Hindawi Publishing Corporation
Journal of Applied Mathematics
Volume 2014, Article ID 942520, 16 pages
http://dx.doi.org/10.1155/2014/942520

Research Article

An Interior Point Method for $L_{1/2}$-SVM and Application to Feature Selection in Classification

Lan Yao,[1] Xiongji Zhang,[2] Dong-Hui Li,[2] Feng Zeng,[3] and Haowen Chen[1]

[1] *College of Mathematics and Econometrics, Hunan University, Changsha 410082, China*
[2] *School of Mathematical Sciences, South China Normal University, Guangzhou 510631, China*
[3] *School of Software, Central South University, Changsha 410083, China*

Correspondence should be addressed to Dong-Hui Li; dhli@scnu.edu.cn

Received 4 November 2013; Revised 12 February 2014; Accepted 18 February 2014; Published 10 April 2014

Academic Editor: Frank Werner

This paper studies feature selection for support vector machine (SVM). By the use of the $L_{1/2}$ regularization technique, we propose a new model $L_{1/2}$-SVM. To solve this nonconvex and non-Lipschitz optimization problem, we first transform it into an equivalent quadratic constrained optimization model with linear objective function and then develop an interior point algorithm. We establish the convergence of the proposed algorithm. Our experiments with artificial data and real data demonstrate that the $L_{1/2}$-SVM model works well and the proposed algorithm is more effective than some popular methods in selecting relevant features and improving classification performance.

1. Introduction

Feature selection plays an important role in solving the classification problems with high dimension features, such as text categorization [1, 2], gene expression array analysis [3–5], and combinatorial chemistry [6, 7]. The advantages of feature selection include (i) ignoring noisy or irrelevant features would prevent overfitting and improve the generalization performance; (ii) a sparse classifier can reduce the computation cost; (iii) a small set of important features is desirable for interpretability.

We address the embedded feature selection methods in the context of linear support vector machines (SVMs). Existing feature selection methods embedded in SVMs fall into three approaches [8]. In the first approach, some greedy search strategies are applied to iteratively adding or removing features from the data. Guyon et al. [3] developed a recursive feature elimination (RFE) algorithm, which has shown good performance on gene selection for microarray data. Beginning with the full feature subset, SVM-RFE trains a SVM at each iteration, and then eliminates the feature that decreases the margin the least. Rakotomamonjy et al. [9] extended this method by using other ranking criteria including the radius margin bound and the span-estimate.

The second approach is to optimize a scaling parameter vector $\sigma \in [0,1]^n$ that indicates the importance of each feature. Weston et al. [10] proposed an iterative method to optimize the scaling parameters by minimizing the bounds on leave-one-out error. Peleg and Meir [11] learned the scaling factors based on the global minimization of a data-dependent generalization error bound.

The third category of approaches is to minimize the number of features by adding a sparsity term to the SVM formulation. Though standard SVM based on $\|w\|_2$ can be solved easily by convex quadratic programming, its solution may not be a desirable sparse solution. A popular way to deal with this problem is the use of L_p regularization technique, which results in a L_p-SVM. It is to minimize $\|w\|_p$, subject to some linear constraints, where $p \in [0,1]$. When $p \in (0,1]$,

$$\|w\|_p = \left(\sum_{j=1}^{m} |w_i|^p \right)^{1/p}. \tag{1}$$

When $p = 0$, $\|w\|_0 = \sum_{j=1}^{m} I_{(w_j \neq 0)}$. The L_0-SVM can find the sparsest classifier by minimizing $\|w\|_0$, the number of nonzero elements in w. However, it is discrete and NP-hard. From computational point of view, it is very difficult

to develop efficient numerical methods to solve the problem. A widely used technique in dealing with the L_0-SVM is to use a smoothing technique so that the discrete model is approached by a smooth problem [4, 12, 13]. However, as the function $\|w\|_0$ is not even continuous, it is not desirable that a smoothing technique based method would work well. Chan et al. [14] explored a convex relaxation to the cardinality constraint and obtained a relaxed convex problem that is close to but different from the previous L_0-SVM. An alternative method is to minimize the convex envelope of the $\|w\|_0$, such as L_1-SVM. The L_1-SVM is a convex problem and can yield sparse solution. It can be equivalent to a linear programming and hence can be solved efficiently. Indeed the L_1 regularization has become quite welcome in SVM [12, 15, 16] and is well known as the LASSO [17] in the statistics literature. However, the L_1 regularization problem often leads to suboptimal sparsity in reality [18]. In many cases, the solutions yielded from L_1-SVM are less sparse than those of L_0-SVM. The L_p problem with $0 < p < 1$ can find sparser solutions than the L_1 problem, which was evidenced in extensive computational [19–21]. It has become a welcome strategy in sparse SVM [22–27].

In this paper, we focus on the $L_{1/2}$ regularization and propose a novel $L_{1/2}$-SVM. Recently, Xu et al. [28] justified that the sparsity-promotion ability of the $L_{1/2}$ problem was strongest among the L_p minimization problems with all $p \in [1/2, 1)$ and similar in $p \in (0, 1/2]$. So the $L_{1/2}$ problem can be taken as a representative of L_p $(0 < p < 1)$ problems. However, as proved by Ge et al. [29], finding the global minimal value of the $L_{1/2}$ problem was still strongly NP-hard. But computing a local minimizer of the problem could be done in polynomial time. Our contributions of this paper are twofold. One is to derive a smooth constrained optimization reformulation to the $L_{1/2}$-SVM. The objective function of the problem is a linear function and the constraints are quadratic and linear. We will establish the equivalence between the constrained problem and the $L_{1/2}$-SVM. We will also show the existence of the KKT condition of the constrained problem. Our second contribution is to develop an interior point method to solve the constrained optimization reformulation and establish its global convergence. We will also test and verify the effectiveness of the proposed method using artificial data and real data.

The rest of this paper is organized as follows. In Section 2, we first briefly introduce the model of the standard SVM (L_2-SVM) and the sparse regularization SVMs. We then reformulate the $L_{1/2}$-SVM into a smooth constrained optimization problem. We propose an interior point method to solve the constrained optimization reformulation and establish its global convergence in Section 3. In Section 4, we do numerical experiments to test the proposed method. Section 5 gives the conclusive remarks.

2. A Smooth Constrained Optimization Reformulation to the $L_{1/2}$-SVM

In this section, after simply reviewing the model of the standard SVM (L_2-SVM) and the sparse regularization SVMs, we derive an equivalent smooth optimization problem to the $L_{1/2}$-SVM model. The smooth optimization problem is to minimize a linear function subject to some simple quadratic constraints and linear constraints.

2.1. Standard SVM. In a two-class classification problem, we are given a training data set $D = (x_i; y_i)_{i=1}^n$, where $x_i \in R^m$ is the feature vector and $y_i \in \{0, 1\}$ is the class label. The linear classifier is to construct the following decision function:

$$f(x) = w^T x + b, \tag{2}$$

where $w = (w_1, w_2, \ldots, w_m)$ is the weight vector and b is the bias. The prediction label is +1 if $f(x) > 0$ and −1 otherwise. The standard SVM (L_2-SVM) [30] aims to find the separating hyperplane $f(x) = 0$ between two classes with maximal margin $2/\|\mathbf{w}\|_2^2$ and minimal training errors, which leads to the following convex optimization problem:

$$\begin{aligned} \min \quad & \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n}\xi_i \\ \text{s.t.} \quad & y_i\left(w^T x_i + b\right) \geq 1 - \xi_i \quad i = 1, \ldots, n,\ \xi_i \geq 0, \end{aligned} \tag{3}$$

where $\|w\|_2 = (\sum_{j=1}^m w_j^2)^{1/2}$ is the L_2 norm of w, ξ_i is the loss function to allow training errors for data that may not be linearly separable, and C is a user-specified parameter to balance the margin and the losses. As the problem is a convex quadratic program, it can be solved by existing methods, such as the interior point method and active set method efficiently.

2.2. Sparse SVM. The L_2-SVM is a nonsparse regularizer in the sense that the learned decision hyperplane often utilizes all the features. In practice, peoples prefer to sparse SVM so that only a few features are used to make a decision. For this purposes, the following L_p-SVM becomes very welcome:

$$\begin{aligned} \min \quad & \|w\|_p^p + C\sum_{i=1}^{n}\xi_i \\ \text{s.t.} \quad & y_i\left(w^T x_i + b\right) \geq 1 - \xi_i \quad i = 1, \ldots, n,\ \xi_i \geq 0, \end{aligned} \tag{4}$$

where $p \in [0, 1]$. $\|w\|_0$ stands for the number of nonzero elements of w, and for $p \in (0, 1]$, $\|w\|_p$ is defined by (1). Problem (4) is obtained by replacing L_2 penalty ($\|w\|_2^2$) by L_p penalty ($\|w\|_p^p$) in (3). The standard SVM (3) corresponds to the model (4) with $p = 2$.

Figure 1 plots the L_p penalty in one dimension. We can see from the figure that the smaller p is, the larger penalties are imposed on the small coefficients ($|w| < 1$). Therefore, the L_p penalties with $p < 1$ may achieve sparser solution than the L_1 penalty. In addition, the L_1 imposes large penalties on large coefficients, which may lead to biased estimation for large coefficients. Consequently, the L_p $(0 < p < 1)$ penalties become attractive due to their good properties in sparsity, unbiasedness [31] and oracle [32]. We are particularly interested in the $L_{1/2}$ penalty. Recently, Xu et al. [21] revealed the representative role of the $L_{1/2}$ penalty in the L_p regularization with $p \in (0, 1)$. We will apply $L_{1/2}$ penalty to SVM to perform feature selection and classification jointly.

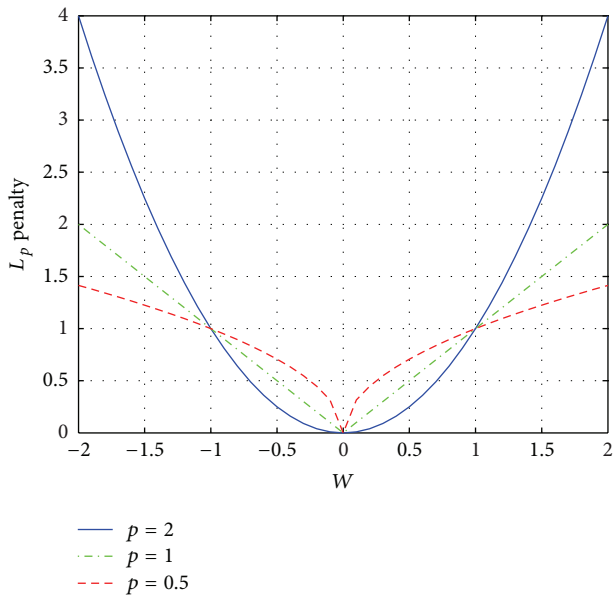

FIGURE 1: L_p penalty in one dimension.

2.3. $L_{1/2}$-SVM Model. We pay particular attention to the $L_{1/2}$-SVM, namely, problem (4) with $p = 1/2$. We will derive a smooth constrained optimization reformulation to the $L_{1/2}$-SVM so that it is relatively easy to design numerical methods. We first specify the $L_{1/2}$-SVM:

$$\begin{aligned} \min \quad & \sum_{j=1}^{m} |w_j|^{1/2} + C\sum_{i=1}^{n}\xi_i \triangleq \phi(w,\xi,b) \\ \text{s.t.} \quad & y_i\left(w^T x_i + b\right) \geq 1 - \xi_i, \quad i = 1,\ldots,n, \\ & \xi_i \geq 0, \quad i = 1,\ldots,n. \end{aligned} \tag{5}$$

Denote by $u = (w,\xi,b)$ and $\mathscr{D}$ the feasible region of the problem; that is,

$$\begin{aligned} \mathscr{D} = \{(w,\xi,b) \mid & \ y_i\left(w^T x_i + b\right) \geq 1 - \xi_i, \\ & \xi_i \geq 0, \ i = 1,\ldots,n\}. \end{aligned} \tag{6}$$

Then the $L_{1/2}$-SVM can be written as an impact form

$$\min \phi(u), \quad u \in \mathscr{D}. \tag{7}$$

It is a nonconvex and non-Lipschitz problem. Due to the existence of the term $|w_i|^{1/2}$, the objective function is not even directionally differentiable at a point with some $w_i = 0$, which makes the problem very difficult to solve. Existing numerical methods that are very efficient for solving smooth problem could not be used directly. One possible way to develop numerical methods for solving (7) is to smoothing the term $|w_j|^{1/2}$ using some smoothing function such as $\phi_\epsilon(w_j) = (w_j^2 + \epsilon^2)^{1/4}$ with some $\epsilon > 0$. However, it is easy to see that the derivative of $\phi_\epsilon(w_j)$ will be unbounded as $w_j \to 0$ and $\epsilon \to 0$. Consequently, it is not desirable that the smoothing function based numerical methods could work well.

Recently, Tian and Yang [33] proposed an interior point $L_{1/2}$-penalty function method to solve general nonlinear programming problems by using a quadratic relaxation scheme for their $L_{1/2}$-lower order penalty problems. We will follow the idea of [33] to develop an interior point method for solving the $L_{1/2}$-SVM. To this end, in the next subsection, we reformulate problem (7) to a smooth constrained optimization problem.

2.4. A Reformulation to the $L_{1/2}$-SVM Model. Consider the following constrained optimization problem:

$$\begin{aligned} \min \quad & \sum_{j=1}^{m} t_j + C\sum_{i=1}^{n}\xi_i \triangleq f(w,\xi,t,b) \\ \text{s.t.} \quad & t_j^2 - w_j \geq 0, \quad j = 1,\ldots,m, \\ & t_j^2 + w_j \geq 0, \quad j = 1,\ldots,m, \\ & t_j \geq 0, \quad j = 1,\ldots,m, \\ & y_i\left(w^T x_i + b\right) \geq 1 - \xi_i, \quad i = 1,\ldots,n, \\ & \xi_i \geq 0, \quad i = 1,\ldots,n. \end{aligned} \tag{8}$$

It is obtained by letting $t_j = |w_j|^{1/2}$ in the objective function and adding constraints $t_j^2 - w_j \geq 0$ and $t_j^2 + w_j \geq 0$, $j = 1,\ldots,m$, in (7). Denote by $\mathscr{F}$ the feasible region of the problem; that is,

$$\begin{aligned} \mathscr{F} = \{(w,\xi,t,b) \mid & \ t_j^2 - w_j \geq 0, \ t_j^2 + w_j \geq 0, \\ & t_j \geq 0, \ j = 1,\ldots,m\} \\ \cap \{(w,\xi,t,b) \mid & \ y_i\left(w^T x_i + b\right) \geq 1 - \xi_i, \\ & \xi_i \geq 0, \ i = 1,\ldots,n\}. \end{aligned} \tag{9}$$

Let $z = (w,\xi,t,b)$. Then the above problem can be written as

$$\min f(z), \quad z \in \mathscr{F}. \tag{10}$$

The following theorem establishes the equivalence between the $L_{1/2}$-SVM and (10).

Theorem 1. *If $u^* = (w^*,\xi^*,b^*) \in R^{m+n+1}$ is a solution of the $L_{1/2}$ − SVM (7), then $z^* = (w^*,\xi^*,|w^*|^{1/2},b^*) \in R^{m+n+m+1}$ is a solution of the optimization problem (10). Conversely, if (w^*,ξ^*,t^*,b^*) is a solution of the optimization problem (10), then (w^*,ξ^*,b^*) is a solution of the $L_{1/2}$-SVM (7).*

Proof. Let $u^* = (w^*,\xi^*,b^*)$ be a solution of the $L_{1/2}$-SVM (7) and let $\bar{z} = (\bar{w},\bar{\xi},\bar{t},\bar{b})$ be a solution of the constrained

optimization problem (10). It is clear that $z^* = (w^*, \xi^*, |w^*|^{1/2}, b^*) \in \mathscr{F}$. Moreover, we have $\bar{t}_j^2 \geq |\bar{w}_j|$, $\forall j = 1, 2, \ldots, m$, and hence

$$\begin{aligned} f(\bar{z}) &= \sum_{j=1}^{m} \bar{t}_j + C \sum_{i=1}^{n} \bar{\xi}_i \geq \sum_{j=1}^{m} \left|\bar{w}_j\right|^{1/2} + C \sum_{i=1}^{n} \bar{\xi}_i \\ &\geq \sum_{j=1}^{m} \left|w_j^*\right|^{1/2} + C \sum_{i=1}^{n} \xi_i^* = \phi(u^*). \end{aligned} \tag{11}$$

Since $z^* \in \mathscr{F}$, we have

$$\phi(u^*) = f(z^*) \geq f(\bar{z}). \tag{12}$$

This together with (11) implies that $f(\bar{z}) = \phi(u^*)$. The proof is complete. □

It is clear that the constraint functions of (10) are convex. Consequently, at any feasible point, the set of all feasible directions is the same as the set of all linearized feasible directions.

As a result, the KKT point exists. The KKT system of the problem (10) can be written as the following system of nonlinear equations:

$$R(w, \xi, t, b, \lambda) = \begin{pmatrix} \lambda^{(1)} - \lambda^{(2)} - X^T Y^T \lambda^{(4)} \\ C * e_n - \lambda^{(4)} - \lambda^{(5)} \\ e_m - 2T\lambda^{(1)} - 2T\lambda^{(2)} - \lambda^{(3)} \\ y^T \lambda^{(4)} \\ \min\left\{t_j^2 - w_j, \lambda_j^{(1)}\right\}_{j=1,2,\ldots,m} \\ \min\left\{t_j^2 + w_j, \lambda_j^{(2)}\right\}_{j=1,2,\ldots,m} \\ \min\left\{t_j, \lambda_j^{(3)}\right\}_{j=1,2,\ldots,m} \\ \min\left\{p_i, \lambda_i^{(4)}\right\}_{i=1,2,\ldots,n} \\ \min\left\{\xi_i, \lambda_i^{(5)}\right\}_{i=1,2,\ldots,n} \end{pmatrix} = 0, \tag{13}$$

where $\lambda = (\lambda^{(1)}, \lambda^{(2)}, \lambda^{(3)}, \lambda^{(4)}, \lambda^{(5)})$ are the Lagrangian multipliers, $X = (x_1, x_2, \ldots, x_n)^T$, $Y = \mathrm{diag}(y)$ is diagonal matrix and $p_i = y_i(w^T x_i + b) - (1 - \xi_i)$, $i = 1, 2, \ldots, n$.

For the sake of simplicity, the properties of the reformulation to $L_{1/2}$-SVM are shown in Appendix A.

3. An Interior Point Method

In this section, we develop an interior point method to solve the equivalent constrained problem (10) of the $L_{1/2}$-SVM (7).

3.1. Auxiliary Function. Following the idea of the interior point method, the constrained problem (10) can be solved by minimizing a sequence of logarithmic barrier functions as follows:

$$\begin{aligned} \min \quad & \Phi_\mu(w, \xi, t, b) \\ & \triangleq \sum_{j=1}^{m} t_j + C \sum_{i=1}^{n} \xi_i \\ & - \mu \sum_{j=1}^{m} \left[\log\left(t_j^2 - w_j\right) + \log\left(t_j^2 + w_j\right) + \log t_j\right] \\ & - \mu \sum_{i=1}^{n} (\log p_i + \log \xi_i) \\ \text{s.t.} \quad & t_j^2 - w_j > 0, \quad j = 1, 2, \ldots, m, \\ & t_j^2 + w_j > 0, \quad j = 1, 2, \ldots, m, \\ & t_j > 0, \quad j = 1, 2, \ldots, m, \\ & p_i = y_i\left(w^T x_i + b\right) + \xi_i - 1 > 0, \quad i = 1, 2, \ldots, n, \\ & \xi_i > 0, \quad i = 1, 2, \cdots, n, \end{aligned} \tag{14}$$

where μ is the barrier parameter, converging to zero from above.

The KKT system of problem (14) is the following system of linear equations:

$$\begin{aligned} \lambda^{(1)} - \lambda^{(2)} - X^T Y^T \lambda^{(4)} &= 0, \\ C * e_n - \lambda^{(4)} - \lambda^{(5)} &= 0, \\ e_m - 2T\lambda^{(1)} - 2T\lambda^{(2)} - \lambda^{(3)} &= 0, \\ y^T \lambda^{(4)} &= 0, \\ \left(T^2 - W\right)\lambda^{(1)} - \mu e_m &= 0, \\ \left(T^2 + W\right)\lambda^{(2)} - \mu e_m &= 0, \\ T\lambda^{(3)} - \mu e_m &= 0, \\ P\lambda^{(4)} - \mu e_n &= 0, \\ \Xi\lambda^{(5)} - \mu e_n &= 0, \end{aligned} \tag{15}$$

where $\lambda = (\lambda^{(1)}, \lambda^{(2)}, \lambda^{(3)}, \lambda^{(4)}, \lambda^{(5)})$ are the Lagrangian multipliers, $T = \mathrm{diag}(t)$, $W = \mathrm{diag}(w)$, $\Xi = \mathrm{diag}(\xi)$, and $P = \mathrm{diag}(Y(Xw + b * e_n) + \xi - e_n)$ are diagonal matrices, and $e_m \in R^m$ and $e_n \in R^n$ stand for the vector whose elements are all ones.

3.2. Newton's Method. We apply Newton's method to solve the nonlinear system (15) in variables w, ξ, t, b, and λ. The subproblem of the method is the following system of linear equations:

$$M = \begin{pmatrix} \Delta w \\ \Delta \xi \\ \Delta t \\ \Delta b \\ \Delta\lambda^{(1)} \\ \Delta\lambda^{(2)} \\ \Delta\lambda^{(3)} \\ \Delta\lambda^{(4)} \\ \Delta\lambda^{(5)} \end{pmatrix} = - \begin{pmatrix} \lambda^{(1)} - \lambda^{(2)} - X^T Y^T \lambda^{(4)} \\ C * e_n - \lambda^{(4)} - \lambda^{(5)} \\ e_m - 2T\lambda^{(1)} - 2T\lambda^{(2)} - \lambda^{(3)} \\ e_n^T Y \lambda^{(4)} \\ \left(T^2 - W\right)\lambda^{(1)} - \mu e_m \\ \left(T^2 + W\right)\lambda^{(2)} - \mu e_m \\ T\lambda^{(3)} - \mu e_m \\ P\lambda^{(4)} - \mu e_n \\ \Xi\lambda^{(5)} - \mu e_n \end{pmatrix}, \tag{16}$$

where M is the Jacobian of the left function in (15) and takes the form

$$M = \begin{pmatrix} 0 & 0 & 0 & 0 & I & -I & 0 & -X^T Y^T & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -I & -I \\ 0 & 0 & -2(D_1 + D_2) & 0 & -2T & -2T & -I & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & y^T & 0 \\ -D_1 & 0 & 2TD_1 & 0 & T^2 - W & 0 & 0 & 0 & 0 \\ D_2 & 0 & 2TD_2 & 0 & 0 & T^2 + W & 0 & 0 & 0 \\ 0 & 0 & D_3 & 0 & 0 & 0 & T & 0 & 0 \\ D_4 YX & D_4 & 0 & D_4 y & 0 & 0 & 0 & P & 0 \\ 0 & D_5 & 0 & 0 & 0 & 0 & 0 & 0 & \Xi \end{pmatrix}, \tag{17}$$

where $D_1 = \operatorname{diag}(\lambda^{(1)})$, $D_2 = \operatorname{diag}(\lambda^{(2)})$, $D_3 = \operatorname{diag}(\lambda^{(3)})$, $D_4 = \operatorname{diag}(\lambda^{(4)})$, and $D_5 = \operatorname{diag}(\lambda^{(5)})$.

We can rewrite (16) as

$$\begin{aligned} &\left(\lambda^{(1)} + \Delta\lambda^{(1)}\right) - \left(\lambda^{(2)} + \Delta\lambda^{(2)}\right) - X^T Y^T \left(\lambda^{(4)} + \Delta\lambda^{(4)}\right) = 0, \\ &\left(\lambda^{(4)} + \Delta\lambda^{(4)}\right) + \left(\lambda^{(5)} + \Delta\lambda^{(5)}\right) = C * e_n, \\ &2(D_1 + D_2)\Delta t + 2T\left(\lambda^{(1)} + \Delta\lambda^{(1)}\right) \\ &\quad + 2T\left(\lambda^{(2)} + \Delta\lambda^{(2)}\right) + \left(\lambda^{(3)} + \Delta\lambda^{(3)}\right) = e_m, \\ &y^T\left(\lambda^{(4)} + \Delta\lambda^{(4)}\right) = 0, \\ &-D_1\Delta w + 2TD_1\Delta t + \left(T^2 - W\right)\left(\lambda^{(1)} + \Delta\lambda^{(1)}\right) = \mu e_m, \\ &D_2\Delta w + 2TD_2\Delta t + \left(T^2 + W\right)\left(\lambda^{(2)} + \Delta\lambda^{(2)}\right) = \mu e_m, \\ &D_3\Delta t + T\left(\lambda^{(3)} + \Delta\lambda^{(3)}\right) = \mu e_m, \\ &D_4 YX\Delta w + D_4\Delta\xi + D_4 y * \Delta b + P\left(\lambda^{(4)} + \Delta\lambda^{(4)}\right) = \mu e_n, \\ &D_5\Delta\xi + \Xi\left(\lambda^{(5)} + \Delta\lambda^{(5)}\right) = \mu e_n. \end{aligned} \tag{18}$$

It follows from the last five equations that vector $\widehat{\lambda} \triangleq \lambda + \Delta\lambda$ can be expressed as

$$\begin{aligned} \widehat{\lambda}^{(1)} &= \left(T^2 - W\right)^{-1}\left(\mu e_m + D_1\Delta w - 2TD_1\Delta t\right), \\ \widehat{\lambda}^{(2)} &= \left(T^2 + W\right)^{-1}\left(\mu e_m - D_2\Delta w - 2TD_2\Delta t\right), \\ \widehat{\lambda}^{(3)} &= T^{-1}\left(\mu e_m - D_3\Delta t\right), \\ \widehat{\lambda}^{(4)} &= P^{-1}\left(\mu e_n - D_4 YX\Delta w - D_4\Delta\xi - D_4 y * \Delta b\right), \\ \widehat{\lambda}^{(5)} &= \Xi^{-1}\left(\mu e_n - D_5\Delta\xi\right). \end{aligned} \tag{19}$$

Substituting (19) into the first four equations of (18), we obtain

$$\begin{aligned} &S\begin{pmatrix} \Delta w \\ \Delta\xi \\ \Delta t \\ \Delta b \end{pmatrix} \\ &= \begin{pmatrix} -\mu\left(T^2 - W\right)^{-1} e_m + \mu\left(T^2 + W\right)^{-1} e_m \\ +X^T Y^T P^{-1} e_n \\ -C * e_n + \mu P^{-1} e_n + \mu\Xi^{-1} e_n \\ -e_m + 2\mu T\left(T^2 - W\right)^{-1} e_m \\ +2\mu T\left(T^2 + W\right)^{-1} e_m + \mu T^{-1} e_m \\ \mu y^T P^{-1} e_n \end{pmatrix}, \end{aligned} \tag{20}$$

where matrix S takes the form

$$S = \begin{pmatrix} S_{11} & S_{12} & S_{13} & S_{14} \\ S_{21} & S_{22} & S_{23} & S_{24} \\ S_{31} & S_{32} & S_{33} & S_{34} \\ S_{41} & S_{42} & S_{43} & S_{44} \end{pmatrix}, \tag{21}$$

with blocks

$$\begin{aligned} &S_{11} = U + V + X^T Y^T P^{-1} D_4 YX, \\ &S_{12} = X^T Y^T P^{-1} D_4, \\ &S_{13} = -2(U - V)T, \qquad S_{14} = X^T Y^T P^{-1} D_4 y, \\ &S_{21} = P^{-1} D_4 YX, \\ &S_{22} = P^{-1} D_4 + \Xi^{-1} D_5, \\ &S_{23} = 0, \qquad S_{24} = P^{-1} D_4 y, \end{aligned}$$

$$S_{31} = -2(U-V)T, \qquad S_{32} = 0,$$

$$S_{33} = 4T(U+V)T + T^{-1}D_3 - 2(D_1+D_2), \qquad S_{34} = 0,$$

$$S_{41} = y^T P^{-1} D_4 YX, \qquad S_{42} = y^T P^{-1} D_4,$$

$$S_{43} = 0, \qquad S_{44} = y^T P^{-1} D_4 y, \tag{22}$$

and $U = (T^2 - W)^{-1} D_1$ and $V = (T^2 + W)^{-1} D_2$.

3.3. The Interior Pointer Algorithm. Let $z = (w, \xi, t, b)$ and $\lambda = (\lambda^{(1)}, \lambda^{(2)}, \lambda^{(3)}, \lambda^{(4)}, \lambda^{(5)})$; we first present the interior pointer algorithm to solve the barrier problem (14), and then discuss the details of the algorithm.

Algorithm 2. The interior pointer algorithm (IPA) is as follows.

Step 0. Given tolerance ϵ_μ, set $\tau_1 \in (0, (1/2))$, $l > 0$, $\gamma_1 > 1$, $\overline{\beta} \in (0,1)$. Let $k = 0$.

Step 1. Stop if KKT condition (15) holds.

Step 2. Compute Δz^k from (20) and $\widehat{\lambda}^{k+1}$ from (19). Compute $z^{k+1} = z^k + \alpha_k \Delta z^k$. Update the Lagrangian multipliers to obtain λ^{k+1}.

Step 3. Let $k := k + 1$. Go to Step 1.

In Step 2, a step length α^k is used to calculate z^{k+1}. We estimate α^k by Armijo line search [34], in which $\alpha^k = \max\{\bar{\beta}^j \mid j = 0, 1, 2, \ldots\}$ for some $\bar{\beta} \in (0,1)$ and satisfies the following inequalities:

$$\left(t^{k+1}\right)^2 - w^{k+1} > 0,$$

$$\left(t^{k+1}\right)^2 + w^{k+1} > 0,$$

$$t^{k+1} > 0,$$

$$\begin{aligned}
&\Phi_\mu\left(z^{k+1}\right) - \Phi_\mu\left(z^k\right) \\
&\quad \le \tau_1 \alpha_k \Big(\nabla w \Phi_\mu \left(z^k\right)^T \Delta w^k \\
&\qquad + \nabla_t \Phi_\mu \left(z^k\right)^T \Delta t^k \\
&\qquad + \nabla_\xi \Phi_\mu \left(z^k\right)^T \Delta \xi^k + \nabla_b \Phi_\mu \left(z^k\right)^T \Delta b^k \Big),
\end{aligned} \tag{23}$$

where $\tau_1 \in (0, 1/2)$.

To avoid ill-conditioned growth of $\widehat{\lambda}^k$ and guarantee the strict dual feasibility, the Lagrangian multipliers λ^k should be sufficiently positive and bounded from above. Following a similar idea of [33], we first update the dual multipliers by

$$\bar{\lambda}_i^{(1)(k+1)} = \begin{cases} \min\left\{l, \dfrac{\mu}{\left(t_i^k\right)^2 - w_i^k}\right\}, & \text{if } \widehat{\lambda}_i^{(1)(k+1)} < \min\left\{l, \dfrac{\mu}{\left(t_i^k\right)^2 - w_i^k}\right\}, \\ \widehat{\lambda}_i^{(1)(k+1)}, & \text{if } \min\left\{l, \dfrac{\mu}{\left(t_i^k\right)^2 - \mathbf{x}_i^k}\right\} \le \widehat{\lambda}_i^{(1)(k+1)} \le \dfrac{\mu\gamma_1}{\left(t_i^k\right)^2 - w_i^k}, \\ \dfrac{\mu\gamma_1}{\left(t_i^k\right)^2 - w_i^k}, & \text{if } \widehat{\lambda}_i^{(1)(k+1)} > \dfrac{\mu\gamma_1}{\left(t_i^k\right)^2 - w_i^k}, \end{cases}$$

$$\bar{\lambda}_i^{(2)(k+1)} = \begin{cases} \min\left\{l, \dfrac{\mu}{\left(t_i^k\right)^2 + w_i^k}\right\}, & \text{if } \widehat{\lambda}_i^{(2)(k+1)} < \min\left\{l, \dfrac{\mu}{\left(t_i^k\right)^2 + w_i^k}\right\}, \\ \widehat{\lambda}_i^{(2)(k+1)}, & \text{if} \min\left\{l, \dfrac{\mu}{\left(t_i^k\right)^2 + w_i^k}\right\} \le \widehat{\lambda}_i^{(2)(k+1)} \le \dfrac{\mu\gamma_1}{\left(t_i^k\right)^2 + w_i^k}, \\ \dfrac{\mu\gamma_1}{\left(t_i^k\right)^2 + w_i^k}, & \text{if } \widehat{\lambda}_i^{(2)(k+1)} > \dfrac{\mu\gamma_1}{\left(t_i^k\right)^2 + w_i^k}, \end{cases}$$

$$\bar{\lambda}_i^{(3)(k+1)} = \begin{cases} \min\left\{l, \dfrac{\mu}{t_i^k}\right\}, & \text{if } \widehat{\lambda}_i^{(3)(k+1)} < \min\left\{l, \dfrac{\mu}{t_i^k}\right\}, \\ \widehat{\lambda}_i^{(3)(k+1)}, & \text{if} \min\left\{l, \dfrac{\mu}{t_i^k}\right\} \le \widehat{\lambda}_i^{(3)(k+1)} \le \dfrac{\mu\gamma_1}{t_i^k}, \\ \dfrac{\mu\gamma_1}{t_i^k}, & \text{if } \widehat{\lambda}_i^{(3)(k+1)} > \dfrac{\mu\gamma_1}{t_i^k}, \end{cases}$$

$$\bar{\lambda}_i^{(4)(k+1)} = \begin{cases} \min\left\{l, \dfrac{\mu}{p_i^k}\right\}, & \text{if } \widehat{\lambda}_i^{(4)(k+1)} < \min\left\{l, \dfrac{\mu}{p_i^k}\right\}, \\ \widehat{\lambda}_i^{(4)(k+1)}, & \text{if} \min\left\{l, \dfrac{\mu}{p_i^k}\right\} \le \widehat{\lambda}_i^{(4)(k+1)} \le \dfrac{\mu\gamma_1}{p_i^k}, \\ \dfrac{\mu\gamma_1}{p_i^k}, & \text{if } \widehat{\lambda}_i^{(4)(k+1)} > \dfrac{\mu\gamma_1}{p_i^k}, \end{cases} \tag{24}$$

where $p_i = y_i(w^T x_i + b) + \xi_i - 1$;

$$\bar{\lambda}_i^{(5)(k+1)} = \begin{cases} \min\left\{l, \dfrac{\mu}{\xi_i^k}\right\}, & \text{if } \widehat{\lambda}_i^{(5)(k+1)} < \min\left\{l, \dfrac{\mu}{\xi_i^k}\right\}, \\ \widehat{\lambda}_i^{(5)(k+1)}, & \text{if } \min\left\{l, \dfrac{\mu}{\xi_i^k}\right\} \\ & \quad \le \widehat{\lambda}_i^{(5)(k+1)} \le \dfrac{\mu\gamma_1}{\xi_i^k}, \\ \dfrac{\mu\gamma_1}{\xi_i^k}, & \text{if } \widehat{\lambda}_i^{(5)(k+1)} > \dfrac{\mu\gamma_1}{\xi_i^k}, \end{cases} \tag{25}$$

where the parameters l and γ_1 satisfy $0 < l, \gamma_1 > 1$.

Since positive definiteness of the matrix S is demanded in this method, the Lagrangian multipliers $\bar{\lambda}^{k+1}$ should satisfy the following condition:

$$\lambda^{(3)} - 2\left(\lambda^{(1)} + \lambda^{(2)}\right) t \ge 0. \tag{26}$$

For the sake of simplicity, the proof is given in Appendix B.

Therefore, if $\bar{\lambda}^{k+1}$ satisfies (26), we let $\lambda^{k+1} = \bar{\lambda}^{k+1}$. Otherwise, we would further update it by the following setting:

$$\begin{aligned} &\lambda^{(1)(k+1)} = \gamma_2 \bar{\lambda}^{(1)(k+1)}, \qquad \lambda^{(2)(k+1)} = \gamma_2 \bar{\lambda}^{(2)(k+1)}, \\ &\lambda^{(3)(k+1)} = \gamma_3 \bar{\lambda}^{(3)(k+1)}, \end{aligned} \tag{27}$$

where constants $\gamma_2 \in (0, 1)$ and $\gamma_3 \ge 1$ satisfy

$$\frac{\gamma_3}{\gamma_2} = \max_{i \in E}\left\{\frac{2t_i^{k+1}\left(\bar{\lambda}_i^{(1)(k+1)} + \bar{\lambda}_i^{(2)(k+1)}\right)}{\bar{\lambda}_i^{(3)(k+1)}}\right\}, \tag{28}$$

with $E = \{1, 2, \ldots, n\}$. It is not difficult to see that the vector $(\bar{\lambda}^{(1)(k+1)}, \bar{\lambda}^{(2)(k+1)}, \bar{\lambda}^{(3)(k+1)})$ determined by (27) satisfies (26).

In practice, the KKT conditions (15) are allowed to be satisfied within a tolerance ϵ_μ. It turns to be that the iterative process stops, while the following inequalities meet:

$$\mathrm{Res}(z, \lambda, \mu) = \left\| \begin{matrix} \lambda^{(1)} - \lambda^{(2)} - X^T Y^T \lambda^{(4)} \\ C * e_n - \lambda^{(4)} - \lambda^{(5)} \\ e_m - 2T\lambda^{(1)} - 2T\lambda^{(2)} - \lambda^{(3)} \\ e_n^T Y \lambda^{(4)} \\ \left(T^2 - W\right)\lambda^{(1)} - \mu e_m \\ \left(T^2 + W\right)\lambda^{(2)} - \mu e_m \\ T\lambda^{(3)} - \mu e_m \\ P\lambda^{(4)} - \mu e_n \\ \Xi\lambda^{(5)} - \mu e_n \end{matrix} \right\| < \epsilon_\mu, \tag{29}$$

$$\lambda \ge -\epsilon_\mu e, \tag{30}$$

where ϵ_μ is related to the current barrier parameter μ, and satisfies $\epsilon_\mu \downarrow 0$ as $\mu \to 0$.

Since function Φ_μ in (14) is convex, we have the following lemma which shows that Algorithm 2 is well defined (the proof is given in Appendix B).

Lemma 3. *Let z^k be strictly feasible for problem* (10). *If $\Delta z^k = 0$, then* (23) *is satisfied for all $\alpha_k \ge 0$. If $\Delta z^k \ne 0$, then there exists a $\bar{\alpha}_k \in (0, 1]$ such that* (23) *holds for all $\alpha_k \in (0, \bar{\alpha}_k]$.*

The proposed interior point method successively solves the barrier subproblem (14) with a decreasing sequence $\{\mu_k\}$. We simply reduce both ϵ_μ and μ by a constant factor $\beta \in (0, 1)$. Finally, we test optimality for problem (10) by means of the residual norm $\|\mathrm{Res}(z, \widehat{\lambda}, 0)\|$.

Here, we present the whole algorithm to solve the $L_{1/2}$-SVM problem(10)

Algorithm 4. Algorithm for solving $L_{1/2}$-SVM problem is as follows.

Step 0. Set $w^0 \in R^m, b^0 \in R, \xi^0 \in R^n$, $\lambda^0 = \widehat{\lambda}^0 > 0$, $t_i^0 \ge \sqrt{|w_i^0|} + (1/2), i \in \{1, 2, \ldots, m\}$. Given constants $\mu_0 > 0, \epsilon_{\mu_0} > 0, \beta \in (0, 1)$ and $\bar{\epsilon} > 0$, let $j = 0$.

Step 1. Stop if $\mathrm{Res}(z^j, \widehat{\lambda}^j, 0) \le \bar{\epsilon}$ and $\widehat{\lambda}^j \ge 0$.

Step 2. Starting from (z^j, λ^j), apply Algorithm 2 to solve (14) with barrier parameter μ_j and stopping tolerance ϵ_{μ_j}. Set $z^{j+1} = z^{j,k}$, $\widehat{\lambda}^{j+1} = \widehat{\lambda}^{j,k}$, and $\lambda^{j+1} = \lambda^{j,k}$.

Step 3. Set $\mu_{j+1} = \beta\mu_j$ and $\epsilon_{\mu_{j+1}} = \beta\epsilon_{\mu_j}$. Let $j := j + 1$ and go to Step 1.

In Algorithm 4, the index j denotes an outer iteration, while k denotes the last inner iteration of Algorithm 2.

The convergence of the proposed interior point method can be proved. We list the theorem here and give the proof in Appendix C.

Theorem 5. *Let $\{(z^k, \lambda^k)\}$ be generated by Algorithm 2. Then, any limit point of the sequence $\{(z^k, \widehat{\lambda}^k)\}$ generated by Algorithm 2 satisfies the first-order optimality conditions* (15).

Theorem 6. *Let $\{(z^j, \widehat{\lambda}^j)\}$ be generated by Algorithm 4 by ignoring its termination condition. Then the following statements are true.*

(i) *The limit point of $\{(z^j, \widehat{\lambda}^j)\}$ satisfies the first order optimality condition* (13).

(ii) *The limit point z^* of the convergent subsequence $\{z^j\}_{\mathcal{J}} \subseteq \{z^j\}$ with unbounded multipliers $\{\widehat{\lambda}^j\}_{\mathcal{J}}$ is a Fritz-John point [35] of problem* (10).

4. Experiments

In this section, we tested the constrained optimization reformulation to the $L_{1/2}$-SVM and the proposed interior point method. We compared the performance of the $L_{1/2}$-SVM with L_2-SVM [30], L_1-SVM [12], and L_0-SVM [12] on artificial data and ten UCI data sets (http://archive.ics.uci.edu/ml/). These four problems were solved in primal, referencing the machine learning toolbox Spider (http://people.kyb.tuebingen.mpg.de/spider/). The L_2-SVM and L_1-SVM were solved directly by quadratic programming and linear programming, respectively. To the NP-hard problem L_0-SVM, a commonly cited approximation method Feature Selection Concave (FSV) [12] was applied and then the FSV

problem was solved by a Successive Linear Approximation (SLA) algorithm. All the experiments were run in the personal computer (1.6 GHz of CPU, 4 GB of RAM) with MATLAB R2010b on 64 bit Windows 7.

In the proposed interior point method (Algorithms 2 and 4), we set the parameters as $\bar{\beta} = 0.6$, $\tau_1 = 0.5$, $l = 0.00001$, $\gamma_1 = 100000$, and $\beta = 0.5$. The balance parameter C was selected by 5-fold cross-validation on training set over the range $\{2^{-10}, 2^{-9}, \ldots, 2^{10}\}$. After training, the weights that did not satisfy the criteria $|w_j|/\max_i(|w_i|) \geq 10^4$ [14] were set to zero. Then the cardinality of the hyperplane was computed as the number of the nonzero weights.

4.1. Artificial Data. First, we took an artificial binary linear classification problem as an example. The problem is similar to that in [13]. The probability of $y = 1$ or -1 is equal. The first 6 features are relevant but redundant. In 70% samples, the first three features $\{x_1, x_2, x_3\}$ were drawn as $x_i = yN(i, 1)$ and the second three features $\{x_4, x_5, x_6\}$ as $x_i = N(0, 1)$. Otherwise, the first three were drawn as $x_i = N(0, 1)$ and the second three as $x_i = yN(i - 3, 1)$. The rest features are noise $x_i = N(0, 20)$, $i = 7, \ldots, m$. Here, m is the dimension of input features. The inputs were scaled to mean zero and standard deviation. In each trial, 500 points were generated for testing and the average results were estimated over 30 trials.

In the first experiment, we consider the cases with the fixed feature size $m = 30$ and different training sample sizes $n = 10, 20, \ldots, 100$. The average results over the 30 trials are shown in Table 1 and Figure 2. Figure 2 (left) plots the average cardinality of each classifier. Since the artificial data sets have 2 relevant and nonredundant features, the ideal average cardinality is 2. Figure 2 (left) shows that the three sparse SVMs, L_1-SVM, $L_{1/2}$-SVM, and L_0-SVM, can achieve sparse solution, while the L_2-SVM almost uses full features in each data set. Furthermore, the solutions of $L_{1/2}$-SVM and L_0-SVM are much sparser than L_1-SVM. As shown in Table 1, the L_1-SVM selects more than 6 features in all cases, which implies that some redundant or irrelevant features are selected. The average cardinalities of $L_{1/2}$-SVM and L_0-SVM are similar and close to 2. However, when $n = 10$ and 20, the L_0-SVM has the average cardinalities of 1.42 and 1.87, respectively. It means that the L_0-SVM sometimes selects only one feature in low sample data set and maybe ignores some really relevant feature. Consequently, with the cardinalities between 2.05 and 2.9, $L_{1/2}$-SVM has the more reliable solution than L_0-SVM. In short, as far as the number of selected features is concerned, the $L_{1/2}$-SVM behaves better than the other three methods.

Figure 2 (right) plots the trend of the prediction accuracy versus the size of the training sample. The classification performance of all methods is generally improved with the increasing of the training sample size n. L_1-SVM has the best prediction performance in all cases and a slightly better than $L_{1/2}$-SVM. $L_{1/2}$-SVM shows more accuracy in classification than L_2-SVM and L_0-SVM, especially in the case of $n = 10, \ldots, 50$. As shown in Table 1, when there are only 10 training samples, the average accuracy of $L_{1/2}$-SVM is 88.05%, while the results of L_2-SVM and L_0-SVM are 84.65% and 77.65%, respectively. Compared with L_2-SVM and L_0-SVM, $L_{1/2}$-SVM has the average accuracy increased by 3.4% and 10.4%, respectively, as can be explained in what follows. To the L_2-SVM, all features are selected without discrimination, and the prediction would be misled by the irrelevant features. To the L_0-SVM, few features are selected, and some relevant features are not included, which would put negative impact on the prediction result. As the tradeoff between L_2-SVM and L_0-SVM, $L_{1/2}$-SVM has better performance than the two.

The average results over ten artificial data sets in the first experiment are shown in the bottom of Table 1. On average, the accuracy of $L_{1/2}$-SVM is 0.87% lower than the L_1-SVM, while the features selected by $L_{1/2}$-SVM are 74.14% less than L_1-SVM. It indicates that the $L_{1/2}$-SVM can achieve much sparser solution than L_1-SVM with little cost of accuracy. Moreover, the average accuracy of $L_{1/2}$-SVM over 10 data sets is 2.22% higher than L_0-SVM with the similar cardinality. To sum up, the $L_{1/2}$-SVM provides the best balance between accuracy and sparsity among the three sparse SVMs.

To further evaluate the feature selection performance of $L_{1/2}$-SVM, we investigate whether the features are correctly selected. For the L_2-SVM is not designed for feature selection, it is not included in this comparison. Since our artificial data sets have 2 best features (x_3, x_6), the best result should have the two features (x_3, x_6) ranking on the top according to their absolute values of weights $|w_j|$. In the experiment, we select the top 2 features with the maximal $|w_j|$ for each method and calculate the frequency that the top 2 features are x_3 and x_6 in 30 runs. The results are listed in Table 2. When the training sample size is too small, it is difficult to discriminate the two most important features for all sparse SVMs. For example, when $n = 10$, the selected frequencies of L_1-SVM, L_0-SVM, and $L_{1/2}$-SVM are 7, 3, and 9, respectively. When n increases, all methods tend to make more correct selection. Moreover, Table 2 shows that the $L_{1/2}$-SVM outperforms the other two methods in all cases. For example, when $n = 100$, the selected frequencies of L_1-SVM andL_0-SVM are 22 and 25, respectively, and the result of $L_{1/2}$-SVM is 27. The L_1-SVM selects too many redundant or irrelevant features, which may influence the ranking in some extent. Therefore, L_1-SVM is not so good as $L_{1/2}$-SVM at distinguishing the critical features. The L_0-SVM has the lower hit frequency than $L_{1/2}$-SVM, which is probably due to the excessive small feature subset it obtained. Above all, Tables 1 and 2 and Figure 2 clearly show that the $L_{1/2}$-SVM is a promising sparsity driven classification method.

In the second simulation, we consider the cases with various dimensions of feature space $m = 20, 40, \ldots, 180, 200$ and the fixed training sample size $n = 100$. The average results over 30 trials are shown in Figure 3 and Table 3. Since there are only 6 relevant features yet, the larger m means the more noisy features. Figure 3 (left) shows that as the dimension increases from 20 to 200, the number of features selected by L_1-SVM increases from 8.26 to 23.1. However, the cardinalities of $L_{1/2}$-SVM and L_0-SVM keep stable (from 2.2 to 2.95). It indicates that the $L_{1/2}$-SVM and L_0-SVM are more suitable for feature selection than L_1-SVM.

Figure 3 (right) shows that with the increasing of the noise features, the accuracy of L_2-SVM drops significantly (from 98.68% to 87.47%). On the contrary, to the other three sparse

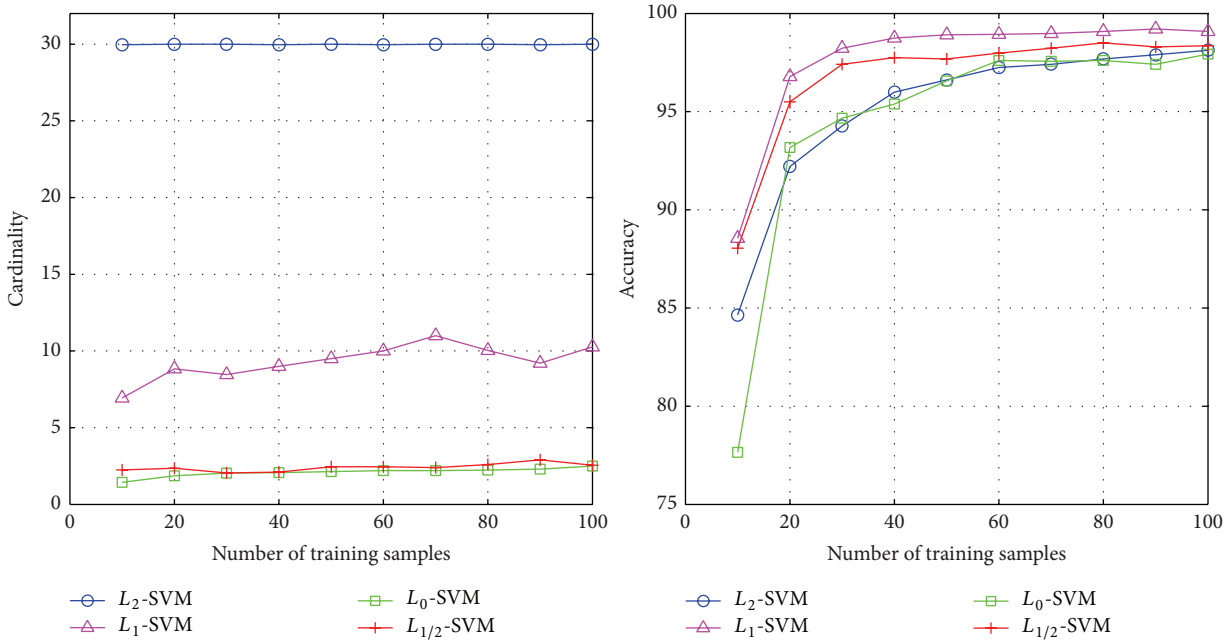

Figure 2: Results comparison on 10 artificial data sets with various training samples.

Table 1: Results comparison on 10 artificial data sets with various training sample sizes.

n	L_2-SVM		L_1-SVM		L_0-SVM		$L_{1/2}$-SVM	
	Card	Acc	Card	Acc	Card	Acc	Card	Acc
10	29.97	84.65	6.93	88.55	1.43	77.65	2.25	88.05
20	30.00	92.21	8.83	96.79	1.87	93.17	2.35	95.50
30	30.00	94.27	8.47	98.23	2.03	94.67	2.05	97.41
40	29.97	95.99	9.00	98.75	2.07	95.39	2.10	97.75
50	30.00	96.61	9.50	98.91	2.13	96.57	2.45	97.68
60	29.97	97.25	10.00	98.94	2.20	97.61	2.45	97.98
70	30.00	97.41	11.00	98.98	2.20	97.56	2.40	98.23
80	30.00	97.68	10.03	99.09	2.23	97.61	2.60	98.50
90	29.97	97.89	9.20	99.21	2.30	97.41	2.90	98.30
100	30.00	98.13	10.27	99.09	2.50	97.93	2.55	98.36
On average	29.99	95.21	9.32	**97.65**	**2.10**	94.56	2.41	96.78

"n" is the number of training samples, "Card" represents the number of selected features, and "Acc" is the classification accuracy.

Table 2: Frequency of the most important features (x_3, x_6) ranked on top 2 in 30 runs.

n	10	20	30	40	50	60	70	80	90	100
L_1-SVM	7	14	16	18	22	23	20	22	21	22
L_0-SVM	3	11	12	15	19	22	22	23	22	25
$L_{1/2}$-SVM	9	16	19	24	24	23	27	26	26	27

Table 3: Average results over 10 artificial data sets with varies dimension.

	L_2-SVM	L_1-SVM	L_0-SVM	$L_{1/2}$-SVM
Cardinality	109.95	15.85	**2.38**	2.41
Accuracy	92.93	**99.07**	97.76	98.26

Table 4: Feature selection performance comparison on UCI data sets (Cardinality).

No.	UCI Data Set	n	m	L_2-SVM		L_1-SVM		L_0-SVM		$L_{1/2}$-SVM	
				Mean	Std	Mean	Std	Mean	Std	Mean	Std
1	Pima diabetes	768	8	8.00	0.00	7.40	0.89	**6.20**	1.10	7.60	1.52
2	Breast Cancer	683	9	9.00	0.00	8.60	1.22	**6.40**	2.07	8.40	1.67
3	Wine (3)	178	13	13.00	0.00	9.73	0.69	**3.73**	0.28	4.07	0.49
4	Image (7)	2100	19	19.00	0.00	8.20	0.53	**3.66**	0.97	5.11	0.62
5	SPECT	267	22	22.00	0.00	17.40	5.27	9.80	4.27	**9**	1.17
6	WDBC	569	30	30.00	0.00	9.80	1.52	9.00	3.03	**5.60**	2.05
7	Ionosphere	351	34	33.20	0.45	25.00	3.35	24.60	10.21	**21.80**	8.44
8	SPECTF	267	44	44.00	0.00	38.80	10.98	**24.00**	11.32	25.60	7.75
9	Sonar	208	60	60.00	0.00	31.40	13.05	27.00	7.18	**13.60**	10.06
10	Musk1	476	166	166.00	0.00	85.80	18.85	48.60	4.28	**41**	14.16
	Average cardinality		40.50	40.42		24.21		16.30		**14.18**	

Table 5: Classification performance comparison on UCI data sets (accuaracy).

No.	UCI Data Set	L_2-SVM		L_1-SVM		L_0-SVM		$L_{1/2}$-SVM	
		Mean	Std	Mean	Std	Mean	Std	Mean	Std
1	Pima diabetes	76.99	3.47	76.73	0.85	76.73	2.94	**77.12**	1.52
2	Breast cancer	96.18	2.23	**97.06**	0.96	96.32	1.38	96.47	1.59
3	Wine (3)	97.71	3.73	**98.29**	1.28	93.14	2.39	97.14	2.56
4	Image (7)	88.94	1.49	91.27	1.77	86.67	4.58	**91.36**	1.94
5	SPECT	**83.40**	2.15	78.49	2.31	78.11	3.38	82.34	5.10
6	WDBC	96.11	2.11	**96.46**	1.31	95.58	1.15	**96.46**	1.31
7	Ionosphere	83.43	5.44	**88.00**	3.70	84.57	5.38	87.43	5.94
8	SPECTF	**78.49**	3.91	74.34	4.50	73.87	3.25	**78.49**	1.89
9	Sonar	73.66	5.29	75.61	5.82	74.15	7.98	**76.10**	4.35
10	Musk1	**84.84**	1.73	82.11	2.42	76.00	5.64	82.32	3.38
	Average accuracy	85.97		85.84		83.51		**86.52**	

SVMs, there is little change in the accuracy. It reveals that SVMs can benefit from the features reduction.

Table 3 shows the average results over all data sets in the second experiment. On average, the solution of $L_{1/2}$-SVM yields much sparser than L_1-SVM and a slightly better accuracy than L_0-SVM.

4.2. UCI Data Sets. We further tested the reformulation and the proposed interior point methods to $L_{1/2}$-SVM on 10 UCI data sets [36]. There are 8 binary classification problems and 2 multiclass problems (wine, image). Each feature of the input data was normalized to zero mean and unit variance, and the instances with missing value were deleted. Then, the data was randomly split into training set (80%) and testing set (20%). For the two multiclass problems, a one-against-rest method was applied to construct a binary classifier for each class. We repeated the training and testing procedure 10 times, and the average results were shown in Tables 4 and 5 and Figure 4.

Tables 4 and 5 summarize the feature selection and classification performance of the numerical experiments on UCI data sets, respectively. Here, n is the numbers of samples and m is the number of the input features. For the two multiclass data sets, the numbers of the classes are marked behind their names. Sparsity is defined as card/m and the small value of sparsity is preferred. The data sets are arranged in descending order according to the dimension. The lowest cardinality and the best accuracy rate for each problem are bolded.

As shown in Tables 4 and 5, the three sparse SVMs can encourage sparsity in all data sets, while remaining roughly identical accuracy with L_2-SVM. Among the three sparse methods, the $L_{1/2}$-SVM has the lowest cardinality (14.18) and the highest classification accuracy (86.52%) on average. While the L_1-SVM has the worst feature selection performance with the highest average cardinality (24.21), and the L_0-SVM has the lowest average classification accuracy (83.51%).

Figure 4 plots the sparsity (left) and classification accuracy (right) of each classifier on UCI data sets. In three data sets (6, 9, 10), the $L_{1/2}$-SVM has the best performance both in feature selection and classification among the three sparse SVMs. Compared with L_1-SVM, $L_{1/2}$-SVM can achieve sparser solution in nine data sets. For example, in the data set "8 SPECTF," the features selected by $L_{1/2}$-SVM are 34%

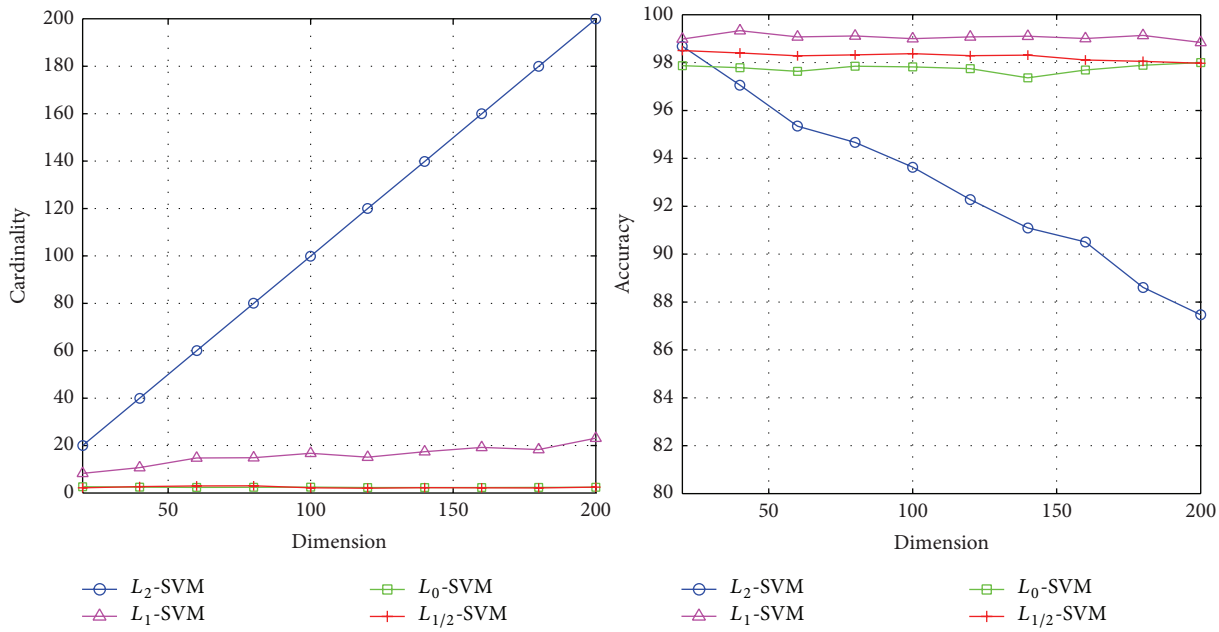

Figure 3: Results comparison on artificial data set with various dimensions.

less than L_1-SVM, at the same time, the accuracy of $L_{1/2}$-SVM is 4.1% higher than L_1-SVM. In most data sets (4, 5, 6, 9, 10), the cardinality of $L_{1/2}$-SVM drops significantly (at least 37%) with the equal or a slightly better result in accuracy. Only in the data set "3 Wine," the accuracy of $L_{1/2}$-SVM is decreased by 1.1%, but the sparsity provided by $L_{1/2}$-SVM leads to 58.2% improvement over L_1-SVM. In the rest three data sets (1, 2, 7), the two methods have similar results in feature selection and classification. As seen above, the $L_{1/2}$-SVM can provide lower dimension representation than L_1-SVM with the competitive prediction performance.

Figure 4 (right) shows that, compared with L_0-SVM, $L_{1/2}$-SVM has the classification accuracy improved in all data sets. For instance, in four data sets (3, 4, 8, 10), the classification accuracy of $L_{1/2}$-SVM is at least 4.0% higher than L_0-SVM. Especially in the data set "10 Musk," $L_{1/2}$-SVM gives a 6.3% rise in accuracy over L_0-SVM. Meanwhile, it can be observed from Figure 4 (left) that $L_{1/2}$-SVM selects fewer feature than L_0-SVM in five data sets (5, 6, 7, 9, 10). For example, in the data sets 6 and 9, the cardinalities of $L_{1/2}$-SVM are 37.8% and 49.6% less than L_0-SVM, respectively. In summary, $L_{1/2}$-SVM presents better classification performance than L_0-SVM, while it is effective in choosing relevant features.

5. Conclusions

In this paper, we proposed a $L_{1/2}$ regularization technique for simultaneous feature selection and classification in the SVM. We have reformulated the $L_{1/2}$-SVM into an equivalent smooth constrained optimization problem. The problem possesses a very simple structure and is relatively easy to develop numerical methods. By the use of this interesting reformulation, we proposed an interior point method and established its convergence. Our numerical results supported the reformulation and the proposed method. The $L_{1/2}$-SVM can get more sparsity solution than L_1-SVM with the comparable classification accuracy. Furthermore, the $L_{1/2}$-SVM can achieve more accuracy classification results than L_0-SVM (FSV).

Inspired by the good performance of the smooth optimization reformulation of $L_{1/2}$-SVM, there are some interesting topics deserving further research. For examples, to develop more efficient algorithms for solving the reformulation, to study nonlinear $L_{1/2}$-SVM, and to explore varies applications of the $L_{1/2}$-SVM and further validate its effective are interesting research topics. Some of them are under our current investigation.

Appendix

A. Properties of the Reformulation to $L_{1/2}$-SVM

Let D be set of all feasible points of the problem (10). For any $z \in D$, we let $FD(z, D)$ and $LFD(z, D)$ be the set of all feasible directions and linearized feasible directions of D at z. Since the constraint functions of (10) are all convex, we immediately have the following lemma.

Lemma A.1. *For any feasible point z of* (10)*, one has*

$$FD(z, D) = LFD(z, D). \tag{A.1}$$

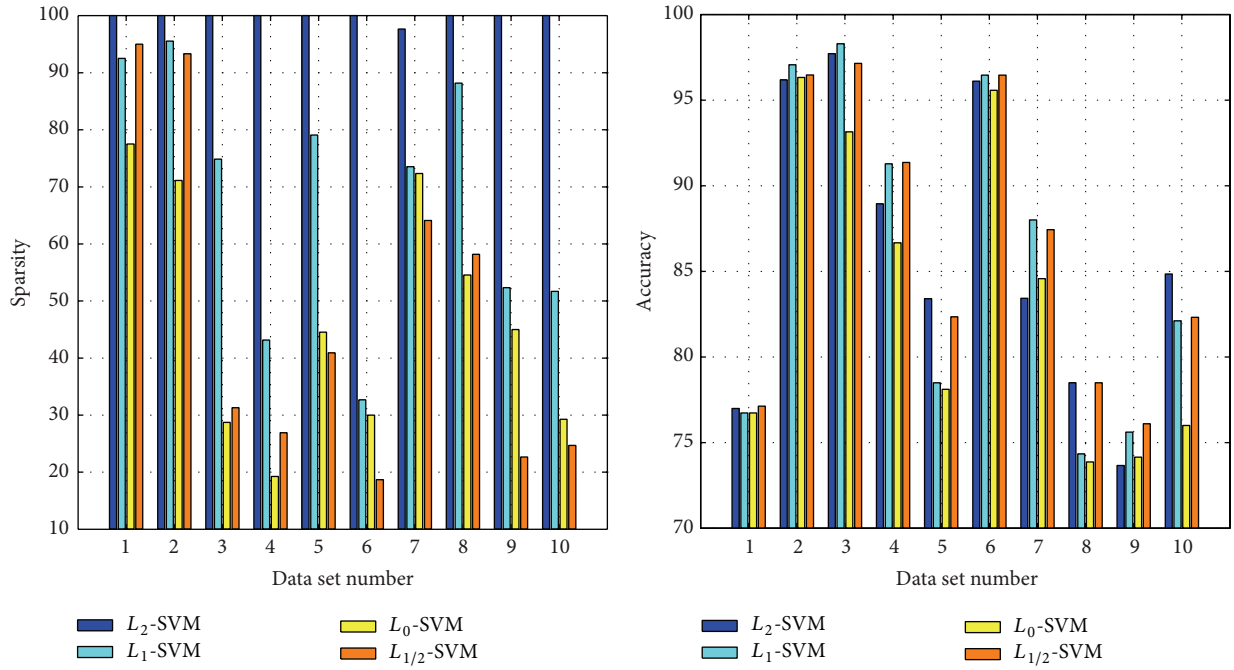

Figure 4: Results comparison on UCI data sets.

Based on the above Lemma, we can easily derive a first order necessary condition for (10). The Lagrangian function of (10) is

$$\begin{aligned} &L(w,t,\xi,b,\lambda,\mu,\nu,p,q) \\ &= C\sum_{i=1}^{n}\xi_i + \sum_{j=1}^{m}t_j - \sum_{j=1}^{m}\lambda_j\left(t_j^2 - w_j\right) - \sum_{j=1}^{m}\mu_j\left(t_j^2 + w_j\right) \\ &- \sum_{j=1}^{m}\nu_j t_j - \sum_{i=1}^{n}p_i\left(y_i w^T x_i + y_i b + \xi_i - 1\right) - \sum_{i=1}^{n}q_i\xi_i, \end{aligned} \tag{A.2}$$

where λ, μ, ν, p, q are the Lagrangian multipliers. By the use of Lemma A.1, we immediately have the following theorem about the first order necessary condition.

Theorem A.2. *Let (w,t,ξ,b) be a local solution of* (10). *Then there are Lagrangian multipliers $(\lambda,\mu,\nu,p,q) \in R^{m+m+m+n+n}$ such that the following KKT conditions hold:*

$$\begin{aligned} &\frac{\partial L}{\partial w} = \lambda_j - \mu_j - p_i\sum_{i=1}^{n}y_i x_i = 0, \quad i = 1,2,\ldots,n, \\ &\frac{\partial L}{\partial t} = 1 - 2\lambda_j t_j - 2\mu_j t_j - \nu_j = 0, \quad j = 1,2,\ldots,m, \\ &\frac{\partial L}{\partial \xi} = C - p_i - q_i = 0, \quad i = 1,2,\ldots,n, \\ &\frac{\partial L}{\partial b} = \sum_{i=1}^{n}p_i y_i = 0, \quad i = 1,2,\ldots,n, \\ &\lambda_j \geq 0, \quad t_j^2 - w_j \geq 0, \quad \lambda_j\left(t_j^2 - w_j\right) = 0, \\ &\qquad j = 1,2,\ldots,m, \\ &\mu_j \geq 0, \quad t_j^2 + w_j \geq 0, \quad \mu_j\left(t_j^2 + w_j\right) = 0, \\ &\qquad j = 1,2,\ldots,m, \\ &\nu_j \geq 0, \quad t_j \geq 0 \quad \nu_j t_j = 0, \quad j = 1,2,\ldots,m, \\ &p_i \geq 0, \quad y_i\left(w^T x_i + b\right) + \xi_i - 1 \geq 0, \\ &p_i\left(y_i\left(w^T x_i + b\right) + \xi_i - 1\right) = 0, \quad i = 1,2,\ldots,n, \\ &q_i \geq 0, \quad \xi_i \geq 0, \quad q_i\xi_i = 0, \quad i = 1,2,\ldots,n. \end{aligned} \tag{A.3}$$

The following theorem shows that the level set will be bounded.

Theorem A.3. *For any given constant $c > 0$, the level set*

$$W_c = z = \{(w,t,\xi) \mid f(z) \leq c\} \cap D \tag{A.4}$$

is bounded.

Proof. For any $(w,t,\xi) \in \Omega_c$, we have

$$\sum_{i}^{m}t_i + C\sum_{j}^{n}\xi_j \leq c. \tag{A.5}$$

Combining with $t_i \geq 0$, $i = 1, \ldots, m$, and $\xi_j \geq 0$, $j = 1, \ldots, n$, we have

$$\begin{aligned} 0 \leq t_i \leq c, &\quad i = 1, \ldots, m, \\ 0 \leq \xi_j \leq \frac{c}{C}, &\quad j = 1, \ldots, n. \end{aligned} \tag{A.6}$$

Moreover, for any $(w, t, \xi) \in D$,

$$\left|w_i\right| \leq t_i^2, \quad i = 1, \ldots, n. \tag{A.7}$$

Consequently, w, t, ξ are bounded. What is more, from the condition

$$y_i \left(w^T x_i + b\right) \geq 1 - \xi_i, \quad i = 1, \ldots, m, \tag{A.8}$$

we have

$$\begin{aligned} w^T x_i + b \geq 1 - \xi_i, &\quad \text{if } y_i = 1,\ i = 1, \ldots, m, \\ w^T x_i + b \leq -1 + \xi_i, &\quad \text{if } y_i = -1,\ i = 1, \ldots, m. \end{aligned} \tag{A.9}$$

Thus if the feasible region D is not empty, then we have

$$\max_{y_i=1} \left(1 - \xi_i - w^T x_i\right) \leq b \leq \min_{y_i=-1} \left(-1 + \xi_i - w^T x_i\right). \tag{A.10}$$

Hence b is also bounded. The proof is complete. □

B. Proof of Lemma 3

Lemma 3 (see Section 3) shows that Algorithm 2 is well defined. We first introduce the following proposition to show that the matrix S defined by (21) is always positive definite which ensures Algorithm 2 to be well defined.

Proposition B.1. *Let the inequality* (26) *hold; then the matrix* S *defined by* (21) *is positive definite.*

Proof. By an elementary deduction, we have for any $v^{(1)} \in R^m$, $v^{(2)} \in R^n$, $v^{(3)} \in R^m$, $v^{(4)} \in R$ with $(v^{(1)}, v^{(2)}, v^{(3)}, v^{(4)}) \neq 0$

$$\begin{aligned} &\left(v^{(1)T}, v^{(2)T}, v^{(3)T}, v^{(4)T}\right) S \begin{pmatrix} v^{(1)} \\ v^{(2)} \\ v^{(3)} \\ v^{(4)} \end{pmatrix} \\ &= v^{(1)T} S_{11} v^{(1)} + v^{(1)T} S_{12} v^{(2)} + v^{(1)T} S_{13} v^{(3)} \\ &\quad + v^{(1)T} S_{14} v^{(4)} + v^{(2)T} S_{21} v^{(1)} + v^{(2)T} S_{22} v^{(2)} \\ &\quad + v^{(2)T} S_{23} v^{(3)} + v^{(2)T} S_{24} v^{(4)} \\ &\quad + v^{(3)T} S_{31} v^{(1)} + v^{(3)T} S_{32} v^{(2)} \\ &\quad + v^{(3)T} S_{33} v^{(3)} + v^{(3)T} S_{34} v^{(4)} \\ &\quad + v^{(4)T} S_{41} v^{(1)} + v^{(4)T} S_{42} v^{(2)} \\ &\quad + v^{(4)T} S_{43} v^{(3)} + v^{(4)T} S_{44} v^{(4)} \\ &= v^{(1)T} \left(U + V + X^T Y^T P^{-1} D_4 Y X\right) v^{(1)} \\ &\quad + v^{(1)T} \left(X^T Y^T P^{-1} D_4\right) v^{(2)} \\ &\quad + v^{(1)T} \left(-2 (U - V) T\right) v^{(3)} \\ &\quad + v^{(1)T} \left(X^T Y^T P^{-1} D_4 y\right) v^{(4)} \\ &\quad + v^{(2)T} P^{-1} D_4 Y X v^{(1)} + v^{(2)T} \left(P^{-1} D_4 + \Xi^{-1} D_5\right) v^{(2)} \\ &\quad + v^{(2)T} P^{-1} D_4 y v^{(4)} \\ &\quad - 2 v^{(3)T} (U - V) T v^{(1)} \\ &\quad + v^{(3)T} \left(4T (U + V) T + T^{-1} D_3 - 2 \left(D_1 + D_2\right)\right) v^{(3)} \\ &\quad + v^{(4)T} y^T P^{-1} D_4 Y X v^{(1)} \\ &\quad + v^{(4)T} y^T P^{-1} D_4 v^{(2)} + v^{(4)T} y^T P^{-1} D_4 y v^{(4)} \\ &= v^{(1)T} U v^{(1)} - 4 v^{(3)T} U T v^{(1)} + 4 v^{(3)T} T U T v^{(3)} \\ &\quad + v^{(1)T} V v^{(1)} + 4 v^{(3)T} V T v^{(1)} + 4 v^{(3)T} T V T v^{(3)} \\ &\quad + v^{(2)T} \Xi^{-1} D_5 v^{(2)} + v^{(3)T} \left(T^{-1} D_3 - 2 D_1 - 2 D_2\right) v^{(3)} \\ &\quad + \left(Y X v^{(1)} + v^{(2)} + v^{(4)} y\right)^T P^{-1} D_4 \\ &\quad \times \left(Y X v^{(1)} + v^{(2)} + v^{(4)} y\right) \\ &= e_m^T U \left(D v_1 - 2 T D v_3\right)^2 e_m + e_m^T V \left(D v_1 + 2 T D v_3\right)^2 e_m \\ &\quad + v^{(2)T} \Xi^{-1} D_5 v^{(2)} \\ &\quad + v^{(3)T} \left(T^{-1} D_3 - 2 D_1 - 2 D_2\right) v^{(3)} \\ &\quad + \left(Y X v^{(1)} + v^{(2)} + v^{(4)} y\right)^T P^{-1} D_4 \left(Y X v^{(1)} + v^{(2)} + v^{(4)} y\right) \\ &> 0, \end{aligned} \tag{B.1}$$

where $Dv_1 = \operatorname{diag}(v^{(1)})$, $Dv_2 = \operatorname{diag}(v^{(2)})$, $Dv_3 = \operatorname{diag}(v^{(3)})$, and $Dv_4 = \operatorname{diag}(v^{(4)})$. □

B.1. Proof of Lemma 3

Proof. It is easy to see that if $\Delta z^k = 0$, which implies that (23) holds for all $\alpha_k \geq 0$.

Suppose $\Delta z^k \neq 0$. We have

$$\begin{aligned}&\nabla_w \Phi_\mu\left(z^k\right)^T \Delta w^k + \nabla_t \Phi_\mu\left(z^k\right)^T \Delta t^k \\ &\quad + \nabla_\xi \Phi_\mu\left(z^k\right)^T \Delta \xi^k + \nabla_b \Phi_\mu\left(z^k\right)^T \Delta b^k \\ &= -\left(\Delta w^{kT}, \Delta \xi^{kT}, \Delta t^{kT}, \Delta b^k\right) S \begin{pmatrix} \Delta w^k \\ \Delta \xi^k \\ \Delta t^k \\ \Delta b^k \end{pmatrix}.\end{aligned} \tag{B.2}$$

Since matrix S is positive definite and $\Delta z^k \neq 0$, the last equation implies

$$\begin{aligned}&\nabla_w \Phi_\mu\left(z^k\right)^T \Delta w^k + \nabla_t \Phi_\mu\left(z^k\right)^T \Delta t^k \\ &\quad + \nabla_\xi \Phi_\mu\left(z^k\right)^T \Delta \xi^k + \nabla_b \Phi_\mu\left(z^k\right)^T \Delta b^k < 0.\end{aligned} \tag{B.3}$$

Consequently, there exists a $\bar{\alpha}_k \in (0, \widehat{\alpha}_k]$ such that the fourth inequality in (23) is satisfied for all $\alpha_k \in (0, \bar{\alpha}_k]$.

On the other hand, since (x^k, t^k) is strictly feasible, the point $(x^k, t^k) + \alpha(\Delta x, t^k, \Delta t^k)$ will be feasible for all $\alpha > 0$ sufficiently small. The proof is complete. □

C. Convergence Analysis

This appendix is devoted to the global convergence of the interior point method. We first show the convergence of Algorithm 2 when a fixed μ is applied to the barrier subproblem (14).

Lemma C.1. *Let $\{(z^k, \lambda^k)\}$ be generated by Algorithm 2. Then $\{z^k\}$ are strictly feasible for problem (10) and the Lagrangian multipliers $\{\lambda^k\}$ are bounded from above.*

Proof. For the sake of convenience, we use $g_i(z)$, $i = 1, 2, \ldots, 3m + 2n$, to denote the constraint functions of the constrained problem (10).

We first show that $\{z^k\}$ are strictly feasible. Suppose on the contrary that there exists an infinite index subset $\mathcal{K}$ and an index $i \in \{1, 2, \ldots, 3m + 2n\}$ such that $\{g_i(z^k)\}_{\mathcal{K}} \downarrow 0$. By the definition of $\Phi_\mu(z)$ and $f(z) = \sum_{j=1}^m t_i + C\sum_{i=1}^n \xi_i$ being bounded from below in the feasible set, it must hold that $\{\Phi_\mu(z^k)\}_{\mathcal{K}} \to \infty$.

However, the line search rule implies that the sequence $\{\Phi_\mu(z^k)\}$ is decreasing. So, we get a contradiction. Consequently, for any $i \in \{1, 2, \ldots, 3m+2n\}$, $\{g_i(x^k, t^k)\}$ is bounded away from zero. The boundedness of $\{\lambda^k\}$ then follows from (24)–(27). □

Lemma C.2. *Let $\{(z^k, \lambda^k)\}$ and $\{(\Delta z^k, \widehat{\lambda}^{k+1})\}$ be generated by Algorithm 2. If $\{z^k\}_{\mathcal{K}}$ is a convergent subsequence of $\{z^k\}$, then the sequence $\{(\Delta z^k, \widehat{\lambda}^{k+1})\}_{\mathcal{K}}$ is bounded.*

Proof. Again, we use $g_i(z), i = 1, 2, \ldots, 3m + 2n$ to denote the constraint functions of the constrained problem (10).

We suppose on the contrary that there exists an infinite subset $\mathcal{K}' \subseteq \mathcal{K}$ such that the subsequence $\{\|(\Delta z^k, \widehat{\lambda}^{k+1})\|\}_{\mathcal{K}'}$ tends to infinity. Let $\{z^k\}_{\mathcal{K}'} \to z^*$. It follows from Lemma C.1 that there is an infinite subset $\bar{\mathcal{K}} \subseteq \mathcal{K}'$ such that $\{\lambda^k\}_{\bar{\mathcal{K}}} \to \lambda^*$ and $g_i(z^*) > 0, \forall i \in \{1, 2, \ldots, 3m + 2n\}$. Thus, we have

$$S^k \longrightarrow S^*, \tag{C.1}$$

as $k \to \infty$ with $k \in \bar{\mathcal{K}}$. By Proposition B.1, S^* is positive definite. Since the right hand size of (18) is bounded and continuous, the unboundedness of $\{(\Delta z^k, \widehat{\lambda}^{k+1})\}_{\bar{\mathcal{K}}}$ implies that the limit of the coefficient matrices of (18) is singular, which yields a contradiction. The proof is complete. □

Lemma C.3. *Let $\{(z^k, \lambda^k)\}$ and $\{(\Delta z^k, \widehat{\lambda}^{k+1})\}$ be generated by Algorithm 2. If $\{z^k\}_{\mathcal{K}}$ is a convergent subsequence of $\{z^k\}$, then one has $\{\Delta z^k\}_{\mathcal{K}} \to 0$.*

Proof. It follows from Lemma C.1 that the sequence $\{\Delta z^k\}_{\mathcal{K}}$ is bounded. Suppose on the contrary that there exists an infinite subset $\mathcal{K}' \subseteq \mathcal{K}$ such that $\{\Delta z^k\}_{\mathcal{K}'} \to \Delta z^* \neq 0$. Since subsequences $\{z^k\}_{\mathcal{K}'}$, $\{\lambda^k\}_{\mathcal{K}'}$ and $\{\widehat{\lambda}^k\}_{\mathcal{K}'}$ are all bounded, there are points z^*, λ^*, and $\widehat{\lambda}^*$, as well as an infinite index set $\bar{\mathcal{K}} \subseteq \mathcal{K}'$ such that $\{z^k\}_{\bar{\mathcal{K}}} \to z^*$, $\{\lambda^k\}_{\bar{\mathcal{K}}} \to \lambda^*$, and $\{\widehat{\lambda}^k\}_{\bar{\mathcal{K}}} \to \widehat{\lambda}^*$. By Lemma C.1, we have $g_i(z^*) > 0$, $\forall i \in \{1, 2, \ldots, 3m + 2n\}$. Similar to the proof of Lemma 3, it is not difficult to get

$$\begin{aligned}&\nabla_w \Phi_\mu(z^*)^T \Delta w^* + \nabla_t \Phi_\mu(z^*)^T \Delta t^* \\ &\quad + \nabla_\xi \Phi_\mu(z^*)^T \Delta \xi^* + \nabla_b \Phi_\mu(z^*)^T \Delta b^* \\ &= -\left(\Delta w^{*T}, \Delta \xi^{*T}, \Delta t^{*T}, \Delta b^*\right) S \begin{pmatrix} \Delta w^* \\ \Delta \xi^* \\ \Delta t^* \\ \Delta b^* \end{pmatrix} < 0.\end{aligned} \tag{C.2}$$

Since $g_i(z^*) > 0$, $\forall i \in \{1, 2, \ldots, 3m + 2n\}$, there exists a $\bar{\alpha} \in (0, 1]$, such that, for all $\alpha \in (0, \bar{\alpha}]$,

$$g_i\left(z^* + \alpha \Delta z^*\right) > 0, \quad \forall i \in \{1, 2, \ldots, 3n\}. \tag{C.3}$$

Taking into account $\tau_1 \in (0, 1/2)$, we claim that there exists a $\widehat{\alpha} \in (0, \bar{\alpha}]$ such that the following inequality holds for all $\alpha \in (0, \widehat{\alpha}]$:

$$\Phi_\mu\left(z^* + \alpha \Delta z^*\right) - \Phi_\mu\left(z^*\right) \leq 1.1\tau_1 \alpha \nabla_z \Phi_\mu(z^*)^T \Delta z^*. \tag{C.4}$$

Let $m_* = \min\{j \mid \bar{\beta}^j \in (0, \widehat{\alpha}], j = 0, 1, \ldots\}$ and $\alpha_* = \bar{\beta}^{m_*}$. It follows from (C.3) that the following inequality is satisfied for all $k \in \bar{\mathcal{K}}$ sufficient large and any $i \in \{1, 2, \ldots, 3m + 2n\}$,

$$g_i\left(z^k + \alpha_* \Delta z^k\right) > 0. \tag{C.5}$$

Moreover, we have

$$\begin{aligned}
&\Phi_\mu\left(z^k+\alpha_*\Delta z^k\right)-\Phi_\mu\left(z^k\right)\\
&\quad=\Phi_\mu\left(z^*+\alpha_*\Delta z^*\right)-\Phi_\mu\left(z^*\right)\\
&\qquad+\Phi_\mu\left(z^k+\alpha_*\Delta z^k\right)-\Phi_\mu\left(z^*+\alpha_*\Delta z^*\right)\\
&\qquad+\Phi_\mu\left(z^*\right)-\Phi_\mu\left(z^k\right)\\
&\quad\le 1.05\tau_1\alpha_*\nabla_z\Phi_\mu\left(z^*\right)^T\Delta z^*\\
&\quad\le \tau_1\alpha_*\nabla_z\Phi_\mu\left(z^k\right)^T\Delta z^k.
\end{aligned}\tag{C.6}$$

By the backtracking line search rule, the last inequality together with (C.5) yields the inequality $\alpha_k \ge \alpha_*$ for all $k \in \tilde{\mathcal{K}}$ large enough. Consequently, when $k \in \tilde{\mathcal{K}}$ is sufficiently large, we have from (23) and (C.3) that

$$\begin{aligned}
\Phi_\mu\left(z^{k+1}\right)&\le \Phi_\mu\left(z^k\right)+\tau_1\alpha_k\nabla_z\Phi_\mu\left(z^k\right)^T\Delta z^k\\
&\le \Phi_\mu\left(z^k\right)+\tau_1\alpha_*\nabla_z\Phi_\mu\left(z^k\right)^T\Delta z^k\\
&\le \Phi_\mu\left(z^k\right)+\frac{1}{2}\tau_1\alpha_*\nabla_z\Phi_\mu\left(z^*\right)^T\Delta z^*.
\end{aligned}\tag{C.7}$$

This shows that $\{\Phi_\mu(z^k)\}_{\tilde{\mathcal{K}}} \to -\infty$, contradicting with the fact that $\{\Phi_\mu(z^k)\}_{\mathcal{K}}$ is bounded from below. The proof is complete. □

Then, we will establish the convergence of Algorithms 2 and 4, that is, Theorems 5 and 6 in Section 3.

C.1. Proof of Theorem 5

Proof. We let $(z^*, \widehat{\lambda}^*)$ be a limit point of $\{(z^k, \widehat{\lambda}^k)\}$ and the subsequence $\{(z^k, \widehat{\lambda}^k)\}_{\mathcal{K}}$ converges to $(z^*, \widehat{\lambda}^*)$.

Recall that $(\Delta z^k, \widehat{\lambda}^k)$ are the solution of (18) with $(z, \lambda) = (z^k, \lambda^k)$. Taking limits in both sides of (18) with $(z, \lambda) = (z^k, \lambda^k)$, as $k \to \infty$ with $k \in \mathcal{K}$, by the use of Lemma C.1, we obtain

$$\begin{aligned}
\widehat{\lambda}^{*(1)}-\widehat{\lambda}^{*(2)}-X^TY^T\widehat{\lambda}^{*(4)}&=0,\\
C*e_n-\widehat{\lambda}^{*(4)}-\widehat{\lambda}^{*(5)}&=0,\\
e_m-2T^*\widehat{\lambda}^{*(1)}-2T^*\widehat{\lambda}^{*(2)}-\widehat{\lambda}^{*(3)}&=0,\\
y^T\widehat{\lambda}^{*(4)}&=0,\\
\left(T^{*2}-W^*\right)\widehat{\lambda}^{*(1)}-\mu e_m&=0,\\
\left(T^{*2}+W^*\right)\widehat{\lambda}^{*(2)}-\mu e_m&=0,\\
T^*\widehat{\lambda}^{*(3)}-\mu e_m&=0,\\
P^*\widehat{\lambda}^{*(4)}-\mu e_n&=0,\\
\Xi^*\widehat{\lambda}^{*(5)}-\mu e_n&=0.
\end{aligned}\tag{C.8}$$

This shows that $\{(z^*, \widehat{\lambda}^*)\}$ satisfies the first-order optimality conditions (15). □

C.2. Proof of Theorem 6

Proof. (i) Without loss of generality, we suppose that the bounded subsequence $\{(z^j, \widehat{\lambda}^j)\}_{\mathcal{J}}$ converges to some point $(z^*, \widehat{\lambda}^*)$. It is clear that z^* is a feasible point of (10). Since $\{\mu_j, \epsilon_{\mu_j}\} \to 0$, it follows from (29) and (30) that

$$\begin{aligned}
\left\{\mathrm{Res}\left(z^j,\widehat{\lambda}^j,\mu_j\right)\right\}_{\mathcal{J}}&\longrightarrow \mathrm{Res}\left(z^*,\widehat{\lambda}^*,0\right)=0,\\
\left\{\widehat{\lambda}^j\right\}_{\mathcal{J}}&\longrightarrow \widehat{\lambda}^*\ge 0.
\end{aligned}\tag{C.9}$$

Consequently, the $(z^*, \widehat{\lambda}^*)$ satisfies the KKT conditions (13).

(ii) Let $\xi_j = \max\{\|\widehat{\lambda}^j\|_\infty, 1\}$ and $\bar{\lambda}^j = \xi_j^{-1}\widehat{\lambda}^j$. Obviously, $\{\bar{\lambda}^j\}$ is bounded. Hence, there exists an infinite subset $\mathcal{J}' \subseteq \mathcal{J}$ such that $\{\bar{\lambda}^j\}_{\mathcal{J}}' \to \bar{\lambda} \ne 0$ and $\|\bar{\lambda}^j\|_\infty = 1$ for large $j \in \mathcal{J}$. From (30) we know that $\bar{\lambda} \ge 0$. Dividing both sides of the first inequality of (18) with $(z, \lambda) = (z^k, \lambda^k)$ by ξ_j, then taking limits as $j \to \infty$ with $j \in \mathcal{J}$, we get

$$\begin{aligned}
\bar{\lambda}^{(1)}-\bar{\lambda}^{(2)}-X^TY^T\bar{\lambda}^{(4)}&=0,\\
\bar{\lambda}^{(4)}+\bar{\lambda}^{(5)}&=0,\\
2T^*\bar{\lambda}^{(1)}+2T^*\bar{\lambda}^{(2)}+\bar{\lambda}^{(3)}&=0,\\
y^T\bar{\lambda}^{(4)}&=0,\\
\left(T^{*2}-W^*\right)\bar{\lambda}^{(1)}&=0,\\
\left(T^{*2}+W^*\right)\bar{\lambda}^{(2)}&=0,\\
T^*\bar{\lambda}^{(3)}&=0,\\
P^*\bar{\lambda}^{(4)}&=0,\\
\Xi^*\bar{\lambda}^{(5)}&=0.
\end{aligned}\tag{C.10}$$

Since z^* is feasible, the above equations has shown that z^* is a Fritz-John point of problem (10). □

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

The authors would like to acknowledge support for this project from the National Science Foundation (NSF Grants 11371154, 11071087, 11271069, and 61103202) of China.

References

[1] Y. M. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," in *Proceedings of the 14th*

International Conference on Machine Learning (ICML '97), vol. 97, pp. 412–420, 1997.

[2] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *The Journal of Machine Learning Research*, vol. 3, pp. 1289–1305, 2003.

[3] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, no. 1–3, pp. 389–422, 2002.

[4] H. H. Zhang, J. Ahn, X. Lin, and C. Park, "Gene selection using support vector machines with non-convex penalty," *Bioinformatics*, vol. 22, no. 1, pp. 88–95, 2006.

[5] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[6] J. Weston, F. Pérez-Cruz, O. Bousquet, O. Chapelle, A. Elisseeff, and B. Schölkopf, "Feature selection and transduction for prediction of molecular bioactivity for drug design," *Bioinformatics*, vol. 19, no. 6, pp. 764–771, 2003.

[7] Y. Liu, "A comparative study on feature selection methods for drug discovery," *Journal of Chemical Information and Computer Sciences*, vol. 44, no. 5, pp. 1823–1828, 2004.

[8] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, Eds., *Feature Extraction: Foundations and Applications*, vol. 207, Springer, 2006.

[9] A. Rakotomamonjy, "Variable selection using SVM-based criteria," *Journal of Machine Learning Research*, vol. 3, no. 7-8, pp. 1357–1370, 2003.

[10] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik, "Feature selection for SVMs," in *Proceedings of the Conference on Neural Information Processing Systems*, vol. 13, pp. 668–674, Denver, Colo, USA, 2000.

[11] D. Peleg and R. Meir, "A feature selection algorithm based on the global minimization of a generalization error bound," in *Proceedings of the Conference on Neural Information Processing Systems*, Vancouver, Canada, 2004.

[12] P. S. Bradley and O. L. Mangasarian, "Feature selection via concave minimization and support vector machines," in *Proceedings of the International Conference on Machine Learning*, pp. 82–90, San Francisco, Calif, USA, 1998.

[13] J. Weston, A. Elisseeff, B. Schölkopf, and M. Tipping, "Use of the zero-norm with linear models and kernel methods," *Journal of Machine Learning Research*, vol. 3, pp. 1439–1461, 2003.

[14] A. B. Chan, N. Vasconcelos, and G. R. G. Lanckriet, "Direct convex relaxations of sparse SVM," in *Proceedings of the 24th International Conference on Machine Learning (ICML '07)*, pp. 145–153, Corvallis, Ore, USA, June 2007.

[15] G. M. Fung and O. L. Mangasarian, "A feature selection Newton method for support vector machine classification," *Computational Optimization and Applications*, vol. 28, no. 2, pp. 185–202, 2004.

[16] J. Zhu, S. Rosset, T. Hastie, and R. Tibshirani, "1-norm support vector machines," *Advances in Neural Information Processing Systems*, vol. 16, no. 1, pp. 49–56, 2004.

[17] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society B: Methodological*, vol. 58, no. 1, pp. 267–288, 1996.

[18] T. Zhang, "Analysis of multi-stage convex relaxation for sparse regularization," *Journal of Machine Learning Research*, vol. 11, pp. 1081–1107, 2010.

[19] R. Chartrand, "Exact reconstruction of sparse signals via nonconvex minimization," *IEEE Signal Processing Letters*, vol. 14, no. 10, pp. 707–710, 2007.

[20] R. Chartrand, "Nonconvex regularizaron for shape preservation," in *Proceedings of the 14th IEEE International Conference on Image Processing (ICIP '07)*, vol. 1, pp. I293–I296, San Antonio, Tex, USA, September 2007.

[21] Z. Xu, H. Zhang, Y. Wang, X. Chang, and Y. Liang, "$L_{1/2}$ regularization," *Science China: Information Sciences*, vol. 53, no. 6, pp. 1159–1169, 2010.

[22] W. J. Chen and Y. J. Tian, "Lp-norm proximal support vector machine and its applications," *Procedia Computer Science*, vol. 1, no. 1, pp. 2417–2423, 2010.

[23] J. Liu, J. Li, W. Xu, and Y. Shi, "A weighted L_q adaptive least squares support vector machine classifiers-Robust and sparse approximation," *Expert Systems with Applications*, vol. 38, no. 3, pp. 2253–2259, 2011.

[24] Y. Liu, H. H. Zhang, C. Park, and J. Ahn, "Support vector machines with adaptive L_q penalty," *Computational Statistics & Data Analysis*, vol. 51, no. 12, pp. 6380–6394, 2007.

[25] Z. Liu, S. Lin, and M. Tan, "Sparse support vector machines with Lp penalty for biomarker identification," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 7, no. 1, pp. 100–107, 2010.

[26] A. Rakotomamonjy, R. Flamary, G. Gasso, and S. Canu, "l_p-l_q penalty for sparse linear and sparse multiple kernel multitask learning," *IEEE Transactions on Neural Networks*, vol. 22, no. 8, pp. 1307–1320, 2011.

[27] J. Y. Tan, Z. Zhang, L. Zhen, C. H. Zhang, and N. Y. Deng, "Adaptive feature selection via a new version of support vector machine," *Neural Computing and Applications*, vol. 23, no. 3-4, pp. 937–945, 2013.

[28] Z. B. Xu, X. Y. Chang, F. M. Xu, and H. Zhang, "$L_{1/2}$ regularization: an iterative half thresholding algorithm," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 7, pp. 1013–1027, 2012.

[29] D. Ge, X. Jiang, and Y. Ye, "A note on the complexity of L_p minimization," *Mathematical Programming*, vol. 129, no. 2, pp. 285–299, 2011.

[30] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, NY, USA, 1995.

[31] J. Fan and H. Peng, "Nonconcave penalized likelihood with a diverging number of parameters," *The Annals of Statistics*, vol. 32, no. 3, pp. 928–961, 2004.

[32] K. Knight and W. Fu, "Asymptotics for lasso-type estimators," *The Annals of Statistics*, vol. 28, no. 5, pp. 1356–1378, 2000.

[33] B. S. Tian and X. Q. Yang, "An interior-point l $L_{1/2}$-penalty-method for nonlinear programming," Technical Report, Department of Applied Mathematics, Hong Kong Polytechnic University, 2013.

[34] L. Armijo, "Minimization of functions having Lipschitz continuous first partial derivatives," *Pacific Journal of Mathematics*, vol. 16, pp. 1–3, 1966.

[35] F. John, "Extremum problems with inequalities as side-conditions," in *Studies and Essays. Courant Anniversary Volume*, pp. 187–1204, 1948.

[36] D. J. Newman, S. Hettich, C. L. Blake, and C. J. Merz, "UCI repository of machine learning databases," Technical Report 9702, Department of Information and Computer Science, Univerisity of California, Irvine, Calif, USA, 1998, http://archive.ics.uci.edu/ml/.