

Subspace Newton Method for Sparse SVM

Shenglong Zhou(shenglong.zhou@soton.ac.uk)*

Abstract

Kernel-based methods for support vector machines (SVM) have seen a great advantage in various applications. However, they may incur prohibitive computational costs when the involved sample size is on a large scale. Therefore, reducing the number of support vectors (or say sample reduction) appears to be necessary, which gives rise to the topic of the sparse SVM. Motivated by this, we aim at solving a sparsity constrained kernel SVM optimization, which is capable of controlling the number of the support vectors. Based on the established optimality conditions associated with the stationary equations, a subspace Newton method is cast to tackle the sparsity constrained problem and enjoys one-step convergence property if the starting point is close to a local region of a stationary point, leading to a super-fast computational speed. Numerical comparisons with some other excellent solvers demonstrate that the proposed method performs exceptionally well, especially for datasets with large numbers of samples, in terms of a much fewer number of support vectors and shorter computational time.

Keywords: sample reduction, support vectors, sparsity constrained optimization, subspace Newton method, one-step convergence

1 Introduction

Support vector machines (SVM) were first introduced by Vapnik and Cortes [5], with wide applications in machine learning, statistic and pattern recognition. The basic idea of SVM is to find a hyperplane in the input space that best separates the training set. In the paper, we consider a binary classification problem that can be described as follows. Suppose we are given a training set $\{(\mathbf{x}_i, y_i)\}_{i=1}^m$, where $\mathbf{x}_i \in \mathbb{R}^n$ is the sample vector and $y_i \in \{-1, 1\}$ is the class. The purpose of SVM is to train a hyperplane $\langle \mathbf{w}, \mathbf{x} \rangle + \mathbf{b} = 0$ with variable $\mathbf{w} \in \mathbb{R}^n$ and bias $\mathbf{b} \in \mathbb{R}$. For any new input vector $\bar{\mathbf{x}}$, we can predict the corresponding class \bar{y} , where $\bar{y} = 1$ if $\langle \mathbf{w}, \bar{\mathbf{x}} \rangle + \mathbf{b} > 0$ and $\bar{y} = -1$ otherwise. In order to find optimal hyperplane, there are two possible cases. The training data is linearly separable and inseparable in the input space. For the latter, the popular approach is to consider the so-called soft-margin SVM optimization,

$$(1.1) \quad \min_{\mathbf{w} \in \mathbb{R}^n} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^m \ell \left[1 - y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + \mathbf{b}) \right],$$

*School of Mathematics, University of Southampton, Southampton SO17 1BJ, United Kingdom.

where $C > 0$ is a penalty parameter, $\|\cdot\|$ is the Euclidean norm and ℓ can be some loss functions. One of the most well-known loss functions is the hinge loss $\ell_h(t) := \max\{0, t\}$. The corresponding dual problem is the following quadratic kernel SVM optimization,

$$(1.2) \quad \begin{aligned} \min_{\boldsymbol{\alpha} \in \mathbb{R}^m} \quad & \frac{1}{2} \|\mathbf{Q}\boldsymbol{\alpha}\|^2 - \langle \boldsymbol{\alpha}, \mathbf{1} \rangle, \\ \text{s.t.} \quad & \langle \boldsymbol{\alpha}, \mathbf{y} \rangle = 0, \quad 0 \leq \alpha_i \leq C, i \in \mathbb{N}_m, \end{aligned}$$

where $\mathbf{Q} := [\mathbf{y}_1 \mathbf{x}_1, \dots, \mathbf{y}_m \mathbf{x}_m] \in \mathbb{R}^{n \times m}$, $\mathbf{y} := (\mathbf{y}_1, \dots, \mathbf{y}_m)^\top \in \mathbb{R}^m$, $\mathbf{1} := (1, \dots, 1)^\top \in \mathbb{R}^m$ and $\mathbb{N}_m := \{1, \dots, m\}$. Note that the matrix $\mathbf{Q}^\top \mathbf{Q}$ involved in (1.2) has the order $m \times m$, which makes it impossible for training on million-size data since the storage of such a scale data requires considerably large memory and the incurred computational cost is extremely expensive. Therefore, sample reduction plays a crucial role in overcoming this drawback. The idea is to use a small portion of samples to train a classifier \mathbf{w} , which can be expressed through

$$(1.3) \quad \mathbf{w} = \mathbf{Q}\boldsymbol{\alpha} = \sum_{i=1}^m \alpha_i \mathbf{y}_i \mathbf{x}_i$$

by the Representer Theorem. The training vectors \mathbf{x}_i corresponding to non-zero α_i are the support vectors. If large numbers of coefficients α_i are zero, then the sample size can be reduced significantly. Because of this, computations and storage for large scale size data are possible. However, solutions to the problem (1.2) are not sparse in general. In order to ensure a solution being sparse enough, various approaches have been developed. Those methods aiming at reducing the number of support vectors can be categorized into the sparse SVM group. We will explore more in the sequel.

1.1 Selective Literature Review

Most SVM methods have something to do with designing a proper loss function ℓ in the model (1.1), which can be summarized into two categories based on the convexity of ℓ . Convex soft margin loss functions include the famous hinge loss [5], the pinball loss [13, 11], the hybrid Huber loss [26, 30, 32], the square loss [29, 33], the insensitive zone pinball loss [11], the exponential loss [9] and log loss [10]. Convexity makes the computations of their corresponding SVM models tractable, but meanwhile induces the unboundedness, which reduces the robustness of those functions to outliers from the training data. In order to overcome such a drawback, authors in [18], [24] set an upper bound and enforce the loss function to stop increasing after a certain point. By doing so, the convex loss functions become non-convex. Other non-convex ones consist of the ramp loss [4], the truncated pinball loss [28], the asymmetrical truncated pinball loss [34], the sigmoid loss [23] and the normalized sigmoid cost loss [17]. Compared with convex margin loss functions, most non-convex ones are less sensitive to the outliers due to their boundedness. However, non-convexity would incur difficulties in computations in practice.

As for the sparse SVM, one of the earliest attempts was in [20], where it is suggested to solve the kernel SVM optimization problem to find a solution first and then seek a sparse approximation through support vector regression. This idea is adopted as a key component of the method developed in [36]. In [1], the ℓ_1 -norm regularization is employed to inherently perform variable/sample selection. In [14], a greedy method is devised, where in each iteration a new training vector is carefully selected into the set of support vectors. In [7], authors introduce a ‘slant-loss’ function, which is analogous to the 0/1 loss ($\ell_{0/1}(\mathbf{t}) = 1$ if $\mathbf{t} \leq 0$ and 0 otherwise) and the hinge loss. Other related methods can be found in [2, 15, 27, 6, 19]. Numerical experiments have demonstrated that those methods perform exceptionally well in terms of reducing the number of support vectors.

1.2 Contributions

Motivated by those work in the sparse SVM, in this paper, we aim at solving the following sparsity constrained kernel SVM optimization problem,

$$(1.4) \quad \begin{aligned} \min_{\boldsymbol{\alpha} \in \mathbb{R}^m} \quad & \frac{1}{2} \|\mathbf{Q}\boldsymbol{\alpha}\|^2 + \sum_{i=1}^m h_{c,C}(\alpha_i) - \langle \mathbf{1}, \boldsymbol{\alpha} \rangle =: \mathbf{D}(\boldsymbol{\alpha}), \\ \text{s.t.} \quad & \langle \boldsymbol{\alpha}, \mathbf{y} \rangle = 0, \quad \|\boldsymbol{\alpha}\|_0 \leq s, \end{aligned}$$

where the given integer $s \in \mathbb{N}_m$ is far less than m and is called the sparsity level, and $\|\boldsymbol{\alpha}\|_0$ counts the number of non-zero elements of $\boldsymbol{\alpha}$. Here,

$$(1.5) \quad h_{c,C}(\mathbf{t}) := \frac{\mathbf{t}^2}{2} \left[\mathbf{C}\rho(\mathbf{t} \geq 0) + \mathbf{c}\rho(\mathbf{t} < 0) \right]^{-1},$$

where $\mathbf{C} \geq \mathbf{c} > 0$ are two parameters and $\rho(\mathbf{t} \in \mathbf{S})$ is an indicator function, returning 1 if $\mathbf{t} \in \mathbf{S}$ and 0 otherwise. The definition of $h_{c,C}$ implies it gives penalty $1/\mathbf{C}$ for $\alpha_i \geq 0$ and $1/\mathbf{c}$ for $\alpha_i < 0$. Therefore, if $\mathbf{c} \rightarrow 0$, then $\alpha_i \geq 0, i \in \mathbb{N}_m$. Compared with (1.2), the problem (1.4) at least has three advantages:

- (i) Its objective function is strongly convex, hence the optimal solutions exist (see Theorem 2.1) and might be unique under mild conditions (see Theorem 2.4). While the objective function of (1.2) is convex but not strongly convex when $\mathbf{n} \leq \mathbf{m}$ since $\mathbf{Q} \in \mathbb{R}^{\mathbf{n} \times \mathbf{m}}$. This means it may have multiple optimal solutions.
- (ii) Since the bounded constraints are absent, computation is much more tractable. Note that when \mathbf{m} is large, those bounded constraints in (1.2) may incur expensive computational costs.
- (iii) Most importantly, the constraint $\|\boldsymbol{\alpha}\|_0 \leq s$ manifests that at most s non-zero elements are contained in $\boldsymbol{\alpha}$, i.e., the number of support vectors is expected to be less than s by (1.3). In this way, the number is able to be controlled and could be small.

Actually, if we remove the sparsity constraint, i.e.,

$$(1.6) \quad \min_{\boldsymbol{\alpha} \in \mathbb{R}^m} D(\boldsymbol{\alpha}), \quad \text{s.t. } \langle \boldsymbol{\alpha}, \mathbf{y} \rangle = 0,$$

then this model is the dual problem of the following soft-margin SVM problem (see Theorem 2.1)

$$(1.7) \quad \min_{\mathbf{w} \in \mathbb{R}^n} \frac{1}{2} \|\mathbf{w}\|^2 + \sum_{i=1}^m \ell_{c,c} \left[1 - \mathbf{y}_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + \mathbf{b}) \right],$$

where $\ell_{c,c}$ is defined by

$$\ell_{c,c}(\mathbf{t}) := \frac{\mathbf{t}^2}{2} \left[C\rho(\mathbf{t} \geq 0) + c\rho(\mathbf{t} < 0) \right].$$

Note that $\ell_{c,c}$ is reduced to the squared hinge loss $(\ell_h(\mathbf{t}))^2$ when $\mathbf{c} = 0$. Now we summarize the main contributions of this paper.

- (C1) As we mentioned above, the sparsity constrained model (1.4) has its advantages. It is able to govern the number of support vectors and hence does sample reduction sufficiently, which ensures fast computation and alleviates the demand for huge volumes of hardware memory.
- (C2) Based on the established optimality conditions, associated with the stationary equations by Theorem 2.5, a subspace Newton method is employed. The method is shown in very low computational complexity and also enjoys one-step convergence property. Namely, the method converges to a stationary point within one step if the chosen starting point is close enough to the stationary point, see Theorem 3.1. It is worth mentioning that since \mathbf{s} is unknown in practice, the selection of $\mathbf{s} \in \mathbb{N}_m$ is somewhat tedious. Fortunately, we manage designing a mechanism to tune the sparsity level \mathbf{s} adaptively.
- (C3) Numerical experiments have demonstrated that the proposed method performs exceptionally well, especially for datasets with $m \gg n$, namely, the number of samples being far greater than the number of features. When comparing with two powerful existing solvers, it takes much shorter computational time due to a tiny number of support vectors being used.

1.3 Preliminaries

We present some notation to be employed throughout the paper. For the sake of easy reference, we list all relevant ones in the following table including those that have been given in the above part of this section.

\mathbb{N}_m	Index set $\{1, 2, \dots, m\}$.
$ \mathbf{T} $	The number of elements of an index set $\mathbf{T} \subseteq \mathbb{N}_m$.
$\bar{\mathbf{T}}$	The complementary set of an index set \mathbf{T} , namely, $\mathbb{N}_m \setminus \mathbf{T}$.
$\boldsymbol{\alpha}_{\mathbf{T}}$	The sub-vector of $\boldsymbol{\alpha}$ indexed on \mathbf{T} and $\boldsymbol{\alpha}_{\mathbf{T}} \in \mathbb{R}^{ \mathbf{T} }$.
$ \boldsymbol{\alpha} $	$:= (\alpha_1 , \dots, \alpha_m)^\top$.
$\ \boldsymbol{\alpha}\ _{[s]}$	The s th largest element of $ \boldsymbol{\alpha} $.
$\ \boldsymbol{\alpha}\ _\infty$	The ℓ_∞ norm of $\boldsymbol{\alpha}$, namely, $\max_i \alpha_i $.
$\ \boldsymbol{\alpha}\ _0$	The ℓ_0 norm of $\boldsymbol{\alpha}$, counting the number of non-zero elements of $\boldsymbol{\alpha}$.
$\text{supp}(\boldsymbol{\alpha})$	The support set of $\boldsymbol{\alpha}$, namely, $\{i \in \mathbb{N}_m : \alpha_i \neq 0\}$.
\mathbf{y}	The labels/classes $(\mathbf{y}_1, \dots, \mathbf{y}_m)^\top \in \mathbb{R}^m$.
\mathbf{X}	The samples data $[\mathbf{x}_1 \dots \mathbf{x}_m]^\top \in \mathbb{R}^{m \times n}$.
\mathbf{Q}	$:= [\mathbf{y}_1 \mathbf{x}_1 \dots \mathbf{y}_m \mathbf{x}_m] \in \mathbb{R}^{n \times m}$.
$\mathbf{Q}_{\mathbf{T}}$	The sub-matrix containing the columns of \mathbf{Q} indexed on \mathbf{T} .
$\mathbf{Q}_{\Gamma, \mathbf{T}}$	The sub-matrix containing the rows of $\mathbf{Q}_{\mathbf{T}}$ indexed on Γ .
\mathbf{P}	$:= \mathbf{I}/C + \mathbf{Q}^\top \mathbf{Q}$.
$\mathbf{1}$	$:= (1, \dots, 1)^\top$ whose dimension varies in the context.
\mathbf{I}	The identity matrix whose dimension varies in the context.
$\ \mathbf{Q}\ $	Spectral norm of the matrix \mathbf{Q} , returning its maximum singular value.
$\mathbf{U}(\boldsymbol{\alpha}, \delta)$	The neighbourhood of $\boldsymbol{\alpha}$ with radius $\delta > 0$, i.e., $\{\mathbf{u} \in \mathbb{R}^m : \ \mathbf{u} - \boldsymbol{\alpha}\ < \delta\}$.

The hard-thresholding operator is denoted as \mathbb{P}_s , defined by

$$(1.8) \quad \mathbb{P}_s(\boldsymbol{\alpha}) = \underset{\mathbf{u}}{\text{argmin}} \left\{ \|\mathbf{u} - \boldsymbol{\alpha}\| : \|\mathbf{u}\|_0 \leq s \right\},$$

which can be obtained by retaining the s largest elements in magnitude from $\boldsymbol{\alpha}$ and setting the remaining to zero. To well characterize the solution of (1.8), we define a useful set by

$$(1.9) \quad \mathbb{T}_s(\boldsymbol{\alpha}) := \left\{ \mathbf{T} \subseteq \mathbb{N}_m : |\mathbf{T}| = s, \mathbf{T} \text{ contains indices of } s \text{ largest elements of } |\boldsymbol{\alpha}| \right\}.$$

Since the s th largest element of $|\boldsymbol{\alpha}|$ may not be unique, $\mathbb{T}_s(\boldsymbol{\alpha})$ might have multiple elements, so does $\mathbb{P}_s(\boldsymbol{\alpha})$. For example, for $\boldsymbol{\alpha} = (1, -1, 0, 0)^\top$, $\mathbb{T}_1(\boldsymbol{\alpha}) = \{\{1\}, \{2\}\}$, $\mathbb{T}_2(\boldsymbol{\alpha}) = \{\{1, 2\}\}$ and $\mathbb{T}_3(\boldsymbol{\alpha}) = \{\{1, 2, 3\}, \{1, 2, 4\}\}$. This definition of \mathbb{T}_s allows us to express \mathbb{P}_s as

$$(1.10) \quad \mathbb{P}_s(\boldsymbol{\alpha}) := \left\{ \begin{bmatrix} \boldsymbol{\alpha}_{\mathbf{T}} \\ 0 \end{bmatrix} : \mathbf{T} \in \mathbb{T}_s(\boldsymbol{\alpha}) \right\}.$$

Finally, we have some observations for the objective function in (1.4),

$$(1.11) \quad D(\boldsymbol{\alpha}) := \frac{1}{2} \|\mathbf{Q}\boldsymbol{\alpha}\|^2 + \sum_{i=1}^m h_{c,C}(\mathbf{u}_i) - \langle \mathbf{1}, \boldsymbol{\alpha} \rangle.$$

Note that its gradient $\nabla D(\boldsymbol{\alpha})$ and Hessian matrix $\mathbf{H}(\boldsymbol{\alpha})$ are

$$(1.12) \quad \begin{aligned} \nabla D(\boldsymbol{\alpha}) &:= \mathbf{H}(\boldsymbol{\alpha})\boldsymbol{\alpha} - \mathbf{1}, \\ \mathbf{H}(\boldsymbol{\alpha}) &:= \mathbf{Q}^\top \mathbf{Q} + \mathbf{E}(\boldsymbol{\alpha}), \end{aligned}$$

where $E(\boldsymbol{\alpha})$ is a diagonal matrix with diagonal elements given by

$$(1.13) \quad E_{ii}(\boldsymbol{\alpha}) := (E(\boldsymbol{\alpha}))_{ii} = \begin{cases} 1/C, & \alpha_i \geq 0, \\ 1/c, & \alpha_i < 0. \end{cases}$$

We need emphasize that $E_{ii}(\boldsymbol{\alpha}) \in [1/C, 1/c]$ if $\alpha_i = 0$ based on the concept of sub-differential [25, Definition 8.3]. For simplicity, we choose $E_{ii}(\boldsymbol{\alpha}) = 1/C$. It is easy to see that the Hessian matrix is positive definite for any $\boldsymbol{\alpha} \in \mathbb{R}^m$ due to

$$(1.14) \quad H(\boldsymbol{\alpha}) \succeq Q^\top Q + I/C = P \succ 0,$$

where $A \succeq 0$ ($A \succ 0$) means A is semi-definite (definite) positive. Here $A \preceq B$ stands for $A - B$ being semi-definite positive. The above condition indicates $D(\boldsymbol{\alpha})$ is a strongly convex function and thus enjoys the property

$$(1.15) \quad \begin{aligned} D(\boldsymbol{\alpha}) &= D(\boldsymbol{\alpha}') + \langle \nabla D(\boldsymbol{\alpha}'), \boldsymbol{\alpha} - \boldsymbol{\alpha}' \rangle + \langle \boldsymbol{\alpha} - \boldsymbol{\alpha}', H(\boldsymbol{\alpha}_t)(\boldsymbol{\alpha} - \boldsymbol{\alpha}') \rangle / 2, \\ &\stackrel{(1.14)}{\geq} D(\boldsymbol{\alpha}') + \langle \nabla D(\boldsymbol{\alpha}'), \boldsymbol{\alpha} - \boldsymbol{\alpha}' \rangle + \langle \boldsymbol{\alpha} - \boldsymbol{\alpha}', P(\boldsymbol{\alpha} - \boldsymbol{\alpha}') \rangle / 2, \end{aligned}$$

for any $\boldsymbol{\alpha}, \boldsymbol{\alpha}' \in \mathbb{R}^m$, where $\boldsymbol{\alpha}_t = \boldsymbol{\alpha} + t(\boldsymbol{\alpha}' - \boldsymbol{\alpha})$ with $t \in [0, 1]$ and the equation is guaranteed by the mean value theorem. For notational convenience, hereafter, let

$$(1.16) \quad \mathbf{z} := \begin{bmatrix} \boldsymbol{\alpha} \\ \mu \end{bmatrix}, \quad \mathbf{z}^* := \begin{bmatrix} \boldsymbol{\alpha}^* \\ \mu^* \end{bmatrix}, \quad \mathbf{z}^k := \begin{bmatrix} \boldsymbol{\alpha}^k \\ \mu^k \end{bmatrix}.$$

Based on which, we denote the following functions

$$(1.17) \quad \begin{aligned} g(\mathbf{z}) &:= \nabla D(\boldsymbol{\alpha}) + \mathbf{y}\mu \stackrel{(1.12)}{=} H(\boldsymbol{\alpha})\boldsymbol{\alpha} - \mathbf{1} + \mathbf{y}\mu, \\ g_T(\mathbf{z}) &:= (g(\mathbf{z}))_T, \quad H_T(\boldsymbol{\alpha}) := (H(\boldsymbol{\alpha}))_{TT}. \end{aligned}$$

So $g(\mathbf{z})$ is a vector and $g_T(\mathbf{z})$ is a sub-vector of $g(\mathbf{z})$. $H(\boldsymbol{\alpha})$ is a matrix and $H_T(\boldsymbol{\alpha})$ is the sub-principal matrix of $H(\boldsymbol{\alpha})$ indexed by T .

1.4 Organization

The rest of the paper is organized as follows. In the next section, we focus on the sparsity constrained model (1.4), establishing its optimality conditions including KKT points and stationary points and also introducing the stationary equations. With the help of the stationary equations, we design the subspace Newton method for SVM (SNSVM) in Section 3, and prove it converges to a stationary point within one step. What is more, a strategy of tuning the sparsity level s is employed into SNSVM to derive SNASVM. Numerical experiments are presented in Section 4, where the implementation of SNASVM as well as its comparisons with two powerful solvers are provided. Concluding remarks are made in the last section.

2 Sparsity Constrained Kernel SVM

We first derive the dual problem (1.6) of (1.7) by the following theorem.

Theorem 2.1 *The dual problem of (1.7) is (1.6) and admits the unique optimal solution denoted as α^* . Furthermore, the optimal solution of the primal model (1.7) is*

$$(2.1) \quad \mathbf{w}^* = Q\alpha^*, \quad \mathbf{b}^* = \frac{1}{m} \langle \mathbf{y}, \mathbf{1} - H(\alpha^*)\alpha^* \rangle.$$

Proof Introducing $\mathbf{u}_i = 1 - \mathbf{y}_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + \mathbf{b})$, (1.7) is rewritten as follows:

$$(2.2) \quad \begin{aligned} \min_{\mathbf{w}, \mathbf{u}} \quad & \frac{1}{2} \|\mathbf{w}\|^2 + \sum_{i=1}^m \ell_{c,C}(\mathbf{u}_i), \\ \text{s.t.} \quad & \mathbf{u}_i + \mathbf{y}_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + \mathbf{b}) = 1, \quad i \in \mathbb{N}_m. \end{aligned}$$

To derive the conclusion, we consider three sub-problems:

$$(2.3) \quad \min_{\mathbf{w}} \quad \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^m \alpha_i \mathbf{y}_i \langle \mathbf{w}, \mathbf{x}_i \rangle = -\frac{1}{2} \|Q\alpha\|^2,$$

where the optimal solution is attained at

$$(2.4) \quad \mathbf{w} = Q\alpha.$$

The second sub-problem is

$$(2.5) \quad \min_{\mathbf{u}} \quad \sum_{i=1}^m \left\{ \ell_{c,C}(\mathbf{u}_i) - \alpha_i \mathbf{u}_i \right\} = -\frac{1}{2C} \|\alpha\|^2.$$

In fact, the optimal solution is attained at the optimality condition

$$(2.6) \quad 0 = \ell'_{c,C}(\mathbf{u}_i) - \alpha_i \iff \alpha_i = \begin{cases} C\mathbf{u}_i, & \mathbf{u}_i \geq 0, \\ c\mathbf{u}_i, & \mathbf{u}_i < 0, \end{cases}$$

which suffices to

$$\begin{aligned} \ell_{c,C}(\mathbf{u}_i) - \alpha_i \mathbf{u}_i &= \begin{cases} \frac{C}{2} \mathbf{u}_i^2 - \alpha_i \mathbf{u}_i, & \mathbf{u}_i \geq 0, \\ \frac{c}{2} \mathbf{u}_i^2 - \alpha_i \mathbf{u}_i, & \mathbf{u}_i < 0, \end{cases} \\ &= \begin{cases} -\frac{\alpha_i^2}{2C}, & \alpha_i \geq 0, \\ -\frac{\alpha_i^2}{2c}, & \alpha_i < 0, \end{cases} = -h_{c,C}(\alpha_i). \end{aligned}$$

The third sub-problem is

$$(2.7) \quad \min_{\mathbf{b}} \quad \sum_{i=1}^m \alpha_i \mathbf{y}_i \mathbf{b} = 0,$$

where the optimal solution is attained at

$$(2.8) \quad \langle \boldsymbol{\alpha}, \mathbf{y} \rangle = 0.$$

These three sub-problems allow us to derive the dual problem by

$$\begin{aligned} & \max_{\boldsymbol{\alpha}} \left\{ \min_{\mathbf{w}, \mathbf{u}} \frac{1}{2} \|\mathbf{w}\|^2 + \sum_{i=1}^m \ell_{c,C}(\mathbf{u}_i) - \sum_{i=1}^m \alpha_i \left(\mathbf{u}_i + \mathbf{y}_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + \mathbf{b}) - 1 \right) \right\} \\ &= \max_{\boldsymbol{\alpha}} \left\{ \min_{\mathbf{w}} \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^m \alpha_i \mathbf{y}_i \langle \mathbf{w}, \mathbf{x}_i \rangle + \sum_{i=1}^m \alpha_i + \right. \\ & \quad \left. \min_{\mathbf{z}} \sum_{i=1}^m \left(\ell_{c,C}(\mathbf{u}_i) - \alpha_i \mathbf{u}_i \right) - \min_{\mathbf{b}} \sum_{i=1}^m \alpha_i \mathbf{y}_i \mathbf{b} \right\} \\ &= \max \left\{ -\frac{1}{2} \|\mathbf{Q}\boldsymbol{\alpha}\|^2 - \sum_{i=1}^m h_{c,C}(\alpha_i) + \sum_{i=1}^m \alpha_i : \langle \boldsymbol{\alpha}, \mathbf{y} \rangle = 0 \right\}. \end{aligned}$$

For the primal model (1.7), \mathbf{w}^* is obtained by (2.4). As for \mathbf{b}^* , it follows from

$$\begin{aligned} \mathbf{y}_i \mathbf{b}^* & \stackrel{(2.2)}{=} 1 - \mathbf{u}_i - \langle \mathbf{w}^*, \mathbf{y}_i \mathbf{x}_i \rangle \\ & \stackrel{(2.4)}{=} 1 - \mathbf{u}_i - \langle \mathbf{Q}\boldsymbol{\alpha}^*, \mathbf{y}_i \mathbf{x}_i \rangle \\ & \stackrel{(2.6)}{=} 1 - \mathbf{E}_{ii}(\boldsymbol{\alpha}^*) \alpha_i^* - \langle \mathbf{Q}\boldsymbol{\alpha}^*, \mathbf{y}_i \mathbf{x}_i \rangle, \end{aligned}$$

for all $i \in \mathbb{N}_m$ that

$$\mathbf{y} \mathbf{b}^* = \mathbf{1} - \mathbf{E}(\boldsymbol{\alpha}^*) \boldsymbol{\alpha}^* - \mathbf{Q}^\top \mathbf{Q} \boldsymbol{\alpha}^* = \mathbf{1} - \mathbf{H}(\boldsymbol{\alpha}^*) \boldsymbol{\alpha}^*.$$

Multiplying both sides of the above equation by \mathbf{y} yields \mathbf{b}^* in (2.1) due to $\langle \mathbf{y}, \mathbf{y} \rangle = \mathbf{m}$, which completes the whole proof. \square

2.1 Optimality

In this part, we focus on the sparsity constrained kernel SVM problem, namely, (1.4). First of all, we can conclude that it admits a global solution/minimizer.

Theorem 2.2 *The global minimizers of (1.4) exist.*

Proof The solution set is non-empty since 0 satisfies the constraints of (1.4). The problem can be written as

$$(2.9) \quad \min_{|\mathbf{T}| \leq s, \mathbf{T} \subseteq \mathbb{N}_m} \left\{ \min_{\boldsymbol{\alpha} \in \mathbb{R}^m} D(\boldsymbol{\alpha}) : \langle \boldsymbol{\alpha}_{\mathbf{T}}, \mathbf{y}_{\mathbf{T}} \rangle = 0 \right\}.$$

It follows from (1.14) that $D(\cdot)$ is strongly convex. So the inner problem is a strongly convex program which admits a unique solution, say $\boldsymbol{\alpha}_{\mathbf{T}}$. In addition, the choices of \mathbf{T} such that $|\mathbf{T}| \leq s, \mathbf{T} \subseteq \mathbb{N}_m$ are finitely many. To derive the global optimal solution, we just pick one \mathbf{T} from those choices making $D(\boldsymbol{\alpha}_{\mathbf{T}})$ the smallest. \square

2.1.1 KKT Points

We say α^* is a KKT point of (1.4) if there is $\mu^* \in \mathbb{R}$ such that

$$(2.10) \quad \begin{cases} g_{s_*}(\mathbf{z}^*) = 0, \\ \langle \alpha^*, \mathbf{y} \rangle = 0, \\ \|\alpha^*\|_0 = s, \end{cases} \quad \text{or} \quad \begin{cases} g(\mathbf{z}^*) = 0, \\ \langle \alpha^*, \mathbf{y} \rangle = 0, \\ \|\alpha^*\|_0 < s. \end{cases}$$

It follows from [21, Theorem 3.2] that the following relationships hold for (1.4).

Theorem 2.3 Consider (1.4) and a point α^* satisfying $\|\alpha^*\|_0 \leq s$ and $\langle \alpha^*, \mathbf{y} \rangle = 0$.

- a) It is a local minimizer if and only if it is a KKT point.
- b) If $\|\alpha^*\|_0 < s$, then the local minimizer, global minimizer and KKT point are identical and unique.

Proof We only prove that the KKT point α^* is unique if $\|\alpha^*\|_0 < s$ since the rest parts can be seen in [21, Theorem 3.2]. If there is an other KKT point $\alpha \neq \alpha^*$, then the strong convexity of $D(\cdot)$ gives rise to

$$\begin{aligned} D(\alpha) &\stackrel{(1.15)}{\geq} D(\alpha^*) + \langle \alpha - \alpha^*, P(\alpha - \alpha^*) \rangle / 2 + \langle \nabla D(\alpha^*), \alpha - \alpha^* \rangle \\ &\stackrel{(1.17)}{=} D(\alpha^*) + \langle \alpha - \alpha^*, P(\alpha - \alpha^*) \rangle / 2 + \langle g(\mathbf{z}^*) - \mathbf{y}\mu^*, \alpha - \alpha^* \rangle \\ &\stackrel{(2.10)}{=} D(\alpha^*) + \langle \alpha - \alpha^*, P(\alpha - \alpha^*) \rangle / 2 - \langle \mathbf{y}\mu^*, \alpha - \alpha^* \rangle \\ &\stackrel{(2.10)}{=} D(\alpha^*) + \langle \alpha - \alpha^*, P(\alpha - \alpha^*) \rangle / 2 \\ &\stackrel{(1.14)}{>} D(\alpha^*), \end{aligned}$$

where the third equation is because KKT points α and α^* satisfy $\langle \alpha^*, \mathbf{y} \rangle = \langle \alpha, \mathbf{y} \rangle = 0$. Since a KKT point is also a global minimizer, it follows $D(\alpha^*) \geq D(\alpha)$, which contradicts with the above inequality. Therefore, α^* is unique. \square

2.1.2 η -Stationary Points

We say α^* is an η -stationary point of (1.4) for some $\eta > 0$ if there is $\mu^* \in \mathbb{R}$ such that

$$(2.11) \quad \begin{cases} \alpha^* \in \mathbb{P}_s[\alpha^* - \eta g(\mathbf{z}^*)], \\ 0 = \langle \alpha^*, \mathbf{y} \rangle. \end{cases}$$

From our notation (1.16) that $\mathbf{z}^* = (\alpha^*; \mu^*)$. We also say \mathbf{z}^* is an η -stationary point of (1.4) if it satisfies the above conditions. Recall that $\mathbb{P}_s(\alpha)$ in (1.10), the above condition can be equivalently written as

$$(2.12) \quad \begin{cases} g_{s_*}(\mathbf{z}^*) = 0, \\ \eta \|g_{\bar{s}_*}(\mathbf{z}^*)\|_\infty \leq \|\alpha^*\|_{[s]}, \\ \langle \alpha^*, \mathbf{y} \rangle = 0, \\ \|\alpha^*\|_0 \leq s. \end{cases}$$

Hereafter, for a given point α^* , we denote

$$(2.13) \quad S_* := \text{supp}(\alpha^*).$$

Note that $\|\alpha^*\|_{[s]}$ is the s th largest element of $|\alpha^*|$. This indicates if $\|\alpha^*\|_0 < s$ then $\|\alpha^*\|_{[s]} = 0$ and hence the above condition is equivalent to the second case $\|\alpha^*\|_0 < s$ in (2.10). In addition, for the case $\|\alpha^*\|_0 = s$, the above condition is clearly stronger than (2.10). Similar to Theorem 2.3, we have the following relationships. These relationships indicate that to seek a local/global minimizer, we instead find an η -stationary point as the latter is more tractable practically.

Theorem 2.4 Consider (1.4) and a point α^* satisfying $\|\alpha^*\|_0 \leq s$ and $\langle \alpha^*, y \rangle = 0$.

- a) An η -stationary point α^* for some $\eta > 0$ is a local minimizer.
- b) A local minimizer α^* is an η -stationary point either for some $\eta > 0$ if $\|\alpha^*\|_0 < s$ or for some $0 < \eta < \eta^*$ if $\|\alpha^*\|_0 = s$, where

$$(2.14) \quad \eta^* := \frac{\|\alpha^*\|_{[s]}}{2\|H(\alpha^*)\alpha^* - \mathbf{1}\|_\infty} > 0.$$

- c) If $\|\alpha^*\|_0 < s$, then the local minimizer, global minimizer and η -stationary point are identical and unique.
- d) If $\|\alpha^*\|_0 = s$, and the point α^* is an η -stationary point for some η such that

$$(2.15) \quad \left[\frac{1}{C} - \frac{1}{\eta} \right] \mathbf{I} + \mathbf{Q}^\top \mathbf{Q} \succeq 0,$$

then it is also a global minimizer. Moreover, it is also unique if the strict \succ holds in the above condition.

Proof a) This is clearly true since an η -stationary point α^* for some $\eta > 0$ satisfying (2.12), which indicates it also satisfies (2.10). Therefore it is a KKT point and a local minimizer from Theorem 2.3 a).

b) If $\|\alpha^*\|_0 < s$, then Theorem 2.3 a) states that a local minimizer α^* satisfies the second condition in (2.10) which is same as (2.12). Therefore, it is also an η -stationary point for some $\eta > 0$. If $\|\alpha^*\|_0 = s$, then $\|\alpha^*\|_{[s]} > 0$, which implies $\eta^* > 0$. A local minimizer satisfies the first condition in (2.10), that

$$g_{S_*}(\mathbf{z}^*) \stackrel{(1.17)}{=} (H(\alpha^*)\alpha^*)_{S_*} - \mathbf{1} + \mathbf{y}_{S_*}\mu^* = 0,$$

which gives rise to

$$|\mu^*| = \|\mathbf{y}_{S_*}\|_\infty |\mu^*| = \|(H(\alpha^*)\alpha^*)_{S_*} - \mathbf{1}\|_\infty$$

because of $|\mathbf{y}| = \mathbf{1}$. In addition $0 < \eta < \eta^*$ gives rise to

$$\begin{aligned}
\|g_{\bar{S}_*}(\mathbf{z}^*)\|_\infty &\stackrel{(1.17)}{=} \|(\mathbf{H}(\boldsymbol{\alpha}^*)\boldsymbol{\alpha}^*)_{\bar{S}_*} - \mathbf{1} + \mathbf{y}_{\bar{S}_*}\boldsymbol{\mu}^*\|_\infty \\
&\leq \|(\mathbf{H}(\boldsymbol{\alpha}^*)\boldsymbol{\alpha}^*)_{\bar{S}_*} - \mathbf{1}\|_\infty + |\boldsymbol{\mu}^*| \|\mathbf{y}_{\bar{S}_*}\|_\infty \\
&= \|(\mathbf{H}(\boldsymbol{\alpha}^*)\boldsymbol{\alpha}^*)_{\bar{S}_*} - \mathbf{1}\|_\infty + |\boldsymbol{\mu}^*| \\
&= \|(\mathbf{H}(\boldsymbol{\alpha}^*)\boldsymbol{\alpha}^*)_{\bar{S}_*} - \mathbf{1}\|_\infty + \|(\mathbf{H}(\boldsymbol{\alpha}^*)\boldsymbol{\alpha}^*)_{S_*} - \mathbf{1}\|_\infty \\
&\leq \|\mathbf{H}(\boldsymbol{\alpha}^*)\boldsymbol{\alpha}^* - \mathbf{1}\|_\infty 2 \\
&= \|\boldsymbol{\alpha}^*\|_{[s]}/\eta^* \\
&\leq \|\boldsymbol{\alpha}^*\|_{[s]}/\eta.
\end{aligned}$$

This verifies the second inequality in (2.12), together with (2.10) claiming the conclusion.

c) An η -stationary point $\boldsymbol{\alpha}^*$ with $\|\boldsymbol{\alpha}^*\|_0 < s$ satisfies the condition (2.12), which is same as the second case $\|\boldsymbol{\alpha}^*\|_0 < s$ in (2.10). Namely, $\boldsymbol{\alpha}^*$ is also a KKT point, which makes the conclusion immediately from Theorem 2.3 b).

d) An η -stationary point $\boldsymbol{\alpha}^*$ with $\|\boldsymbol{\alpha}^*\|_0 = s$ satisfies (2.11), which means for any feasible point $\|\boldsymbol{\alpha}\| \leq s$ and $\langle \boldsymbol{\alpha}, \mathbf{y} \rangle = 0$, we have

$$\|\boldsymbol{\alpha}^* - (\boldsymbol{\alpha}^* - \eta g(\mathbf{z}^*))\|^2 \leq \|\boldsymbol{\alpha} - (\boldsymbol{\alpha}^* - \eta g(\mathbf{z}^*))\|^2.$$

This suffices to

$$(2.16) \quad -\|\boldsymbol{\alpha}^* - \boldsymbol{\alpha}\|^2 \leq 2\eta \langle \boldsymbol{\alpha} - \boldsymbol{\alpha}^*, g(\mathbf{z}^*) \rangle.$$

The strong and quadratic convexity of $D(\cdot)$ gives rise to

$$\begin{aligned}
2D(\boldsymbol{\alpha}) - 2D(\boldsymbol{\alpha}^*) &\stackrel{(1.15)}{\geq} \langle \boldsymbol{\alpha} - \boldsymbol{\alpha}^*, \mathbf{P}(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*) \rangle + 2\langle \nabla D(\boldsymbol{\alpha}^*), \boldsymbol{\alpha} - \boldsymbol{\alpha}^* \rangle \\
&\stackrel{(1.17)}{=} \langle \boldsymbol{\alpha} - \boldsymbol{\alpha}^*, \mathbf{P}(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*) \rangle + 2\langle g(\mathbf{z}^*) - \mathbf{y}\boldsymbol{\mu}^*, \boldsymbol{\alpha} - \boldsymbol{\alpha}^* \rangle \\
&\stackrel{(2.12)}{=} \langle \boldsymbol{\alpha} - \boldsymbol{\alpha}^*, \mathbf{P}(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*) \rangle + 2\langle g(\mathbf{z}^*), \boldsymbol{\alpha} - \boldsymbol{\alpha}^* \rangle \\
&\stackrel{(2.16)}{\geq} \langle \boldsymbol{\alpha} - \boldsymbol{\alpha}^*, \mathbf{P}(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*) \rangle - \|\boldsymbol{\alpha}^* - \boldsymbol{\alpha}\|^2/\eta \\
&= \langle \boldsymbol{\alpha} - \boldsymbol{\alpha}^*, (\mathbf{P} - \mathbf{I}/\eta)(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*) \rangle \\
&\geq 0,
\end{aligned}$$

where the last inequity follows from (2.16) and

$$\mathbf{P} - \frac{\mathbf{I}}{\eta} = \left[\frac{1}{C} - \frac{1}{\eta} \right] \mathbf{I} + \mathbf{Q}^\top \mathbf{Q} \succeq 0.$$

Therefore, $\boldsymbol{\alpha}^*$ is a global minimizer. If there is another global minimizer $\hat{\boldsymbol{\alpha}} \neq \boldsymbol{\alpha}^*$, then the strictness \succ in above condition leads to a contradiction,

$$0 = D(\hat{\boldsymbol{\alpha}}) - D(\boldsymbol{\alpha}^*) \geq \langle \hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*, (\mathbf{P} - \mathbf{I}/\eta)(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*) \rangle > 0.$$

Hence $\boldsymbol{\alpha}^*$ is unique. The whole proof is completed. \square

2.2 Stationary Equations

Based on (1.10), we characterize an η -stationary point of (1.4) as an equation system.

Theorem 2.5 *A point \mathbf{z}^* is an η -stationary point of (1.4) for some $\eta > 0$ if and only if there is a $\mathbf{T}_* \in \mathbb{T}_s(\boldsymbol{\alpha}^* - \eta \mathbf{g}(\mathbf{z}^*))$ such that*

$$(2.17) \quad \mathbf{F}(\mathbf{z}^*; \mathbf{T}_*) := \begin{bmatrix} \mathbf{g}_{\mathbf{T}_*}(\boldsymbol{\alpha}^*) \\ \boldsymbol{\alpha}_{\mathbf{T}_*}^* \\ \langle \boldsymbol{\alpha}_{\mathbf{T}_*}^*, \mathbf{y}_{\mathbf{T}_*} \rangle \end{bmatrix} = \begin{bmatrix} \mathbf{H}_{\mathbf{T}_*}(\boldsymbol{\alpha}^*) \boldsymbol{\alpha}_{\mathbf{T}_*}^* - \mathbf{1} + \mathbf{y}_{\mathbf{T}_*} \mu^* \\ \boldsymbol{\alpha}_{\mathbf{T}_*}^* \\ \langle \boldsymbol{\alpha}_{\mathbf{T}_*}^*, \mathbf{y}_{\mathbf{T}_*} \rangle \end{bmatrix} = 0.$$

Proof It follows from \mathbf{z}^* being an η -stationary point and (2.11) that $\langle \boldsymbol{\alpha}^*, \mathbf{y} \rangle = 0$ and

$$\boldsymbol{\alpha}^* \in \mathbb{P}_s(\boldsymbol{\alpha}^* - \eta \mathbf{g}(\mathbf{z}^*)) \stackrel{(1.10)}{=} \left\{ \begin{bmatrix} \boldsymbol{\alpha}_{\mathbf{T}}^* - \eta \mathbf{g}_{\mathbf{T}}(\mathbf{z}^*) \\ 0 \end{bmatrix} : \mathbf{T} \in \mathbb{T}_s(\boldsymbol{\alpha}^* - \eta \mathbf{g}(\mathbf{z}^*)) \right\},$$

which is equivalent to that there is a $\mathbf{T}_* \in \mathbb{T}_s(\boldsymbol{\alpha}^* - \eta \mathbf{g}(\mathbf{z}^*))$ satisfying $\boldsymbol{\alpha}_{\mathbf{T}_*}^* = 0$ and

$$0 = \mathbf{g}_{\mathbf{T}_*}(\mathbf{z}^*) \stackrel{(1.17)}{=} \mathbf{H}_{\mathbf{T}_*}(\boldsymbol{\alpha}^*) \boldsymbol{\alpha}_{\mathbf{T}_*}^* - \mathbf{1} + \mathbf{y}_{\mathbf{T}_*} \mu^*.$$

This concludes the conclusion immediately. \square

We call (2.17) the stationary equations. Comparing with those conditions in (2.11), equations (2.17) allow us to employ the Newton method. Moreover, it is worth mentioning that if a point \mathbf{z} is an η -stationary point of (1.4) for some $\eta > 0$, then for any fixed $\mathbf{T} \in \mathbb{T}_s(\boldsymbol{\alpha} - \eta \mathbf{g}(\mathbf{z}))$, the following Jacobian matrix of \mathbf{F} is always non-singular,

$$(2.18) \quad \nabla \mathbf{F}(\mathbf{z}; \mathbf{T}) = \begin{bmatrix} \mathbf{H}_{\mathbf{T}}(\boldsymbol{\alpha}) & 0 & \mathbf{y}_{\mathbf{T}} \\ 0 & \mathbf{I} & 0 \\ \mathbf{y}_{\mathbf{T}}^\top & 0 & 0 \end{bmatrix} \succ 0.$$

This is because $\nabla \mathbf{F}(\mathbf{z}; \mathbf{T})$ is congruent to a non-singular matrix

$$(2.19) \quad \begin{bmatrix} \mathbf{H}_{\mathbf{T}}(\boldsymbol{\alpha}) & 0 & \mathbf{y}_{\mathbf{T}} \\ 0 & \mathbf{I} & 0 \\ 0 & 0 & \mathbf{y}_{\mathbf{T}}^\top (\mathbf{H}_{\mathbf{T}}(\boldsymbol{\alpha}))^{-1} \mathbf{y}_{\mathbf{T}} \end{bmatrix},$$

since $\mathbf{H}_{\mathbf{T}}(\boldsymbol{\alpha})$ is positive semi-definite and $\langle \mathbf{y}_{\mathbf{T}}, (\mathbf{H}_{\mathbf{T}}(\boldsymbol{\alpha}))^{-1} \mathbf{y}_{\mathbf{T}} \rangle > 0$ for any \mathbf{T} .

3 Subspace Newton Method

This section applies the Newton method to solve the equation (2.17). Let \mathbf{z}^k be defined in (1.16) and the current approximation to a solution of (2.17). Choose one $\mathbf{T}_s(\boldsymbol{\alpha}^k - \eta \mathbf{g}(\mathbf{z}^k))$. Then Newton's method for (2.17) takes the following form to get the direction $\mathbf{d}^k \in \mathbb{R}^{m+1}$:

$$(3.1) \quad \nabla \mathbf{F}(\mathbf{z}^k; \mathbf{T}_k) \mathbf{d}^k = -\mathbf{F}(\mathbf{z}^k; \mathbf{T}_k).$$

Substituting (2.17) and (2.18) into (3.1) derives

$$(3.2) \quad \begin{bmatrix} H_{T_k}(\boldsymbol{\alpha}^k) & 0 & \mathbf{y}_{T_k} \\ 0 & I & 0 \\ \mathbf{y}_{T_k}^\top & 0 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{d}_{T_k}^k \\ \mathbf{d}_{T_k}^k \\ \mathbf{d}_{m+1}^k \end{bmatrix} = - \begin{bmatrix} g_{T_k}(\mathbf{z}^k) \\ \boldsymbol{\alpha}_{T_k}^k \\ \langle \boldsymbol{\alpha}_{T_k}^k, \mathbf{y}_{T_k} \rangle \end{bmatrix}.$$

After we get the direction, the full Newton step size is taken and brings out

$$(3.3) \quad \mathbf{z}^{k+1} = \mathbf{z}^k + \mathbf{d}^k = \begin{bmatrix} \boldsymbol{\alpha}_{T_k}^k \\ \boldsymbol{\alpha}_{T_k}^k \\ \mu^k \end{bmatrix} + \begin{bmatrix} \mathbf{d}_{T_k}^k \\ \mathbf{d}_{T_k}^k \\ \mathbf{d}_{m+1}^k \end{bmatrix} = \begin{bmatrix} \boldsymbol{\alpha}_{T_k}^k + \mathbf{d}_{T_k}^k \\ 0 \\ \mu^k + \mathbf{d}_{m+1}^k \end{bmatrix}.$$

Now we summarize the whole framework in Algorithm 1. Note that $\mathbf{d}_{T_k}^k$ can be derived directly and the new point is still sparse due to $\|\boldsymbol{\alpha}^{k+1}\|_0 \leq |T_k| = s$. So the major computation is from the part on T_k . Since the space indexed on $T_k \in \mathbb{N}_m$ can be treated as a subspace of the whole space and only has a very small dimension $|T_k| = s$ comparing to m , we call the method subspace Newton method.

Algorithm 1 SNSVM: Subspace Newton method for SVM

Give parameters $C, c > 0, \eta > 0, s \in \mathbb{N}_m, \text{ToI}$ and MaxIt .
Initialize \mathbf{z}^0 , pick $T_0 \in \mathbb{T}_s(\boldsymbol{\alpha}^0 - \eta g(\mathbf{z}^0))$ and set $k := 0$.
while $\|F(\mathbf{z}^k; T_k)\| \geq \text{ToI}$ and $k \leq \text{MaxIt}$ **do**
 Update \mathbf{d}^k by solving (3.2).
 Update \mathbf{z}^{k+1} by (3.3).
 Update $T_{k+1} \in \mathbb{T}_s(\boldsymbol{\alpha}^{k+1} - \eta g(\mathbf{z}^{k+1}))$ and set $k := k + 1$.
end while
return the solution \mathbf{z}^k .

3.1 Complexity Analysis

To derive the Newton direction in (3.2), we need to address the following equations

$$(3.4) \quad \begin{aligned} \mathbf{d}_{m+1}^k &= -\frac{\langle \mathbf{y}_{T_k}, \Theta^{-1} g_{T_k}(\mathbf{z}^k) - \boldsymbol{\alpha}_{T_k}^k \rangle}{\langle \mathbf{y}_{T_k}, \Theta^{-1} \mathbf{y}_{T_k}^k \rangle}, \\ \mathbf{d}_{T_k}^k &= -\Theta^{-1} [g_{T_k}(\mathbf{z}^k) + \mathbf{d}_{m+1}^k \mathbf{y}_{T_k}], \\ \mathbf{d}_{T_k}^k &= -\boldsymbol{\alpha}_{T_k}^k, \end{aligned}$$

where $\Theta := H_{T_k}(\boldsymbol{\alpha}^k)$. Regarding the computational complexity of SNSVM in Algorithm 1, one can observe that Θ^{-1} and \mathbb{T}_s dominate the whole computation. Recall the definition (1.12) of $H(\cdot)$ that

$$\Theta = (E(\boldsymbol{\alpha}^k))_{T_k T_k} + Q_{T_k}^\top Q_{T_k}.$$

The complexity of computing Θ is about $O(n s^2)$ since $|T_k| = s$. And computing its inverse takes complexity about $O(s^\kappa)$ where $\kappa \in (2, 3)$. Therefore, the complexity is

$$(3.5) \quad O\left(\min\{n, s\}s^2\right).$$

If $s \ll m$, the above complexity is relatively low. Otherwise, one could take advantage of the conjugate gradient method to solve the following equations

$$\begin{bmatrix} H_{T_k}(\alpha^k) & y_{T_k} \\ y_{T_k}^\top & 0 \end{bmatrix} \begin{bmatrix} d_{T_k}^k \\ d_{m+1}^k \end{bmatrix} = - \begin{bmatrix} g_{T_k}(z^k) \\ \langle \alpha_{T_k}^k, y_{T_k} \rangle \end{bmatrix},$$

which still possesses a low computational complexity. To pick T_{k+1} from T_s , we need to compute $g(z^{k+1})$ and select the k largest elements of $|\alpha^{k+1} - \eta g(z^{k+1})|$. The complexity of computing the former is about $O(mn)$ and the latter is $O(m + sn)$. Here, we benefit from a MATLAB built-in function `mink` to select the s largest elements. Overall, the total complexity in each step is

$$(3.6) \quad O(mn + \min\{n, s\}s^2).$$

3.2 Convergence Analysis

Before the main convergence property, we define some constants

$$(3.7) \quad \gamma := 2 \max \left\{ 1 + \eta/c + \eta \|Q\|^2, \eta \sqrt{m} \right\},$$

$$(3.8) \quad \eta^* := \begin{cases} \|\alpha^*\|_{[s]} \|g(z^*)\|_\infty^{-1}, & \text{if } \|\alpha^*\|_0 = s, \\ +\infty, & \text{if } \|\alpha^*\|_0 < s, \end{cases}$$

$$(3.9) \quad \delta^* := \begin{cases} \gamma^{-1} (\|\alpha^*\|_{[s]} - \eta \|g(z^*)\|_\infty), & \text{if } \|\alpha^*\|_0 = s, \\ \gamma^{-1} \min_{i \in S_*} |\alpha_i^*|, & \text{if } 0 < \|\alpha^*\|_0 < s, \\ +\infty, & \text{if } \|\alpha^*\|_0 = 0. \end{cases}$$

Based on which, we first present some properties regarding an η -stationary point of (1.4).

Lemma 3.1 *Let z^* be an η -stationary point of (1.4) for some $0 < \eta < \eta^*$, and η^* and δ^* be given by (3.8) and (3.9). Then for any $z \in \mathcal{U}(z^*, \delta^*)$, we have the following results.*

a) *The parameters $\eta^* > 0$ and $\delta^* > 0$.*

b) *For any $T \in T_s(\alpha - \eta g(z))$ and any $T_* \in T_s(\alpha^* - \eta g(z^*))$, it holds*

$$(3.10) \quad \begin{cases} S_* = T_* = T = \text{supp}(\alpha), & \text{if } \|\alpha^*\|_0 = s, \\ S_* \subseteq (T_* \cap T \cap \text{supp}(\alpha)), & \text{if } \|\alpha^*\|_0 < s. \end{cases}$$

c) *For any $T \in T_s(\alpha - \eta g(z))$, it holds*

$$(3.11) \quad F(z^*; T) = 0.$$

Proof a) If $\|\alpha^*\|_0 = s$, then $\|\alpha^*\|_{[s]} > 0$ and so is $\eta^* > 0$, which suffices to $\delta^* = \|\alpha^*\|_{[s]} - \eta\|g(\mathbf{z}^*)\|_\infty > 0$ due to $0 < \eta < \eta^*$. If $\|\alpha^*\|_0 < s$, the claim is true obviously.
b) It follows from Theorem 2.5 and \mathbf{z}^* being an η -stationary point of (1.4) that

$$(3.12) \quad F(\mathbf{z}^*; \mathbb{T}_*) = \begin{bmatrix} g_{\mathbb{T}_*}(\mathbf{z}^*) \\ \alpha_{\mathbb{T}_*}^* \\ \langle \alpha_{\mathbb{T}_*}^*, \mathbf{y}_{\mathbb{T}_*} \rangle \end{bmatrix} = 0.$$

for any $\mathbb{T}_* \in \mathbb{T}_s(\alpha^* - \eta g(\mathbf{z}^*))$. We first derive the following fact

$$(3.13) \quad E(\alpha^*)\alpha^* = E(\alpha)\alpha^*.$$

If $\alpha^* = 0$, the above equation is true clearly. If $\alpha^* \neq 0$, we have

$$(3.14) \quad \alpha_i^* > 0 \implies \alpha_i > 0, \quad \alpha_i^* < 0 \implies \alpha_i < 0.$$

If (3.14) is not true, then there is an j that violates one of the above relations, namely α_j^* and α_j have different signs. As a consequence, we get a contradiction

$$\delta^* \stackrel{(3.9)}{\leq} |\alpha_j^*| < |\alpha_j^* - \alpha_j| \leq \|\alpha^* - \alpha\| < \delta^*.$$

Therefore, we must have (3.14). Recall that $E(\alpha)$ is a diagonal matrix with diagonal elements given by (1.13). It follows

$$\begin{aligned} \left[(E(\alpha^*) - E(\alpha))\alpha^* \right]_i &= \left[E_{ii}(\alpha^*) - E_{ii}(\alpha) \right] \alpha_i^* \\ &= \begin{cases} (1/C - 1/c)\alpha_i^*, & \alpha_i^* > 0 \\ (1/c - 1/C)\alpha_i^*, & \alpha_i^* < 0 \\ (E_{ii}(\alpha^*) - E_{ii}(\alpha))0, & \alpha_i^* = 0 \end{cases} \\ &= 0. \end{aligned}$$

Therefore, the main task is to prove (3.14), before which we present two facts. For any two vectors \mathbf{a} and \mathbf{b} , we have

$$(3.15) \quad \begin{aligned} |\mathbf{a}_i - \mathbf{b}_i| + |\mathbf{a}_j - \mathbf{b}_j| &\leq \sqrt{2} \left[|\mathbf{a}_i - \mathbf{b}_i|^2 + |\mathbf{a}_j - \mathbf{b}_j|^2 \right]^{1/2} \\ &\leq \sqrt{2} \|\mathbf{a} - \mathbf{b}\|. \end{aligned}$$

In addition, it holds

$$\begin{aligned} \|g(\mathbf{z}^*) - g(\mathbf{z})\| &\stackrel{(1.17)}{=} \left\| H(\alpha^*)\alpha^* - H(\alpha)\alpha + (\mu^* - \mu)\mathbf{y} \right\| \\ &\leq \|H(\alpha^*)\alpha^* - H(\alpha)\alpha\| + |\mu^* - \mu| \cdot \|\mathbf{y}\| \\ &\stackrel{(1.12)}{=} \left\| (E(\alpha^*) + Q^\top Q)\alpha^* - (E(\alpha) + Q^\top Q)\alpha \right\| + |\mu^* - \mu| \cdot \|\mathbf{y}\| \\ &\leq \|E(\alpha^*)\alpha^* - E(\alpha)\alpha\| + \|Q\|^2 \|\alpha^* - \alpha\| + |\mu^* - \mu| \cdot \|\mathbf{y}\| \end{aligned}$$

$$\begin{aligned}
& \stackrel{(3.14)}{=} \|E(\boldsymbol{\alpha})(\boldsymbol{\alpha}^* - \boldsymbol{\alpha})\| + \|Q\|^2 \|\boldsymbol{\alpha}^* - \boldsymbol{\alpha}\| + |\mu^* - \mu| \cdot \|\mathbf{y}\| \\
(3.16) \quad & \stackrel{(1.13)}{\leq} \left[1/c + \|Q\|^2\right] \|\boldsymbol{\alpha}^* - \boldsymbol{\alpha}\| + \sqrt{m} |\mu^* - \mu|.
\end{aligned}$$

Now we claim b) by two case: $\|\boldsymbol{\alpha}^*\|_0 = s$ and $\|\boldsymbol{\alpha}^*\|_0 < s$.

Case i) $\|\boldsymbol{\alpha}^*\|_0 = s$. Since $|T_*| = s$ and $\boldsymbol{\alpha}_{T_*}^* = 0$, it holds

$$(3.17) \quad T_* = \text{supp}(\boldsymbol{\alpha}^*) = S_*.$$

We first check $S_* \subseteq \text{supp}(\boldsymbol{\alpha})$. In fact, if it is not true, then there is an $i \in S_*$ but $i \notin \text{supp}(\boldsymbol{\alpha})$, which incurs the following contradiction

$$(3.18) \quad \delta^* \leq \|\boldsymbol{\alpha}^*\|_{[s]} \leq |\alpha_i^*| = |\alpha_i^* - \alpha_i| \leq \|\boldsymbol{\alpha}^* - \boldsymbol{\alpha}\| \leq \|\mathbf{z}^* - \mathbf{z}\| < \delta^*.$$

So $S_* \subseteq \text{supp}(\boldsymbol{\alpha})$, together with $|S_*| = |\text{supp}(\boldsymbol{\alpha})| = s$ yielding $S_* = \text{supp}(\boldsymbol{\alpha})$. We next show $T_* = T$ for any $T \in \mathbb{T}_s(\boldsymbol{\alpha} - \eta g(\mathbf{z}))$. If $T_* \neq T$, owing to $|T_*| = |T| = s$, there is an $i \in T_*, i \notin T$ and a $j \notin T_*, j \in T$. The definition (1.9) of \mathbb{T}_s indicates

$$\begin{aligned}
& |\alpha_j - \eta g_j(\mathbf{z})| \geq |\alpha_i - \eta g_i(\mathbf{z})| \\
(3.19) \quad & |\alpha_i^*| \stackrel{(3.12)}{=} |\alpha_i^* - \eta g_i(\mathbf{z}^*)| \geq |\alpha_j^* - \eta g_j(\mathbf{z}^*)| \stackrel{(3.12)}{=} \eta |g_j(\mathbf{z}^*)|.
\end{aligned}$$

Direct calculation yields the following chain of inequalities,

$$\begin{aligned}
\phi : &= |\alpha_i^* - \eta g_i(\mathbf{z}^*)| - |\alpha_i - \eta g_i(\mathbf{z})| + |\alpha_j - \eta g_j(\mathbf{z})| - |\alpha_j^* - \eta g_j(\mathbf{z}^*)| \\
&\leq \left| \alpha_i^* - \eta g_i(\mathbf{z}^*) - (\alpha_i - \eta g_i(\mathbf{z})) \right| + \left| \alpha_j^* - \eta g_j(\mathbf{z}^*) - (\alpha_j - \eta g_j(\mathbf{z})) \right| \\
&\leq |\alpha_i^* - \alpha_i| + \eta |g_i(\mathbf{z}^*) - g_i(\mathbf{z})| + |\alpha_j^* - \alpha_j| + \eta |g_j(\mathbf{z}^*) - g_j(\mathbf{z})| \\
&\stackrel{(3.15)}{\leq} \sqrt{2} \|\boldsymbol{\alpha}^* - \boldsymbol{\alpha}\| + \eta \sqrt{2} \|g(\mathbf{z}^*) - g(\mathbf{z})\| \\
&\stackrel{(3.16)}{\leq} \sqrt{2} \left[1 + \eta/c + \eta \|Q\|^2\right] \|\boldsymbol{\alpha}^* - \boldsymbol{\alpha}\| + \eta \sqrt{2m} |\mu^* - \mu| \\
&\leq \sqrt{2} \max \left\{1 + \eta/c + \eta \|Q\|^2, \eta \sqrt{m}\right\} \left[\|\boldsymbol{\alpha}^* - \boldsymbol{\alpha}\| + |\mu^* - \mu|\right] \\
&\stackrel{(3.7)}{=} \left[\gamma/\sqrt{2}\right] \left[\|\boldsymbol{\alpha}^* - \boldsymbol{\alpha}\| + |\mu^* - \mu|\right] \\
(3.20) \quad &\stackrel{(3.15)}{\leq} \gamma \|\mathbf{z}^* - \mathbf{z}\| \leq \gamma \delta^*.
\end{aligned}$$

These give rise to the following contradiction,

$$\begin{aligned}
\|\boldsymbol{\alpha}^*\|_{[s]} - \eta \|g(\mathbf{z}^*)\|_\infty &\leq |\alpha_i^*| - \eta |g_j(\mathbf{z}^*)| \\
&\stackrel{(3.19)}{=} |\alpha_i^* - \eta g_i(\mathbf{z}^*)| - |\alpha_j^* - \eta g_j(\mathbf{z}^*)| \\
&\stackrel{(3.19)}{\leq} \phi \stackrel{(3.20)}{\leq} \gamma \delta^* \\
&\stackrel{(3.9)}{<} \|\boldsymbol{\alpha}^*\|_{[s]} - \eta \|g(\mathbf{z}^*)\|_\infty.
\end{aligned}$$

Case ii) $\|\alpha^*\|_0 < s$. The fact $\alpha_{T_*}^* = 0$ from (3.12) indicates $S_* \subseteq T_*$. We next show $S_* \subseteq T \cap \text{supp}(\alpha)$. If $\alpha^* = 0$, then $S_* = \emptyset \subseteq T$ clearly. Therefore, we focus on $\alpha^* \neq 0$. Similar reasoning (3.18) also enables us to show $S_* \subseteq \text{supp}(\alpha)$. Now, we verify $S_* \subseteq T$. Since $\|\alpha^*\|_0 < s$, $\|\alpha^*\|_{[s]} = 0$, which together with (2.12) derives

$$(3.21) \quad g(\mathbf{z}^*) = 0.$$

If $S_* \not\subseteq T$, then the fact $|S_*| < s = |T|$ also indicates that there is an $i \in S_*, i \notin T$ and a $j \notin S_*, j \in T$. This together with the definition (1.9) of T_s results in

$$(3.22) \quad \begin{aligned} |\alpha_j - \eta g_j(\mathbf{z})| &\geq |\alpha_i - \eta g_i(\mathbf{z})|, \\ |\alpha_i^*| &\stackrel{(3.21)}{=} |\alpha_i^* - \eta g_i(\mathbf{z}^*)| > 0 \stackrel{(3.21)}{=} |\alpha_j^* - \eta g_j(\mathbf{z}^*)|, \end{aligned}$$

which leads to the following contradiction

$$\begin{aligned} \min_{i \in S_*} |\alpha_i^*| &\leq |\alpha_i^*| \\ &\stackrel{(3.22)}{=} |\alpha_i^* - \eta g_i(\mathbf{z}^*)| - |\alpha_j^* - \eta g_j(\mathbf{z}^*)| \\ &\stackrel{(3.22)}{\leq} \phi \stackrel{(3.20)}{\leq} \gamma \delta^* \stackrel{(3.9)}{<} \min_{i \in S_*} |\alpha_i^*|. \end{aligned}$$

c) To prove $F(\mathbf{z}^*; T) = 0$, we need to show

$$(3.23) \quad F(\mathbf{z}^*; T) = \begin{bmatrix} g_T(\mathbf{z}^*) \\ \alpha_T^* \\ \langle \alpha_T^*, \mathbf{y}_T \rangle \end{bmatrix} = 0.$$

If $\|\alpha^*\|_0 = s$, then $T = T_*$ by b), which shows the result by (3.12) immediately. If $\|\alpha^*\|_0 < s$, then $g_T(\mathbf{z}^*) = 0$ by (3.21). Again from b), $S_* \subseteq (T \cap T_*)$ means $\bar{T} \subseteq \bar{S}_*$, which indicates $\alpha_{\bar{T}}^* = 0$ due to $\alpha_{\bar{S}_*}^* = 0$. Finally,

$$\begin{aligned} \langle \alpha_T^*, \mathbf{y}_T \rangle &= \langle \alpha_{S_*}^*, \mathbf{y}_{S_*} \rangle + \langle \alpha_{T \setminus S_*}^*, \mathbf{y}_{T \setminus S_*} \rangle = \langle \alpha_{S_*}^*, \mathbf{y}_{S_*} \rangle \\ &= \langle \alpha_{T_*}^*, \mathbf{y}_{T_*} \rangle - \langle \alpha_{T_* \setminus S_*}^*, \mathbf{y}_{T_* \setminus S_*} \rangle \stackrel{(3.12)}{=} 0. \end{aligned}$$

The whole proof is finished. \square

The main convergence result is stated by the following theorem, where one can discern that SNSVM will terminate at the next step if the current point falls into a local area of an η -stationary point. This means if the starting point by chance is chosen within the local area, then SNSVM will take one step to terminate. Hence, it enjoys a very fast convergence property. It is worth mentioning that such convergence property is much better than the quadratic convergence property.

Theorem 3.1 (One step convergence) *Let \mathbf{z}^* be an η -stationary point of (1.4) for some $0 < \eta < \eta^*$, and η^* and δ^* be given by (3.8) and (3.9). Let $\{\mathbf{z}^k\}$ be the sequence generated by SNSVM. As long as there exists one k such that $\mathbf{z}^k \in \mathcal{U}(\mathbf{z}^*, \delta^*)$, then we have*

$$\mathbf{z}^{k+1} = \mathbf{z}^*, \quad \|F(\mathbf{z}^{k+1}, T_{k+1})\| = 0.$$

Namely, SNSVM terminates at the $th(k+1)$ step.

Proof Consider a point $\mathbf{z}_t^k = \mathbf{z}^* + t(\mathbf{z}^k - \mathbf{z}^*)$ with $t \in [0, 1]$. Since $\mathbf{z}^k \in \mathcal{U}(\mathbf{z}^*, \delta^*)$, it also holds $\mathbf{z}_t^k \in \mathcal{U}(\mathbf{z}^*, \delta^*)$ because of

$$\|\mathbf{z}_t^k - \mathbf{z}^*\| = t\|\mathbf{z}^k - \mathbf{z}^*\| \leq \|\mathbf{z}^k - \mathbf{z}^*\| < \delta^*.$$

We first prove that

$$(3.24) \quad E(\boldsymbol{\alpha}^k) = E(\boldsymbol{\alpha}_t^k).$$

In fact, if $\boldsymbol{\alpha}^* = 0$, then $\boldsymbol{\alpha}_t^k = t\boldsymbol{\alpha}^k$, which means $\boldsymbol{\alpha}_t^k$ and $\boldsymbol{\alpha}^k$ have the same signs. This together with the definition (1.13) of $E(\cdot)$ shows (3.24) immediately. If $\boldsymbol{\alpha}^* \neq 0$, then same reasoning proving (3.14) also derives that

$$\begin{aligned} \alpha_i^* > 0 &\implies \alpha_i^k > 0, \quad (\alpha_t^k)_i = (1-t)\alpha_i^* + t\alpha_i^k > 0, \\ \alpha_i^* < 0 &\implies \alpha_i^k < 0, \quad (\alpha_t^k)_i = (1-t)\alpha_i^* + t\alpha_i^k < 0, \\ \alpha_i^* = 0 &\implies (\alpha_t^k)_i = t\alpha_i^k. \end{aligned}$$

These also mean $\boldsymbol{\alpha}_t^k$ and $\boldsymbol{\alpha}^k$ have the same signs. So (3.24) is true and brings out

$$H(\boldsymbol{\alpha}^k) \stackrel{(1.12)}{=} E(\boldsymbol{\alpha}^k) + Q^\top Q \stackrel{(3.24)}{=} E(\boldsymbol{\alpha}_t^k) + Q^\top Q = H(\boldsymbol{\alpha}_t^k).$$

Then for any $T_k \in \mathbb{T}_s(\boldsymbol{\alpha}^k - \eta g(\mathbf{z}^k))$, the above equation contributes to

$$\nabla F(\mathbf{z}_t^k; T_k) = \begin{bmatrix} H_{T_k}(\boldsymbol{\alpha}_t^k) & 0 & \mathbf{y}_{T_k} \\ 0 & I & 0 \\ \mathbf{y}_{T_k}^\top & 0 & 0 \end{bmatrix} = \begin{bmatrix} H_{T_k}(\boldsymbol{\alpha}^k) & 0 & \mathbf{y}_{T_k} \\ 0 & I & 0 \\ \mathbf{y}_{T_k}^\top & 0 & 0 \end{bmatrix} = \nabla F(\mathbf{z}^k; T_k).$$

It follows from the mean value theorem that there exists a \mathbf{z}_t^k satisfying

$$\begin{aligned} F(\mathbf{z}^k; T_k) &\stackrel{(3.11)}{=} F(\mathbf{z}^k; T_k) - F(\mathbf{z}^*; T_k) \\ &= \nabla F(\mathbf{z}_t^k; T_k)(\mathbf{z}^k - \mathbf{z}^*) \\ &= \nabla F(\mathbf{z}^k; T_k)(\mathbf{z}^k - \mathbf{z}^*), \end{aligned}$$

which together with $\nabla F(\mathbf{z}^k; T_k)$ being always non-singular because of (2.18) suffices to

$$\begin{aligned} \mathbf{z}^* &= \mathbf{z}^k - (\nabla F(\mathbf{z}^k; T_k))^{-1} F(\mathbf{z}^k; T_k) \\ &\stackrel{(3.1)}{=} \mathbf{z}^k + \mathbf{d}^k \stackrel{(3.3)}{=} \mathbf{z}^{k+1}. \end{aligned}$$

Finally, for any $T_{k+1} \in \mathbb{T}_s(\boldsymbol{\alpha}^{k+1} - \eta g(\mathbf{z}^{k+1})) = \mathbb{T}_s(\boldsymbol{\alpha}^* - \eta g(\mathbf{z}^*))$, it follows from \mathbf{z}^* being an η -stationary point that

$$\|F(\mathbf{z}^{k+1}, T_{k+1})\| = \|F(\mathbf{z}^*, T_{k+1})\| \stackrel{(2.17)}{=} 0.$$

The whole proof is completed. \square

3.3 The Sparsity Level Tuning

One major issue we encounter is that the sparsity level s in (1.4) usually is unknown beforehand. And it plays two important roles: (i) The larger s , the better classifications since more samples are taken into consideration. (ii) However, the smaller s , the faster computational speed of the method in Algorithm 1 as the complexity in (3.6) relies on s . Moreover, from (1.3) that the number of support vectors is smaller than $\|\alpha\|_0 \leq s$. Because of this, the smaller s , the smaller the number of support vectors.

Therefore, to balance these two, we design the following rule to update the unknown s . Start with small integer and then increase it until to satisfy some conditions. We thus derive **SNASVM** (subspace Newton method with adaptively tuning the sparsity level s for SVM) in Algorithm 2, where the halting condition becomes

$$\left| \text{ACC}(\alpha^k) - \max\{\text{ACC}(\alpha^1), \dots, \text{ACC}(\alpha^{k-1})\} \right| \leq 10^{-4} \quad \text{and} \quad \|F(\mathbf{z}^k; T_k)\| \leq \text{To1}$$

and the classification accuracy is defined by

$$(3.25) \quad \text{ACC}(\alpha) := \left[1 - \frac{1}{m} \|\text{sgn}(X\alpha + \mathbf{b}) - \mathbf{y}\|_0 \right] \times 100\%,$$

with \mathbf{b} being derived by (2.1) and $\text{sgn}(\mathbf{t}) = 1$ if $\mathbf{t} > 0$ and -1 otherwise. The above halting condition means that α^k is almost an η -stationary point due to the small value of $\|F(\mathbf{z}^k; T_k)\|$, while the classification accuracy does not increase significantly even though s still ascends. Therefore, it is unreasonable to keep s rising to achieve a better solution since the bigger s would cause more computational costs. So it is suggested to stop it if α^k satisfies such a halt condition. Our numerical experiments demonstrate that **SNASVM** under this rule works very well.

Algorithm 2 SNASVM: Subspace Newton method with adaptively tuning s for SVM

Give parameters $C, c > 0, \eta > 0, r > 1, \text{MaxACC} = 0, s_0 \in \mathbb{N}_m, \text{To1}$ and MaxIt .

Initialize \mathbf{z}^0 , pick $T_0 \in \mathbb{T}_s(\alpha^0 - \eta g(\mathbf{z}^0))$ and set $k := 0$.

while ($\|F(\mathbf{z}^k; T_k)\| \geq \text{To1}$ or $|\text{ACC}(\alpha^k) - \text{MaxACC}| > 10^{-4}$) and ($k \leq \text{MaxIt}$) **do**

 Update \mathbf{d}^k by solving (3.2).

 Update \mathbf{z}^{k+1} by (3.3).

 Update $\text{MaxACC} = \max\{\text{ACC}(\alpha^1), \dots, \text{ACC}(\alpha^{k-1})\}$.

 Update $s_{k+1} = rs_k$ if k is a multiple of 10 and $s_{k+1} = s_k$ otherwise.

 Update $T_{k+1} \in \mathbb{T}_{s_{k+1}}(\alpha^{k+1} - \eta g(\mathbf{z}^{k+1}))$ and set $k := k + 1$.

end while

return the solution \mathbf{z}^k .

4 Numerical Experiments

This part conducts numerical experiments of **SNASVM** in Algorithm 2 by using MATLAB (R2019a) on a laptop of 32GB memory and Inter(R) Core(TM) i9-9880H 2.3Ghz CPU.

4.1 Implementation

The starting point \mathbf{z}^0 is initialized as $\boldsymbol{\alpha}^0 = 0$ and $\boldsymbol{\mu}^0 = \text{sgn}(\langle \mathbf{y}, \mathbf{1} \rangle)$. Parameters are tuned as follows. The maximum number of iteration and the tolerance are set as $\text{maxIt} = 1000$, $\text{tol} = \max\{\sqrt{mn}\}10^{-6}$. Empirically numerical experience has demonstrated that the involved parameters such as C, c or η are suggested to be selected through the cross validation for better results. However, for simplicity, we fix them as $C = 1, c = 0.01, \eta = 1/m$ and $r = 1.15$. The last parameter s_0 is chosen as $s_0 = 100 \log_{10}(m)$ for synthetic datasets and $s_0 = \beta \log_{10}(m)$ for real datasets, where

$$\beta = \begin{cases} 1 + 10^{-3}n, & \text{if } m/n < 100, \\ 10^{-2}n, & \text{if } 100 \leq m/n < 60000, \\ 50n, & \text{if } m/n \geq 60000. \end{cases}$$

4.2 Testing Examples

We first consider a two-dimensional ($n = 2$) example with synthetic data, where the features come from Gaussian distributions.

Example 4.1 (Synthetic data in \mathbb{R}^2 [32, 12]) *In this example, samples \mathbf{x}_i with positive labels $y_i = +1$ are drawn from the normal distribution with mean $(0.5, -3)^\top$ and variance Σ and samples \mathbf{x}_j with negative labels $y_j = -1$ are drawn from the normal distribution with mean $(-0.5, 3)^\top$ and variance Λ , where Σ and Λ are diagonal matrices with $\Sigma_{11} = \Lambda_{11} = 0.2, \Sigma_{22} = \Lambda_{22} = 3$. We generate m samples with two classes having equal numbers and then evenly split all samples into a training set and a testing set. Finally, we randomly flip $r\%$ percentage of labels in the training data, which means $r\%$ percentage of samples are treated as outliers.*

Example 4.2 (Real data in higher dimensions) *We select 21 datasets with $m \gg n$ from the libraries: libsvm*, uci† and kiggle‡. All datasets are feature-wisely scaled to $[-1, 1]$ and all the classes being not 1 are treated as -1 . Their details are presented in Table 1. For each of those without testing data, we split it into two parts. The first part containing 90% of samples is treated as the training data and the rest is the testing data.*

To compare the performance of all selected methods, let $\boldsymbol{\alpha}$ be the solution/classifier generated by one method. We report the CPU time (TIME), the training classification accuracy (ACC) by (3.25) where \mathbf{X} and \mathbf{y} are the training samples and classes, the testing classification accuracy (TACC) by (3.25) where \mathbf{X} and \mathbf{y} are the testing samples and classes, and the number of support vectors (NSV).

*<https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/>

†<http://archive.ics.uci.edu/ml/datasets.php>

‡<https://www.kaggle.com/datasets>

Table 1: Data sets with less training samples and more features, namely $m > n$.

Data	Datasets	Source	n	Train	Test
mrpe	malware analysis datasets: raw pe	kaggle	1024	51959	0
dhrb	hospital readmissions binary	kaggle	17	59557	0
aips	airline passenger satisfaction	kaggle	22	103904	25976
sctp	santander customer transaction	kaggle	200	200000	0
skin	skin_nonskin	libsvm	3	245056	0
ccfd	credit card fraud detection	kaggle	28	284807	0
rlc1		uci	9	574914	0
rlc2		uci	9	574914	0
rlc3		uci	9	574914	0
rlc4		uci	9	574914	0
rlc5	record linkage	uci	9	574914	0
rlc6	comparison patterns	uci	9	574914	0
rlc7		uci	9	574914	0
rlc8		uci	9	574914	0
rlc9		uci	9	574914	0
rlc10		uci	9	574914	0
covt	covtype.binary	libsvm	54	581012	0
retb	real time bidding	kaggle	88	1000000	0
susy	susy	uci	18	5000000	0
hepm	hepmass	uci	28	7000000	3500000
higg	higgs	uci	28	11000000	0

4.3 Numerical Comparisons

(a) **Benchmark methods.** There are numerous excellent methods have been developed to tackle the SVM [29, 3, 22, 16, 14, 31, 11, 35]. Those methods perform extremely well, especially for datasets in small or mediate size. Some of them calculate the kernel matrix $Q^T Q$ with size $m \times m$, and thus require a huge volume of hardware memory if m is large (e.g., $m \geq 10^5$). Note that most datasets in Table 2 have at least 10^5 samples. Therefore, we only select a Matlab built-in solver `fitclinear`[§] and `liblinear`[¶] [8] for comparisons since they are very fast to deal with those datasets. For the former, we set `Learner = 'svm'` and `Solver = 'dual'` in order to obtain the number of the support vectors. For the same reason, the dual problem of the ℓ_2 -regularized ℓ_1 -loss support vector classification is chosen to compute in `liblinear`. To do so, we set `-s 3` in the function `train` from `liblinear`. It may be much faster if we set `-s 2`, but such a setting suits for computing the primal model and does not render the number of support vectors.

[§]<https://mathworks.com/help/stats/fitclinear.html>

[¶]<https://www.csie.ntu.edu.tw/~cjlin/liblinear/>

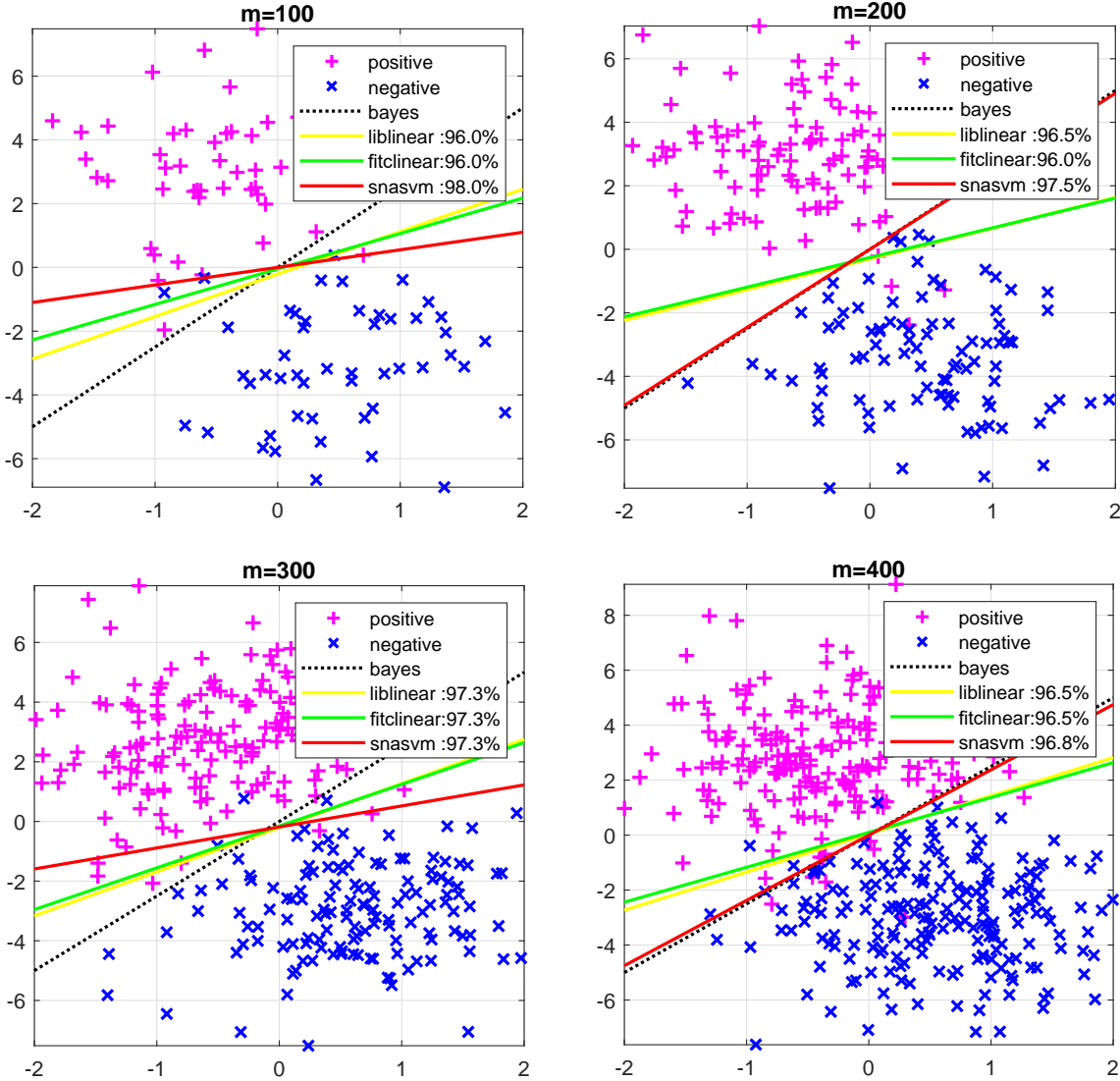


Figure 1: Classifiers by three solvers for Example 4.1.

(b) Comparisons for Example 4.1. We first apply three solvers to solve some datasets with small scales, where $m = 100, 200, 300, r = 0$ and depict there classifiers in Figure 1, where the Bayes classifier for this example is $w_2 = 0.25w_1$, see the black dotted line. Basically, all solvers classify the dataset well, and clearly, `snasvm` gets the best classification accuracy. When it comes to the large scales of samples from 10^3 to 10^8 , the picture is significantly different. For this experiment, we choose 10% (i.e., $r = 10$) training samples as outliers. Since data is randomly generated, the average results of 20 instances for each method are recorded in Figure 2, where `liblinear` runs too long time when $m > 10^6$ and hence its results are omitted. Generally speaking, the training classifications accuracy from `snasvm` and `liblinear` are similar, being better than those gotten by `fitclinear`. However, it can be clearly seen that `snasvm` runs the fastest and

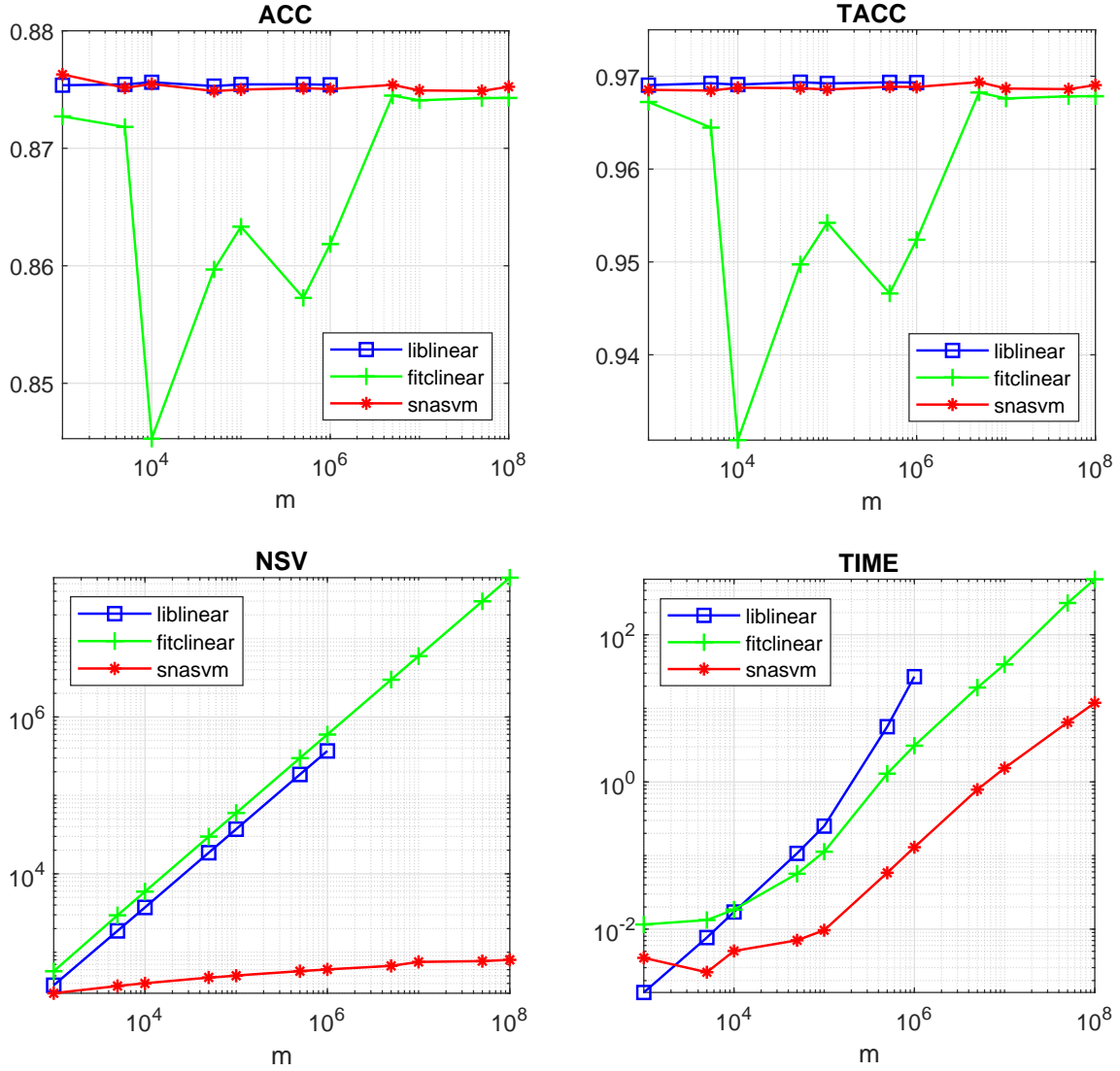


Figure 2: Average results of three solvers for Example 4.1.

uses the much fewer numbers of support vectors. Moreover, the bigger m is, the more obvious advantage **snasvm** has.

(c) Comparisons for Example 4.2. Results of three solvers are reported in Table 2. Again **snasvm** runs the fastest and generates the smallest numbers of support vectors. Taking the dataset **higg** as an instance, the other two solvers respectively take 2938 and 46.86 seconds to classify the data, by contrast, our proposed method only needs 2.551 seconds, much shorter than the one from **liblinear**. Moreover, the number 274229 of support vectors is less than 10% percent of those (2899291 and 8922973) from the other two solvers. For the classification accuracy (training or testing accuracies), there is no significant difference between these three solvers.

Table 2: Results of three solvers libsvm:liblinear, fitsvm:fitlinear and snasvm:SNASVM for Example 4.2.

Data	ACC(%)			TACC(%)			TIME(in seconds)			NSV		
	libsvm	fitsvm	snasvm	libsvm	fitsvm	snasvm	libsvm	fitsvm	snasvm	libsvm	fitsvm	snasvm
mrpe	95.01	94.78	93.07	95.25	94.90	92.57	31.751	2.2828	1.4842	42745	29120	9353
dhrrb	82.70	82.71	82.96	83.43	83.43	83.81	0.1668	0.0966	0.0220	45760	41217	137
aips	87.66	87.26	85.63	87.41	87.08	85.00	0.8731	0.2143	0.0662	34746	61160	243
sctp	89.95	78.01	90.55	90.00	77.80	90.62	13.483	1.0879	0.7188	76482	115807	21022
skin	92.90	92.73	93.79	92.66	92.45	93.56	0.4214	0.3342	0.1089	48372	62609	2405
ccfd	99.94	99.94	99.89	99.92	99.92	99.86	2.6611	0.5232	0.0910	1583	2884	425
r1c1	100.0	100.0	99.98	100.0	100.0	99.97	0.3928	0.6560	0.0617	191	342	47
r1c2	100.0	100.0	99.98	100.0	100.0	99.98	0.3865	0.5099	0.0601	142	295	47
r1c3	100.0	100.0	99.99	100.0	100.0	99.99	0.4902	0.5595	0.0752	164	337	47
r1c4	100.0	100.0	99.99	100.0	100.0	99.99	0.3764	0.6736	0.0621	154	333	47
r1c5	100.0	100.0	99.98	100.0	100.0	99.97	0.4006	0.5701	0.0909	166	331	47
r1c6	100.0	100.0	99.99	100.0	100.0	99.99	0.4961	0.5055	0.0758	140	317	47
r1c7	100.0	100.0	100.0	100.0	100.0	99.99	0.3795	0.5163	0.0472	163	305	47
r1c8	100.0	100.0	99.99	100.0	100.0	99.99	0.6826	0.6734	0.0607	154	332	47
r1c9	100.0	100.0	99.96	100.0	100.0	99.95	0.6600	0.5422	0.0929	175	345	47
r1c10	100.0	100.0	99.99	100.0	100.0	99.98	0.3880	0.5455	0.0607	187	340	47
covt	76.29	75.73	75.14	76.15	75.47	75.03	13.228	2.0480	0.4031	306358	382638	1668
retb	99.81	99.81	99.81	99.80	99.80	99.80	8.9880	4.0662	0.9898	130224	421118	4611
susy	78.44	78.55	78.72	78.39	78.52	78.69	596.31	16.685	2.9198	2305606	2632113	107783
hepm	78.36	83.31	83.54	78.33	83.23	83.52	1688.9	28.392	1.6050	2682838	3698522	268328
higg	47.01	63.84	64.06	46.98	63.79	64.00	2938.7	46.863	2.5512	2899291	8922973	274229

References

- [1] J. Bi, K. Bennett, M. Embrechts, C. Breneman, and M. Song, Dimensionality reduction via sparse support vector machines, *Journal of Machine Learning Research*, 3, 1229-1243, 2003.
- [2] C. Burges and B. Schölkopf, Improving the accuracy and speed of support vector machines, *In Advances in neural information processing systems*, 375381, 1997.
- [3] C. Chang and C. Lin, LIBSVM: A library for support vector machines, *ACM transactions on intelligent systems and technology*, 2(3), 1-27, 2011.
- [4] R. Collobert, F. Sinz, J. Weston, and L. Bottou, Large scale transductive SVMs, *Journal of Machine Learning Research*, 7(1), 1687-1712, 2006.
- [5] C. Cortes and V. Vapnik, Support vector networks, *Machine learning*, 20(3), 273-297, 1995.
- [6] O. Dekel, S. Shalev-Shwartz and Y. Singer, The Forgetron: A kernel-based perceptron on a fixed budget. *In Advances in neural information processing systems*, 259-266, 2006.
- [7] A. Cotter, S. Shalev-Shwartz and N. Srebro, Learning optimally sparse support vector machines, *In International Conference on Machine Learning*, 266-274, 2013.
- [8] R. Fan, K. Chang, C. Hsieh, X. Wang and C. Lin, Liblinear: A library for large linear classification, *Journal of machine learning research*, 9, 1871-1874, 2008.
- [9] Y. Freund and R. E. Schapire, A decision theoretic generalization of on-line learning and an application to boosting, *Journal of Computer and System Sciences*, 55(1), 119-139, 1997.
- [10] J. Friedman, T. Hastie, and R. Tibshirani, Additive logistic regression: a statistical view of boosting, *Annals of statistics*, 28(2), 337-374, 2000.
- [11] X. Huang, L. Shi and A.K. Suykens, Support vector machine classifier with pinball loss, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(5), 984-997, 2014.
- [12] X. Huang, L. Shi and J. Suykens, Solution path for pin-svm classifiers with positive and negative values, *IEEE transactions on neural networks and learning systems*, 28(7), 1584-1593, 2016.
- [13] V. Jumutc, X. Huang and A. Suykens, Fixed-size Pegasos for hinge and pinball loss SVM, *International Joint Conference on Neural Networks*, 2013.

- [14] S. Keerthi, O. Chapelle, and D. DeCoste, Building support vector machines with reduced classifier complexity, *Journal of Machine Learning Research*, 7, 1493-1515, 2006.
- [15] Y. Lee and O. Mangasarian, RSVM: Reduced support vector machines, *In Proceedings of the 2001 SIAM International Conference on Data Mining*, 1-17, 2001.
- [16] K. Lin and C. Lin., A study on reduced support vector machines, *IEEE transactions on Neural Networks*, 14(6), 1449-1459, 2003.
- [17] L. Mason, J. Baxter, P. Bartlett and M. Frean, Boosting algorithms as gradient descent, *In Advances in neural information processing systems*, 1999.
- [18] L. Mason, P. Bartlett and J. Baxter, Improved generalization through explicit optimization of margins, *Machine Learning*, 38(3), 243-255, 2000.
- [19] D. Nguyen, K. Matsumoto, Y. Takishima and K. Hashimoto, Condensed vector machines: learning fast machine for large data, *IEEE transactions on neural networks*, 21(12), 1903-1914, 2010.
- [20] E. Osuna and G. Federico, Reducing the run-time complexity of support vector machines, *In International Conference on Pattern Recognition*, 1998.
- [21] L. Pan, J. Fan and N. Xiu, Optimality conditions for sparse nonlinear programming, *Science China Mathematics*, 60(5), 759-776, 2017.
- [22] K. Pelckmans, J. Suykens, T. Gestel, J. Brabanter, L. Lukas, B. Hamers, B. Moor and J. Vandewalle, A matlab/c toolbox for least square support vector machines, *ESATSCD-SISTA Technical Report*, 02-145, 2002.
- [23] F. Pérez-Cruz, A. Navia-Vazquez, P. Alarcón-Dian and A. Artes-Rodriguez, Support vector classifier with hyperbolic tangent penalty function, *Acoustics, Speech, and Signal Processing*, 2000.
- [24] F. Pérez-Cruz, A. Navia-Vazquez, A. Figueiras-Vidal and A. Artes-Rodriguez. Empirical risk minimization for support vector classifiers, *IEEE Transactions on Neural Networks*, 14(2), 296-303, 2003.
- [25] R. Rockafellar, and R. Wets, Variational analysis, *Springer Science & Business Media*, 317, 2009.
- [26] S. Rosset and J. Zhu, Piecewise linear regularized solution paths, *The Annals of Statistics*, 35(3), 1012-1030, 2007.
- [27] M. Wu, B. Schölkopf and G. Bakir, Building sparse large margin classifiers, *In Proceedings of the 22nd international conference on Machine learning*, 996-1003, 2005.

-
- [28] X. Shen, L. Niu, Z. Qi, and Y. Tian, Support vector machine classifier with truncated pinball loss, *Pattern Recognition*, 68, 2017.
 - [29] A. Suykens and J. Vandewalle, Least squares support vector machine classifiers, *Neural Processing Letters*, 9(3), 293-300, 1999.
 - [30] L. Wang, J. Zhu, and H. Zou, Hybrid huberized support vector machines for microarray classification, *Bioinformatics*, 24(3), 412-419, 2008.
 - [31] Y. Wu and Y. Liu, Robust truncated hinge loss support vector machines, *Journal of the American Statistical Association*, 102(479), 974-983, 2007.
 - [32] Y. Xu, I. Akrotirianakis, and A. Chakraborty, Proximal gradient method for huberized support vector machine, *Pattern Analysis and Applications*, 19(4), 989-1005, 2016.
 - [33] X. Yang, L. Tan and L. He, A robust least squares support vector machine for regression and classification with noise, *Neurocomputing*, 140, 41-52, 2014.
 - [34] L. Yang and H. Dong, Support vector machine with truncated pinball loss and its application in pattern recognition, *Chemometrics and Intelligent Laboratory Systems*, 177, 89-99, 2018.
 - [35] J. Yin and Q. Li, A semismooth Newton method for support vector classification and regression, *Computational Optimization and Applications*, 73(2), 477-508, 2019.
 - [36] Y. Zhan and D. Shen, Design efficient support vector machine for fast classification, *Pattern Recognition*, 38(1), 157-161, 2005.