Journal of Machine Learning Research 11 (2010) 2423-2455 Submitted 12/09; Published 9/10

Sparse Semi-supervised Learning Using Conjugate Functions

Shiliang Sun SHILIANGSUN@GMAIL.COM
Department of Computer Science and Technology
East China Normal University
500 Dongchuan Road, Shanghai 200241, P. R. China

John Shawe-Taylor JST@CS.UCL.AC.UK
Department of Computer Science
University College London
Gower Street, London WC1E 6BT, United Kingdom

Editor: Tony Jebara

Abstract

In this paper, we propose a general framework for sparse semi-supervised learning, which concerns using a small portion of unlabeled data and a few labeled data to represent target functions and thus has the merit of accelerating function evaluations when predicting the output of a new example. This framework makes use of Fenchel-Legendre conjugates to rewrite a convex insensitive loss involving a regularization with unlabeled data, and is applicable to a family of semi-supervised learning methods such as multi-view co-regularized least squares and single-view Laplacian support vector machines (SVMs). As an instantiation of this framework, we propose sparse multi-view SVMs which use a squared ε-insensitive loss. The resultant optimization is an inf-sup problem and the optimal solutions have arguably saddle-point properties. We present a globally optimal iterative algorithm to optimize the problem. We give the margin bound on the generalization error of the sparse multi-view SVMs, and derive the empirical Rademacher complexity for the induced function class. Experiments on artificial and real-world data show their effectiveness. We further give a sequential training approach to show their possibility and potential for uses in large-scale problems and provide encouraging experimental results indicating the efficacy of the margin bound and empirical Rademacher complexity on characterizing the roles of unlabeled data for semi-supervised learning.

Keywords: semi-supervised learning, Fenchel-Legendre conjugate, representer theorem, multi-view regularization, support vector machine, statistical learning theory

1. Introduction

Semi-supervised learning, considering how to estimate a target function from a few labeled examples and a large quantity of unlabeled examples, is one of currently active research directions. If the unlabeled data are properly used, it can get a superior performance over the counterpart supervised learning approaches. For an overview of semi-supervised learning methods, refer to Chapelle et al. (2006) and Zhu (2008).

Although semi-supervised learning was largely motivated by different real-world applications where obtaining labels is expensive or time-consuming, a lot of theoretical outcomes have also been accomplished. Typical applications of semi-supervised learning include natural image classification and text classification, where it is inexpensive to collect large numbers of images and texts by

automatic programs, but needs a high cost to label them manually. Theoretical results on semi-supervised learning include PAC-analysis (Balcan and Blum, 2005), manifold regularization (Belkin et al., 2006), and multi-view regularization theories (Sindhwani and Rosenberg, 2008), etc.

Among the methods proposed for semi-supervised learning, a family of them, for example, Laplacian regularized least squares (RLS), Laplacian support vector machines (SVMs), Co-RLS, Co-Laplacian RLS, Co-Laplacian SVMs, and manifold co-regularization (Belkin et al., 2006; Sindhwani et al., 2005; Sindhwani and Rosenberg, 2008), make use of the following representer theorem (Kimeldorf and Wahba, 1971) to represent the target function in a reproducing kernel Hilbert space (RKHS).

Theorem 1 (Representer theorem) *Let $\mathcal{H}$ be an RKHS with kernel $k : \mathcal{X} \times \mathcal{X} \to \boldsymbol{R}$. Fix any function $V : \boldsymbol{R}^n \to \boldsymbol{R}$ and any nondecreasing function $\Psi : \boldsymbol{R} \to \boldsymbol{R}$. Define*

$$J(f) = V(f(x_1), ..., f(x_n)) + \Psi(\|f\|^2),$$

and linear space $\mathcal{L} = span\{k(x_1, \cdot), ..., k(x_n, \cdot)\}$. Then for any $f \in \mathcal{H}$ we have $J(f_{\mathcal{L}}) \leq J(f)$ with $f_{\mathcal{L}}$ being the projection of f onto $\mathcal{L}$ in the following form

$$f_{\mathcal{L}} = \sum_{i=1}^{n} \alpha_i k(x_i, \cdot).$$

Thus if $J^* = \min_f J(f)$ exists, this minimum is attained for some $f \in \mathcal{L}$. Moreover, if Ψ is strictly increasing, each minimizer of $J(f)$ over $\mathcal{H}$ must be contained in $\mathcal{L}$.

Generally, in the objective function $J(f)$ of these semi-supervised learning methods, labeled examples are used to calculate an empirical loss of the target function and simultaneously unlabeled examples are used for some regularization purpose. By the representer theorem, the target function would involve kernel evaluations on all the labeled and unlabeled examples. This is computationally undesirable, because for semi-supervised learning usually a considerably large number of unlabeled examples are available. Consequently, sparsity in the number of unlabeled data used to represent target functions is crucial, which constitutes the focus of this paper.

However, little work has been done on this theme. In particular, there is no unified framework proposed yet to deal with this sparsity concern. While the sparse Laplacian core vector machines (Tsang and Kwok, 2007) touched this problem, it has a complicated optimization and is not generic enough to generalize to other similar semi-supervised learning methods. In contrast with this, the technique developed in this paper, based on Fenchel-Legendre conjugates, is computationally simple and widely applicable.

As far as multi-view learning is concerned there has been work that introduces sparsity of the unlabeled data into the representation of the classifiers (Szedmak and Shawe-Taylor, 2007). This builds on the ideas developed for two view learning known as the SVM-2K (Farquhar et al., 2006). The approach adopted is the use of an ε-insensitive loss function for the similarity constraint between the two functions from two views. Unfortunately the resulting optimization is somewhat unmanageable and only scales to small-scale data sets despite interesting theoretical bounds that show the improvement gained using the unlabeled data.

The work by Szedmak and Shawe-Taylor (2007) forms the starting point for the current paper which aims to develop related methods that are possible to be scaled to very large data sets. Our approach is to go back to consider l_2 loss between the outputs of the classifiers arising from two

views and shows that this problem can be solved implicitly with variables only indexed by the labeled data. To compute the value of this function on new data would still require a non-sparse dual representation in terms of the unlabeled data. However, we show that through optimizing weights of the unlabeled data the solution of the l_2 problem converges to the solution of an ε-insensitive problem ensuring that we subsequently obtain sparsity in the unlabeled data. Furthermore, we develop the generalization analysis of Szedmak and Shawe-Taylor (2007) to this case giving computable expressions for the corresponding empirical Rademacher complexity.

To show the application of Fenchel-Legendre conjugates, in Section 2 we propose a novel sparse semi-supervised learning approach: sparse multi-view SVMs, where the conjugate functions play a central role in reformulating the optimization problem. The dual optimization of a subroutine of the sparse multi-view SVMs is converted to a quadratic programming problem in Section 3 whose scale only depends on the number of labeled examples, indicating the advantages of using conjugate functions. The generalization error of the sparse multi-view SVMs is given in Section 4 in terms of Rademacher complexity theory, followed by a derivation of empirical Rademacher complexity of the class of functions induced by this new method in Section 5. Section 6 reports experimental results of the sparse multi-view SVMs, comparisons with related methods, and the possibility and potential for large-scale applications through sequential training. Extensions of the use of conjugate functions to a general convex loss and other related semi-supervised learning approaches are discussed in Section 7. Finally, Section 8 concludes this paper.

2. Sparse Multi-view SVMs

Multi-view semi-supervised learning, an important branch of semi-supervised learning, combines different sets of properties of an example to learn a target function. These different sets of properties are often referred to as views. Typical applications of multi-view learning are web-page categorization and content-based multimedia information retrieval. In web-page categorization, each web-page can be simultaneously described by disparate properties such as main text, inbound and outbound hyper-links. In content-based multimedia information retrieval, a multimedia segment can include both audio and video components. For such scenarios learning with multiple views is usually very beneficial. Even for problems with no natural multiple views, artificially generated views can still work favorably (Nigam and Ghani, 2000).

A useful assumption for multi-view learning is that features from each view are sufficient to train a good learner (Blum and Mitchell, 1998; Balcan et al., 2005; Farquhar et al., 2006). Making good use of this assumption through collaborative training or regularization between views can remove many false hypotheses from the hypothesis space, and thus facilitates effective learning.

For multi-view learning, an input x consists of multiple components from different views, for example, $x = (x^1, \ldots, x^m)$ for an m-view representation. A function f_j defined on view j only depends on x^j, that is $f_j(x) := f_j(x^j)$. Suppose we have a set of ℓ labeled examples $\{(x_i, y_i)\}_{i=1}^{\ell}$ with $y_i \in \{1, -1\}$, and a set of u unlabeled examples $\{x_i\}_{i=\ell+1}^{\ell+u}$. The objective function of our sparse multi-view SVMs in the case of two views is given as follows, which can be readily extended to

more than two views.

$$\min_{f_1\in\mathcal{H}_1, f_2\in\mathcal{H}_2} \quad \frac{1}{2\ell}\sum_{i=1}^{\ell}[(1-y_i f_1(x_i))_+ + (1-y_i f_2(x_i))_+] + \gamma_n(\|f_1\|^2 + \|f_2\|^2) + \gamma_v \sum_{i=1}^{\ell+u}(|f_1(x_i) - f_2(x_i)| - \varepsilon)_+^2, \tag{1}$$

where nonnegative scalars γ_n, γ_v are respectively norm regularization and multi-view regularization coefficients, and the last term is an ε-insensitive loss between two views with function $(\cdot)_+ := \max(0,\cdot)$ being the hinge loss. The final classifier for predicting the label of a new example is

$$f_c(x) = \operatorname{sgn}\left(\frac{f_1(x) + f_2(x)}{2}\right). \tag{2}$$

In the rest of this section, we will show that the use of the ε-insensitive loss indeed enforces sparsity, and Fenchel-Legendre conjugates can be adopted to reformulate the optimization problem. We also show the saddle-point properties for optimal solutions and give a (globally optimal) iterative optimization algorithm.

2.1 Sparsity

In order to show the role of the ε-insensitive loss for sparsity pursuit, here we represent $f_1(x)$ and $f_2(x)$ in feature spaces as

$$f_1(x) = \mathbf{w}_1^\top \phi_1(x) + b_1, \quad f_2(x) = \mathbf{w}_2^\top \phi_2(x) + b_2,$$

where $\phi_i(x)$ $(i = 1,2)$ is the image of x in feature spaces. Problem (1) can be rewritten as

$$\begin{aligned} \min_{\mathbf{w}_1,\mathbf{w}_2,\xi_1,\xi_2,b_1,b_2} \quad & P_0 = \frac{1}{2\ell}\sum_{i=1}^{\ell}(\xi_1^i + \xi_2^i) + \gamma_n(\|\mathbf{w}_1\|^2 + \|\mathbf{w}_2\|^2) + \\ & \gamma_v \sum_{i=1}^{\ell+u}(|\mathbf{w}_1^\top \phi_1(x_i) + b_1 - \mathbf{w}_2^\top \phi_2(x_i) - b_2| - \varepsilon)_+^2 \\ \text{s.t.} \quad & \begin{cases} y_i(\mathbf{w}_1^\top \phi_1(x_i) + b_1) \geq 1 - \xi_1^i, \\ y_i(\mathbf{w}_2^\top \phi_2(x_i) + b_2) \geq 1 - \xi_2^i, \\ \xi_1^i,\ \xi_2^i \geq 0, \quad i = 1,\ldots,\ell, \end{cases} \end{aligned}$$

where $\xi_1 := [\xi_1^1,\ldots,\xi_1^\ell]$ and $\xi_2 := [\xi_2^1,\ldots,\xi_2^\ell]$.

The Lagrangian is

$$\begin{aligned} L = \quad & P_0 - \sum_{i=1}^{\ell}[\lambda_1^i(y_i(\mathbf{w}_1^\top \phi_1(x_i) + b_1) - 1 + \xi_1^i) + \\ & \lambda_2^i(y_i(\mathbf{w}_2^\top \phi_2(x_i) + b_2) - 1 + \xi_2^i) + \nu_1^i \xi_1^i + \nu_2^i \xi_2^i], \end{aligned}$$

where $\lambda_1^i, \lambda_2^i, \nu_1^i, \nu_2^i \geq 0$ $(i = 1,\ldots,\ell)$ are Lagrange multipliers.

Suppose $\mathbf{w}_{1*}$, $\mathbf{w}_{2*}$ are the optimal solutions. By the KKT conditions, the optimal solutions $\mathbf{w}_{1*}$ should satisfy $\frac{\partial L}{\partial \mathbf{w}_{1*}} = 0$. Therefore, we get

$$\mathbf{w}_{1*} = -\frac{\gamma_v}{\gamma_n}\sum_{i=1}^{\ell+u}(|\mathbf{w}_1^\top \phi_1(x_i) + b_1 - \mathbf{w}_2^\top \phi_2(x_i) - b_2| - \varepsilon)_+ \tilde{\phi}_i + \frac{1}{2\gamma_n}\sum_{i=1}^{\ell}\lambda_1^i y_i \phi_1(x_i),$$

where we suppose the derivative exists everywhere and $\tilde{\phi}_i := \text{sgn}\{\mathbf{w}_1^\top \phi_1(x_i) + b_1 - \mathbf{w}_2^\top \phi_2(x_i) - b_2\}\phi_1(x_i)$. Now we can assess the sparsity of problem (1). From the above equation, $\mathbf{w}_{1*}$ is the linear combination of labeled examples with $\lambda_1^i > 0$, and those unlabeled examples on which the difference of predictions from two views exceeds ε. In this sense, we can get sparse solutions by providing a non-zero ε. Similar analysis applies to $\mathbf{w}_{2*}$. Therefore, function $f(x)$ is sparse in the number of used unlabeled examples. This analysis on sparsity is also well justified by the representer theorem.

2.2 Reformulation Using Conjugate Functions

Define $t_i := [f_1(x_i) - f_2(x_i)]^2$. Then the ε-insensitive loss term can be written as

$$f_\varepsilon(\mathbf{t}) = \sum_{i=1}^{\ell+u} (\sqrt{t_i} - \varepsilon)_+^2, \tag{3}$$

where vector $\mathbf{t} := [t_1, \ldots, t_{\ell+u}]^\top$. We give a theorem affirming the convexity of function $f_\varepsilon(\mathbf{t})$.

Theorem 2 *Function $f_\varepsilon(\boldsymbol{t})$ defined by (3) is convex.*

Proof First, we show that $f_\varepsilon(t_i) := (\sqrt{t_i} - \varepsilon)_+^2$ with a convex domain $[0, +\infty)$ is convex. When $t_i \in (\varepsilon^2, +\infty)$, the second derivative $\nabla^2 f_\varepsilon(t_i) = \frac{1}{2}\varepsilon t_i^{-3/2} \geq 0$. Thus, function $f_\varepsilon(t_i)$ is convex for $t_i \in (\varepsilon^2, +\infty)$. Moreover, the value of function $f_\varepsilon(t_i)$ for $t_i \in (\varepsilon^2, +\infty)$ is larger than 0 which is the value of $f_\varepsilon(t_i)$ for $t_i \in [0, \varepsilon^2]$, and function $f_\varepsilon(t_i)$ with domain $[0, +\infty)$ is continuous at ε^2. Hence, $f_\varepsilon(t_i)$ is convex on the domain $[0, +\infty)$.

Then, being a nonnegative weighted sum of convex functions, $f_\varepsilon(\mathbf{t})$ is indeed convex. ■

Define conjugate vector $z = [z_1, \ldots, z_{\ell+u}]^\top$ with entries being conjugate variables. The Fenchel-Legendre conjugate (which is also often called convex conjugate or conjugate function) $f_\varepsilon^*(z)$ is

$$f_\varepsilon^*(z) = \sup_{\mathbf{t} \in \mathbf{dom}\, f_\varepsilon} (\mathbf{t}^\top z - f_\varepsilon(\mathbf{t})) = \sup_{\mathbf{t}} \sum_{i=1}^{\ell+u} [z_i t_i - (\sqrt{t_i} - \varepsilon)_+^2] = \sum_{i=1}^{\ell+u} \sup_{t_i} [z_i t_i - (\sqrt{t_i} - \varepsilon)_+^2].$$

The domain of the conjugate function consists of $z \in \mathbf{R}^{\ell+u}$ for which the supremum is finite (i.e., bounded above) (Boyd and Vandenberghe, 2004). Define

$$f_\varepsilon^*(z_i) = \sup_{t_i} [z_i t_i - (\sqrt{t_i} - \varepsilon)_+^2]. \tag{4}$$

Then, $f_\varepsilon^*(z) = \sum_{i=1}^{\ell+u} f_\varepsilon^*(z_i)$. As a pointwise supremum of a family of affine functions, $f_\varepsilon^*(z_i)$ is convex. Being a nonnegative weighted sum of convex functions, $f_\varepsilon^*(z)$ is also convex. Below we derive the formulation of $f_\varepsilon^*(z_i)$.

Theorem 3 *Function $f_\varepsilon^*(z_i)$ defined by (4) has the following form*

$$f_\varepsilon^*(z_i) = \begin{cases} \frac{z_i \varepsilon^2}{1 - z_i}, & \text{for } 0 < z_i < 1 \\ 0, & \text{for } z_i \leq 0\,. \end{cases} \tag{5}$$

Proof By definition, we have

$$f_\varepsilon^*(z_i) = \max_{t_i} \left\{ \sup_{0\le t_i\le \varepsilon^2} z_i t_i, \sup_{t_i>\varepsilon^2} [z_i t_i - (\sqrt{t_i}-\varepsilon)^2] \right\}. \tag{6}$$

The value of function $\sup_{0\le t_i\le\varepsilon^2} z_i t_i$ is simple to characterize. We now characterize the second term $\sup_{t_i>\varepsilon^2}[z_i t_i - (\sqrt{t_i}-\varepsilon)^2] = \sup_{t_i>\varepsilon^2}(z_i t_i - t_i - \varepsilon^2 + 2\varepsilon\sqrt{t_i})$. For $0 < z_i < 1$, we let the first derivative equal to zero to find the supremum. For $z_i \le 0$ or $z_i \ge 1$ the derivative does not exist and thus we use function values at end points to find the supremum. As a result, we have

$$\sup_{t_i>\varepsilon^2}[z_i t_i - (\sqrt{t_i}-\varepsilon)^2] = \begin{cases} \frac{z_i\varepsilon^2}{1-z_i}, & \text{for } 0 < z_i < 1 \\ z_i\varepsilon^2, & \text{for } z_i \le 0 \\ +\infty, & \text{for } z_i \ge 1\,. \end{cases}$$

According to (6) and further removing the range where $f_\varepsilon^*(z_i)$ is unbounded above, we reach the conjugate given in (5). ■

Now the Fenchel-Legendre conjugate $f_\varepsilon^*(z)$ can be represented by $\sum_{i=1}^{\ell+u} f_\varepsilon^*(z_i)$, which is also well justified by the following theorem.

Theorem 4 (Boyd and Vandenberghe, 2004) *If $\varphi(u,v) = \varphi_1(u) + \varphi_2(v)$, where φ_1 and φ_2 are independent convex functions (independent means they are functions of different variables) with conjugates φ_1^* and φ_2^*, respectively, then*

$$\varphi^*(\omega, z) = \varphi_1^*(\omega) + \varphi_2^*(z).$$

A nice property of the conjugate function is on the conjugate of the conjugate, which is central to the reformulation of our optimization problem. This property is stated by Lemma 5.

Lemma 5 (Rifkin and Lippert, 2007) *If function f is closed, convex, and proper, then the conjugate function of the conjugate is itself, that is, $f^{**} = f$, where we have defined function f is closed if its epigraph is closed, and f is proper if $\mathbf{dom} f \neq \emptyset$ and $f > -\infty$.*

It is true that function $f_\varepsilon(\mathbf{t})$ is closed and proper. Moreover, we have proved the convexity of $f_\varepsilon(\mathbf{t})$ in Theorem 2. Therefore, we can use Lemma 5 to get the following equality

$$f_\varepsilon(\mathbf{t}) = \sup_z (z^\top \mathbf{t} - f_\varepsilon^*(z)). \tag{7}$$

That is

$$\sum_{i=1}^{\ell+u} (\sqrt{t_i}-\varepsilon)_+^2 = \sup_z (z^\top \mathbf{t} - f_\varepsilon^*(z)) = \sup_z \sum_{i=1}^{\ell+u} [z_i t_i - f_\varepsilon^*(z_i)].$$

By (7), we have

$$\sum_{i=1}^{\ell+u} (|f_1(x_i) - f_2(x_i)| - \varepsilon)_+^2 = \sum_{i=1}^{\ell+u} (\sqrt{t_i}-\varepsilon)_+^2 = \sup_z \sum_{i=1}^{\ell+u} [z_i t_i - f_\varepsilon^*(z_i)].$$

Therefore, the objective function for sparse multi-view SVMs becomes

$$\min_{f_1\in\mathcal{H}_1,\, f_2\in\mathcal{H}_2} \quad \frac{1}{2\ell}\sum_{i=1}^{\ell}[(1-y_i f_1(x_i))_+ + (1-y_i f_2(x_i))_+] + \gamma_n(\|f_1\|^2+\|f_2\|^2) + \gamma_v \sup_z \sum_{i=1}^{\ell+u}\{z_i[f_1(x_i)-f_2(x_i)]^2 - f_\varepsilon^*(z_i)\}. \tag{8}$$

As an application of Theorem 1, the solution to problem (8) has the following form

$$f_1(x)=\sum_{i=1}^{\ell+u}\alpha_1^i k_1(x_i,x),\ f_2(x)=\sum_{i=1}^{\ell+u}\alpha_2^i k_2(x_i,x). \tag{9}$$

Applying the reproducing properties of kernels, we get

$$\|f_1\|^2=\alpha_1^\top K_1\alpha_1,\ \|f_2\|^2=\alpha_2^\top K_2\alpha_2,$$

where K_1 and K_2 are $(\ell+u)\times(\ell+u)$ Gram matrices from two views $\mathcal{V}^1$ and $\mathcal{V}^2$, respectively, and vector $\alpha_1=(\alpha_1^1,...,\alpha_1^{\ell+u})^\top$, $\alpha_2=(\alpha_2^1,...,\alpha_2^{\ell+u})^\top$. Moreover, we have

$$\mathbf{f}_1=K_1\alpha_1,\ \mathbf{f}_2=K_2\alpha_2,$$

with $\mathbf{f}_1 := (f_1(x_1),...,f_1(x_{\ell+u}))^\top$, $\mathbf{f}_2 := (f_2(x_1),...,f_2(x_{\ell+u}))^\top$. Define diagonal matrix $U=diag(z_1,\ldots,z_{\ell+u})$ with every element taking values in the range $[0,1)$. Problem (8) can be reformulated as

$$\begin{aligned}\min_{\alpha_1,\alpha_2,\xi_1,\xi_2,b_1,b_2}\sup_z \quad & \frac{1}{2\ell}\sum_{i=1}^{\ell}(\xi_1^i+\xi_2^i)+\gamma_n(\alpha_1^\top K_1\alpha_1+\alpha_2^\top K_2\alpha_2)+\\ & \gamma_v[(K_1\alpha_1-K_2\alpha_2)^\top U(K_1\alpha_1-K_2\alpha_2)-\sum_{i=1}^{\ell+u}f_\varepsilon^*(z_i)]\\ \text{s.t.} \quad & \begin{cases} y_i(\sum_{j=1}^{\ell+u}\alpha_1^j k_1(x_j,x_i)+b_1)\geq 1-\xi_1^i,\\ y_i(\sum_{j=1}^{\ell+u}\alpha_2^j k_2(x_j,x_i)+b_2)\geq 1-\xi_2^i,\\ \xi_1^i,\ \xi_2^i\geq 0,\quad i=1,\ldots,\ell\,. \end{cases}\end{aligned} \tag{10}$$

2.3 Saddle-Point Property

We present a theorem concerning the convexity and concavity of optimization problem (10).

Theorem 6 *The objective function in problem (10) is convex with respect to* α_1,α_2, ξ_1, ξ_2, b_1, *and* b_2, *and concave with respect to z.*

Proof First, we show the convexity. The standard form of this optimization problem is

$$\begin{aligned}\min_{\alpha_1,\alpha_2,\xi_1,\xi_2,b_1,b_2}\sup_z \quad & \frac{1}{2\ell}\sum_{i=1}^{\ell}(\xi_1^i+\xi_2^i)+\gamma_n(\alpha_1^\top K_1\alpha_1+\alpha_2^\top K_2\alpha_2)+\\ & \gamma_v[(K_1\alpha_1-K_2\alpha_2)^\top U(K_1\alpha_1-K_2\alpha_2)-\sum_{i=1}^{\ell+u}f_\varepsilon^*(z_i)]\\ \text{s.t.} \quad & \begin{cases} -y_i(\sum_{j=1}^{\ell+u}\alpha_1^j k_1(x_j,x_i)+b_1)+1-\xi_1^i\leq 0,\\ -y_i(\sum_{j=1}^{\ell+u}\alpha_2^j k_2(x_j,x_i)+b_2)+1-\xi_2^i\leq 0,\\ -\xi_1^i,\ -\xi_2^i\leq 0,\quad i=1,\ldots,\ell\,. \end{cases}\end{aligned}$$

This problem involves one objective function and three sets of inequality constraint functions (on the left hand side of each inequality). Clearly, the domain of each objective and constraint function is a convex set. Now it suffices to prove the convexity of this problem by assessing the convexity of these functions. As all constraint functions are affine, they are convex. Then, we use the second-order condition, positive semidefinite property of a function's Hessian or second derivative to judge the convexity of the objective function (Boyd and Vandenberghe, 2004). According to this condition, the first two items of the objective function are clear to be convex. The third part can be rewritten as

$$(K_1\alpha_1 - K_2\alpha_2)^\top U(K_1\alpha_1 - K_2\alpha_2) = \|U^{1/2}\begin{pmatrix} K_1 & -K_2 \end{pmatrix}\begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}\|^2,$$

which is a convex function $\|\cdot\|^2$ composed with an affine mapping and thus also convex (Boyd and Vandenberghe, 2004). Being a nonnegative weighted sums of convex functions, the objective function is therefore convex with respect to $\alpha_1, \alpha_2, \xi_1, \xi_2, b_1, b_2$.

Then, we show the concavity using (8). As $f_\varepsilon^*(z_i)$ is convex, $z_i[f_1(x_i) - f_2(x_i)]^2 - f_\varepsilon^*(z_i)$ is concave with respect to z_i. The concavity of $\sum_{i=1}^{\ell+u}\{z_i[f_1(x_i) - f_2(x_i)]^2 - f_\varepsilon^*(z_i)\}$ follows from the fact that a nonnegative weighted sum of concave functions is concave (Boyd and Vandenberghe, 2004). Hence, the objective function in problem (10) is concave with respect to z. ■

Let θ denote the parameters $\alpha_1, \alpha_2, \xi_1, \xi_2, b_1, b_2$. We can simply denote the above optimization problem as

$$\inf_\theta \sup_z f(\theta, z) \tag{11}$$

associated with constraints on the labeled examples, where $f(\theta, z)$ is convex with respect to θ, and concave with respect to z. We give the following theorem on the equivalence of swapping the infimum and supremum for our optimization problem and include a proof for completeness.

Theorem 7 (Boyd and Vandenberghe, 2004) *If $f(\theta, \mathbf{z})$ with domain Θ and Z is convex with respect to $\theta \in \Theta$, and concave with respect to $\mathbf{z} \in Z$, the following equality holds*

$$\inf_\theta \sup_{\mathbf{z}} f(\theta, \mathbf{z}) = \sup_{\mathbf{z}} \inf_\theta f(\theta, \mathbf{z})$$

under some slight assumptions.

Proof The idea is first to represent the left-hand side as a value of a convex function, and then show that the conjugate of its conjugate is equal to the right-hand side when the same input value is plugged in Boyd and Vandenberghe (2004).

The left-hand side can be expressed as $p(0)$, where

$$p(\mathbf{u}) = \inf_\theta \sup_z [f(\theta, z) + \mathbf{u}^\top z] .$$

It is not difficult to show that p is a convex function. Being a pointwise supremum of convex function $f(\theta, z) + \mathbf{u}^\top z$, $\sup_z[f(\theta, z) + \mathbf{u}^\top z]$ is a convex function of $(\theta, \mathbf{u})$. Because $\sup_z[f(\theta, z) + \mathbf{u}^\top z]$ is convex with respect to $(\theta, \mathbf{u})$, we have $p(\mathbf{u})$ is convex.

The Fenchel-Legendre conjugate of $p(\mathbf{u})$ is

$$p^*(\mathbf{v}) = \sup_{\mathbf{u}}[\mathbf{u}^\top \mathbf{v} - \inf_\theta \sup_z (f(\theta, z) + \mathbf{u}^\top z)] ,$$

which would be $+\infty$ if $z \neq \mathbf{v}$. Therefore,

$$p^*(\mathbf{v}) = \begin{cases} -\inf_\theta f(\theta, \mathbf{v}), & \text{for } \mathbf{v} \in Z \\ +\infty, & \text{otherwise .} \end{cases}$$

The conjugate of $p^*(z)$ is given by

$$p^{**}(\mathbf{u}) = \sup_{z \in Z}(\mathbf{u}^\top z - p^*(z)) = \sup_{z \in Z}(\mathbf{u}^\top z + \inf_\theta f(\theta, z)) = \sup_z \inf_\theta [f(\theta, z) + \mathbf{u}^\top z] .$$

Suppose $0 \in \mathbf{dom} p(\mathbf{u})$ and $p(\mathbf{u})$ is closed and proper. Then by Lemma 5 we have $p(0) = p^{**}(0)$ which completes the proof. ■

Now we give a theorem showing that the optimal pair $\tilde{\theta}$, $\tilde{z}$ is a saddle-point.

Theorem 8 *If the following equality holds for function* $f(\theta, z)$

$$\inf_\theta \sup_z f(\theta, \boldsymbol{z}) = \sup_{\boldsymbol{z}} \inf_\theta f(\theta, \boldsymbol{z}) = f(\tilde{\theta}, \tilde{\boldsymbol{z}}) ,$$

then the optimal pair $\tilde{\theta}$, $\tilde{\boldsymbol{z}}$ *forms a saddle-point.*

Proof From the given equality, we have

$$\inf_\theta \sup_z f(\theta, z) = \sup_z f(\tilde{\theta}, z) = f(\tilde{\theta}, \tilde{z}) ,$$

and

$$\sup_z \inf_\theta f(\theta, z) = \inf_\theta f(\theta, \tilde{z}) = f(\tilde{\theta}, \tilde{z}) ,$$

Therefore,

$$f(\tilde{\theta}, z) \leq f(\tilde{\theta}, \tilde{z}) \leq f(\theta, \tilde{z}),$$

which indeed satisfies the definition of a saddle-point. The proof is completed. ■

2.4 Iterative Optimization Algorithm

To solve the optimization problem $\sup_z \inf_\theta f(\theta, z)$ which is respectively concave and convex with respect to z and θ, we give an algorithm with guaranteed convergence by the following theorem.

Theorem 9 *Given an initial value* $\boldsymbol{z}_0$ *for* $\boldsymbol{z}$*, solve* $\inf_\theta f(\theta, \boldsymbol{z}_0)$ *and obtain the global optimal point* θ_0*. Then we find* $\arg\max_{\boldsymbol{z}} f(\theta_0, \boldsymbol{z})$ *to get* $\boldsymbol{z}_1$ *from which we can get* θ_1 *as a result of optimize* $\inf_\theta f(\theta, \boldsymbol{z}_1)$*. Repeat this process until a convergence point* $(\hat{\theta}, \hat{\boldsymbol{z}})$ *is reached. Suppose* $\tilde{\theta}$, $\tilde{\boldsymbol{z}}$ *is a saddle point. We have* $f(\hat{\theta}, \hat{\boldsymbol{z}}) = f(\tilde{\theta}, \tilde{\boldsymbol{z}})$*. That is, we got the optimal values of the objective function. If* f *is strictly concave and strictly convex with respect to the variables, we further have* $\hat{\theta} = \tilde{\theta}$ *and* $\hat{\boldsymbol{z}} = \tilde{\boldsymbol{z}}$*.*

Proof According to the properties of function f and the algorithm procedure, we know that the convergence point is a saddle point. Thus, we have

$$f(\tilde{\theta}, \hat{z}) \geq f(\hat{\theta}, \hat{z}) \geq f(\hat{\theta}, \tilde{z}) .$$

By the saddle-point property of $(\tilde{\theta},\tilde{z})$, we have

$$f(\tilde{\theta},\tilde{z}) \geq f(\tilde{\theta},\hat{z}) ,$$

and

$$f(\tilde{\theta},\tilde{z}) \leq f(\hat{\theta},\tilde{z}) .$$

Therefore, the above inequalities should hold with equalities and we have $f(\tilde{\theta},\tilde{z}) = f(\hat{\theta},\hat{z})$. Furthermore, if f is strictly concave and strictly convex with respect to the variables, it is true that $\hat{\theta} = \tilde{\theta}$ and $\hat{z} = \tilde{z}$. ■

On solving $\arg\max_{\mathbf{z}} f(\theta,\mathbf{z})$ required in Theorem 9, we can maximize the term related to $\mathbf{z}$, namely $z^\top \mathbf{t} - f_\varepsilon^*(z) = \sum_{i=1}^{\ell+u}[z_i t_i - f_\varepsilon^*(z_i)]$. For this purpose, we have the following theorem.

Theorem 10

$$\sup_{z_i \in \mathbf{dom}\, f_\varepsilon^*(z_i)} [z_i t_i - f_\varepsilon^*(z_i)] = (\sqrt{t_i} - \varepsilon)_+^2,$$

and

$$\arg \sup_{z_i \in \mathbf{dom}\, f_\varepsilon^*(z_i)} [z_i t_i - f_\varepsilon^*(z_i)] = \begin{cases} 1 - \frac{\varepsilon}{\sqrt{t_i}}, & \text{for } t_i > \varepsilon^2 \\ 0, & \text{for } 0 \leq t_i \leq \varepsilon^2 . \end{cases}$$

Without loss of generality, we can confine the range of z_i to $[0,1)$.

Proof We have

$$\sup_{z_i \in \mathbf{dom}\, f_\varepsilon^*(z_i)} [z_i t_i - f_\varepsilon^*(z_i)] = \max_{z_i} \{ \sup_{0<z_i<1} z_i t_i - \frac{z_i \varepsilon^2}{1-z_i}, \sup_{z_i \leq 0} z_i t_i \} .$$

The first supremum can be solved by setting the derivative with respect to z_i to zero. We have

$$\sup_{0<z_i<1} z_i t_i - \frac{z_i \varepsilon^2}{1-z_i} = (\sqrt{t_i} - \varepsilon)^2$$

where $t_i > \varepsilon^2$, and the supremum is attained with $z_i = 1 - \frac{\varepsilon}{\sqrt{t_i}}$.

When $t_i < 0$, $\sup_{z_i \leq 0} z_i t_i$ is unbounded above. When $0 \leq t_i \leq \varepsilon^2$, $\sup_{z_i \leq 0} z_i t_i = 0$ with the supremum attained at $z_i = 0$. Therefore, $\max_{z_i}\{\sup_{0<z_i<1} z_i t_i - \frac{z_i \varepsilon^2}{1-z_i}, \sup_{z_i \leq 0} z_i t_i\} = (\sqrt{t_i} - \varepsilon)_+^2$ with the supremum attained when $z_i \in [0,1)$, which completes the proof. ■

For sparsity pursuit, during each iteration we remove those unlabeled examples whose corresponding z_i's are zero. By the representer theorem, this would not influence the value of the objective function. For Theorem 9, this means that the element of z whose values are zero in the last iteration will remain zero for the next iteration. When there are no unlabeled examples eligible for elimination, the iteration will terminate and the convergence point $(\hat{\theta},\hat{\mathbf{z}})$ is reached.

3. Dual Optimization

According to the iterative optimization algorithm, when optimizing problem (10), we start from an initial value z_0 and then solve θ_0. In this section, we show how to solve this subroutine with fixed z.

Now the optimization problem is equivalent to

$$\begin{aligned}\min_{\alpha_1,\alpha_2,\xi_1,\xi_2,b_1,b_2} \quad & F_0 = \frac{1}{2\ell}\sum_{i=1}^{\ell}(\xi_1^i+\xi_2^i)+\gamma_n(\alpha_1^\top K_1\alpha_1+\alpha_2^\top K_2\alpha_2)+ \\ & \gamma_v(K_1\alpha_1-K_2\alpha_2)^\top U(K_1\alpha_1-K_2\alpha_2) \\ \text{s.t.} \quad & \begin{cases} y_i(\sum_{j=1}^{\ell+u}\alpha_1^j k_1(x_j,x_i)+b_1)\geq 1-\xi_1^i, \\ y_i(\sum_{j=1}^{\ell+u}\alpha_2^j k_2(x_j,x_i)+b_2)\geq 1-\xi_2^i, \\ \xi_1^i,\ \xi_2^i\geq 0, \quad i=1,\ldots,\ell\,. \end{cases}\end{aligned} \tag{12}$$

3.1 Lagrange Dual Function

We will solve problem (12) through optimizing its dual problem which is simpler to solve. Now we derive its Lagrange dual function.

Suppose $\lambda_1^i,\lambda_2^i,\nu_1^i,\nu_2^i\geq 0$ $(i=1,\ldots,\ell)$ be the Lagrange multipliers associated with the inequality constraints. Define $\lambda_j=[\lambda_j^1,\ldots,\lambda_j^\ell]^\top$ and $\nu_j=[\nu_j^1,\ldots,\nu_j^\ell]^\top$ $(j=1,2)$. The Lagrangian $L(\alpha_1,\alpha_2,\xi_1,\xi_2,b_1,b_2,\lambda_1,\lambda_2,\nu_1,\nu_2)$ can be written as

$$\begin{aligned}L= \quad & F_0-\sum_{i=1}^{\ell}[\lambda_1^i(y_i(\sum_{j=1}^{\ell+u}\alpha_1^j k_1(x_j,x_i)+b_1)-1+\xi_1^i)+ \\ & \lambda_2^i(y_i(\sum_{j=1}^{\ell+u}\alpha_2^j k_2(x_j,x_i)+b_2)-1+\xi_2^i)+\nu_1^i\xi_1^i+\nu_2^i\xi_2^i].\end{aligned}$$

Note that

$$\begin{aligned}& (K_1\alpha_1-K_2\alpha_2)^\top U(K_1\alpha_1-K_2\alpha_2) \\ = \quad & \alpha_1^\top K_1UK_1\alpha_1-2\alpha_1^\top K_1UK_2\alpha_2+\alpha_2^\top K_2UK_2\alpha_2.\end{aligned}$$

To obtain the Lagrangian dual function, L has to be minimized with respect to the primal variables $\alpha_1,\alpha_2,\xi_1,\xi_2,b_1,b_2$. To eliminate these variables, we compute the corresponding partial derivatives and set them to 0, obtaining the following conditions

$$2J_1\alpha_1-2\gamma_v K_1UK_2\alpha_2=\Lambda_1, \tag{13}$$

$$2J_2\alpha_2-2\gamma_v K_2UK_1\alpha_1=\Lambda_2, \tag{14}$$

$$\lambda_1^i+\nu_1^i=\frac{1}{2\ell}, \tag{15}$$

$$\lambda_2^i+\nu_2^i=\frac{1}{2\ell}, \tag{16}$$

$$\sum_{i=1}^{\ell}\lambda_1^i y_i=0,$$

$$\sum_{i=1}^{\ell}\lambda_2^i y_i=0, \tag{17}$$

where

$$\begin{aligned} J_1 &:= \gamma_n K_1 + \gamma_v K_1 U K_1, \\ J_2 &:= \gamma_n K_2 + \gamma_v K_2 U K_2, \\ \Lambda_1 &:= \sum_{i=1}^{\ell} \lambda_1^i y_i K_1(:,i), \\ \Lambda_2 &:= \sum_{i=1}^{\ell} \lambda_2^i y_i K_2(:,i), \end{aligned}$$

with $K_1(:,i)$ and $K_2(:,i)$ being the ith column of the corresponding Gram matrices.

Substituting (13)$\sim$(17) into L results in the following expression of the Lagrangian dual function $g_L(\lambda_1,\lambda_2,\nu_1,\nu_2)$

$$\begin{aligned} g_L &= \gamma_n(\alpha_1^\top K_1\alpha_1 + \alpha_2^\top K_2\alpha_2) + \gamma_v(\alpha_1^\top K_1 U K_1\alpha_1 - 2\alpha_1^\top K_1 U K_2\alpha_2 + \\ &\quad \alpha_2^\top K_2 U K_2\alpha_2) - \alpha_1^\top\Lambda_1 - \alpha_2^\top\Lambda_2 + \sum_{i=1}^{\ell}(\lambda_1^i + \lambda_2^i) \\ &= \frac{1}{2}\alpha_1^\top\Lambda_1 + \frac{1}{2}\alpha_2^\top\Lambda_2 - \alpha_1^\top\Lambda_1 - \alpha_2^\top\Lambda_2 + \sum_{i=1}^{\ell}(\lambda_1^i + \lambda_2^i) \\ &= -\frac{1}{2}\alpha_1^\top\Lambda_1 - \frac{1}{2}\alpha_2^\top\Lambda_2 + \sum_{i=1}^{\ell}(\lambda_1^i + \lambda_2^i). \end{aligned} \tag{18}$$

We obtain the following from (13) and (14)

$$\alpha_1 = \frac{1}{2}J_1^{-1}(\Lambda_1 + 2\gamma_v K_1 U K_2\alpha_2) \tag{19}$$

$$\alpha_2 = \frac{1}{2}J_2^{-1}(\Lambda_2 + 2\gamma_v K_2 U K_1\alpha_1). \tag{20}$$

From (13) and (20), we have

$$(2J_1 - 2\gamma_v^2 K_1 U K_2 J_2^{-1} K_2 U K_1)\alpha_1 = \Lambda_1 + \gamma_v K_1 U K_2 J_2^{-1}\Lambda_2.$$

Define $M_1 = 2J_1 - 2\gamma_v^2 K_1 U K_2 J_2^{-1} K_2 U K_1$. Suppose the above linear system is well-posed (if ill-posed we can employ approximate numerical analysis techniques). We get

$$\alpha_1 = M_1^{-1}(\Lambda_1 + \gamma_v K_1 U K_2 J_2^{-1}\Lambda_2).$$

From (14) and (19), we have

$$(2J_2 - 2\gamma_v^2 K_2 U K_1 J_1^{-1} K_1 U K_2)\alpha_2 = \Lambda_2 + \gamma_v K_2 U K_1 J_1^{-1}\Lambda_1.$$

Define $M_2 = 2J_2 - 2\gamma_v^2 K_2 U K_1 J_1^{-1} K_1 U K_2$. Thus we get

$$\alpha_2 = M_2^{-1}(\Lambda_2 + \gamma_v K_2 U K_1 J_1^{-1}\Lambda_1).$$

Now with α_1 and α_2 substituted into (18), the Lagrange dual function $g_L(\lambda_1,\lambda_2,\nu_1,\nu_2)$ is

$$\begin{aligned} g_L &= \inf_{\alpha_1,\alpha_2,\xi_1,\xi_2,b_1,b_2} L = -\frac{1}{2}\alpha_1^\top \Lambda_1 - \frac{1}{2}\alpha_2^\top \Lambda_2 + \sum_{i=1}^{\ell}(\lambda_1^i+\lambda_2^i) \\ &= -\frac{1}{2}(\Lambda_1+\gamma_v K_1 U K_2 J_2^{-1}\Lambda_2)^\top M_1^{-1}\Lambda_1 - \frac{1}{2}(\Lambda_2 + \\ &\quad \gamma_v K_2 U K_1 J_1^{-1}\Lambda_1)^\top M_2^{-1}\Lambda_2 + \sum_{i=1}^{\ell}(\lambda_1^i+\lambda_2^i). \end{aligned}$$

3.2 Solving the Dual Problem

The Lagrange dual problem is given by

$$\begin{array}{ll} \max\limits_{\lambda_1,\lambda_2} & g_L \\ \text{s.t.} & \left\{ \begin{array}{ll} 0 \le \lambda_1^i \le \frac{1}{2\ell}, & i=1,\ldots,\ell \\ 0 \le \lambda_2^i \le \frac{1}{2\ell}, & i=1,\ldots,\ell \\ \sum_{i=1}^{\ell}\lambda_1^i y_i = 0, & \\ \sum_{i=1}^{\ell}\lambda_2^i y_i = 0. & \end{array} \right. \end{array}$$

As Lagrange dual functions are always concave (Boyd and Vandenberghe, 2004), we can formulate the above problem as a convex optimization problem

$$\begin{array}{ll} \min\limits_{\lambda_1,\lambda_2} & -g_L \\ \text{s.t.} & \left\{ \begin{array}{ll} 0 \le \lambda_1^i \le \frac{1}{2\ell}, & i=1,\ldots,\ell \\ 0 \le \lambda_2^i \le \frac{1}{2\ell}, & i=1,\ldots,\ell \\ \sum_{i=1}^{\ell}\lambda_1^i y_i = 0, & \\ \sum_{i=1}^{\ell}\lambda_2^i y_i = 0. & \end{array} \right. \end{array} \tag{21}$$

Define matrix $Y = diag(y_1,\ldots,y_\ell)$. Then, $\Lambda_1 = K_{\ell 1} Y \lambda_1$ and $\Lambda_2 = K_{\ell 2} Y \lambda_2$ with $K_{\ell 1} = K_1(:,1:\ell)$ and $K_{\ell 2} = K_2(:,1:\ell)$. We have

$$\begin{aligned} -g_L &= \frac{1}{2}(\Lambda_1+\gamma_v K_1 U K_2 J_2^{-1}\Lambda_2)^\top M_1^{-1}\Lambda_1 + \frac{1}{2}(\Lambda_2 + \\ &\quad \gamma_v K_2 U K_1 J_1^{-1}\Lambda_1)^\top M_2^{-1}\Lambda_2 - \sum_{i=1}^{\ell}(\lambda_1^i+\lambda_2^i) \\ &= \frac{1}{2}(\lambda_1^\top \ \lambda_2^\top)\begin{pmatrix} A & B \\ C & D \end{pmatrix}\begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} - \mathbf{1}^\top(\lambda_1+\lambda_2), \end{aligned}$$

where

$$\begin{aligned} A &:= Y K_{\ell 1}^\top M_1^{-1} K_{\ell 1} Y, \\ B &:= \gamma_v Y K_{\ell 1}^\top J_1^{-1} K_1 U K_2 M_2^{-1} K_{\ell 2} Y, \\ C &:= \gamma_v Y K_{\ell 2}^\top J_2^{-1} K_2 U K_1 M_1^{-1} K_{\ell 1} Y, \\ D &:= Y K_{\ell 2}^\top M_2^{-1} K_{\ell 2} Y, \end{aligned}$$

and $\mathbf{1} = (1,\ldots,1_{(\ell)})^\top$.

Substituting M_1 and M_2 into the expressions of B and C, we can prove that $B = C^\top$. In addition, because of the convexity of function $-g$, we affirm that matrix $\begin{pmatrix} A & B \\ C & D \end{pmatrix}$ is positive semi-definite.

Hence, the optimization problem in (21) can be rewritten as

$$\begin{array}{ll} \min\limits_{\lambda_1,\lambda_2} & \frac{1}{2}(\lambda_1^\top \lambda_2^\top)\begin{pmatrix} A & B \\ C & D \end{pmatrix}\begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} - \mathbf{1}^\top(\lambda_1+\lambda_2) \\ \text{s.t.} & \begin{cases} 0 \preceq \lambda_1 \preceq \frac{1}{2l}\mathbf{1}, \\ 0 \preceq \lambda_2 \preceq \frac{1}{2l}\mathbf{1}, \\ \lambda_1^\top \mathbf{y} = 0, \\ \lambda_2^\top \mathbf{y} = 0, \end{cases} \end{array}$$

where $\mathbf{y} = (y_1,\ldots,y_\ell)^\top$. After solving this problem using standard software, we then obtain ν_1^i and ν_2^i by (15) and (16).

We now state the advantages of optimizing this dual problem over optimizing the primal problem (12):

- Less optimization variables as for typical semi-supervised learning $\ell \ll u$, and
- Simpler constraint functions.

The solution of bias terms b_1 and b_2 can be obtained through support vectors. Due to KKT conditions, the following equalities hold

$$\begin{aligned} &\lambda_1^i(y_i(\sum_{j=1}^{\ell+u}\alpha_1^j k_1(x_j,x_i)+b_1)-1+\xi_1^i)=0, \\ &\lambda_2^i(y_i(\sum_{j=1}^{\ell+u}\alpha_2^j k_2(x_j,x_i)+b_2)-1+\xi_2^i)=0, \\ &\nu_1^i\xi_1^i=0, \\ &\nu_2^i\xi_2^i=0, \quad i=1,\ldots,\ell. \end{aligned}$$

For support vectors x_i, we have $\nu_j^i > 0$ (and thus $\xi_j^i = 0$) and $\lambda_j^i > 0$ $(j = 1,2)$. Therefore, we can resolve the bias terms by averaging $y_i(\sum_{j=1}^{\ell+u}\alpha_1^j k_1(x_j,x_i)+b_1)-1=0$ and $y_i(\sum_{j=1}^{\ell+u}\alpha_2^j k_2(x_j,x_i)+b_2)-1=0$ over all support vectors.

3.3 Advantages of Using Conjugate Functions

In this subsection, we show the direct optimization of problem (1) without the use of conjugate functions is of large scale and time-consuming, which justifies the advantages of using conjugate functions.

The primal problem can be rewritten as

$$\min_{\alpha_1,\alpha_2,\xi_1,\xi_2,b_1,b_2,\delta_i} D_0 = \frac{1}{2\ell}\sum_{i=1}^{\ell}(\xi_1^i+\xi_2^i)+\gamma_n(\alpha_1^\top K_1\alpha_1+\alpha_2^\top K_2\alpha_2)+\gamma_v\sum_{i=1}^{\ell+u}\delta_i^2$$

$$\text{s.t.}\begin{cases} y_i(\sum_{j=1}^{\ell+u}\alpha_1^j k_1(x_j,x_i)+b_1)\geq 1-\xi_1^i, \quad i=1,\ldots,\ell\,,\\ y_i(\sum_{j=1}^{\ell+u}\alpha_2^j k_2(x_j,x_i)+b_2)\geq 1-\xi_2^i, \quad i=1,\ldots,\ell\,,\\ \xi_1^i,\ \xi_2^i\geq 0, \quad i=1,\ldots,\ell\,,\\ (\sum_{j=1}^{\ell+u}\alpha_1^j k_1(x_j,x_i)+b_1)-(\sum_{j=1}^{l+u}\alpha_2^j k_2(x_j,x_i)+b_2)\geq -\delta_i-\varepsilon, \quad i=1,\ldots,\ell+u\,,\\ (\sum_{j=1}^{\ell+u}\alpha_1^j k_1(x_j,x_i)+b_1)-(\sum_{j=1}^{l+u}\alpha_2^j k_2(x_j,x_i)+b_2)\leq \delta_i+\varepsilon, \quad i=1,\ldots,\ell+u\,, \end{cases} \tag{22}$$

where $y_i\in\{1,-1\}$, $\gamma_n,\gamma_v\geq 0$.

We will solve problem (22) through optimizing its dual problem which can be simpler to solve. Suppose $\lambda_1^i,\lambda_2^i,\nu_1^i,\nu_2^i\geq 0$ $(i=1,\ldots,\ell)$ and μ_1^i,μ_2^i $(i=1,\ldots,\ell+u)$ are the Lagrange multipliers associated with the inequality constraints of problem (22). Define $\delta=[\delta_1,\ldots,\delta_{\ell+u}]^\top$, $\lambda_j=[\lambda_j^1,\ldots,\lambda_j^\ell]^\top$, $\nu_j=[\nu_j^1,\ldots,\nu_j^\ell]^\top$, and $\mu_j=[\mu_j^1,\ldots,\mu_j^{\ell+u}]^\top$ $(j=1,2)$. The Lagrangian $L(\alpha_1,\alpha_2,\xi_1,\xi_2,b_1,b_2,\delta,\lambda_1,\lambda_2,\nu_1,\nu_2,\mu_1,\mu_2)$ can be written as

$$\begin{aligned} L= \quad & D_0-\sum_{i=1}^{\ell}[\lambda_1^i(y_i(\sum_{j=1}^{\ell+u}\alpha_1^j k_1(x_j,x_i)+b_1)-1+\xi_1^i)+\\ & \lambda_2^i(y_i(\sum_{j=1}^{\ell+u}\alpha_2^j k_2(x_j,x_i)+b_2)-1+\xi_2^i)+\nu_1^i\xi_1^i+\nu_2^i\xi_2^i]-\\ & \sum_{i=1}^{\ell+u}\mu_1^i[\sum_{j=1}^{\ell+u}\alpha_1^j k_1(x_j,x_i)+b_1-\sum_{j=1}^{\ell+u}\alpha_2^j k_2(x_j,x_i)-b_2+\delta_i+\varepsilon]+\\ & \sum_{i=1}^{\ell+u}\mu_2^i[\sum_{j=1}^{\ell+u}\alpha_1^j k_1(x_j,x_i)+b_1-\sum_{j=1}^{\ell+u}\alpha_2^j k_2(x_j,x_i)-b_2-\delta_i-\varepsilon]. \end{aligned}$$

To obtain the Lagrangian dual function, L has to be minimized with respect to the primal variables $\alpha_1,\alpha_2,\xi_1,\xi_2,b_1,b_2,\delta$. To eliminate these variables, we compute the corresponding partial

derivatives and set them to 0, obtaining the following conditions

$$
\begin{aligned}
&2\gamma_n K_1\alpha_1 = \sum_{i=1}^{\ell}\lambda_1^i y_i K_1(:,i) + \sum_{i=1}^{\ell+u}(\mu_1^i - \mu_2^i)K_1(:,i),\\
&2\gamma_n K_2\alpha_2 = \sum_{i=1}^{\ell}\lambda_2^i y_i K_2(:,i) - \sum_{i=1}^{\ell+u}(\mu_1^i - \mu_2^i)K_2(:,i),\\
&\lambda_1^i + \nu_1^i = \frac{1}{2\ell}, \quad i = 1,\ldots,\ell\\
&\lambda_2^i + \nu_2^i = \frac{1}{2\ell}, \quad i = 1,\ldots,\ell\\
&-\sum_{i=1}^{\ell}\lambda_1^i y_i - \sum_{i=1}^{\ell+u}\mu_1^i + \sum_{i=1}^{\ell+u}\mu_2^i = 0,\\
&-\sum_{i=1}^{\ell}\lambda_2^i y_i + \sum_{i=1}^{\ell+u}\mu_1^i - \sum_{i=1}^{\ell+u}\mu_2^i = 0,\\
&2\gamma_v\delta_i - \mu_1^i - \mu_2^i = 0, \quad i = 1,\ldots,\ell+u.
\end{aligned}
$$

Substituting these equations into the Lagrangian as what was done in Section 3.1, it is clear that finally L is a quadratic function involving $\lambda_1, \lambda_2, \mu_1, \mu_2$. The dual optimization problem would be a quadratic optimization involving $2\ell + 2(\ell+u)$ parameters. Now we see this direct optimization is indeed of large-scale and time-consuming.

4. Generalization Error

In this section, we analyze the generalization performance of the sparse multi-view SVMs making use of Rademacher complexity theory and the margin bound.

4.1 Rademacher Complexity Theory

Important background on Rademacher complexity theory (Bartlett and Mendelson, 2002; Shawe-Taylor and Cristianini, 2004) is introduced below.

Definition 11 *For a sample $S = \{x_1,\ldots,x_\ell\}$ generated by a distribution $\mathcal{D}$ on a set $\mathcal{X}$ and a real-valued function class $\mathcal{F}$ with domain $\mathcal{X}$, the empirical Rademacher complexity of $\mathcal{F}$ is the random variable*

$$\hat{R}_\ell(\mathcal{F}) = \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\left|\frac{2}{\ell}\sum_{i=1}^{\ell}\sigma_i f(x_i)\right| \middle| x_1,\ldots,x_\ell\right],$$

where $\sigma = \{\sigma_1,\ldots,\sigma_\ell\}$ are independent uniform $\{\pm 1\}$-valued (Rademacher) random variables. The Rademacher complexity of $\mathcal{F}$ is

$$R_\ell(\mathcal{F}) = \mathbb{E}_S[\hat{R}_\ell(\mathcal{F})] = \mathbb{E}_{S\sigma}\left[\sup_{f\in\mathcal{F}}\left|\frac{2}{\ell}\sum_{i=1}^{\ell}\sigma_i f(x_i)\right|\right].$$

Lemma 12 *Fix $\delta \in (0,1)$ and let $\mathcal{F}$ be a class of functions mapping from an input space $\tilde{\mathcal{X}}$ ($\tilde{\mathcal{X}} = \mathcal{X}\times\mathcal{Y}$ or $\tilde{\mathcal{X}} = \mathcal{X}$) to $[0,1]$. Let $(\tilde{x}_i)_{i=1}^{\ell}$ be drawn independently according to a probability distribution*

$\mathcal{D}$. Then with probability at least $1-\delta$ over random draws of samples of size ℓ, every $f \in \mathcal{F}$ satisfies

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}[f(\tilde{x})] &\leq \hat{\mathbb{E}}[f(\tilde{x})] + R_\ell(\mathcal{F}) + \sqrt{\frac{\ln(2/\delta)}{2\ell}} \\ &\leq \hat{\mathbb{E}}[f(\tilde{x})] + \hat{R}_\ell(\mathcal{F}) + 3\sqrt{\frac{\ln(2/\delta)}{2\ell}},\end{aligned}$$

where $\hat{\mathbb{E}}[f(\tilde{x})]$ is the empirical error averaged on the ℓ examples.

4.2 Margin Bound for Sparse Multi-view SVMs

By (2), the prediction function of the sparse multi-view SVMs is derived from the average of predictions from two views. Define the soft prediction function as

$$g(x) = \frac{1}{2}(f_1(x) + f_2(x)).$$

We obtain the following margin bound regarding the generalization error of sparse multi-view SVMs. This bound is widely applicable to multi-view SVMs, for example, Szedmak and Shawe-Taylor (2007) independently provided a similar bound for the SVM-2K method.

Theorem 13 *Fix $\delta \in (0,1)$ and let $\mathcal{F}$ be the class of functions mapping from $\tilde{\mathcal{X}} = \mathcal{X} \times \mathcal{Y}$ to $\mathbf{R}$ given by $\tilde{f}(x,y) = -yg(x)$ where $g = \frac{1}{2}(f_1 + f_2) \in \mathcal{G}$ and $\tilde{f} \in \mathcal{F}$. Let $S = \{(x_1,y_1), \cdots, (x_\ell,y_\ell)\}$ be drawn independently according to a probability distribution $\mathcal{D}$. Then with probability at least $1-\delta$ over samples of size ℓ, every $g \in \mathcal{G}$ satisfies*

$$P_{\mathcal{D}}(y \neq sgn(g(\mathbf{x}))) \leq \quad \frac{1}{2\ell}\sum_{i=1}^{\ell}(\xi_1^i + \xi_2^i) + 2\hat{R}_\ell(\mathcal{G}) + 3\sqrt{\frac{\ln(2/\delta)}{2\ell}},$$

where $\xi_1^i := (1 - y_i f_1(x_i))_+$, $\xi_2^i := (1 - y_i f_2(x_i))_+$. Function $y_i f_1(x_i)$ and $y_i f_2(x_i)$ are called margins.

Proof Let $H(\cdot)$ be the Heaviside function that returns 1 if its argument is greater than 0 and zero otherwise. We have

$$P_{\mathcal{D}}(y \neq sgn(g(\mathbf{x}))) = \mathbb{E}_{\mathcal{D}}[H(-yg(\mathbf{x}))]. \tag{23}$$

Consider a loss function $\mathcal{A} : \mathbf{R} \to [0,1]$, given by

$$\mathcal{A}(a) = \begin{cases} 1, & \text{if } a \geq 0; \\ 1+a, & \text{if } -1 \leq a \leq 0; \\ 0, & \text{otherwise.} \end{cases}$$

By Lemma 12 and since function $\mathcal{A} - 1$ dominates $H - 1$, we have (Shawe-Taylor and Cristianini, 2004)

$$\begin{aligned}&\mathbb{E}_{\mathcal{D}}[H(\tilde{f}(x,y)) - 1] \leq \mathbb{E}_{\mathcal{D}}[\mathcal{A}(\tilde{f}(x,y)) - 1] \\ &\leq \hat{\mathbb{E}}[\mathcal{A}(\tilde{f}(x,y)) - 1] + \hat{R}_\ell((\mathcal{A} - 1) \circ \mathcal{F}) + 3\sqrt{\frac{\ln(2/\delta)}{2\ell}}.\end{aligned}$$

Therefore,

$$\mathbb{E}_{\mathcal{D}}[H(\tilde{f}(x,y))] \leq \quad \hat{\mathbb{E}}[\mathcal{A}(\tilde{f}(x,y))] + \hat{R}_\ell((\mathcal{A}-1)\circ\mathcal{F}) + 3\sqrt{\frac{\ln(2/\delta)}{2\ell}}.$$

In addition, we have

$$\begin{aligned}
\hat{E}[\mathcal{A}(\tilde{f}(x,y))] &\leq \frac{1}{\ell}\sum_{i=1}^{\ell}(1-y_i g(x_i))_+ \\
&= \frac{1}{2\ell}\sum_{i=1}^{\ell}(1-y_i f_1(x_i)+1-y_i f_2(x_i))_+ \\
&\leq \frac{1}{2\ell}\sum_{i=1}^{\ell}[(1-y_i f_1(x_i))_+ + (1-y_i f_2(x_i))_+] \\
&= \frac{1}{2\ell}\sum_{i=1}^{\ell}(\xi_1^i + \xi_2^i),
\end{aligned}$$

where ξ_1^i denotes the amount by which function f_1 fails to achieve margin 1 for (x_i, y_i) and ξ_2^i applies similarly to function f_2.

Since $(\mathcal{A}-1)(0) = 0$, we can apply the Lipschitz condition (Bartlett and Mendelson, 2002) of function $(\mathcal{A}-1)$ to get

$$\hat{R}_\ell((\mathcal{A}-1)\circ\mathcal{F}) \leq 2\hat{R}_\ell(\mathcal{F}).$$

It remains to bound the empirical Rademacher complexity of the class $\mathcal{F}$.

With $y_i \in \{1,-1\}$, we have

$$\begin{aligned}
\hat{R}_\ell(\mathcal{F}) &= \mathbb{E}_\sigma[\sup_{f\in\mathcal{F}}|\frac{2}{\ell}\sum_{i=1}^{\ell}\sigma_i\tilde{f}(x_i,y_i)|] \\
&= \mathbb{E}_\sigma[\sup_{g\in\mathcal{G}}|\frac{2}{\ell}\sum_{i=1}^{\ell}\sigma_i y_i g(x_i)|] \\
&= \mathbb{E}_\sigma[\sup_{g\in\mathcal{G}}|\frac{2}{\ell}\sum_{i=1}^{\ell}\sigma_i g(x_i)|] \\
&= \hat{R}_\ell(\mathcal{G}). \qquad (24)
\end{aligned}$$

Finally, combining (23)∼(24) completes the proof. ■

5. Empirical Rademacher Complexity

Our optimization algorithm iteratively updates z to solve θ. In this section, we first derive the empirical Rademacher complexity of $\hat{R}_\ell(\mathcal{G})$ for the function class induced after one iteration with an initially fixed z, and then give its formulation applicable for any number of subsequent iterations including the termination case. This Rademacher complexity is crucial for Theorem 13 when analyzing the performance of the corresponding classifiers obtained by our iterative optimization algorithm. Specifically, for the empirical Rademacher complexity we give the following theorem

Theorem 14 *Suppose* $\mathcal{S} = \frac{1}{\gamma_n}(K_{1\ell}K_1^{-1}K_{1\ell}^\top + K_{2\ell}K_2^{-1}K_{2\ell}^\top)$, $\Theta = \frac{1}{\gamma_n}U_u^{1/2}(K_{1u}K_1^{-1}K_{1u}^\top + K_{2u}K_2^{-1}K_{2u}^\top)U_u^{1/2}$, $\mathcal{J} = \frac{1}{\gamma_n}U_u^{1/2}(K_{1u}K_1^{-1}K_{1\ell}^\top - K_{2u}K_2^{-1}K_{2\ell}^\top)$, *where* $K_{1\ell}$ *and* $K_{2\ell}$ *are respectively the first* ℓ *rows of the Gram matrices* K_1 *and* K_2, K_{1u} *and* K_{2u} *are respectively the last* u *rows of matrix* K_1 *and* K_2, *and* U_u *is the diagonal matrix including the last* u *diagonal elements (initially fixed* $z_{\ell+1},\ldots,z_{\ell+u}$*) of* U. *Then the empirical Rademacher complexity* $\hat{R}_\ell(\mathcal{G})$ *is bounded as* $\frac{\mathcal{U}}{\sqrt{2\ell}} \le \hat{R}_\ell(\mathcal{G}) \le \frac{\mathcal{U}}{\ell}$, *where* $\mathcal{U}^2 = tr(\mathcal{S}) - \gamma_v tr(\mathcal{J}^\top(I+\gamma_v\Theta)^{-1}\mathcal{J})$ *for the first iteration of sparse multi-view SVMs, and* $\mathcal{U}^2 = tr(\mathcal{S})$ *for subsequent iterations.*

The remainder of this section before Section 5.4 completes the proof of this theorem, which was partially inspired by Rosenberg and Bartlett (2007) for analyzing co-regularized least squares.

We use problem (8) to reason about $\hat{R}_\ell(\mathcal{G})$. As a result of fixed z, we can remove $f_\varepsilon^*(z_i)$ without loss of generality to resolve f_1 and f_2. It is true that the loss function $\hat{L} : \mathcal{H}^1 \times \mathcal{H}^2 \to [0,\infty)$ with $\hat{L} := \frac{1}{2\ell}\sum_{i=1}^{\ell}[(1-y_i f_1(x_i))_+ + (1-y_i f_2(x_i))_+]$ satisfies

$$\hat{L}(0,0) = 1.$$

Let $Q(f_1,f_2)$ denote the objective function in (8) with $f_\varepsilon^*(z_i)$ removed. Substituting in the trivial predictors $f_1 \equiv 0$ and $f_2 \equiv 0$ gives the following upper bound

$$\min_{f_1,f_2\in\mathcal{H}^1\times\mathcal{H}^2} Q(f_1,f_2) \le Q(0,0) = \hat{L}(0,0) = 1.$$

Since all terms of $Q(f_1,f_2)$ are nonnegative, we conclude that any (f_1^*, f_2^*) minimizing $Q(f_1,f_2)$ is contained in

$$\hat{\mathcal{H}} = \{(f_1,f_2) : \gamma_n(\|f_1\|^2 + \|f_2\|^2) + \gamma_v \sum_{i=\ell+1}^{\ell+u} z_i[f_1(x_i) - f_2(x_i)]^2 \le 1\}. \tag{25}$$

Therefore, the final predictor is chosen from the function class

$$\mathcal{G} = \{x \to \frac{1}{2}[f_1(x) + f_2(x)] : (f_1,f_2) \in \hat{\mathcal{H}}\}.$$

The complexity $\hat{R}_\ell(\mathcal{G})$ is

$$\hat{R}_\ell(\mathcal{G}) = \mathbb{E}_\sigma\left[\sup_{(f_1,f_2)\in\hat{\mathcal{H}}}\left|\frac{1}{\ell}\sum_{i=1}^{\ell}\sigma_i(f_1(x_i)+f_2(x_i))\right|\right]. \tag{26}$$

As it only depends on the values of function $f_1(\cdot)$ and $f_2(\cdot)$ on the ℓ labeled examples, by the reproducing kernel property which says the projection of function f onto a closed subspace containing $k(x,\cdot)$ has the same value at x as f itself does (Rosenberg and Bartlett, 2007) we can restrict the function class $\hat{\mathcal{H}}$ to the span of labeled and unlabeled data and thus write it as

$$\begin{aligned}\hat{\mathcal{H}} &= \{(f_1,f_2) : \gamma_n(\alpha_1^\top K_1\alpha_1 + \alpha_2^\top K_2\alpha_2) + \\ &\quad \gamma_v(K_{1u}\alpha_1 - K_{2u}\alpha_2)^\top U_u(K_{1u}\alpha_1 - K_{2u}\alpha_2) \le 1\} \\ &= \{(f_1,f_2) : (\alpha_1^\top\ \alpha_2^\top)N\begin{pmatrix}\alpha_1\\ \alpha_2\end{pmatrix} \le 1\},\end{aligned}$$

where K_{1u} and K_{2u} are respectively the last u rows of matrix K_1 and K_2, U_u is the diagonal matrix including the last u diagonal elements of U, and

$$N := \gamma_n\begin{pmatrix}K_1 & \mathbf{0}\\ \mathbf{0} & K_2\end{pmatrix} + \gamma_v\begin{pmatrix}K_{1u}^\top\\ -K_{2u}^\top\end{pmatrix}U_u(K_{1u}\ \ -K_{2u}). \tag{27}$$

5.1 Evaluating the Supremum in Euclidean Space

Since $(f_1, f_2) \in \hat{\mathcal{H}}$ implies $(-f_1, -f_2) \in \hat{\mathcal{H}}$, we can drop the absolute sign in (26). Now we can write

$$\begin{aligned}\hat{R}_\ell(\mathcal{G}) &= \frac{1}{\ell}\mathbb{E}_\sigma \sup_{\alpha_1,\alpha_2\in\mathcal{R}^{\ell+u}} \{\sigma^\top K_{1\ell}\alpha_1 + \sigma^\top K_{2\ell}\alpha_2 : (\alpha_1^\top\ \alpha_2^\top)N\begin{pmatrix}\alpha_1\\ \alpha_2\end{pmatrix} \le 1\} \\ &= \frac{1}{\ell}\mathbb{E}_\sigma \sup_{\alpha_1,\alpha_2\in\mathcal{R}^{\ell+u}} \{\sigma^\top (K_{1\ell}\ K_{2\ell})\begin{pmatrix}\alpha_1\\ \alpha_2\end{pmatrix} : (\alpha_1^\top\ \alpha_2^\top)N\begin{pmatrix}\alpha_1\\ \alpha_2\end{pmatrix} \le 1\}, \qquad (28)\end{aligned}$$

where $K_{1\ell}, K_{2\ell}$ represent the first ℓ rows of the Gram matrices K_1 and K_2, respectively.

For a symmetric positive definite matrix M, it is simple to show that (Rosenberg and Bartlett, 2007)

$$\sup_{\alpha:\alpha^\top M\alpha\le 1} \mathbf{v}^\top\alpha = \|M^{-1/2}\mathbf{v}\|.$$

Without loss of generality, suppose positive semi-definite matrix N in (28) is positive definite and thus has full rank. If N does not have full rank, we can use subspace decomposition to rewrite $\hat{R}_\ell(\mathcal{G})$ to obtain a similar representation. Thus, we can evaluate the supremum as described above to get

$$\hat{R}_\ell(\mathcal{G}) = \frac{1}{\ell}\mathbb{E}_\sigma\|N^{-1/2}\begin{pmatrix}K_{1\ell}^\top\\ K_{2\ell}^\top\end{pmatrix}\sigma\|.$$

5.2 Bounding $\hat{R}_\ell(\mathcal{G})$ above and below

We make use of the Kahane-Khintchine inequality (Latala and Oleszkiewicz, 1994), stated here for convenience, to bound $\hat{R}_\ell(\mathcal{G})$.

Lemma 15 *For any vectors $\boldsymbol{a}_1, \cdots, \boldsymbol{a}_n$ in a Hilbert space and independent Rademacher random variables $\sigma_1, \cdots, \sigma_n$, we have*

$$\frac{1}{2}\mathbb{E}\|\sum_{i=1}^n \sigma_i\boldsymbol{a}_i\|^2 \le (\mathbb{E}\|\sum_{i=1}^n \sigma_i\boldsymbol{a}_i\|)^2 \le \mathbb{E}\|\sum_{i=1}^n \sigma_i\boldsymbol{a}_i\|^2.$$

By Lemma 15 we have

$$\frac{\mathcal{U}}{\sqrt{2}\ell} \le \hat{R}_\ell(\mathcal{G}) \le \frac{\mathcal{U}}{\ell}, \qquad (29)$$

where

$$\begin{aligned}\mathcal{U}^2 &= \mathbb{E}_\sigma\|N^{-1/2}\begin{pmatrix}K_{1\ell}^\top\\ K_{2\ell}^\top\end{pmatrix}\sigma\|^2 \\ &= \mathbb{E}_\sigma tr[(K_{1\ell}\ K_{2\ell})N^{-1}\begin{pmatrix}K_{1\ell}^\top\\ K_{2\ell}^\top\end{pmatrix}\sigma\sigma^\top] \\ &= tr[(K_{1\ell}\ K_{2\ell})N^{-1}\begin{pmatrix}K_{1\ell}^\top\\ K_{2\ell}^\top\end{pmatrix}].\end{aligned}$$

Recall that

$$N = \gamma_n\begin{pmatrix}K_1 & \mathbf{0}\\ \mathbf{0} & K_2\end{pmatrix} + \gamma_v\begin{pmatrix}K_{1u}^\top\\ -K_{2u}^\top\end{pmatrix}U_u(K_{1u}\ \ -K_{2u}).$$

Define

$$\Sigma = \gamma_n \begin{pmatrix} K_1 & \mathbf{0} \\ \mathbf{0} & K_2 \end{pmatrix}, \quad R = \begin{pmatrix} K_{1u}^\top \\ -K_{2u}^\top \end{pmatrix} U_u^{1/2}.$$

Using the Sherman-Morrison-Woodbury formula (Golub and Loan, 1996), we expand N^{-1} as

$$N^{-1} = \Sigma^{-1} - \gamma_v \Sigma^{-1} R (I + \gamma_v R^\top \Sigma^{-1} R)^{-1} R^\top \Sigma^{-1}.$$

Define $\Omega = (K_{1\ell}\, K_{2\ell})$. We get

$$\mathcal{U}^2 = tr(\Omega\Sigma^{-1}\Omega^\top) - \gamma_v tr[\Omega\Sigma^{-1} R (I + \gamma_v R^\top \Sigma^{-1} R)^{-1} R^\top \Sigma^{-1} \Omega^\top].$$

Define

$$\begin{aligned} \mathcal{S} &= \Omega\Sigma^{-1}\Omega^\top = \frac{1}{\gamma_n}(K_{1\ell}K_1^{-1}K_{1\ell}^\top + K_{2\ell}K_2^{-1}K_{2\ell}^\top), \\ \Theta &= R^\top\Sigma^{-1}R = \frac{1}{\gamma_n}U_u^{1/2}(K_{1u}K_1^{-1}K_{1u}^\top + K_{2u}K_2^{-1}K_{2u}^\top)U_u^{1/2}, \\ \mathcal{J} &= R^\top\Sigma^{-1}\Omega^\top = \frac{1}{\gamma_n}U_u^{1/2}(K_{1u}K_1^{-1}K_{1\ell}^\top - K_{2u}K_2^{-1}K_{2\ell}^\top). \end{aligned} \tag{30}$$

Putting expressions together, we get

$$\mathcal{U}^2 = tr(\mathcal{S}) - \gamma_v tr(\mathcal{J}^\top (I + \gamma_v \Theta)^{-1} \mathcal{J}). \tag{31}$$

5.2.1 REGULARIZATION TERM ANALYSIS

From (29) and (31), it is clear to see the roles the regularization parameters γ_n and γ_v play in the empirical Rademacher complexity $\hat{R}_l(\mathcal{G})$.

The amount of reduction in the Rademacher complexity brought by γ_v is

$$\Delta(\gamma_v) = \gamma_v tr(\mathcal{J}^\top (I + \gamma_v \Theta)^{-1} \mathcal{J}).$$

This term has the property shown by the following lemma given by Rosenberg and Bartlett (2007) when analyzing co-regularized least squares. Here the meanings of $\mathcal{J}$ and Θ are different from Rosenberg and Bartlett (2007).

Lemma 16 (Rosenberg and Bartlett, 2007) *$\Delta(0) = 0$, $\Delta(\gamma_v)$ is nondecreasing on $\gamma_v \geq 0$, and given that Θ is positive definite, we have*

$$\lim_{\gamma_v \to \infty} \Delta(\gamma_v) = tr(\mathcal{J}^\top \Theta^{-1} \mathcal{J}).$$

5.3 Extending to Iterative Optimization

As our sparse multi-view SVMs employ an iterative optimization procedure for sparsity pursuit, the former outcome for empirical Rademacher complexity would not apply if we use more than one iteration to update z. However, we can extend the former analysis to suit this case.

Recall that $z_i \in [0,1)$ $(i = \ell+1,\ldots,\ell+u)$ and $U_u = diag(z_{\ell+1},\ldots,z_{\ell+u})$. During iterations, it is possible that U_u becomes a zero matrix or other arbitrary matrix with diagonal elements in the range $[0,1)$. In any case, the resultant function class can be covered by

$$\hat{\mathcal{H}} \quad = \quad \{(f_1,f_2) : \gamma_n(\|f_1\|^2 + \|f_2\|^2) \leq 1\},$$

which is obtained by omitting the term containing z_i in (25). Following a similar derivation, the matrix N in (27) would be

$$N = \gamma_n \begin{pmatrix} K_1 & \mathbf{0} \\ \mathbf{0} & K_2 \end{pmatrix}.$$

Finally, we can obtain a bound on the empirical Rademacher complexity $\hat{R}_l(\mathcal{G})$ identical to (29) but now $\mathcal{U}^2 = tr(\mathcal{S})$ with $\mathcal{S}$ defined in (30). The proof of Theorem 14 is completed.

5.4 Examining $\hat{R}_l(\mathcal{G})$

Here, we examine the role $\hat{R}_l(\mathcal{G})$ plays in the margin bound. Since $K_{1\ell}$ and $K_{2\ell}$ are the first ℓ rows of K_1 and K_2, the formulation of $tr(\mathcal{S})$ can be simplified as

$$tr(\mathcal{S}) \quad = \quad \frac{1}{\gamma_n} tr(K_{1\ell}K_1^{-1}K_{1\ell}^\top + K_{2\ell}K_2^{-1}K_{2\ell}^\top) = \frac{1}{\gamma_n}\sum_{i=1}^{\ell}(K_1(i,i) + K_2(i,i)). \tag{32}$$

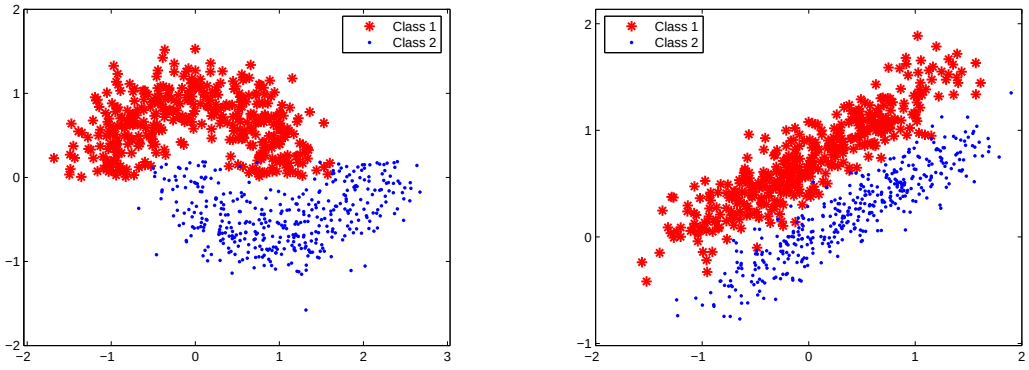
Now, we see that for iterative optimization of sparse multi-view SVMs, the empirical Rademacher complexity $\hat{R}_l(\mathcal{G})$ with $\mathcal{U}^2 = tr(\mathcal{S})$ only depends on the ℓ labeled examples and the chosen kernel functions. Consequently, the margin bound does not rely on the unlabeled training sets. In this case the margin bound is quite straightforward to reason.

If we do not use iterative optimization, the empirical Rademacher complexity $\hat{R}_l(\mathcal{G})$ will involve other two terms Θ and $\mathcal{J}$. By a similar technique as in (32), we can show that Θ only depends on the unlabeled data and the kernel functions, while $\mathcal{J}$ encodes the interaction between labeled and unlabeled data. As a result, the margin bound relies on both labeled and unlabeled data. For this case, we will give an evaluation of the margin bound with different sizes of unlabeled sets in Section 6.4.

6. Experiments

We performed experiments on artificial data and real-world data to evaluate the proposed sparse multi-view SVMs (SpMvSVMs). For SpMvSVMs with $\varepsilon > 0$, the entries of $\mathbf{z}$ were fixed as 1 for labeled data and initialized as 0.995 for unlabeled data. The termination condition for iterative optimization is either no unlabeled examples can be removed or the maximum iteration number surpasses 50. Comparisons are made with supervised SVMs, and the unsupervised SVM-2K method. Each accuracy/error reported in this paper is an averaged accuracy/error value over ten random splits of data into labeled, unlabeled and test data.

Later in this section, we also provide a sequential training strategy for SpMvSVMs, which shows an accuracy improvement over the gradual adding of unlabeled data while with roughly linear and sub-linear increases of running time. This indicates the possibility and potential of applying SpMvSVMs to large-scale data sets. At the end, margin bound evaluation results are reported.



(a) View 1 (b) View 2

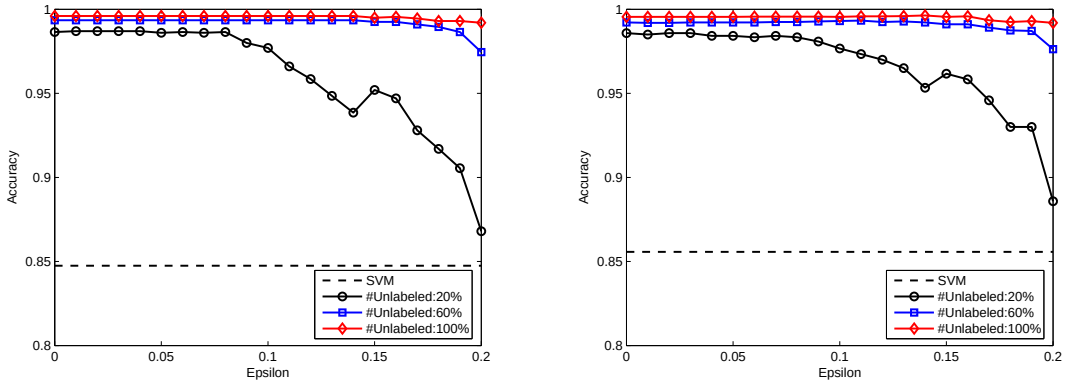
Figure 1: Examples in the two-moons-two-lines data set.

6.1 Artificial Data

This two-moons-two-lines synthetic data set was generated according to Sindhwani et al. (2005). Examples in two classes scatter like two moons in one view and two parallel lines in the other. To link the two views, points on one moon were enforced to associate at random with points on one line. Each class has 400 examples and a total of 800 examples were generated as shown in Figure 1. For SpMvSVMs, the numbers of examples in the labeled training set, unlabeled training set and test set were fixed as four, 596, and 200, respectively. Gaussian kernel with bandwidth 0.35 and the linear kernel were used for view 1 and view 2, respectively. The parameters γ_n and γ_v were selected from a small grid $\{10^{-6}, 10^{-4}, 10^{-2}, 1, 10, 100\}$ by five-fold cross validation on the whole data set. The chosen values are $\gamma_n = 10^{-4}$ and $\gamma_v = 1$. In this paper, γ_v is normalized by the number of labeled and unlabeled examples involved in the multi-view regularization term. For supervised SVMs, which concatenated features from the two views, we also found the regularization coefficient from this grid by five-fold cross validation.

To evaluate SpMvSVMs, we varied the size of the unlabeled training set from 20%, 60% to 100% of the total number of unlabeled data, and used different values for the insensitive parameter ε, which ranged from 0 to 0.2 with an interval 0.01 (when ε is zero, sparsity is not considered). The test accuracies and transductive accuracies (on the corresponding unlabeled set) are given in Figure 2(a) and Figure 2(b), respectively. It should be noted that the numbers of data used to calculate transductive accuracies are different for the three curves in Figure 2(b). The numbers of removed unlabeled examples for different ε values are shown in Figure 3.

From Figure 2 and Figure 3, we find that with the increase of ε, more and more unlabeled data are removed, and the remove of a small number of unlabeled data can hardly decrease the performance of the resultant classifiers, especially when the original size of unlabeled set is large. Therefore, we can find a good balance between sparsity and accuracy using an appropriate ε. In addition, more unlabeled data can benefit the performance of the learned classifiers with the same ε.



(a) Test accuracy (b) Transductive accuracy

Figure 2: Classification accuracies of SpMvSVMs with different sizes of unlabeled set and ε values on the artificial data. The accuracies of SVMs are also shown.

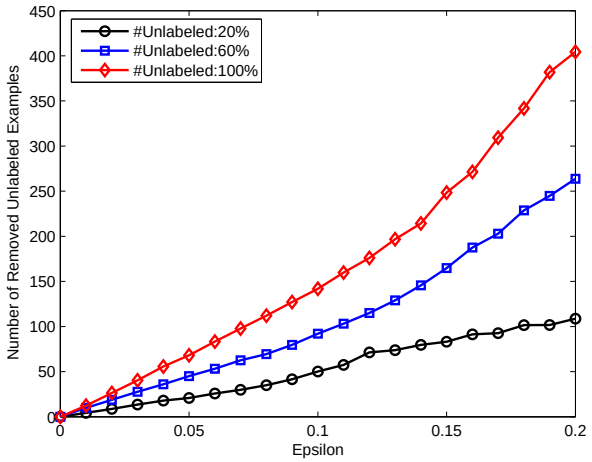
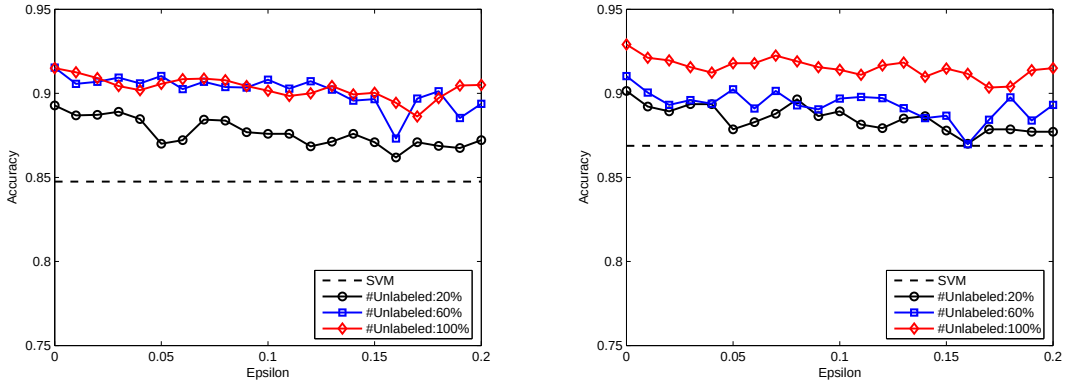


Figure 3: The numbers of unlabeled examples removed by SpMvSVMs for different ε values on the artificial data.

6.2 Text Classification

We applied the SpMvSVMs to the WebKB text classification task studied in Blum and Mitchell (1998); Sindhwani et al. (2005); Sun (2008). The data set consists of 1051 two-view web pages collected from the computer science department web sites at four U.S. universities: Cornell, University of Washington, University of Wisconsin, and University of Texas. The task is to predict whether a web page is a course home page or not. This problem has an unbalanced class distribution since there are a total of 230 course home pages (positive examples). The first view of the data is the words appearing on the web page itself, whereas the second view is the underlined words



(a) Test accuracy

(b) Transductive accuracy

Figure 4: Classification accuracies of SpMvSVMs with different sizes of unlabeled set and ε values on text classification. The accuracies of SVMs are also shown.

in all links pointing to the web page from other pages. We preprocessed each view by removing stop words, punctuation and numbers and then applied Porter's stemming to the text (Porter, 1980). In addition, words that occur in five or fewer documents were ignored. This resulted in 2332 and 87-dimensional vectors in the first and second view, respectively. Finally, document vectors were normalized to $tf.idf$ (the product of term frequency and inverse document frequency) features (Salton and Buckley, 1988).

For SpMvSVMs, the numbers of examples in the labeled training set, unlabeled training set and test set were fixed as 32, 699, and 320, respectively. In the training set and test set, the numbers of negative examples are three times of those of positive examples to reflect the overall proportion of positive and negative examples. The linear kernel was used for both views. The parameters γ_n and γ_v for SpMvSVMs and the regularziation coefficient for SVMs were selected using the same method as in Section 6.1. The chosen values for SpMvSVMs are $\gamma_n = 10^{-6}$ and $\gamma_v = 0.01$.

To evaluate SpMvSVMs, we also varied the size of the unlabeled training set from 20%, 60% to 100% of the total number of unlabeled data, and used different values for the insensitive parameter ε ranging from 0 to 0.2 with an interval 0.01. The test accuracies and transductive accuracies are given in Figure 4(a) and Figure 4(b), respectively. The numbers of removed unlabeled examples for different ε values are shown in Figure 5.

From Figure 5, we find that with the increase of ε, more and more unlabeled data can be removed. Reflected by Figure 4, the remove of unlabeled data only slightly decrease the performance of the resultant classifiers, and this decrease is less when more unlabeled data is used. We draw a same conclusion as before: an appropriate ε can be adopted to keep a good balance between sparsity and accuracy. We also observe a different phenomenon, that is, using 60% and 100% unlabeled data result in similar test accuracies as shown in Figure 4(a). This is reasonable because the performance improvement of any classifier is always bounded no matter how many data are used.

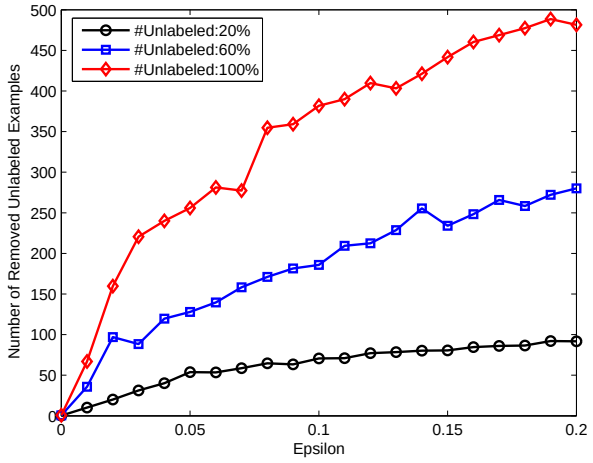


Figure 5: The numbers of unlabeled examples removed by SpMvSVMs for different ε values on text classification.

6.3 Comparison with SVM-2K, and Sequential Training

The SVM-2K method proposed by Szedmak and Shawe-Taylor (2007) can exploit unlabeled data for multi-view learning. Similar to SpMvSVMs, it also combines the maximum margin and multi-view regularization principles. However, it adopts l_1 norm for multi-view regularization, and this regularization only uses unlabeled data. Specifically, the SVM-2K method has the following optimization for classifier parameters $\mathbf{w}_1$, $\mathbf{w}_2$, b_1, and b_2 in two views

$$\begin{aligned}
\min \quad & \frac{1}{2}\|\mathbf{w}_1\|^2+\frac{1}{2}\|\mathbf{w}_2\|^2+C_1\sum_{i=1}^{\ell}\xi_1^i+C_2\sum_{i=1}^{\ell}\xi_2^i+C_\eta\sum_{j=\ell+1}^{\ell+u}\eta_j \\
\text{s.t.} \quad & \begin{cases} |\mathbf{w}_1^\top\phi_1(x_j)+b_1-\mathbf{w}_2^\top\phi_2(x_j)-b_2|\le\eta_j+\varepsilon \\ y_i(\mathbf{w}_1^\top\phi_1(x_i)+b_1)\ge 1-\xi_1^i \\ y_i(\mathbf{w}_2^\top\phi_2(x_i)+b_2)\ge 1-\xi_2^i \\ \xi_1^i\ge 0,\xi_2^i\ge 0,\eta_j\ge 0 \text{ for all } i=1,\ldots,\ell, \text{and } j=\ell+1,\ldots,\ell+u\,, \end{cases}
\end{aligned}$$

where an ε-insensitive parameter is used to relax the prediction consistency between views. In this subsection, we carry out an empirical comparison between SVM-2K and our SpMvSVMs for semi-supervised learning with identical data splits.

Our first comparison takes the ε-insensitive parameter in both SpMvSVMs and SVM-2K as zero and uses the above two data sets with different sizes of unlabeled training sets, namely, from 20% to 60% to 100%. For SVM-2K, we adopted the same parameter selection approach as in Szedmak and Shawe-Taylor (2007) through five-fold cross-validation. That is, the values of C_1 and C_2 were fixed to 1 and C_η were selected from the range $\{0.01\times 2^i\}$ $(i=1,\ldots,10)$. The experimental results are listed in Table 1, from which we see that both the test accuracies and transductive accuracies of SpMvSVMs are superior to those counterparts of SVM-2K.

The second comparison considers sequential training of SpMvSVMs and SVM-2K. The purpose is to show the relationship between running time, classification accuracies and the number of

# Unlabeled	SVM-2K		SpMvSVMs	
	Test Acc.	Transductive Acc.	Test Acc.	Transductive Acc.
20%	95.70	97.75	98.65	98.58
60%	98.50	99.19	99.35	99.22
100%	98.35	99.18	99.60	99.55
20%	84.72	84.71	89.28	90.14
60%	85.88	84.79	91.53	91.02
100%	85.84	87.85	91.50	92.90

Table 1: Test and transductive accuracies (%) of SVM-2K and SpMvSVMs with different sizes of unlabeled training sets on the artificial data (the first three lines) and text classification data (the last three lines).

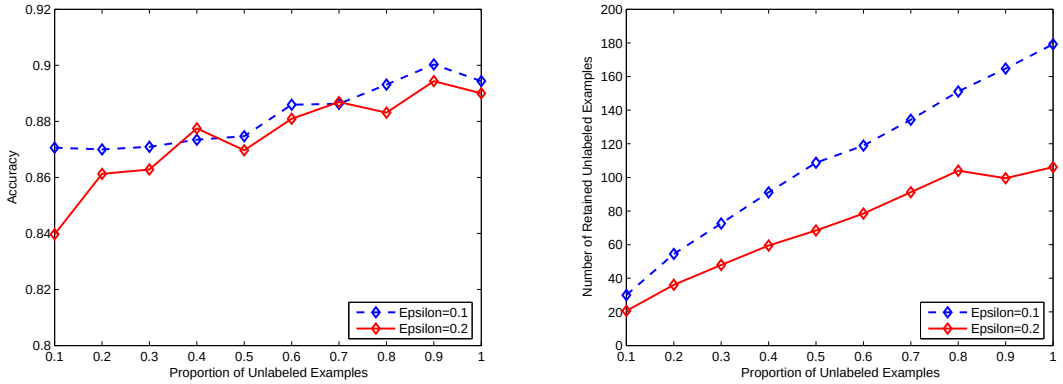
gradually added unlabeled examples, and thus evaluate the possibility and potential of applying the methods to large-scale problems. The text classification data were used where all the unlabeled training data were divided into ten equal sizes. For sequential training of SpMvSVMs, we adopted two different ε values 0.1 and 0.2. The procedure is as follows. First, we train SpMvSVMs using the labeled data and the first portion of unlabeled data. Then, we combine the retained unlabeled data from the last training with the next portion of unlabeled data together to train SpMvSVMs (with the original labeled data). We repeat this progress for ten times to complete the whole procedure.

The test accuracies and total numbers of retained unlabeled data after each step are shown in Figure 6(a) and Figure 6(b). The averaged classifier training time is given in Figure 7. Figure 6(a) indicates the effectiveness of sequential training, which is reflected by the fact that the test accuracies have an overall increasing tendency. Figure 6(b) shows that SpMvSVMs obtain sparse solutions in the sense that the number of retained unlabeled data is small compared to the number of all the added unlabeled data. Figure 7 shows that when $\varepsilon = 0.1$ the running time is roughly linear with respect to the gradual adding of unlabeled data, and when $\varepsilon = 0.2$ the relationship is roughly sub-linear. In fact, the running time can be further reduced if larger ε values rather than the given values are properly used. In practice, we can also vary the value of ε during the sequential training process.

Though the SVM-2K method was not initially proposed for sparse semi-supervised learning, we find that with $\varepsilon > 0$ it can reduce the number of unlabeled data used for representing classifiers. For this reason, we attempted to explore its possibility on sequential training using this data set under the same setting with the sequential training of SpMvSVMs. However, for different ε values we did not observe an improvement of test accuracies with the gradual adding of unlabeled data. Actually, for this data set when $\varepsilon > 0.05$ SVM-2K would not use any unlabeled data at all. This indicates that the roles of ε for SpMvSVMs and SVM-2K are quantitatively different.

6.4 Margin Bound Evaluation

To evaluate the margin bound in Theorem 13 for SpMvSVMs, we carried out experiments on the text classification data with a priori fixed regularization values $\gamma_n = 10^{-5}$ and $\gamma_v = 0.1$. This choice of parameters did not intend to be optimal in terms of test errors, but attempted to show the relationship, if any, between the generalization bound and the test error. The empirical Rademacher complexity



(a) Test accuracy (b) # Retained Unlabeled Examples

Figure 6: Classification accuracies and numbers of retained unlabeled examples for sequential training of SpMvSVMs. Parameter ε varies in $\{0.1, 0.2\}$.

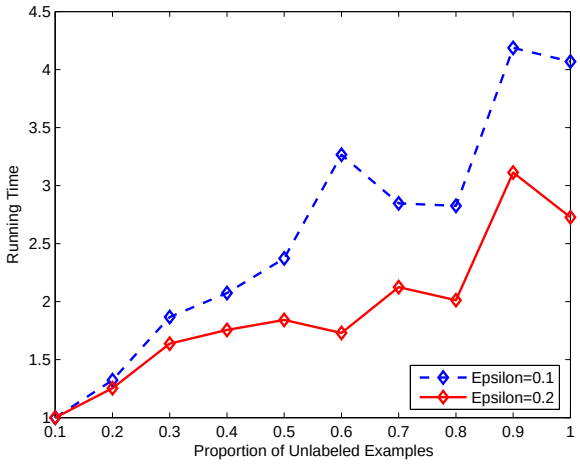
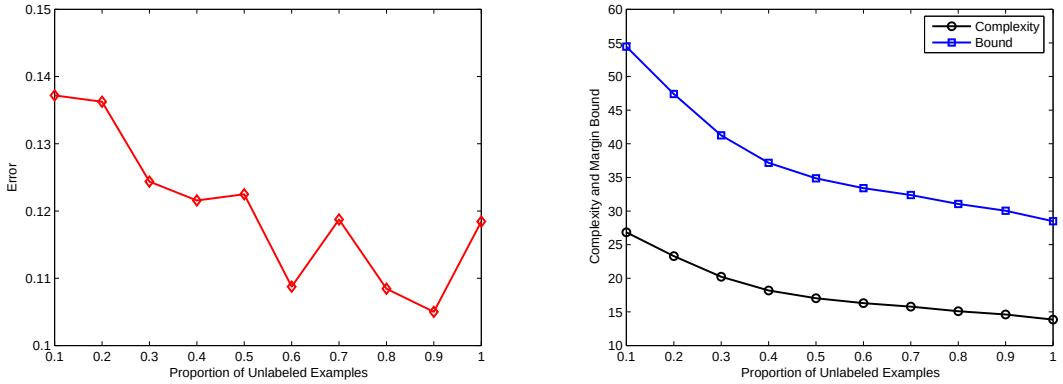


Figure 7: Running time for sequential training of SpMvSVMs. Parameter ε varies in $\{0.1, 0.2\}$. We have normalized the running times with 10% unlabeled examples to be 1.

$\hat{R}_\ell(\mathcal{G})$ in the margin bound is replaced by its upper bound $\mathcal{U}/\ell$ with $\mathcal{U}^2 = tr(\mathcal{S}) - \gamma_v tr(\mathcal{J}^\top (I + \gamma_v \Theta)^{-1} \mathcal{J})$.

For SpMvSVMs, only one iteration was performed in order to apply the margin bound. In other words, we learned classifiers only with the initially provided conjugate vector $\mathbf{z}$ whose entries were fixed as 1 for labeled data and 0.995 for unlabeled data. We used the same data split as in Section 6.2, but varied the size of unlabeled training set from $\{10\%, 20\%, \ldots, 100\%\}$ of all the available unlabeled data. To compute the margin bound, the confidence level in Theorem 13 is fixed as 95% ($\delta = 0.05$). The test error rate, empirical Rademacher complexity, and margin bound are



(a) Test error rate

(b) Complexity and margin bound

Figure 8: Classification error rates, empirical Rademacher complexity, and the margin bound of SpMvSVMs with different sizes of unlabeled sets.

shown in Figure 8. The overall decrease of error rates is well explained by the drop of the margin bound and empirical Rademacher complexity brought by the regularization role of more and more unlabeled data. Figure 8(b) also indicates that after adding a certain number of unlabeled data, including more unlabeled data will only improve the performance marginally. This phenomenon is observed in Figure 8(a) as well.

7. Extensions

In this section, we discuss possible extensions of using conjugate functions for sparse semi-supervised learning. In particular, the ε-insensitive loss term can be replaced by a somewhat general convex function. We also briefly discuss two sparse variants for Co-RLS and Laplacian SVMs using the same approach as sparse multi-view SVMs.

7.1 Arbitrary Convex Loss

In Section 2.2, for each example the ε-insensitive loss used is $f_\varepsilon(t_i) = (\sqrt{t_i} - \varepsilon)_+^2$ with $t_i = [f_1(x_i) - f_2(x_i)]^2$. This can be relaxed to a general class of user-designed losses that can be defined as a convex function of t_i, for example, using existing convex functions or compositions of convex functions with some good properties (Boyd and Vandenberghe, 2004).

Provided the ε-insensitive loss conforms the slight assumptions closed, convex, and proper listed in Lemma 5, the methodology used for sparse multi-view SVMs and the advantages of using Fenchel-Legendre conjugates apply well to the new optimization problem. This is an important contribution of this paper, which gives a framework for solving problems involving different ε-insensitive loss functions. Also, this framework applies to problems with a single view or more than two views, as long as the objective function is convex with respect to θ (parameters of classifiers or regressors) as in (11).

7.2 A Sparse Variant for Co-RLS

The objective function of Co-RLS in the case of two views is given as follows (Sindhwani et al., 2005; Brefeld et al., 2006)

$$\min_{f_1\in\mathcal{H}_1, f_2\in\mathcal{H}_2} \quad \frac{1}{2\ell}\sum_{i=1}^{\ell}[(f_1(x_i)-y_i)^2+(f_2(x_i)-y_i)^2]+ \\ \gamma_n(\|f_1\|^2+\|f_2\|^2)+\gamma_v\sum_{i=1}^{\ell+u}(f_1(x_i)-f_2(x_i))^2,$$

where nonnegative scalars γ_n, γ_v are respectively norm regularization and multi-view regularization coefficients. This optimization problem is indeed convex with respect to expansion coefficients α_1 and α_2 which have the same meanings as in (9).

Replacing the last term with $\sum_{i=1}^{\ell+u}(|f_1(x_i)-f_2(x_i)|-\varepsilon)_+^2$ results in the sparse Co-RLS algorithm

$$\min_{f_1\in\mathcal{H}_1, f_2\in\mathcal{H}_2} \quad \frac{1}{2\ell}\sum_{i=1}^{\ell}[(f_1(x_i)-y_i)^2+(f_2(x_i)-y_i)^2]+ \\ \gamma_n(\|f_1\|^2+\|f_2\|^2)+\gamma_v\sum_{i=1}^{\ell+u}(|f_1(x_i)-f_2(x_i)|-\varepsilon)_+^2.$$

This ε-insensitive loss is identical to that used in sparse multi-view SVMs, and therefore we can directly use the technique developed in this paper to solve this optimization.

7.3 A Sparse Variant for Laplacian SVMs

By including a penalty term on the intrinsic manifold smoothness, Belkin et al. (2006) proposed the Laplacian SVMs as an extension of SVMs by solving the following problem in an RKHS

$$\min_{f\in\mathcal{H}}\frac{1}{\ell}\sum_{i=1}^{\ell}(1-y_i f(x_i))_+ +\gamma_A\|f\|^2+\gamma_I\sum_{i,j=1}^{\ell+u}W_{ij}(f(x_i)-f(x_j))^2, \tag{33}$$

where $\mathcal{H}$ is the RKHS induced by a kernel, γ_A and γ_I are respectively ambient and intrinsic regularization coefficients, and $W_{ij}\geq 0$ are entries of the weight matrix W of the graph representing the manifold. The last term can be rewritten as

$$\begin{aligned}\sum_{i,j=1}^{\ell+u}W_{ij}(f(x_i)-f(x_j))^2 &= 2\left[\sum_{i=1}^{\ell+u}(\sum_{j=1}^{\ell+u}W_{ij})f^2(x_i)-\sum_{i,j=1}^{\ell+u}W_{ij}f(x_i)f(x_j)\right]\\ &= 2\mathbf{f}^\top(V-W)\mathbf{f}=2\mathbf{f}^\top\mathcal{L}\mathbf{f},\end{aligned}$$

where $\mathbf{f}=[f(x_1),\ldots,f(x_{\ell+u})]^\top$, matrix V is diagonal with the ith diagonal entry $V_{ii}=\sum_{j=1}^{\ell+u}W_{ij}$, and $\mathcal{L}$ is the positive semi-definite graph Laplacian.

This optimization problem is also convex with respect to expansion coefficient α, slack variable ξ and bias b if we formulate it as in (12). Replacing the last term in (33) with $\sum_{i,j=1}^{\ell+u}W_{ij}(|f(x_i)-f(x_j)|-\varepsilon)_+^2$ results in the sparse Laplacian SVMs

$$\min_{f\in\mathcal{H}}\frac{1}{\ell}\sum_{i=1}^{\ell}(1-y_i f(x_i))_+ +\gamma_A\|f\|^2+\gamma_I\sum_{i,j=1}^{\ell+u}W_{ij}(|f(x_i)-f(x_j)|-\varepsilon)_+^2,$$

This ε-insensitive loss has a similar form with that used in sparse multi-view SVMs, and thus facilitates an extension of the technique developed in this paper to solve this optimization.

8. Conclusion

In this paper, we proposed a sparse semi-supervised learning framework using Fenchel-Legendre conjugates. It is extendable to a wide range of semi-supervised learning methods. In particular, we formulated and solved the sparse multi-view SVMs, which incorporate an ε-insensitive multi-view regularization term. By rewriting this regularization in terms of conjugate functions, we obtained an inf-sup optimization problem whose globally optimal solutions can be found by our proposed iterative algorithm. We also showed that the quadratic program involved in each iteration only depends on the size of the labeled set, which would be very efficient for semi-supervised learning problems. For sparse multi-view SVMs, we characterized their generalization error in terms of the margin bound and derived the empirical Rademacher complexity of the considered function class. The empirical Rademacher complexity has two different forms depending on whether the iterative algorithm iterates only once or multiple steps.

Experimental results on sparse multi-view SVMs with different ε values showed that it is unnecessary to retain all the unlabeled data to represent target functions and using sparse semi-supervised learning can effectively reach a good balance between classifier performance and the number of unlabeled examples retained. This would be beneficial to speed up function evaluations during the classification of new examples. Comparisons with SVM-2K showed the superiority of our proposed method both on classification accuracies and the possibility and potential to be applied to large-scale problems when a sequential training strategy is adopted. As in this paper we only concern the possibility and potential for large-scale applications, we employed a moderate data set. It leaves as future work to apply the approach to much larger data sets. We also performed experiments to validate the usefulness of the margin bound and empirical Rademacher complexity, which explain well the regularization role unlabeled data play for multi-view learning.

Acknowledgments

We would like to thank Dr. Sandor Szedmak for providing the code of SVM-2K and related discussions. This work is supported in part by the National Natural Science Foundation of China under Project 61075005, Shanghai Educational Development Foundation under Project 2007CG30, and the PASCAL network of excellence.

References

M. Balcan and A. Blum. A PAC-style model for learning from labeled and unlabeled data. In *Proceedings of the 18th Annual Conference on Computational Learning Theory*, pages 111–126, 2005.

M. Balcan, A. Blum, and K. Yang. Co-training and expansion: Towards bridging theory and practice. *Advances in Neural Information Processing Systems*, 17:89–96, 2005.

P. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled exampls. *Journal of Machine Learning Research*, 7:2399–2434, 2006.

A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, pages 92–100, 1998.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, England, 2004.

U. Brefeld, T. Gärtner, T. Sheffer, and S. Wrobel. Efficient co-regularized least squares regression. In *Proceedings of the 23th International Conference on Machine Learning*, pages 137–144, 2006.

O. Chapelle, B. Schölkopf, and A. Zien. *Semi-supervised Learning*. MIT Press, Cambridge, MA, 2006.

J. Farquhar, D. Hardoon, H. Meng, J. Shawe-Taylor, and S. Szedmak. Two view learning: SVM-2K, theory and practice. *Advances in Neural Information Processing Systems*, 18:355–362, 2006.

G.H. Golub and C.F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, USA, 1996.

G. Kimeldorf and G. Wahba. Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, 33(1):82–95, 1971.

R. Latala and K. Oleszkiewicz. On the best constant in the Khintchine-Kahane inequality. *Studia Mathematica*, 109(1):101–104, 1994.

K. Nigam and R. Ghani. Analyzing the effectiveness and applicability of co-training. In *Proceedings of the 9th International Conference on Information and Knowledge Management*, pages 86–93, 2000.

M. F. Porter. An algorithm for suffix stripping. *Program*, 14:130–137, 1980.

R. Rifkin and R. Lippert. Value regularization and Fenchel duality. *Journal of Machine Learning Research*, 8:441–479, 2007.

D. Rosenberg and P. Bartlett. The Rademacher complexity of co-regularized kernel classes. *Journal of Machine Learning Research Workshop and Conference Proceedings*, 2:396–403, 2007.

G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24:513–523, 1988.

J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge, England, 2004.

V. Sindhwani and D. Rosenberg. An RKHS for multi-view learning and manifold co-regularization. In *Proceedings of the 25th International Conference on Machine Learning*, pages 976–983, 2008.

V. Sindhwani, P. Niyogi, and M. Belkin. A co-regularization approach to semi-supervised learning with multiple views. In *Proceedings of the Workshop on Learning with Multiple Views, 22nd ICML*, 2005.

S. Sun. Semantic features for multi-view semi-supervised and active learning of text classification. In *Proceedings of the IEEE International Conference on Data Mining Workshops*, pages 731–735, 2008.

S. Szedmak and J. Shawe-Taylor. Synthesis of maximum margin and multiview learning using unlabeled data. *Neurocomputing*, 70(7-9):1254–1264, 2007.

I. Tsang and J. Kwok. Large-scale sparsified manifold regularization. *Advances in Neural Information Processing Systems*, 19:1401–1408, 2007.

X. Zhu. Semi-supervised learning literature survey. Technical Report 1530, Department of Computer Sciences, University of Wisconsin Madison, 2008.