Semi-Supervised Optimal Margin Distribution Machines*

Teng Zhang and **Zhi-Hua Zhou**
National Key Lab for Novel Software Technology, Nanjing University, Nanjing 210023, China
{zhangt, zhouzh}@lamda.nju.edu.cn

Abstract

Semi-supervised support vector machines is an extension of standard support vector machines with unlabeled instances, and the goal is to find a label assignment of the unlabeled instances, so that the decision boundary has the maximal *minimum margin* on both the original labeled instances and unlabeled instances. Recent studies, however, disclosed that maximizing the minimum margin does not necessarily lead to better performance, and instead, it is crucial to optimize the *margin distribution*. In this paper, we propose a novel approach ssODM (Semi-Supervised Optimal margin Distribution Machine), which tries to assign the labels to unlabeled instances and to achieve optimal margin distribution simultaneously. Specifically, we characterize the margin distribution by the first- and second-order statistics, i.e., the margin mean and variance, and extend a stochastic mirror prox method to solve the resultant saddle point problem. Extensive experiments on twenty UCI data sets show that ssODM is significantly better than compared methods, which verifies the superiority of optimal margin distribution learning.

1 Introduction

Traditional supervised learning uses only labeled instances to train the classifiers. However, labeled instances are often difficult or expensive to obtain since they require the efforts of experienced human annotators. On the other hand, unlabeled instances are universal and relatively easy to collect. To exploit the value of them, many semi-supervised learning algorithms have been proposed, among which a very popular type of algorithms is the semi-supervised support vector machines (S3VMs). Examples include the semi-supervised SVM [Bennett and Demiriz, 1999], the transductive SVM (TSVM) [Joachims, 1999], the Laplacian SVM [Belkin *et al.*, 2006], and the S3VM using label mean (MeanS3VM) [Li *et al.*, 2009a], just to name a few. Bennett and Demiriz's S3VM and the TSVM are built upon the cluster assumption and use the unlabeled instances to regularize the decision boundary. Specifically, these methods prefer the decision boundary that passes through low-density regions [Chapelle and Zien, 2005]. The Laplacian SVM is a S3VM that exploits the instance's manifold structure via the graph Laplacian. It encodes both the labeled and unlabeled instances by a connected graph, where each instance is represented as a vertex and two vertices are connected by an edge if they are quite similar to each other. The goal is to find labels for the unlabeled instances such that their inconsistencies with both the labeled instances and the underlying graph structure are minimized. The MeanS3VM bases on the observation that S3VMs with knowledge of the means of the labels of the unlabeled instances is closely related to the supervised SVM with known labels on all the unlabeled instances. So the approach is consist of two steps, i.e., first estimate the label means of the unlabeled instances, and then build a SVM model. Moreover, several works also try to extend S3VMs for other learning settings, e.g., the CS4VM [Li *et al.*, 2010] considers the situation where different misclassification errors are associated with unequal costs, and the safe S3VM (S4VM) [Li and Zhou, 2011] tries to exploit many candidate low-density separators simultaneously to reduce the risk of identifying only one poor separator with unlabeled instances, so it always performed better than S3VMs.

Aforementioned all kinds of S3VMs are all based on the large margin principle, i.e., try to maximize the minimum margin of training instances. However, recent studies on margin theory [Gao and Zhou, 2013] disclosed that maximizing the minimum margin does not necessarily lead to better performance, and instead, it is crucial to optimize the margin distribution. Inspired by this recognition, Zhang and Zhou (2014) proposed ODMs (optimal margin distribution machines) which can achieve better generalization performance than large margin based methods. Later, Zhang and Zhou (2017; 2018) extends the idea to multi-class learning and clustering. The success of optimal margin distribution learning suggests that there may still exist large space to further enhance for S3VMs.

In this paper, we propose a novel approach ssODM (semi-supervised optimal margin distribution machines), which tries to learn the label assignments of unlabeled instances and to achieve optimal margin distribution simultaneously.

*This research was supported by National Key R&D Program of China (2018YFB1004300) and NSFC (61751306) and the Program B for outstanding PhD candidate of Nanjing University.

Specifically, we characterize the margin distribution by the first- and second-order statistics, i.e., the margin mean and variance, and then apply the minimax convex relaxation proposed in [Li *et al.*, 2009b], which is proven to be tighter than SDP relaxations [Xu *et al.*, 2005], to get a convex reformulation. For the optimization of the resultant saddle point problem, we propose a stochastic mirror prox method which can converge more quickly in practise than the general subgradient descent for non-smooth problem. Extensive experiments on twenty UCI data sets show that ssODM is significantly better than compared methods, which verifies the superiority of optimal margin distribution learning.

The rest of this paper is organized as follows. We first introduce some preliminaries and then present the ssODM method. Next we show the experimental studies. Finally we conclude this paper with future work.

2 Preliminaries

We start with the traditional supervised learning. Denote $\mathcal{X}$ as the instance space and $\mathcal{Y} = \{+1, -1\}$ as the label set. Let $\mathcal{D}$ be an unknown (underlying) distribution over $\mathcal{X} \times \mathcal{Y}$. A training set $\mathcal{S} = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\} \in (\mathcal{X} \times \mathcal{Y})^m$ is drawn identically and independently (i.i.d.) according to $\mathcal{D}$. Let $\phi : \mathcal{X} \mapsto \mathbb{H}$ be a feature mapping associated to some positive definite kernel $\kappa : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$. The hypothesis is defined based on the linear model $h(\boldsymbol{x}) = \boldsymbol{w}^\top \phi(\boldsymbol{x})$ and the predicted label of instance $\boldsymbol{x}$ is the sign of $h(\boldsymbol{x})$, then the decision function naturally leads to the definition of margin for a labeled instance, i.e., $\gamma(\boldsymbol{x}, y) = y\boldsymbol{w}^\top \phi(\boldsymbol{x})$ [Cristianini and Shawe-Taylor, 2000]. Thus the higher the margin value, the more confidence we will have that $\boldsymbol{x}$'s label is y, and h misclassifies $(\boldsymbol{x}, y)$ if and only if it produces a negative margin. Given a hypothesis set $\mathcal{H}$ of functions mapping $\mathcal{X}$ to $\mathcal{Y}$ and the labeled training set $\mathcal{S}$, our goal is to learn a function $h \in \mathcal{H}$ such that the generalization error $R(h) = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}[1_{\mathrm{sign}(h(\boldsymbol{x}))\neq y}]$ is small, where $1_{(\cdot)}$ is the indicator function that returns 1 when the argument holds, and 0 otherwise.

2.1 Optimal Margin Distribution Machine

It is well known that SVM employs the large margin principle to select h and tries to maximize the minimum margin of training instance, i.e., the smallest distance from the instances to the decision boundary. As a result, the solution of SVM just consists of a small amount of instance, that is support vectors (SV), and the rest (non-SVs) are totally ignored, which may be misleading in some situations. See Figure 1 for an illustration [Zhou, 2014].

A more robust strategy is to consider the whole instances, i.e., optimizing the margin distribution. To characterize the distribution, the two most straightforward statistics are the first- and second-order statistics, that is, the margin mean and variance. Moreover, a recent study [Gao and Zhou, 2013] on margin theory proved that maximizing the margin mean and minimizing the margin variance simultaneously can yield a tighter generalization bound, so we arrive at the following formulation,

$$\min_{\boldsymbol{w},\bar{\gamma},\xi_i,\epsilon_i} \frac{1}{2}\|\boldsymbol{w}\|_{\mathbb{H}}^2 - \eta\bar{\gamma} + \frac{\lambda}{m}\sum_{i=1}^{m}(\xi_i^2 + \epsilon_i^2),$$

$$\text{s.t. } \gamma(\boldsymbol{x}_i, y_i) \geq \bar{\gamma} - \xi_i,$$
$$\gamma(\boldsymbol{x}_i, y_i) \leq \bar{\gamma} + \epsilon_i, \ \forall i,$$

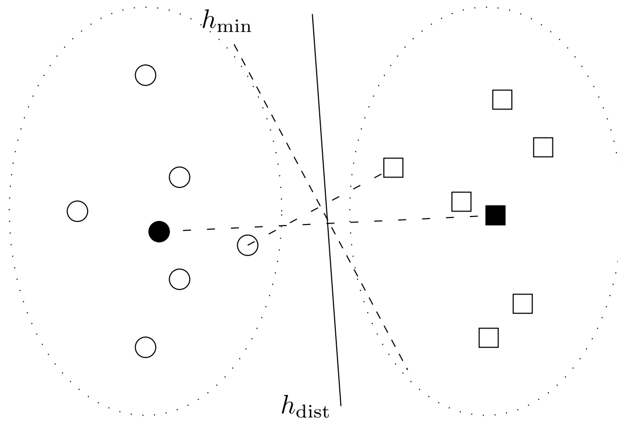

Figure 1: A simple illustration of linear separators optimizing the minimum margin and margin distribution, respectively. Dotted ellipses are two underlying distributions, from which circles and squares are instances sampled. Solid circle and square are mean instances (not necessarily real instances in the training data). $h_{\min}$ and h_{dist} are decision hyperplanes achieved by optimizing the minimum margin and margin distribution, respectively.

where $\bar{\gamma}$ is the margin mean, η and λ are trading-off parameters, ξ_i and ϵ_i are the deviation of $\gamma(\boldsymbol{x}_i, y_i)$ to the margin mean. It's evident that $\sum_{i=1}^m (\xi_i^2 + \epsilon_i^2)/m$ is exactly the margin variance.

First, by scaling $\boldsymbol{w}$ which doesn't affect the final classification results, the margin mean can be fixed as 1, then the deviation of $\gamma(\boldsymbol{x}_i, y_i)$ to the margin mean is $|y_i\boldsymbol{w}^\top\phi(\boldsymbol{x}_i) - 1|$. Secondly, the hyperplane $y_i\boldsymbol{w}^\top\phi(\boldsymbol{x}_i) = 1$ divides the feature space into two parts and for each instance, no matter which part it lies in, it will suffer a loss which is quadratic with the deviation. So it is more reasonable to set different weights for the two kinds of deviations because the instances lie in $\{\boldsymbol{x} \mid y\boldsymbol{w}^\top\phi(\boldsymbol{x}) < 1\}$ are much easier to be misclassified than the other. Thirdly, according to representer theorem [Schölkopf and Smola, 2001], the optimal solution is spanned only by SVs. To achieve a sparse solution, we introduce a θ-insensitive loss like SVR, i.e., the instances whose deviation is smaller than θ are tolerated and only those whose deviation is larger than θ will suffer a loss. Finally, we obtain the formulation of ODM,

$$\begin{aligned} \min_{\boldsymbol{w},\xi_i,\epsilon_i} \ & \frac{1}{2}\|\boldsymbol{w}\|_{\mathbb{H}}^2 + \frac{\lambda}{m}\sum_{i=1}^{m}\frac{\xi_i^2 + \nu\epsilon_i^2}{(1-\theta)^2}, \\ \text{s.t. } & y_i\boldsymbol{w}^\top\phi(\boldsymbol{x}_i) \geq 1 - \theta - \xi_i, \\ & y_i\boldsymbol{w}^\top\phi(\boldsymbol{x}_i) \leq 1 + \theta + \epsilon_i, \ \forall i. \end{aligned} \tag{1}$$

where ν is a parameter for trading-off different kinds of deviations, θ is a parameter for controlling the sparsity of the solution, and $(1-\theta)^2$ in the denominator is to scale the second term to be a surrogate loss for 0-1 loss.

It can be found that the least square loss is a special case of ODM's loss fucntion by ignoring the asymmetry, so it is more sensitive to outliers and can only be suitable for some specific distribution. In addition, without θ-insensitivity, all the instances will be support vectors, which can result in unnecessary computation cost. In other words, ODM is more flexible and can characterize the margin distribution more adaptively.

3 ssODM

In semi-supervised learning setting, not all the training labels are known. Let $S_L = \{\boldsymbol{x}_i, y_i\}_{i=1}^l$ and $S_U = \{\boldsymbol{x}_j\}_{j=l+1}^m$ be the sets of labeled and unlabeled instances, respectively. $L = \{1, \ldots, l\}$ and $U = \{l+1, \ldots, m\}$ are the index sets of the labeled and unlabeled instances. In semi-supervised learning, unlabeled data are typically much more abundant than labeled data, that is, $m - l \gg l$. Hence, one can obtain a trivially "optimal" solution with infinite margin by assigning all the unlabeled examples to the same label. To prevent such a useless solution, Joachims (1999) introduced the balance constraint:

$$\frac{\boldsymbol{e}^\top \hat{\boldsymbol{y}}_U}{m-l} = \frac{\boldsymbol{e}^\top \boldsymbol{y}_L}{l}$$

where $\hat{\boldsymbol{y}}^\top = [\hat{y}_1, \ldots, \hat{y}_m]$ is the vector of learned labels on both labeled and unlabeled examples, $\boldsymbol{y}_L^\top = [y_1, \ldots, y_l]$, $\hat{\boldsymbol{y}}_U^\top = [\hat{y}_{l+1}, \ldots, \hat{y}_m]$, and $\boldsymbol{e}$ stands for the all-one vector. The basic idea of ssODM is to minimize the objective function in Eq. (1) w.r.t. both the labeling $\hat{\boldsymbol{y}}$ and decision function parameter $\boldsymbol{w}$, ξ_i, ϵ_i. Hence, Eq. (1) is extended to

$$\begin{aligned} \min_{\hat{\boldsymbol{y}} \in \mathcal{B}} \min_{\boldsymbol{w}, \xi_i, \epsilon_i} \ & \frac{1}{2}\|\boldsymbol{w}\|_{\mathbb{H}}^2 + \sum_{i=1}^m \lambda_i \frac{\xi_i^2 + \nu\epsilon_i^2}{(1-\theta)^2} \\ \text{s.t. } & \hat{y}_i \boldsymbol{w}^\top \phi(\boldsymbol{x}_i) \geq 1 - \theta - \xi_i, \\ & \hat{y}_i \boldsymbol{w}^\top \phi(\boldsymbol{x}_i) \leq 1 + \theta + \epsilon_i, \ \forall i, \end{aligned} \tag{2}$$

where $\mathcal{B} = \{\hat{\boldsymbol{y}} \mid \hat{\boldsymbol{y}} = [\hat{\boldsymbol{y}}_L; \hat{\boldsymbol{y}}_U], \hat{\boldsymbol{y}}_L = \boldsymbol{y}_L, \hat{\boldsymbol{y}}_U \in \{0,1\}^{m-l}, \frac{\boldsymbol{e}^\top \hat{\boldsymbol{y}}_U}{m-l} = \frac{\boldsymbol{e}^\top \boldsymbol{y}_L}{l}\}$ is a set of candidate label assignments. $\lambda_i = \frac{\lambda_1(m-l) - \lambda_2 l}{l(m-l)} 1_{i \in L} + \frac{\lambda_2}{m-l}$, and λ_1, λ_2 trade off empirical losses on the labeled and unlabeled data, respectively.

To avoid the curse of dimensionality, the inner minimization problem of Eq. (2) is usually cast in the dual form. Denote $\boldsymbol{X}$ as the data matrix whose i-th column is $\phi(\boldsymbol{x}_i)$, i.e., $\boldsymbol{X} = [\phi(\boldsymbol{x}_1), \ldots, \phi(\boldsymbol{x}_m)]$, and introduce the dual variables $\boldsymbol{\alpha} \succeq \boldsymbol{0}$, the Lagrangian of Eq. (2) leads to

$$\begin{aligned} \min_{\hat{\boldsymbol{y}} \in \mathcal{B}} \max_{\boldsymbol{\alpha} \succeq \boldsymbol{0}} & -\frac{1}{2}\boldsymbol{\alpha}^\top \begin{bmatrix} \boldsymbol{K} \odot \hat{\boldsymbol{y}}\hat{\boldsymbol{y}}^\top & -\boldsymbol{K} \odot \hat{\boldsymbol{y}}\hat{\boldsymbol{y}}^\top \\ -\boldsymbol{K} \odot \hat{\boldsymbol{y}}\hat{\boldsymbol{y}}^\top & \boldsymbol{K} \odot \hat{\boldsymbol{y}}\hat{\boldsymbol{y}}^\top \end{bmatrix} \boldsymbol{\alpha} \\ & - \frac{(1-\theta)^2}{4}\boldsymbol{\alpha}^\top \begin{bmatrix} \boldsymbol{I} \oslash \boldsymbol{\lambda} & \boldsymbol{0} \\ \boldsymbol{0} & \frac{1}{\nu}\boldsymbol{I} \oslash \boldsymbol{\lambda} \end{bmatrix} \boldsymbol{\alpha} - \begin{bmatrix} (\theta-1)\boldsymbol{e} \\ (\theta+1)\boldsymbol{e} \end{bmatrix}^\top \boldsymbol{\alpha}, \end{aligned} \tag{3}$$

where $\boldsymbol{K} = \boldsymbol{X}^\top \boldsymbol{X}$ is the kernel matrix, $\boldsymbol{\lambda}^\top = [\lambda_1, \ldots, \lambda_m]$, $\odot$ and $\oslash$ denotes the element-wise product and division, respectively. Note that the objective function is a negative definite quadratic form whose stationary point can't locate at the infinity, so we can replace the constraint $\{\boldsymbol{\alpha} \mid \boldsymbol{\alpha} \succeq \boldsymbol{0}\}$ by a bounded box $\mathcal{A} = \{\boldsymbol{\alpha} \mid \boldsymbol{0} \preceq \boldsymbol{\alpha} \preceq \tau\boldsymbol{e}\}$, where the auxiliary parameter τ is introduced for the sake of mathematical soundness. For a sufficiently large τ, the new problem is equal to the original problem.

To overcome the difficulty of this mixed-integer programming, many relaxations have been proposed, among which the minimax convex relaxation proposed in [Li *et al.*, 2009b; 2013] is proven to be the tightest. So in this paper, we also employ this method to deal with the mixed-integer problem, i.e., interchanging the order of $\max_{\boldsymbol{\alpha} \in \mathcal{A}}$ and $\min_{\hat{\boldsymbol{y}} \in \mathcal{B}}$, then we can obtain

$$\max_{\boldsymbol{\alpha} \in \mathcal{A}} \min_{\hat{\boldsymbol{y}} \in \mathcal{B}} \ G(\boldsymbol{\alpha}, \hat{\boldsymbol{y}}),$$

where $G(\boldsymbol{\alpha}, \hat{\boldsymbol{y}})$ is the objective function of Eq. (3), and this can be further transformed into

$$\max_{\boldsymbol{\alpha} \in \mathcal{A}} \min_{\delta} \ (-\delta) \quad \text{s.t. } G(\boldsymbol{\alpha}, \hat{\boldsymbol{y}}_k) \geq \delta, \ \forall \hat{\boldsymbol{y}}_k \in \mathcal{B}. \tag{4}$$

For the inner optimization in Eq. (4), introduce the dual variables $\boldsymbol{\mu}^\top = [\mu_1, \ldots, \mu_{|\mathcal{B}|}] \succeq \boldsymbol{0}$, the Lagrangian leads to

$$\max_{\boldsymbol{\mu} \succeq \boldsymbol{0}} \min_{\delta} \ \left\{ -\delta - \sum_{k: \hat{\boldsymbol{y}}_k \in \mathcal{B}} \mu_k (G(\boldsymbol{\alpha}, \hat{\boldsymbol{y}}_k) - \delta) \right\},$$

By setting the partial derivative of δ to zero, we can obtain $\sum_{k: \hat{\boldsymbol{y}}_k \in \mathcal{B}} \mu_k = 1$ and the dual turns to

$$\max_{\boldsymbol{\mu} \in \mathcal{M}} \ \left\{ -\sum_{k: \hat{\boldsymbol{y}}_k \in \mathcal{B}} \mu_k G(\boldsymbol{\alpha}, \hat{\boldsymbol{y}}_k) \right\}, \tag{5}$$

where $\mathcal{M} = \{\boldsymbol{\mu} \in \mathbb{R}_+^{|\mathcal{B}|} \mid \boldsymbol{e}^\top \boldsymbol{\mu} = 1\}$ is the simplex in $\mathbb{R}^{|\mathcal{B}|}$. By substituting Eq. (5) into Eq. (4) and denoting $\varphi(\boldsymbol{\mu}, \boldsymbol{\alpha}) = \sum_{k: \hat{\boldsymbol{y}}_k \in \mathcal{B}} \mu_k G(\boldsymbol{\alpha}, \hat{\boldsymbol{y}}_k)$, Eq. (4) can be rewritten as

$$\max_{\boldsymbol{\alpha} \in \mathcal{A}} \min_{\boldsymbol{\mu} \in \mathcal{M}} \ \varphi(\boldsymbol{\mu}, \boldsymbol{\alpha}).$$

Note that $\varphi(\boldsymbol{\mu}, \boldsymbol{\alpha})$ is a convex combination of negative definite quadratic forms, so it's convex in $\boldsymbol{\mu}$ and concave in $\boldsymbol{\alpha}$. According to Sion's minimax theorem [Sion, 1958], there exists a saddle point $(\boldsymbol{\mu}^\star, \boldsymbol{\alpha}^\star) \in \mathcal{M} \times \mathcal{A}$ such that

$$\begin{aligned} \min_{\boldsymbol{\mu} \in \mathcal{M}} \max_{\boldsymbol{\alpha} \in \mathcal{A}} \varphi(\boldsymbol{\mu}, \boldsymbol{\alpha}) & \leq \max_{\boldsymbol{\alpha} \in \mathcal{A}} \varphi(\boldsymbol{\mu}^\star, \boldsymbol{\alpha}) = \varphi(\boldsymbol{\mu}^\star, \boldsymbol{\alpha}^\star) \\ & = \min_{\boldsymbol{\mu} \in \mathcal{M}} \varphi(\boldsymbol{\mu}, \boldsymbol{\alpha}^\star) \leq \max_{\boldsymbol{\alpha} \in \mathcal{A}} \min_{\boldsymbol{\mu} \in \mathcal{M}} \varphi(\boldsymbol{\mu}, \boldsymbol{\alpha}), \end{aligned} \tag{6}$$

By combining with the following minimax inequality [Kim and Boyd, 2008],

$$\max_{\boldsymbol{\alpha} \in \mathcal{A}} \min_{\boldsymbol{\mu} \in \mathcal{M}} \varphi(\boldsymbol{\mu}, \boldsymbol{\alpha}) \leq \min_{\boldsymbol{\mu} \in \mathcal{M}} \max_{\boldsymbol{\alpha} \in \mathcal{A}} \varphi(\boldsymbol{\mu}, \boldsymbol{\alpha}),$$

we can realize that all the equalities hold in Eq. (6) and arrive at the final formulation of ssODM:

$$\min_{\boldsymbol{\mu} \in \mathcal{M}} \max_{\boldsymbol{\alpha} \in \mathcal{A}} \ \varphi(\boldsymbol{\mu}, \boldsymbol{\alpha}). \tag{7}$$

4 Optimization

In this section, we commence with a simple introduction to minimax problem, followed by a stochastic mirror prox method to find the saddle point.

4.1 Minimax Problem

Since $\varphi(\cdot, \boldsymbol{\alpha})$ is convex and $\varphi(\boldsymbol{\mu}, \cdot)$ is concave, according to the convex inequality, for any pair $(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\alpha}}) \in \mathcal{M} \times \mathcal{A}$ we have

$$\varphi(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\alpha}}) - \varphi(\boldsymbol{\mu}, \bar{\boldsymbol{\alpha}}) \leq \partial_{\boldsymbol{\mu}} \varphi(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\alpha}})^\top (\bar{\boldsymbol{\mu}} - \boldsymbol{\mu}),\ \forall \boldsymbol{\mu} \in \mathcal{M},$$
$$\varphi(\bar{\boldsymbol{\mu}}, \boldsymbol{\alpha}) - \varphi(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\alpha}}) \leq -\partial_{\boldsymbol{\alpha}} \varphi(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\alpha}})^\top (\bar{\boldsymbol{\alpha}} - \boldsymbol{\alpha}),\ \forall \boldsymbol{\alpha} \in \mathcal{A}.$$

By adding the above two inequalities together we have

$$\varphi(\bar{\boldsymbol{\mu}}, \boldsymbol{\alpha}) - \varphi(\boldsymbol{\mu}, \bar{\boldsymbol{\alpha}}) \leq g(\boldsymbol{u})^\top (\boldsymbol{u} - \boldsymbol{w}),\ \forall \boldsymbol{\mu}, \boldsymbol{\alpha}, \tag{8}$$

where $\boldsymbol{w} = (\boldsymbol{\mu}, \boldsymbol{\alpha}), \boldsymbol{u} = (\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\alpha}}) \in \mathcal{M} \times \mathcal{A}$, and $g(\boldsymbol{u}) = (\partial_{\boldsymbol{\mu}} \varphi(\boldsymbol{u}), -\partial_{\boldsymbol{\alpha}} \varphi(\boldsymbol{u}))$, which plays a similar role as gradient in general convex optimization. Note that Eq. (8) holds for any $\boldsymbol{\mu}$ and $\boldsymbol{\alpha}$, in particular we have

$$\max_{\boldsymbol{\alpha} \in \mathcal{A}} \varphi(\bar{\boldsymbol{\mu}}, \boldsymbol{\alpha}) - \min_{\boldsymbol{\mu} \in \mathcal{M}} \varphi(\boldsymbol{\mu}, \bar{\boldsymbol{\alpha}}) \leq g(\boldsymbol{u})^\top (\boldsymbol{u} - \boldsymbol{w}). \tag{9}$$

The left hand side is referred to as the "duality gap", which can be decomposed into two parts, i.e.,

$$\begin{aligned}
&\max_{\boldsymbol{\alpha} \in \mathcal{A}} \varphi(\bar{\boldsymbol{\mu}}, \boldsymbol{\alpha}) - \min_{\boldsymbol{\mu} \in \mathcal{M}} \varphi(\boldsymbol{\mu}, \bar{\boldsymbol{\alpha}}) \\
&= \max_{\boldsymbol{\alpha} \in \mathcal{A}} \varphi(\bar{\boldsymbol{\mu}}, \boldsymbol{\alpha}) - \varphi(\boldsymbol{\mu}^\star, \boldsymbol{\alpha}^\star) + \varphi(\boldsymbol{\mu}^\star, \boldsymbol{\alpha}^\star) - \min_{\boldsymbol{\mu} \in \mathcal{M}} \varphi(\boldsymbol{\mu}, \bar{\boldsymbol{\alpha}}) \\
&= \underbrace{\max_{\boldsymbol{\alpha} \in \mathcal{A}} \varphi(\bar{\boldsymbol{\mu}}, \boldsymbol{\alpha}) - \min_{\boldsymbol{\mu} \in \mathcal{M}} \max_{\boldsymbol{\alpha} \in \mathcal{A}} \varphi(\boldsymbol{\mu}, \boldsymbol{\alpha})}_{\text{primal gap}} \\
&\qquad + \underbrace{\max_{\boldsymbol{\alpha} \in \mathcal{A}} \min_{\boldsymbol{\mu} \in \mathcal{M}} \varphi(\boldsymbol{\mu}, \boldsymbol{\alpha}) - \min_{\boldsymbol{\mu} \in \mathcal{M}} \varphi(\boldsymbol{\mu}, \bar{\boldsymbol{\alpha}})}_{\text{dual gap}}.
\end{aligned}$$

As can be seen, the primal gap and the dual gap are both non-negative and the more closer to the saddle point, the smaller both gaps. So duality gap can be viewed as a measure to evaluate the closeness of current point $(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\alpha}})$ to the saddle point $(\boldsymbol{\mu}^\star, \boldsymbol{\alpha}^\star)$.

4.2 Stochastic Mirror Prox Method

For ssODM, the feasible set of $\boldsymbol{\mu}$ and $\boldsymbol{\alpha}$ are simplex and bounded box, respectively, so the most suitable optimization method is mirror descent [Beck and Teboulle, 2003], and the corresponding mirror maps are $\Phi_{\mathcal{M}}(\boldsymbol{\mu}) = \sum_k \mu_k \log \mu_k$ and $\Phi_{\mathcal{A}}(\boldsymbol{\alpha}) = \|\boldsymbol{\alpha}\|_2^2/2$, respectively. Further note that the objective of inner optimization is smooth function, mirror descent can be accelerated to the rate $O(1/t)$ by applying the mirror prox technique in [Nemirovski, 2005].

Introduce the joint map $\Phi(\boldsymbol{w}) = a\Phi_{\mathcal{M}}(\boldsymbol{\mu}) + b\Phi_{\mathcal{A}}(\boldsymbol{\alpha})$, where $a = 1/\sqrt{\log |\mathcal{B}|}$ and $b = \sqrt{2}/\tau\sqrt{m}$. It can be shown that $\nabla \Phi_{\mathcal{M}}(\boldsymbol{\mu}) = \log \boldsymbol{\mu} + \boldsymbol{e}$, $\nabla \Phi_{\mathcal{A}}(\boldsymbol{\alpha}) = \boldsymbol{\alpha}$ and $\nabla \Phi(\boldsymbol{w}) = (a \log \boldsymbol{\mu} + a\boldsymbol{e}, b\boldsymbol{\alpha})$. At the t-th iteration, we first map $\boldsymbol{w}_t = (\boldsymbol{\mu}_t, \boldsymbol{\alpha}_t)$ into the dual space $\nabla \Phi(\boldsymbol{w}_t) = (a \log \boldsymbol{\mu}_t + a\boldsymbol{e}, b\boldsymbol{\alpha}_t)$, followed by one step of stochastic gradient descent in the dual space,

$$\begin{aligned}
&\nabla \Phi(\boldsymbol{u}_t) = \nabla \Phi(\boldsymbol{w}_t) - \eta \widetilde{g}(\boldsymbol{w}_t) \\
&= (a \log \boldsymbol{\mu}_t + a\boldsymbol{e} - \eta \partial_{\boldsymbol{\mu}} \widetilde{\varphi}(\boldsymbol{\mu}_t, \boldsymbol{\alpha}_t), b\boldsymbol{\alpha}_t + \eta \partial_{\boldsymbol{\alpha}} \widetilde{\varphi}(\boldsymbol{\mu}_t, \boldsymbol{\alpha}_t))
\end{aligned}$$

where $\partial_{\boldsymbol{\mu}} \widetilde{\varphi}$, $\partial_{\boldsymbol{\alpha}} \widetilde{\varphi}$ and $\widetilde{g}$ are the noisy unbiased estimation of $\partial_{\boldsymbol{\mu}} \varphi$, $\partial_{\boldsymbol{\alpha}} \varphi$ and g, respectively, and η is the step size. Next, we map $\nabla \Phi(\boldsymbol{u}_t)$ back to the primal space, i.e., to find $\boldsymbol{u}_t = (\bar{\boldsymbol{\mu}}_t, \bar{\boldsymbol{\alpha}}_t)$ such that

$$\begin{aligned}
a \log \bar{\boldsymbol{\mu}}_t + a\boldsymbol{e} &= a \log \boldsymbol{\mu}_t + a\boldsymbol{e} - \eta \partial_{\boldsymbol{\mu}} \widetilde{\varphi}(\boldsymbol{\mu}_t, \boldsymbol{\alpha}_t), \\
b\bar{\boldsymbol{\alpha}}_t &= b\boldsymbol{\alpha}_t + \eta \partial_{\boldsymbol{\alpha}} \widetilde{\varphi}(\boldsymbol{\mu}_t, \boldsymbol{\alpha}_t),
\end{aligned}$$

which implies that $\bar{\boldsymbol{\mu}}_t = \boldsymbol{\mu}_t \exp(-\eta \partial_{\boldsymbol{\mu}} \widetilde{\varphi}(\boldsymbol{\mu}_t, \boldsymbol{\alpha}_t)/a)$ and $\bar{\boldsymbol{\alpha}}_t = \boldsymbol{\alpha}_t + \eta \partial_{\boldsymbol{\alpha}} \widetilde{\varphi}(\boldsymbol{\mu}_t, \boldsymbol{\alpha}_t)/b$. Finally, we project $(\bar{\boldsymbol{\mu}}_t, \bar{\boldsymbol{\alpha}}_t)$ back to $\mathcal{M} \times \mathcal{A}$ based on Kullback-Leibler divergence and Euclidean distance, respectively, i.e., we solve the following two optimization problems:

$$\begin{aligned}
\bar{\boldsymbol{\mu}}_{t+1} &= \operatorname{argmin}_{\boldsymbol{\mu} \in \mathcal{M}} \boldsymbol{\mu}^\top \log \frac{\boldsymbol{\mu}}{\bar{\boldsymbol{\mu}}_t}, \\
\bar{\boldsymbol{\alpha}}_{t+1} &= \operatorname{argmin}_{\boldsymbol{\alpha} \in \mathcal{A}} \|\boldsymbol{\alpha} - \bar{\boldsymbol{\alpha}}_t\|_2^2,
\end{aligned}$$

Fortunately, both problems have a closed-form solution. The latter is to project $\bar{\boldsymbol{\alpha}}_t$ onto the bounded box, so we have $\bar{\boldsymbol{\alpha}}_{t+1} = \max\{\min\{\bar{\boldsymbol{\alpha}}_t, \tau\boldsymbol{e}\}, \boldsymbol{0}\}$. For the former, the Lagrangian function leads to $\boldsymbol{\mu}^\top \log(\boldsymbol{\mu}/\bar{\boldsymbol{\mu}}_t) + \zeta(\boldsymbol{e}^\top \boldsymbol{\mu} - 1)$, where ζ is the dual variable. By setting the partial derivative of $\boldsymbol{\mu}$ to zero, i.e., $\log(\boldsymbol{\mu}/\bar{\boldsymbol{\mu}}_t) + \boldsymbol{e} + \zeta\boldsymbol{e} = \boldsymbol{0}$, we have $\bar{\boldsymbol{\mu}}_{t+1} = \bar{\boldsymbol{\mu}}_t \exp(-1 - \zeta)$. Since $\bar{\boldsymbol{\mu}}_{t+1}$ belongs to a simplex, hence $1 = \boldsymbol{e}^\top \bar{\boldsymbol{\mu}}_{t+1} = \boldsymbol{e}^\top \bar{\boldsymbol{\mu}}_t \exp(-1 - \zeta) = \|\bar{\boldsymbol{\mu}}_t\|_1 \exp(-1 - \zeta)$, which implies that $\exp(-1 - \zeta) = 1/\|\bar{\boldsymbol{\mu}}_t\|_1$, thus we have $\bar{\boldsymbol{\mu}}_{t+1} = \bar{\boldsymbol{\mu}}_t/\|\bar{\boldsymbol{\mu}}_t\|_1$. Once we have $\boldsymbol{y}_{t+1} = (\bar{\boldsymbol{\mu}}_{t+1}, \bar{\boldsymbol{\alpha}}_{t+1})$, start the above procedures from $\boldsymbol{w}_t$ again, but this time using the gradient evaluated at $\boldsymbol{y}_{t+1}$ instead of $\boldsymbol{w}_t$. In words, one iteration of mirror prox consists of two steps of mirror descent starting from the same point, but using gradients evaluated at different points. Figure 2 illustrates one iteration of stochastic mirror prox method.

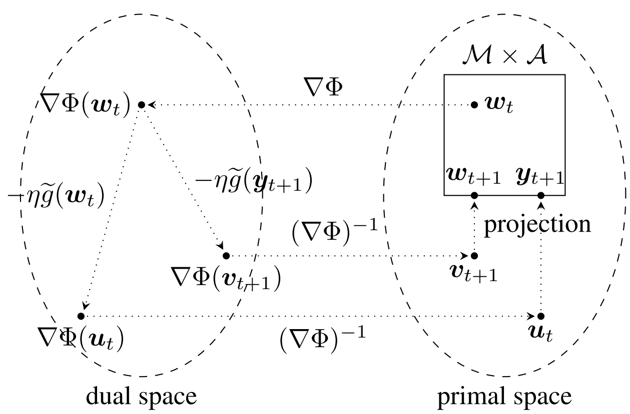

Figure 2: Illustration of one iteration of stochastic mirror prox method.

The remaining question is how to find the stochastic gradient $\partial_{\boldsymbol{\mu}} \widetilde{\varphi}(\boldsymbol{\mu}_t, \boldsymbol{\alpha}_t)$ and $\partial_{\boldsymbol{\alpha}} \widetilde{\varphi}(\boldsymbol{\mu}_t, \boldsymbol{\alpha}_t)$. Note that $\varphi(\boldsymbol{\mu}, \boldsymbol{\alpha}) = \sum_{k:\hat{\boldsymbol{y}}_k \in \mathcal{B}} \mu_k G(\boldsymbol{\alpha}, \hat{\boldsymbol{y}}_k)$, so we have

$$\begin{aligned}
\partial_{\boldsymbol{\mu}} \varphi(\boldsymbol{\mu}_t, \boldsymbol{\alpha}_t) &= [G(\boldsymbol{\alpha}_t, \hat{\boldsymbol{y}}_1), \ldots, G(\boldsymbol{\alpha}_t, \hat{\boldsymbol{y}}_{|\mathcal{B}|})], \\
\partial_{\boldsymbol{\alpha}} \varphi(\boldsymbol{\mu}_t, \boldsymbol{\alpha}_t) &= [\partial_{\boldsymbol{\alpha}} G(\boldsymbol{\alpha}_t, \hat{\boldsymbol{y}}_1), \ldots, \partial_{\boldsymbol{\alpha}} G(\boldsymbol{\alpha}_t, \hat{\boldsymbol{y}}_{|\mathcal{B}|})]\boldsymbol{\mu}_t.
\end{aligned}$$

By uniformly choosing an index i_t from $\{1, 2, \ldots, |\mathcal{B}|\}$, we can obtain $\partial_{\boldsymbol{\mu}} \widetilde{\varphi}(\boldsymbol{\mu}_t, \boldsymbol{\alpha}_t, i_t) = [0, \ldots, |\mathcal{B}| G(\boldsymbol{\alpha}_t, \hat{\boldsymbol{y}}_{i_t}) \ldots, 0]$.

Algorithm 1 Stochastic mirror prox for ssODM

Input: maximum iteration number T, ODM parameters λ_1, λ_2, ν, θ, upper bound τ, stopping criteria ι.
Initialize $\boldsymbol{\mu}_0 \leftarrow [1/|\mathcal{B}|, \dots, 1/|\mathcal{B}|]$, $\boldsymbol{\alpha}_0 \leftarrow \mathbf{0}$, $t \leftarrow 0$.
while $t < T$ **do**
  Uniformly select i_t, i'_t from $\{1, 2, \dots, |\mathcal{B}|\}$.
  $\partial_{\boldsymbol{\mu}} \widetilde{\varphi} \leftarrow [0, \dots, |\mathcal{B}| G(\boldsymbol{\alpha}_t, \hat{\boldsymbol{y}}_{i_t}) \dots, 0]$.
  Select j_t from $\{1, 2, \dots, |\mathcal{B}|\}$ according to $\boldsymbol{\mu}_t$.
  $\partial_{\boldsymbol{\alpha}} \widetilde{\varphi} \leftarrow \partial_{\boldsymbol{\alpha}} G(\boldsymbol{\alpha}_t, \hat{\boldsymbol{y}}_{j_t})$.
  $\bar{\boldsymbol{\mu}}_t \leftarrow \boldsymbol{\mu}_t \exp(-\eta \partial_{\boldsymbol{\mu}} \widetilde{\varphi}/a)$.
  $\bar{\boldsymbol{\alpha}}_t \leftarrow \boldsymbol{\alpha}_t + \eta \partial_{\boldsymbol{\alpha}} \widetilde{\varphi}/b$.
  $\bar{\boldsymbol{\mu}}_{t+1} \leftarrow \bar{\boldsymbol{\mu}}_t / \|\bar{\boldsymbol{\mu}}_t\|_1$.
  $\bar{\boldsymbol{\alpha}}_{t+1} \leftarrow \max\{\min\{\bar{\boldsymbol{\alpha}}_t, \tau \boldsymbol{e}\}, \mathbf{0}\}$.
  $\partial_{\boldsymbol{\mu}} \widetilde{\varphi} \leftarrow [0, \dots, |\mathcal{B}| G(\bar{\boldsymbol{\alpha}}_{t+1}, \hat{\boldsymbol{y}}_{i'_t}) \dots, 0]$.
  Select j'_t from $\{1, 2, \dots, |\mathcal{B}|\}$ according to $\bar{\boldsymbol{\mu}}_{t+1}$.
  $\partial_{\boldsymbol{\alpha}} \widetilde{\varphi} \leftarrow \partial_{\boldsymbol{\alpha}} G(\bar{\boldsymbol{\alpha}}_{t+1}, \hat{\boldsymbol{y}}_{j'_t})$.
  $\bar{\boldsymbol{\mu}}_{t+1} \leftarrow \boldsymbol{\mu}_t \exp(-\eta \partial_{\boldsymbol{\mu}} \widetilde{\varphi}/a)$.
  $\bar{\boldsymbol{\alpha}}_{t+1} \leftarrow \boldsymbol{\alpha}_t + \eta \partial_{\boldsymbol{\alpha}} \widetilde{\varphi}/b$.
  $\boldsymbol{\mu}_{t+1} \leftarrow \bar{\boldsymbol{\mu}}_{t+1} / \|\bar{\boldsymbol{\mu}}_{t+1}\|_1$.
  $\boldsymbol{\alpha}_{t+1} \leftarrow \max\{\min\{\bar{\boldsymbol{\alpha}}_{t+1}, \tau \boldsymbol{e}\}, \mathbf{0}\}$.
  $t \leftarrow t + 1$.
  if duality gap is smaller than ι **then**
    Break.
  end if
end while
Output: $\boldsymbol{\mu}$, $\boldsymbol{\alpha}$.

On the other hand, by randomly sampling an index j_t according to the distribution $\boldsymbol{\mu}_t$ on $\{1, 2, \dots, |\mathcal{B}|\}$, we can obtain $\partial_{\boldsymbol{\alpha}} \widetilde{\varphi}(\boldsymbol{\mu}_t, \boldsymbol{\alpha}_t, j_t) = \partial_{\boldsymbol{\alpha}} G(\boldsymbol{\alpha}_t, \hat{\boldsymbol{y}}_{j_t})$. It can be shown that

$$\mathbb{E}[\partial_{\boldsymbol{\mu}} \widetilde{\varphi}(\boldsymbol{\mu}_t, \boldsymbol{\alpha}_t, i_t) \mid \boldsymbol{\mu}_t, \boldsymbol{\alpha}_t] = \partial_{\boldsymbol{\mu}} \varphi(\boldsymbol{\mu}_t, \boldsymbol{\alpha}_t),$$
$$\mathbb{E}[\partial_{\boldsymbol{\alpha}} \widetilde{\varphi}(\boldsymbol{\mu}_t, \boldsymbol{\alpha}_t, j_t) \mid \boldsymbol{\mu}_t, \boldsymbol{\alpha}_t] = \partial_{\boldsymbol{\alpha}} \varphi(\boldsymbol{\mu}_t, \boldsymbol{\alpha}_t),$$

and $\widetilde{g}(\boldsymbol{w}_t) = (\partial_{\boldsymbol{\mu}} \widetilde{\varphi}(\boldsymbol{\mu}_t, \boldsymbol{\alpha}_t, i_t), -\partial_{\boldsymbol{\alpha}} \widetilde{\varphi}(\boldsymbol{\mu}_t, \boldsymbol{\alpha}_t, j_t))$ is an unbiased estimation of $g(\boldsymbol{w}_t)$. Algorithm 1 summarizes the pseudo-code of ssODM.

4.3 Recovering the Label Assignment

Once the saddle point $(\boldsymbol{\mu}^\star, \boldsymbol{\alpha}^\star)$ is found, we can obtain the label assignment of unlabeled instances according to $\mathrm{sign}(\sum_{k:\hat{\boldsymbol{y}}_k \in \mathcal{B}} \mu_k^\star \hat{\boldsymbol{y}}_k)$.

5 Empirical Study

In this section, we empirically evaluate the proposed method on twenty UCI data sets. Table 1 summarizes the statistics of these data sets. As can be seen, the number of instance is ranged from 62 to 72,309, and the dimensionality is ranged from 6 to 20,958, covering a broad range of properties.

ID	Data set	#Instance	#Feature
1	*echocardiogram*	62	8
2	*house*	232	16
3	*heart*	270	9
4	*heart-statlog*	270	13
5	*haberman*	306	14
6	*live-discorders*	345	6
7	*ionosphere*	351	33
8	*vehicle*	435	16
9	*house-votes*	435	16
10	*clean1*	476	166
11	*wdbc*	569	14
12	*isolet*	600	51
13	*austra*	690	15
14	*australian*	690	42
15	*diabetes*	768	8
16	*german*	1,000	59
17	*optdigits*	1,143	42
18	*krvskp*	3,196	36
19	*sick*	2,643	28
20	*real-sim*	72,309	20,958

Table 1: Characteristics of experimental data sets.

5.1 Setting

For each UCI data set, 75% of the examples are randomly chosen for training, and the rest for testing. We investigate the performance of each approach with varying amount of labeled data (namely, 5%, 10% of the labeled data). The whole setup is repeated 10 times and the average accuracies as well as the standard deviations on the test set are recorded.

We compare our method with 1) the standard SVM (using labeled data only) [Cortes and Vapnik, 1995], and four state-of-the-art S3VMs, namely 2) Transductive SVM (TSVM) [Joachims, 1999]; 3) Laplacian SVM [Belkin *et al.*, 2006]; 4)UniverSVM (USVM) [Collobert *et al.*, 2006]; and 5) S4VM [Li and Zhou, 2011]. Note that TSVM and USVM adopt the same objective but with different optimization strategies (local search and constrained convex-concave procedure, respectively), so they may converge to different local minimum.

For all the methods, the parameters C, λ_1, λ_2 are selected from $\{1, 10, 100, 1000\}$. For ssODM, ν and θ are selected from $[0.2, 0.4, 0.6, 0.8]$. For all data sets, both the linear and Gaussian kernels are used. In particular, the width σ of Gaussian kernel is picked from $\{0.25\sqrt{\gamma}, 0.5\sqrt{\gamma}, \sqrt{\gamma}, 2\sqrt{\gamma}, 4\sqrt{\gamma}\}$, where γ is the average distance between instances. All the experiments are repeated 10 times and the average performance is reported with the best parameter setting.

5.2 Performance

Table 2 summarizes the results on the twenty UCI data sets. As can be seen, for the two settings, i.e., 5% labeled data and 10% labeled data, ssODM achieves the best performance on 14 and 13 data sets, respectively. In addition, in comparing with S3VMs which do not consider margin distribution, the win/tie/loss counts show that ssODM is always better or comparable, almost never worse than S3VMs.

6 Conclusions

Semi-supervised support vector machines (S3VMs), which employs the large margin heuristic from support vector machines, have achieved more accurate results than other semi-

Data set	Label	SVM	TSVM	LapSVM	USVM	S4VM	ssODM
echocardiogram	5%	.800±.071•	.741±.082•	.644±.221•	.801±.061•	.804±.078•	**.819±.011**
	10%	.812±.077•	.761±.087•	.684±.201•	.821±.063•	.824±.073•	**.839±.015**
house	5%	.900±.041•	.903±.056•	.906±.067•	.903±.068•	.909±.073•	**.917±.014**
	10%	.912±.047•	.921±.057•	.918±.171•	.911±.053•	**.924±.066**	.923±.018
heart	5%	.700±.080•	.752±.062•	.733±.063•	.762±.063•	.772±.061•	**.783±.054**
	10%	.751±.049•	.783±.047•	.756±.041•	.779±.053•	.784±.056	**.798±.016**
heart-statlog	5%	.730±.010•	.762±.061•	.753±.068•	.781±.053•	**.791±.061**	.789±.054
	10%	.751±.008•	.792±.062•	.793±.058•	.791±.041•	**.831±.056**	.824±.045
haberman	5%	.651±.071•	.614±.053•	.577±.112•	**.743±.123**	.732±.121	.737±.141
	10%	.683±.067•	.634±.047•	.601±.098•	**.794±.113**	.787±.111	.788±.011
live-discorders	5%	.568±.051•	.555±.053•	.556±.055•	**.590±.052**	.530±.071•	.588±.048
	10%	.583±.067•	.584±.047•	.601±.088•	**.642±.103**	.590±.091•	.639±.008
ionosphere	5%	.678±.061•	**.822±.113**○	.656±.058•	.770±.064•	.701±.064•	.791±.041
	10%	.691±.057•	**.861±.047**○	.681±.058•	.791±.043•	.761±.054•	.831±.015
vehicle	5%	.748±.041•	.751±.083•	.773±.052•	.741±.069•	.789±.062	**.791±.041**
	10%	.761±.035•	.772±.062•	.791±.046•	.760±.060•	.819±.067	**.823±.035**
house-votes	5%	.888±.031•	.891±.043•	.899±.032•	.901±.053•	.912±.041	**.925±.035**
	10%	.897±.026•	.899±.031•	.901±.032•	.910±.045•	.925±.035	**.929±.031**
clean1	5%	.580±.061•	.621±.074•	.641±.065•	.623±.065•	.641±.041•	**.661±.045**
	10%	.591±.054•	.641±.054•	.649±.055•	.634±.542•	.651±.038•	**.671±.035**
wdbc	5%	.813±.064•	.803±.034•	.808±.048•	.820±.061•	.821±.048	**.829±.039**
	10%	.831±.054•	.823±.031•	.818±.044•	.825±.060•	.829±.043	**.835±.033**
isolet	5%	.970±.029•	.976±.031•	.980±.038•	.987±.037•	.988±.048	**.991±.040**
	10%	.973±.025•	.978±.030•	.985±.032•	.989±.035•	.989±.043	**.992±.045**
austra	5%	.770±.059•	**.820±.030**	.766±.033•	.781±.045•	.775±.041•	.813±.040
	5%	.790±.051•	**.850±.032**	.796±.031•	.788±.037•	.799±.036•	.843±.034
australian	5%	.672±.081•	.681±.034•	.782±.031•	.789±.041•	.788±.042•	**.799±.035**
	10%	.681±.075•	.692±.038•	.791±.030•	.792±.040•	.797±.041•	**.805±.039**
diabetes	5%	.679±.080•	.683±.039•	.761±.069•	.771±.049•	.768±.048•	**.790±.049**
	10%	.703±.071•	.703±.034•	.770±.063•	.776±.040•	.771±.044•	**.798±.040**
german	5%	.700±.030•	.703±.010•	.708±.021•	.710±.029•	.715±.041•	**.726±.045**
	10%	.700±.030•	.708±.010•	.709±.016•	.713±.023•	.718±.043•	**.731±.037**
optdigits	5%	**.922±.020**	.891±.090•	.902±.050•	.913±.069•	.918±.041	.919±.042
	10%	**.925±.023**	.894±.091•	.912±.055•	.917±.064•	.913±.040	.921±.040
krvskp	5%	.911±.023•	.899±.091•	.921±.054•	.916±.068•	.926±.042•	**.932±.040**
	10%	.919±.020•	.903±.090•	.927±.050•	.925±.069•	.929±.040•	**.936±.045**
sick	5%	.941±.021	.932±.090•	.939±.034•	.935±.048•	.929±.045•	**.948±.048**
	10%	.946±.020	.938±.088•	.941±.033•	.939±.041•	.934±.041•	**.955±.043**
real-sim	5%	.901±.022•	.913±.065•	.922±.035•	.930±.043•	.924±.049•	**.922±.033**
	10%	.909±.020•	.919±.062•	.929±.031•	.933±.040•	.928±.043•	**.941±.028**
ssODM: w/t/l	5%	18/2/0	18/1/1	20/0/0	18/2/0	13/7/0	
	10%	18/2/0	18/1/1	20/0/0	18/2/0	11/9/0	

Table 2: Accuracies on the various data sets with 5% and 10% labeled instances on twenty UCI data sets. The best performance on each data set is bolded. •/○ indicates ssODM is significantly better/worse than compared methods (paired t-tests at 95% significance level). The win/tie/loss counts for ssODM are summarized in the last two rows.

supervised methods. Recent studies disclosed that instead of minimum margin, it is more crucial to optimize the margin distribution for SVM-style learning algorithms. Inspired by this recognition, we propose a novel approach ssODM for semi-supervised learning by optimizing the margin distribution. To conquer the resultant saddle point problem, we extend a stochastic mirror prox method which can converge more quickly in practise than general sub-gradient descent for non-smooth problem. Experimental results in various data sets show that our method achieves promising performance, which further verifies the superiority of optimal margin distribution learning. In the future, we will apply importance sampling [Schmidt *et al.*, 2015] and variance reduction techniques [Johnson and Zhang, 2013] to further accelerate our method and extend it to other learning settings, i.e., multi-instance multi-label learning.

References

[Beck and Teboulle, 2003] A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

[Belkin *et al.*, 2006] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, pages 2399–2434, 2006.

[Bennett and Demiriz, 1999] K. P. Bennett and A. Demiriz. Semi-supervised support vector machines. In *Proceedings of the 1998 Conference on Advances in Neural Information Processing Systems II*, pages 368–374, 1999.

[Chapelle and Zien, 2005] O. Chapelle and A. Zien. Semi-supervised classification by low density separation. In *Proceedings of the 8th International Conference on Artificial Intelligence and Statistics*, pages 57–64, 2005.

[Collobert *et al.*, 2006] R. Collobert, F. Sinz, J. Weston, and L. Bottou. Large scale transductive svms. *Journal of Machine Learning Research*, 7:1687–1712, 2006.

[Cortes and Vapnik, 1995] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[Cristianini and Shawe-Taylor, 2000] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, Cambridge, UK, 2000.

[Gao and Zhou, 2013] W. Gao and Z.-H. Zhou. On the doubt about margin explanation of boosting. *Artificial Intelligence*, 203:1–18, 2013.

[Joachims, 1999] T. Joachims. Transductive inference for text classification using support vector machines. pages 200–209. Morgan Kaufmann, 1999.

[Johnson and Zhang, 2013] R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 315–323. Curran Associates, Inc., 2013.

[Kim and Boyd, 2008] Seung-Jean Kim and Stephen Boyd. A minimax theorem with applications to machine learning, signal processing, and finance. *SIAM Journal on Optimization*, 19(3):1344–1367, 2008.

[Li and Zhou, 2011] Y.-F. Li and Z.-H. Zhou. Towards making unlabeled data never hurt. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pages 1081–1088, 2011.

[Li *et al.*, 2009a] Y.-F. Li, James T. Kwok, and Z.-H. Zhou. Semi-supervised learning using label mean. In *International Conference on Machine Learning*, pages 633–640, 2009.

[Li *et al.*, 2009b] Y.-F. Li, Ivor W. Tsang, James T. Kwok, and Z.-H. Zhou. Tighter and convex maximum margin clustering. In *In Proceeding of the 12th International Conference on Artificial Intelligence and Statistics*, volume 5, pages 344–351, 2009.

[Li *et al.*, 2010] Y.-F. Li, J. T. Kwok, and Z.-H. Zhou. Cost-sensitive semi-supervised support vector machine. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence*, pages 500–505. AAAI Press, 2010.

[Li *et al.*, 2013] Yu-Feng Li, Ivor W. Tsang, James T. Kwok, and Zhi-Hua Zhou. Convex and scalable weakly labeled svms. *Journal of Machine Learning Research*, 14(1):2151–2188, jan 2013.

[Nemirovski, 2005] A. Nemirovski. *Prox-Method with Rate of Convergence O(1/t) for Variational Inequalities with Lipschitz Continuous Monotone Operators and Smooth Convex-Concave Saddle Point Problems*. Society for Industrial and Applied Mathematics, 2005.

[Schmidt *et al.*, 2015] Mark Schmidt, Reza Babanezhad, Mohamed Ahmed, Aaron Defazio, Ann Clifton, and Anoop Sarkar. Non-uniform stochastic average gradient method for training conditional random fields. In Guy Lebanon and S. V. N. Vishwanathan, editors, *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pages 819–828, San Diego, CA, 09–12 May 2015. PMLR.

[Schölkopf and Smola, 2001] B. Schölkopf and A. J. Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT Press, Cambridge, MA, 2001.

[Sion, 1958] Maurice Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958.

[Xu *et al.*, 2005] L. Xu, J. Neufeld, B. Larson, and D. Schuurmans. Maximum margin clustering. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 1537–1544. MIT Press, 2005.

[Zhang and Zhou, 2014] T. Zhang and Z.-H. Zhou. Large margin distribution machine. In *Proceedings of the 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 313–322, New York, NY, 2014.

[Zhang and Zhou, 2017] T. Zhang and Z.-H. Zhou. Multi-class optimal distribution machine. In *Proceedings of the 34th International Conference on Machine Learning*, pages 4063–4071, Sydney, NSW, Australia, 2017.

[Zhang and Zhou, 2018] T. Zhang and Z.-H. Zhou. Optimal margin distribution clustering. In *Proceedings of the 20th National Conference on Artificial Intelligence*, AAAI'18. AAAI Press, 2018.

[Zhou, 2014] Z.-H. Zhou. Large margin distribution learning. In *Proceedings of the 6th IAPR International Workshop on Artificial Neural Networks in Pattern Recognition*, pages 1–11, Montreal, Canada, 2014.