for Induction and Transduction, in *Proceedings of ICML99*, 474-483.

[23] Zhang, T., and Oles, F. (2000). A probability analysis on the value of unlabeled data for classification problems. *ICML2000.*

[24] Zhu, X., Ghahramani, Z. and Lafferty, J. (2003). Semi-supervised learning using gaussian fields and harmonic functions. *ICML2003*.