Scaling SVM and Least Absolute Deviations via Exact Data Reduction

Jie Wang   JIE.WANG.USTC@ASU.EDU
Arizona State University, Tempe, AZ 85287 USA

Peter Wonka   PETER.WONKA@ASU.EDU
Arizona State University, Tempe, AZ 85287 USA

Jieping Ye   JIEPING.YE@ASU.EDU
Arizona State University, Tempe, AZ 85287 USA

Abstract

The support vector machine (SVM) is a widely used method for classification. Although many efforts have been devoted to develop efficient solvers, it remains challenging to apply SVM to large-scale problems. A nice property of SVM is that the non-support vectors have no effect on the resulting classifier. Motivated by this observation, we present fast and efficient screening rules to discard non-support vectors by analyzing the **d**ual problem of SVM via **v**ariational **i**nequalities (DVI). As a result, the number of data instances to be entered into the optimization can be substantially reduced. Some appealing features of our screening method are: (1) DVI is safe in the sense that the vectors discarded by DVI are guaranteed to be non-support vectors; (2) the data set needs to be scanned only once to run the screening, whose computational cost is negligible compared to that of solving the SVM problem; (3) DVI is independent of the solvers and can be integrated with any existing efficient solvers. We also show that the DVI technique can be extended to detect non-support vectors in the least absolute deviations regression (LAD). To the best of our knowledge, there are currently no screening methods for LAD. We have evaluated DVI on both synthetic and real data sets. Experiments indicate that DVI significantly outperforms the existing state-of-the-art screening rules for SVM, and is very effective in discarding non-support vectors for LAD. The speedup gained by DVI rules can be up to two orders of magnitude.

1. Introduction

The support vector machine is one of the most popular classification tools in machine learning. Many efforts have been devoted to develop efficient solvers for SVM (Hastie et al., 2004; Joachims, 2006; Shalev-Shwartz et al., 2007; Hsieh et al., 2008; Fan et al., 2008). However, the applications of SVM to large-scale problems still pose significant challenges. To address this issue, one promising approach is by "*screening*". The key idea of screening is motivated by the well known feature of SVM, that is, the resulting classifier is determined only by the so called "*support vectors*". If we first identify the non-support vectors via screening, and then remove them from the optimization, it may lead to substantial savings in the computational cost and memory. Another useful tool in machine learning and statistics is the least absolute deviations regression (LAD) (Powell, 1984; Wang et al., 2006; Chen et al., 2008; Rao et al., 2008) or ℓ_1 method. When the protection against outliers is a major concern, LAD provides a useful and plausible alternative to the classical least squares or ℓ_2 method for linear regression. In this paper, we study both SVM and LAD under a unified framework.

The idea of screening has been successfully applied to a large class of ℓ_1-regularized problems (El Ghaoui et al., 2012; Xiang et al., 2011; Tibshirani et al., 2012; Wang et al., 2013; Liu et al., 2014), including Lasso, ℓ_1-regularized logistic regression, elastic net, and more general convex problems. Those methods are able to discard a large portion of "inactive" features, which has 0 coefficients in the optimal solution, and the speedup can be several orders of magnitude.

Recently, Ogawa et al. (2013) proposed a "**s**afe **s**creening" rule to identify **n**on-**s**upport **v**ectors for SVM; in this paper, we refer to this method as SSNSV for convenience. Notice that, the former approaches for ℓ_1-regularized problems aim to discard inactive "*features*", while SSNSV is

to identify the non-support "*vectors*". This essential difference makes SSNSV a nontrivial extension of the existing feature screening methods. Although there exist many methods for data reduction for SVM (Achlioptas et al., 2002; Yu et al., 2003; Cao & Boley, 2006), they are not safe, in the sense that the resulting classification model may be different. To the best of our knowledge, SSNSV is the only existing safe screening method (Ogawa et al., 2013) to identify the non-support vectors for SVM. However, in order to run the screening, SSNSV needs to iteratively determine an appropriate parameter value and an associated feasible solution, which can be very time consuming.

In this paper, we develop novel efficient and effective screening rules, called "DVI", for a class of supervised learning problems including SVM and LAD (Buchinsky, 1998; Jin et al., 2001). The proposed method, DVI, shares the same advantage as SSNSV (Ogawa et al., 2013), that is, both rules are safe in the sense that the discarded vectors are guaranteed to be non-support vectors. The proposed DVI identifies the non-support vectors by estimating a lower bound of the inner product between each vector and the optimal solution, which is unknown. The more accurate the estimation is, the more non-support vectors can be detected. However, the estimation turns out to be highly non-trivial since the optimal solution is not available. To overcome this difficulty, we propose a novel framework to accurately estimate the optimal solution via the estimation of the "dual optimal solution", as the primal and dual optimal solutions can be related by the KKT conditions (Güler, 2010). Our main technical contribution is to estimate the dual optimal solution via the so called "variational inequalities" (Güler, 2010). Our experiments on both synthetic and real data demonstrate that DVI can identify far more non-support vectors than SSNSV. Moreover, by using the same technique, that is, variational inequalities, we can strictly improve SSNSV in identifying the non-support vectors. Our results also show that DVI is very effective in discarding non-support vectors for LAD. The speedup gained by DVI rules can be up to two orders of magnitude.

The rest of this paper is organized as follows. In Section 2, we study the SVM and LAD problems under a unified framework. We then introduce our DVI rules in detail for the general formulation in Sections 3 and 4. In Sections 5 and 6, we extend the DVI rules derived in Section 4 to SVM and LAD respectively. In Section 7, we evaluate our DVI rules for SVM and LAD using both synthetic and real data. We conclude this paper in Section 8.

Notations: Throughout this paper, we use $\langle \mathbf{x}, \mathbf{y} \rangle = \sum_i x_i y_i$ to denote the inner product of vectors $\mathbf{x}$ and $\mathbf{y}$, and $\|\mathbf{x}\|^2 = \langle \mathbf{x}, \mathbf{x} \rangle$. For vector $\mathbf{x}$, let $[\mathbf{x}]_i$ be the ith component of $\mathbf{x}$. If $\mathbf{M}$ is a matrix, $\mathbf{m}_i$ is the ith column of $\mathbf{M}$ and $[\mathbf{M}]_{i,j}$ is the $(i,j)^{th}$ entry of $\mathbf{M}$. Given a scalar x, we de-

note $\max\{x, 0\}$ by $[x]_+$. For the index set $\mathcal{I} := \{1, \ldots, l\}$, let $\mathcal{J} := \{j_1, \ldots, j_k\} \subseteq \mathcal{I}$ and $\mathcal{J}^c := \mathcal{I} \setminus \mathcal{J}$. For a vector $\mathbf{x}$ or a matrix $\mathbf{M}$, let $[\mathbf{x}]_{\mathcal{J}} = ([\mathbf{x}]_{j_1}, \ldots, [\mathbf{x}]_{j_k})^T$ and $[\mathbf{M}]_{\mathcal{J}} = (\mathbf{m}_{j_1}, \ldots, \mathbf{m}_{j_k})$. Moreover, let $\Gamma_0(\Re^n)$ be the class of proper and lower semicontinuous convex functions from $\Re^n$ to $(-\infty, \infty]$. The conjugate of $f \in \Gamma_0(\Re^n)$ is the function $f^* \in \Gamma_0(\Re^n)$ given by

$$f^* : \Re^n \to (-\infty, \infty] : \theta \mapsto \sup_{\mathbf{x} \in \Re^n} \mathbf{x}^T \theta - f(\mathbf{x}). \quad (1)$$

The biconjugate of $f \in \Gamma_0(\Re^n)$ is the function $f^{**} \in \Gamma_0(\Re^n)$ given by

$$f^{**} : \Re^n \to (-\infty, \infty] : \mathbf{x} \mapsto \sup_{\theta \in \Re^n} \mathbf{x}^T \theta - f^*(\theta). \quad (2)$$

2. Basics and Motivations

In this section, we study the SVM and LAD problems under a unified framework. Then, we motivate the general screening rules via the KKT conditions. Consider the convex optimization problems of the following form:

$$\min_{\mathbf{w} \in \Re^n} \frac{1}{2} \|\mathbf{w}\|^2 + C\Phi(\mathbf{w}), \quad (3)$$

where $\Phi : \Re^n \to \Re$ is a convex function but not necessarily differentiable and $C > 0$ is a regularization parameter. Notice that, the function Φ is generally referred to as the empirical loss. More specifically, suppose we have a set of observations $\{\mathbf{x}_i, y_i\}_{i=1}^l$, where $\mathbf{x}_i \in \Re^n$ and $y_i \in \Re$ are the ith data instance and the corresponding response. We focus on the following function class:

$$\Phi(\mathbf{w}) = \sum_{i=1}^l \varphi\left(\mathbf{w}^T(a_i \mathbf{x}_i) + b_i y_i\right), \quad (4)$$

where $\varphi : \Re \to \Re_+$ is a nonconstant continuous sublinear function, and a_i, b_i are scalars. We provide the definition of sublinear function as follows.

Definition 1. *(Hiriart-Urruty & Lemaréchal, 1993) A function $\sigma : \Re^n \to (-\infty, \infty]$ is said to be sublinear if it is convex, and positively homogeneous, i.e.,*

$$\sigma(tx) = t\sigma(x), \ \forall x \in \Re^n \text{and } t > 0. \quad (5)$$

We will see that SVM and LAD are both special cases of problem (3) in Sections 5.1 and 6, respectively. A nice property of the function φ is that the biconjugate φ^{**} is exactly φ itself, as stated in Lemma 1.

Lemma 1. *For the function $\varphi : \Re \to \Re_+$ which is continuous and sublinear, we have $\varphi \in \Gamma_0(\Re)$, and thus $\varphi^{**} = \varphi$.*

It is straightforward to check the statement in Lemma 1 by verifying the requirements of the function class $\Gamma_0(\Re)$. For

self-completeness, we provide a proof in the supplement. According to Lemma 1, problem (3) can be rewritten as

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{l} \varphi^{**}\left(\mathbf{w}^T(a_i\mathbf{x}_i) + b_iy_i\right) \quad (6)$$

$$= \min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{l}\left\{\sup_{\theta_i\in\Re}\theta_i\left[\mathbf{w}^T(a_i\mathbf{x}_i) + b_iy_i\right] - \varphi^*(\theta_i)\right\}$$

$$= \sup_{\theta}\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{l}\left\{\theta_i\left[\mathbf{w}^T(a_i\mathbf{x}_i) + b_iy_i\right] - \varphi^*(\theta_i)\right\}$$

$$= \sup_{\theta}-C\sum_{i=1}^{l}\varphi^*(\theta_i) + \min_{\mathbf{w}\in\Re^n}\frac{1}{2}\|\mathbf{w}\|^2 + C\langle\mathbf{Z}\mathbf{w} + \bar{\mathbf{y}}, \theta\rangle,$$

where $\theta = (\theta_1, \ldots, \theta_l)^T$, $\mathbf{Z} = (a_i\mathbf{x}_i, \ldots, a_l\mathbf{x}_l)^T$ and $\bar{\mathbf{y}} = (b_1y_1, \ldots, b_ly_l)^T$. Let $\ell(\mathbf{w}) := \frac{1}{2}\|\mathbf{w}\|^2 + C\langle\mathbf{Z}\mathbf{w} + \bar{\mathbf{y}}, \theta\rangle$. The reason we can exchange the order of min and sup in Eq. (6) is due to the strong duality of problem (3) (Boyd & Vandenberghe, 2004).

By setting $\frac{\partial\ell(\mathbf{w})}{\partial\mathbf{w}} = 0$, we have

$$\mathbf{w}^* = -C\mathbf{Z}^T\theta, \quad (7)$$

and thus

$$\min_{\mathbf{w}}\ell(\mathbf{w}) = \ell(\mathbf{w}^*) = -\frac{C^2}{2}\|\mathbf{Z}^T\theta\|^2 + C\langle\bar{\mathbf{y}}, \theta\rangle. \quad (8)$$

Hence, Eq. (6) becomes

$$\sup_{\theta\in\Re^l}-C\sum_{i=1}^{l}\varphi^*(\theta_i) - \frac{C^2}{2}\|\mathbf{Z}^T\theta\|^2 + C\langle\bar{\mathbf{y}}, \theta\rangle. \quad (9)$$

Moreover, because $\varphi \in \Gamma_0(\Re)$ is sublinear by Lemma 1, we know that φ^* is the indicator function for a closed convex set. In fact, we have the following result:

Lemma 2. *For the nonconstant continuous sublinear function $\varphi : \Re \to \Re_+$, there exists a nonempty closed interval $I_\varphi = [\alpha, \beta]$ with $\alpha, \beta \in \Re$ and $\alpha < \beta$ such that*

$$\varphi^*(t) := \iota_{[\alpha,\beta]} = \begin{cases} 0, & \text{if } t \in [\alpha, \beta], \\ \infty, & \text{otherwise.} \end{cases} \quad (10)$$

Let $I_\varphi^l = [\alpha, \beta]^l$. We can rewrite problem (9) as

$$\sup_{\theta\in I_\varphi^l}-\frac{C^2}{2}\|\mathbf{Z}^T\theta\|^2 + C\langle\bar{\mathbf{y}}, \theta\rangle. \quad (11)$$

Problem (11) is in fact the dual problem of (3). Moreover, the "sup" in problem (11) can be replaced by "max" due to the strong duality (Boyd & Vandenberghe, 2004) of problem (3). Since $C > 0$, problem (11) is equivalent to

$$\min_{\theta\in I_\varphi^l}\frac{C}{2}\|\mathbf{Z}^T\theta\|^2 - \langle\bar{\mathbf{y}}, \theta\rangle. \quad (12)$$

Let $\mathbf{w}^*(C)$ and $\theta^*(C)$ be the optimal solutions of (3) and (11) respectively. Eq. (7) implies that

$$\mathbf{w}^*(C) = -C\mathbf{Z}^T\theta^*(C). \quad (13)$$

The KKT conditions[1] of problem (12) are

$$[\theta^*(C)]_i \in \begin{cases} \beta, & \text{if } -\langle\mathbf{w}^*(C), a_i\mathbf{x}_i\rangle < b_iy_i; \\ [\alpha, \beta], & \text{if } -\langle\mathbf{w}^*(C), a_i\mathbf{x}_i\rangle = b_iy_i; \\ \alpha, & \text{if } -\langle\mathbf{w}^*(C), a_i\mathbf{x}_i\rangle > b_iy_i; \end{cases} \quad (14)$$
$$i = 1, \ldots, l.$$

For notational convenience, let

$$\mathcal{R} = \{i : -\langle\mathbf{w}^*(C), a_i\mathbf{x}_i\rangle > b_iy_i\},$$
$$\mathcal{E} = \{i : -\langle\mathbf{w}^*(C), a_i\mathbf{x}_i\rangle = b_iy_i\},$$
$$\mathcal{L} = \{i : -\langle\mathbf{w}^*(C), a_i\mathbf{x}_i\rangle < b_iy_i\}.$$

We call the vectors in the set $\mathcal{E}$ as "support vectors". All the other vectors in $\mathcal{R}$ and $\mathcal{L}$ are called "non-support vectors". The KKT conditions in (14) imply that, if some of the data instances are known to be members of $\mathcal{R}$ and $\mathcal{L}$, then the corresponding components of $\theta^*(C)$ can be set accordingly and we only need the other components of $\theta^*(C)$. More precisely, we have the following result:

Lemma 3. *Given index sets $\hat{\mathcal{R}} \subseteq \mathcal{R}$ and $\hat{\mathcal{L}} \subseteq \mathcal{L}$, we have*

1. $[\theta^*(C)]_{\hat{\mathcal{R}}} = \alpha$ and $[\theta^*(C)]_{\hat{\mathcal{L}}} = \beta$.

2. Let $\hat{\mathcal{S}} = \hat{\mathcal{R}}\bigcup\hat{\mathcal{L}}$, $|\hat{\mathcal{S}}^c|$ be the cardinality of the set $\hat{\mathcal{S}}^c$, $\hat{\mathbf{G}}_{11} = [\mathbf{Z}^T]_{\hat{\mathcal{S}}^c}^T[\mathbf{Z}^T]_{\hat{\mathcal{S}}^c}$, $\hat{\mathbf{G}}_{12} = [\mathbf{X}^T]_{\hat{\mathcal{S}}^c}^T[\mathbf{X}^T]_{\hat{\mathcal{S}}}$ and $\hat{\mathbf{y}} = \mathbf{y}_{\hat{\mathcal{S}}^c} - C\hat{\mathbf{G}}_{12}[\theta^*(C)]_{\hat{\mathcal{S}}}$. Then, $[\theta^*(C)]_{\hat{\mathcal{S}}^c}$ can be computed by solving the following problem:

$$\min_{\hat{\theta}\in\Re^{|\hat{\mathcal{S}}^c|}}\frac{C}{2}\hat{\theta}^T\hat{\mathbf{G}}_{11}\hat{\theta} - \hat{\mathbf{y}}^T\hat{\theta}, \text{ s.t. } \hat{\theta} \in [\alpha, \beta]^{|\hat{\mathcal{S}}^c|}. \quad (15)$$

Clearly, if $|\hat{\mathcal{S}}|$ is large compared to $|\mathcal{I}| = l$, the computational cost for solving problem (15) can be much cheaper than solving the full problem (12). To determine the membership of the data instances, Eq. (13) and (14) imply that

$$C\langle\mathbf{Z}^T\theta^*(C), a_i\mathbf{x}_i\rangle > b_iy_i \Rightarrow [\theta^*(C)]_i = \alpha \Leftrightarrow i \in \mathcal{R}; \quad \text{(R1)}$$

$$C\langle\mathbf{Z}^T\theta^*(C), a_i\mathbf{x}_i\rangle < b_iy_i \Rightarrow [\theta^*(C)]_i = \beta \Leftrightarrow i \in \mathcal{L}. \quad \text{(R2)}$$

However, (R1) and (R2) are generally not applicable since $\theta^*(C)$ is unknown. To overcome this difficulty, we can estimate a region Θ such that $\theta^*(C) \in \Theta$. As a result, we obtain the relaxed version of (R1) and (R2):

$$\min_{\theta\in\Theta}C\langle\mathbf{Z}^T\theta, a_i\mathbf{x}_i\rangle > b_iy_i \Rightarrow [\theta^*(C)]_i = \alpha \Leftrightarrow i \in \mathcal{R}; \quad \text{(R1}'\text{)}$$

$$\max_{\theta\in\Theta}C\langle\mathbf{Z}^T\theta, a_i\mathbf{x}_i\rangle < b_iy_i \Rightarrow [\theta^*(C)]_i = \beta \Leftrightarrow i \in \mathcal{L}. \quad \text{(R2}'\text{)}$$

---

[1] Please refer to the supplement for details.

Notice that, (R1$'$) and (R2$'$) serve as the foundation of the proposed DVI rules and the method in Ogawa et al. (2013). In the subsequent sections, we first estimate the region Θ which includes $\theta^*(C)$, and then derive the screening rules based on (R1$'$) and (R2$'$).

Method to solve problem (15) It is known that, problem (15) can be efficiently solved by the dual coordinate descent method (Hsieh et al., 2008). Briefly speaking, the algorithm updates the components of $\hat{\theta}$ one at a time, which is equivalent to minimizing a 1D quadratic function over a compact interval. For self-completeness, we provide a more detailed review in the supplement.

In Section 3, we first give an accurate estimation of the set Θ which includes $\theta^*(C)$ as in (R1$'$) and (R2$'$) via the variational inequalities. Then in Section 4, we present the novel DVI rules for problem (3) in detail.

3. Estimation of the Dual Optimal Solution

For problem (12), suppose we are given two parameter values $0 < C_0 < C$ and $\theta^*(C_0)$ is known. Then, Theorem 5 shows that $\theta^*(C)$ can be effectively bounded in terms of $\theta^*(C_0)$. The main technique we use is the so called variational inequalities. For self-completeness, we cite the definition of variational inequalities as follows.

Theorem 4. *(Güler, 2010) Let $A \subseteq \Re^n$ be a convex set, and let h be a Gâteaux differentiable function on an open set containing A. If $\mathbf{x}^*$ is a local minimizer of h on A, then*

$$\langle \nabla h(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle \geq 0, \quad \forall \mathbf{x} \in A. \tag{16}$$

Via the variational inequalities, the following theorem shows that $\theta^*(C)$ can estimated in terms of $\theta^*(C_0)$.

Theorem 5. *For problem (12), let $C > C_0 > 0$. Then*

$$\|\mathbf{Z}^T \theta^*(C) - \tfrac{C_0+C}{2C} \mathbf{Z}^T \theta^*(C_0)\| \leq \tfrac{C-C_0}{2C} \|\mathbf{Z}^T \theta^*(C_0)\|.$$

Proof. Let $g(\theta)$ be the objective function of problem (12). The variational inequality implies that

$$\langle \nabla g(\theta^*(C_0)), \theta - \theta^*(C_0) \rangle \geq 0, \ \forall \theta \in [\alpha, \beta]^l; \tag{17}$$

$$\langle \nabla g(\theta^*(C)), \theta - \theta^*(C) \rangle \geq 0, \ \forall \theta \in [\alpha, \beta]^l. \tag{18}$$

Notice that $\nabla g(\theta) = C\mathbf{Z}\mathbf{Z}^T\theta - \bar{\mathbf{y}}$, and $\theta^*(C_0) \in [\alpha, \beta]^l$ and $\theta^*(C) \in [\alpha, \beta]^l$. Plugging $\nabla g(\theta^*(C))$ and $\nabla g(\theta^*(C_0))$ into (17) and (18) leads to

$$\langle C_0\mathbf{Z}\mathbf{Z}^T\theta^*(C_0) - \bar{\mathbf{y}}, \theta^*(C) - \theta^*(C_0) \rangle \geq 0; \tag{19}$$

$$\langle C\mathbf{Z}\mathbf{Z}^T\theta^*(C) - \bar{\mathbf{y}}, \theta^*(C_0) - \theta^*(C) \rangle \geq 0. \tag{20}$$

We can see that the inequality in (20) is equivalent to

$$\langle \bar{\mathbf{y}} - C\mathbf{Z}\mathbf{Z}^T\theta^*(C), \theta^*(C) - \theta^*(C_0) \rangle \geq 0. \tag{21}$$

Then the statement follows by adding the inequalities in (19) and (21) together. $\square$

4. The Proposed DVI Rules

Given $C > C_0 > 0$ and $\theta^*(C_0)$, we can estimate $\theta^*(C)$ via Theorem 5. Combining (R1$'$), (R2$'$) and Theorem (5), we develop the basic screening rule for problem (3) as summarized in the following theorem:

Theorem 6. (**DVI**) *For problem (12), suppose we are given $\theta^*(C_0)$. Then, for any $C > C_0$, we have $[\theta^*(C)]_i = \alpha$, i.e., $i \in \mathcal{R}$, if the following holds*

$$\tfrac{C+C_0}{2}\langle \mathbf{Z}^T \theta^*(C_0), a_i\mathbf{x}_i \rangle - \tfrac{C-C_0}{2}\|\mathbf{Z}^T \theta^*(C_0)\|\|a_i\mathbf{x}_i\| > b_iy_i.$$

Similarly, we have $[\theta^*(C)]_i = \beta$, i.e., $i \in \mathcal{L}$, if

$$\tfrac{C+C_0}{2}\langle \mathbf{Z}^T \theta^*(C_0), a_i\mathbf{x}_i \rangle + \tfrac{C-C_0}{2}\|\mathbf{Z}^T \theta^*(C_0)\|\|a_i\mathbf{x}_i\| < b_iy_i.$$

Please refer to the supplement for the proof of Theorem 6. In real applications, the optimal parameter value of C is unknown and we need to estimate it. Commonly used model selection strategies such as cross validation and stability selection need to solve the optimization problems over a grid of turning parameters $0 < C_1 < C_2 < \ldots < C_\mathcal{K}$ to determine an appropriate value for C. This procedure is usually very time consuming, especially for large scale problems. To this end, we propose a sequential version of the proposed DVI below.

Corollary 7. (**DVI$_s^*$**) *For problem (12), suppose we are given a sequence of parameters $0 < C_1 < C_2 < \ldots < C_\mathcal{K}$. Assume $\theta^*(C_k)$ is known for an arbitrary integer $1 \leq k < \mathcal{K}$. Then, for C_{k+1}, we have $[\theta^*(C_{k+1})]_i = \alpha$, i.e., $i \in \mathcal{R}$, if the following holds*

$$\tfrac{C_{k+1}+C_k}{2}\langle \mathbf{Z}^T \theta^*(C_k), a_i\mathbf{x}_i \rangle - \tfrac{C_{k+1}-C_k}{2}\|\mathbf{Z}^T \theta^*(C_k)\|\|a_i\mathbf{x}_i\| > b_iy_i.$$

Similarly, we have $[\theta^*(C_{k+1})]_i = \beta$, i.e., $i \in \mathcal{L}$, if

$$\tfrac{C_{k+1}+C_k}{2}\langle \mathbf{Z}^T \theta^*(C_k), a_i\mathbf{x}_i \rangle + \tfrac{C_{k+1}-C_k}{2}\|\mathbf{Z}^T \theta^*(C_k)\|\|a_i\mathbf{x}_i\| < b_iy_i.$$

The main computational cost of **DVI$_s^*$** is due to the evaluation of $\langle \mathbf{Z}^T \theta^*(C_k), a_i\mathbf{x}_i \rangle$, $\|\mathbf{Z}^T \theta^*(C_k)\|$ and $\|a_i\mathbf{x}_i\|$. Let $\mathbf{G} = \mathbf{Z}\mathbf{Z}^T$. It is easy to see that

$$\langle \mathbf{Z}^T \theta^*(C_k), a_i\mathbf{x}_i \rangle = \mathbf{g}_i^T\theta^*(C_k),$$
$$\|\mathbf{Z}^T \theta^*(C_k)\|^2 = \theta^*(C_k)^T\mathbf{G}\theta^*(C_k),$$
$$\|\bar{\mathbf{x}}_i\|^2 = [\mathbf{G}]_{i,i}.$$

where $\mathbf{g}_i$ is the ith column of $\mathbf{G}$. Since $\mathbf{G}$ is independent of C_k, it can be computed only once and thus the computational cost of **DVI$_s^*$** reduces to $O(l^2)$ to scan the entire data set. Indeed, by noting Eq. (13), we can reconstruct DVI rules without the explicit computation of $\mathbf{G}$.

Corollary 8. (**DVI$_s$**) *For problem (3), suppose we are given a sequence of parameters $0 < C_1 < C_2 < \ldots < C_\mathcal{K}$. Assume $\mathbf{w}^*(C_k)$ is known for an arbitrary integer $1 \leq k <$*

$\mathcal{K}$. Then, for C_{k+1}, we have $[\theta^*(C_{k+1})]_i = \alpha$, i.e., $i \in \mathcal{R}$, if the following holds

$$-\frac{C_k + C_{k+1}}{2C_k}\langle \mathbf{w}^*(C_k), a_i\mathbf{x}_i\rangle - \frac{C_{k+1} - C_k}{2C_k}\|\mathbf{w}^*(C_k)\|\|a_i\mathbf{x}_i\| > b_iy_i.$$

Similarly, we have $[\theta^*(C_{k+1})]_i = \beta$, i.e., $i \in \mathcal{L}$, if

$$-\frac{C_k + C_{k+1}}{2C_k}\langle \mathbf{w}^*(C_k), a_i\mathbf{x}_i\rangle + \frac{C_{k+1} - C_k}{2C_k}\|\mathbf{w}^*(C_k)\|\|a_i\mathbf{x}_i\| < b_iy_i.$$

5. Screening Rules for SVM

In Section 5.1, we first present the sequential DVI rules for SVM based on the results in Section 4. Then, in Section 5.2, we show how to strictly improve SSNSV (Ogawa et al., 2013) by the same technique used in DVI.

5.1. DVI rules for SVM

Given a set of observations $\{\mathbf{x}_i, y_i\}_{i=1}^l$, where $\mathbf{x}_i$ and $y_i \in \{1, -1\}$ are the ith data instance and the corresponding class label, the SVM takes the form of:

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^l \left[1 - \mathbf{w}^T(y_i\mathbf{x}_i)\right]_+. \quad (22)$$

It is easy to see that, if we set $\varphi(t) = [t]_+$ and $-a_i = b_i = y_i$, problem (3) becomes the SVM problem. To construct the DVI rules for SVM by Corollaries 7 and 8, we only need to find α and β. In fact, we have the following result:

Lemma 9. Let $\varphi(t) = [t]_+$, then $\alpha = 0$ and $\beta = 1$, i.e.,

$$\varphi^*(s) = \iota_{[0,1]}. \quad (23)$$

We omit the proof of Lemma 9 since it is a direct application of Eq. (1). Then, we immediately have the following screening rules for the SVM problem. (For notational convenience, let $\bar{\mathbf{x}}_i = y_i\mathbf{x}_i$ and $\overline{\mathbf{X}} = (\bar{\mathbf{x}}_1, \ldots, \bar{\mathbf{x}}_l)^T$.)

Corollary 10. (DVI$_s^*$ for SVM) For problem (22), suppose we are given a sequence of parameters $0 < C_1 < C_2 < \ldots < C_{\mathcal{K}}$. Assume $\theta^*(C_k)$ is known for an arbitrary integer $1 \le k < \mathcal{K}$. Then, for C_{k+1}, we have $[\theta^*(C_{k+1})]_i = 0$, i.e., $i \in \mathcal{R}$, if the following holds

$$\frac{C_{k+1} + C_k}{2}\langle \overline{\mathbf{X}}^T\theta^*(C_k), \bar{\mathbf{x}}_i\rangle - \frac{C_{k+1} - C_k}{2}\|\overline{\mathbf{X}}^T\theta^*(C_k)\|\|\bar{\mathbf{x}}_i\| > 1.$$

Similarly, we have $[\theta^*(C_{k+1})]_i = 1$, i.e., $i \in \mathcal{L}$, if

$$\frac{C_{k+1} + C_k}{2}\langle \overline{\mathbf{X}}^T\theta^*(C_k), \bar{\mathbf{x}}_i\rangle + \frac{C_{k+1} - C_k}{2}\|\overline{\mathbf{X}}^T\theta^*(C_k)\|\|\bar{\mathbf{x}}_i\| < 1.$$

Corollary 11. (DVI$_s$ for SVM) For problem (22), suppose we are given a sequence of parameters $0 < C_1 < C_2 < \ldots < C_{\mathcal{K}}$. Assume $\mathbf{w}^*(C_k)$ is known for an arbitrary integer $1 \le k < \mathcal{K}$. Then, for C_{k+1}, we have $[\theta^*(C_{k+1})]_i = 0$, i.e., $i \in \mathcal{R}$, if the following holds

$$\frac{C_k + C_{k+1}}{2C_k}\langle \mathbf{w}^*(C_k), \bar{\mathbf{x}}_i\rangle - \frac{C_{k+1} - C_k}{2C_k}\|\mathbf{w}^*(C_k)\|\|\bar{\mathbf{x}}_i\| > 1.$$

Similarly, we have $[\theta^*(C_{k+1})]_i = 1$, i.e., $i \in \mathcal{L}$, if

$$\frac{C_k + C_{k+1}}{2C_k}\langle \mathbf{w}^*(C_k), \bar{\mathbf{x}}_i\rangle + \frac{C_{k+1} - C_k}{2C_k}\|\mathbf{w}^*(C_k)\|\|\bar{\mathbf{x}}_i\| < 1.$$

5.2. Improving the existing method

We briefly describe how to strictly improve SSNSV (Ogawa et al., 2013) by using the same technique used in DVI rules (please refer to the supplement for more details). In view of Eq. (13), (R1$'$) and (R2$'$) can be rewritten as:

$$\min_{\mathbf{w} \in \Omega}\langle \mathbf{w}, \bar{\mathbf{x}}_i\rangle > 1 \Rightarrow [\theta^*(C)]_i = 0 \Leftrightarrow i \in \mathcal{R}, \quad (\text{R1}'')$$

$$\max_{\mathbf{w} \in \Omega}\langle \mathbf{w}, \bar{\mathbf{x}}_i\rangle < 1 \Rightarrow [\theta^*(C)]_i = 1 \Leftrightarrow i \in \mathcal{L}, \quad (\text{R2}'')$$

where Ω is a set which includes $\mathbf{w}^*(C)$ (notice that, we have already set $-a_i = b_i = y_i$, $\alpha = 0$ and $\beta = 1$). It is easy to see that, the smaller Ω is, the tighter the bounds are in (R1$''$) and (R2$''$). Thus, more data instances' membership can be identified.

Estimation of $\mathbf{w}^*$ in SSNSV Ogawa et al. (2013) consider the following equivalent formulation of SVM:

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2, \text{ s.t. } \sum_{i=1}^l [1 - y_i\mathbf{w}^T\mathbf{x}_i]_+ \le s \quad (24)$$

Let $\mathcal{F}_s = \{\mathbf{w} : \sum_{i=1}^l [1 - y_i\mathbf{w}^T\mathbf{x}_i]_+ \le s\}$. Suppose we have two scalars $s_a > s_b > 0$, and $\mathcal{F}_{s_b} \ne \emptyset$, $\hat{\mathbf{w}}(s_b) \in \mathcal{F}_{s_b}$. Then for $s \in [s_b, s_a]$, $\mathbf{w}^*(s)$ is inside the following region:

$$\Omega_{[s_b, s_a]} := \left\{\mathbf{w} : \begin{array}{l}\langle \mathbf{w}^*(s_a), \mathbf{w} - \mathbf{w}^*(s_a)\rangle \ge 0, \\ \|\mathbf{w}\|^2 \le \|\hat{\mathbf{w}}(s_b)\|^2\end{array}\right\} \quad (25)$$

Estimation of $\mathbf{w}^*$ via VI By using the same technique as in DVI, we can conclude that $\mathbf{w}^*(s)$ is inside the region:

$$\Omega'_{[s_b, s_a]} := \left\{\mathbf{w} : \begin{array}{l}\langle \mathbf{w}^*(s_a), \mathbf{w} - \mathbf{w}^*(s_a)\rangle \ge 0, \\ \|\mathbf{w} - \frac{1}{2}\hat{\mathbf{w}}(s_b)\| \le \frac{1}{2}\|\hat{\mathbf{w}}(s_b)\|\end{array}\right\} \quad (26)$$

We can see that $\Omega'_{[s_b, s_a]} \subset \Omega_{[s_b, s_a]}$, and thus SSNSV can be strictly improved by the estimation in (26). The rule based on $\Omega'[s_b, s_a]$ is presented in Theorem 18 in the supplement, which is call the "enhanced" SSNSV (ESSNSV).

6. Screening Rules for LAD

In this section, we extend DVI rules in Section 4 to the least absolute deviations regression (LAD). Suppose we have a training set $\{\mathbf{x}_i, y_i\}_{i=1}^l$, where $\mathbf{x}_i \in \Re^n$ and $y_i \in \Re$. The LAD problem takes the form of

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^l |y_i - \mathbf{w}^T\mathbf{x}_i|. \quad (27)$$

We can see that, if we set $\varphi(t) = |t|$ and $-a_i = b_i = 1$, problem (3) becomes the LAD problem. To construct the DVI rules for LAD based on Corollaries 7 and 8, we need to find α and β. Indeed, we have the following result:

Lemma 12. Let $\varphi(t) = |t|$, then $\alpha = -1$ and $\beta = 1$, i.e.,

$$\varphi^*(s) = \iota_{[-1,1]}. \quad (28)$$

We again omit the proof of Lemma 12 since it is a direct application of Eq. (1). Then, it is straightforward to derive the sequential DVI rules for the LAD problem.

Corollary 13. (**DVI**$_s^*$ for **LAD**) *For problem (27), suppose we are given a sequence of parameter values* $0 < C_1 < C_2 < \ldots < C_{\mathcal{K}}$. *Assume* $\theta^*(C_k)$ *is known for an arbitrary integer* $1 \leq k < \mathcal{K}$. *Then, for* C_{k+1}, *we have* $[\theta^*(C_{k+1})]_i = -1$ *or* 1, *i.e.,* $i \in \mathcal{R}$ *or* $i \in \mathcal{L}$, *if the following holds respectively*

1. $\frac{C_{k+1}+C_k}{2}\langle \mathbf{X}^T\theta^*(C_k), \mathbf{x}_i\rangle - \frac{C_{k+1}-C_k}{2}\|\mathbf{X}^T\theta^*(C_k)\|\|\mathbf{x}_i\| > y_i.$
2. $\frac{C_{k+1}+C_k}{2}\langle \mathbf{X}^T\theta^*(C_k), \mathbf{x}_i\rangle + \frac{C_{k+1}-C_k}{2}\|\mathbf{X}^T\theta^*(C_k)\|\|\mathbf{x}_i\| < y_i.$

Corollary 14. (**DVI**$_s$ for **LAD**) *For problem (27), suppose we are given a sequence of parameter values* $0 < C_1 < C_2 < \ldots < C_{\mathcal{K}}$. *Assume* $\mathbf{w}^*(C_k)$ *is known for an arbitrary integer* $1 \leq k < \mathcal{K}$. *Then, for* C_{k+1}, *we have* $[\theta^*(C_{k+1})]_i = -1$ *or* 1, *i.e.,* $i \in \mathcal{R}$ *or* $i \in \mathcal{L}$, *if the following holds respectively*

1. $\frac{C_{k+1}+C_k}{2C_k}\langle \mathbf{w}^*(C_k), \mathbf{x}_i\rangle - \frac{C_{k+1}-C_k}{2C_k}\|\mathbf{w}^*(C_k)\|\|\mathbf{x}_i\| > y_i,$
2. $\frac{C_{k+1}+C_k}{2C_k}\langle \mathbf{w}^*(C_k), \mathbf{x}_i\rangle + \frac{C_{k+1}-C_k}{2C_k}\|\mathbf{w}^*(C_k)\|\|\mathbf{x}_i\| < y_i.$

To the best of our knowledge, ours are the first screening rules for LAD.

7. Experiments

We evaluate DVI rules on both synthetic and real data sets. To measure the performance of the screening rules, we compute the rejection rate, that is, the ratio of the number of data instances whose membership can be identified by the rules to the total number of data instances. We test the rules along a sequence of 100 parameters of $C \in [10^{-2}, 10]$ equally spaced in the logarithmic scale. In Section 7.1, we compare the performance of DVI rules with SSNSV (Ogawa et al., 2013), which is the only existing method for identifying non-support vectors in SVM. Notice that, both of DVI rules and SSNSV are safe in the sense that no support vectors will be mistakenly discarded. We then evaluate DVI rules for LAD in Section 7.2.

7.1. DVI for SVM

In this experiment, we first apply DVI$_s$ to three simple 2D synthetic data sets to illustrate the effectiveness of the proposed screening methods. Then we compare the performance of DVI$_s$, SSNSV and ESSNSV on: (a) IJCNN1 data set (Prokhorov, 2001); (b) Wine Quality data set (Cortez et al., 2009); (c) Forest Covertype data set (Hettich & Bay, 1999). The original Forest Covertype data set includes 7 classes. We randomly pick two of the seven classes to construct the data set used in this paper.

Synthetic Data Sets In this experiment, we show that DVI$_s$ are very effective in discarding non-support vectors even

for largely overlapping classes. We evaluate DVI$_s$ rules on three synthetic data sets, i.e., Toy1, Toy2 and Toy3, plotted in the first row of Fig. 1. For each data set, we generate two classes. Each class has 1000 data points and is generated from $N(\{\mu, \mu\}^T, 0.75^2\mathbf{I})$, where $\mathbf{I} \in \Re^{2\times 2}$ is the identity matrix. For the positive classes (the red dots), $\mu = 1.5, 0.75, 0.5$, for Toy1, Toy2 and Toy 3, respectively; and $\mu = -1.5, -0.75, -0.5$, for the negative classes (the blue dots). From the plots, we can observe that when $|\mu|$ decreases, the two classes increasingly overlap and thus the number of data instances belong to the set $\mathcal{L}$ increases.

Table 1. Running time (in seconds) for solving the SVM problems with 100 parameter values by (a) "Solver" (solver without screening); (b) "Solver+DVI$_s$" (solver combined with DVI$_s$). "DVI$_s$" is the total running time (in seconds) of the rule. "Init." is the running time to solve SVM with the smallest parameter value.

	Solver	Solver+DVI$_s$	DVI$_s$	Init.	Speedup
Toy1	11.83	0.20	0.02	0.12	**59.15**
Toy2	13.68	0.52	0.03	0.15	**26.31**
Toy3	15.35	0.61	0.03	0.16	**25.16**

The second row of Fig. 1 presents the stacked area charts of the rejection rates. For convenience, let $\widetilde{\mathcal{R}}$ and $\widetilde{\mathcal{L}}$ be the indices of data instances which are identified by DVI$_s$ as members of $\mathcal{R}$ and $\mathcal{L}$, respectively. Then, the blue and red regions present the ratios of $|\widetilde{\mathcal{R}}|/l$ and $|\widetilde{\mathcal{L}}|/l$ (recall that, l is the number of data instances, which is 2000 for this experiment). We can see that, for Toy1, the two classes are clearly apart from each other and thus most of the data instances belong to the set $\mathcal{R}$. The first chart in the second row of Fig. 1 indicates that the proposed DVI$_s$ can identify almost all of the non-support vectors and thus the speedup is almost 60 times compared to the solver without screening (please refer to Table 1). When the two classes have a large overlap, e.g., Toy3, the number of data instances in $\mathcal{L}$ significantly increases. This will generally impose great challenge for the solver. But even for this challenging case, DVI$_s$ is still able to identify a large portion of the non-support vectors as indicated by the last charts in the second row of Fig. 1. Notice that, for Toy3, $|\widetilde{\mathcal{L}}|$ is comparable to $|\widetilde{\mathcal{R}}|$. Table 1 shows that the speedup gained by DVI$_s$ is about 25 times for this challenging case. It is worthwhile to mention that the running time of "Solver+DVI$_s$" in Table 1 includes the running time (the 5th column of Table 1) for solving SVM with the smallest parameter value.

Real Data Sets In this experiment, we compare the performance of SSNSV, ESSNSV and DVI$_s$ in terms of the rejection ratio, that is, the ratio of the number of data instances identified as members of $\mathcal{R}$ or $\mathcal{L}$ by the screening rules to the number of total data instances. Fig. 2 shows the rejection ratios of the three methods on three real data sets. We can observe that DVI$_s$ rules identify far more
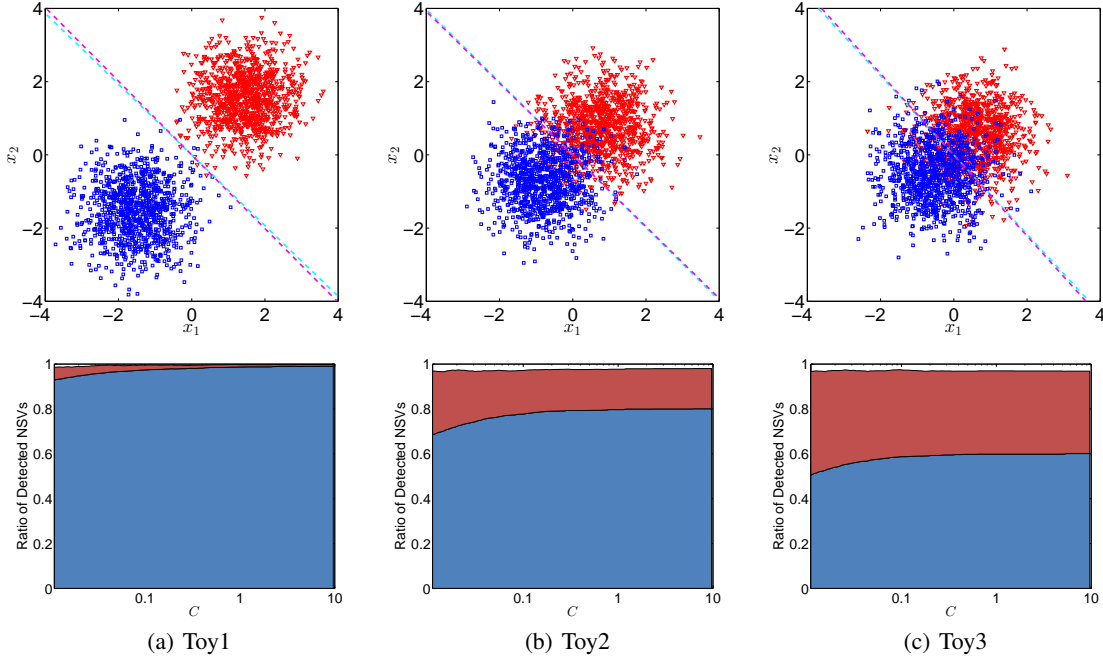
Figure 1. DVI$_s$ for three 2D synthetic data sets. The first row shows the plots of the data. Cyan and magenta dotted lines are the resulting decision functions at $C = 10^{-2}$ and $C = 10$, respectively. The second row presents the rejection rates of DVI$_s$ with the given 100 parameter values.
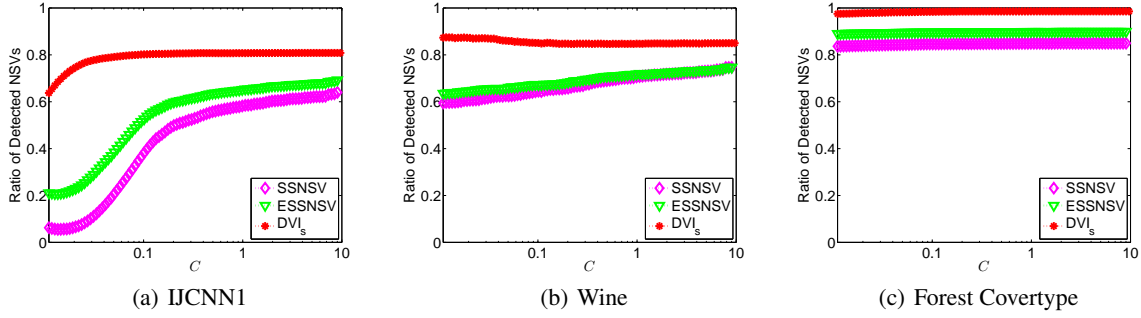

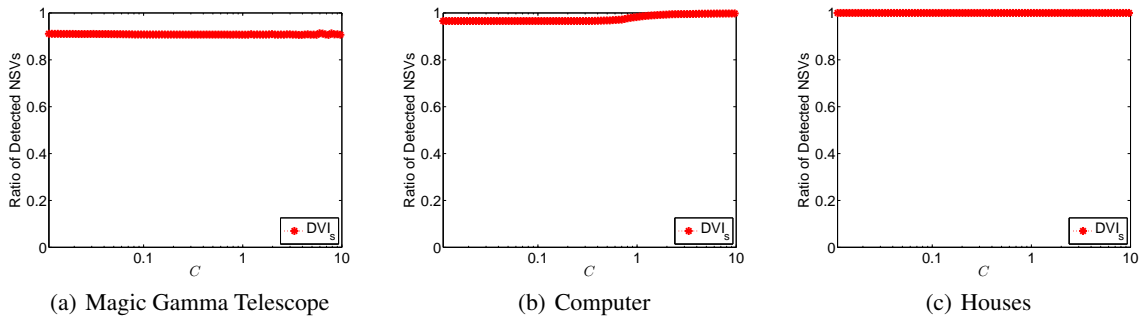*Figure 2.* Comparison of the performance of SSNSV, ESSNSV and DVIs for SVM on three real data sets.

Figure 3. Rejection ratio of DVIs for LAD on three real data sets.

non-support vectors than SSNSV and ESSNSV. For IJCN-N1, about 80% of the data instances are identified as non-support vectors by DVI$_s$. Therefore, as indicated by Table 2 the speedup gained by DVI$_s$ is about 5 times. For the Wine data set, more than 80% of the data instances are i-dentified to belong to $\mathcal{R}$ or $\mathcal{L}$ by DVI$_s$. As indicated in Table 2, the speedup is about 6 times gained by DVI$_s$. For

the Forest Covertype data set, almost all of data instances' membership can be determined by DVI$_s$. Table 2 shows that the speedup gained by DVI$_s$ is almost 80 times, which is much higher than that of SSNSV and ESSNSV. More-over, Fig. 2 indicates that ESSNSV is more effective in i-dentifying non-support vectors than SSNSV, which is con-sistent with our analysis.

Table 2. Running time (in seconds) for solving the SVM problems along the 100 parameter values on three real data sets. In "Solver+SSNSV" and "Solver+ESSNSV", "Init." reports the running time for solving SVM at the smallest and the largest parameter values since they are required to run SSNSV and ESSNSV. In "Solver+DVI$_s$", "Init." reports the running time for solving SVM at the smallest parameter value which is sufficient to run DVI$_s$. The running time reported by Init. is included in the total running time of the solver equipped with the screening methods.

IJCNN1 ($l = 49990, n = 22$)			Speedup
Solver	Total	4669.14	-
Solver+SSNSV	SSNSV	2.08	2.31
	Init.	92.45	
	Total	2018.55	
Solver+ESSNSV	ESSNSV	2.09	3.01
	Init.	91.33	
	Total	1552.72	
Solver+DVI$_s$	DVI$_s$	0.99	**5.64**
	Init.	42.67	
	Total	828.02	

Wine ($l = 6497, n = 12$)			Speedup
Solver	Total	76.52	-
Solver+SSNSV	SSNSV	0.02	3.50
	Init.	1.56	
	Total	21.85	
Solver+ESSNSV	ESSNSV	0.03	4.47
	Init.	1.60	
	Total	17.17	
Solver+DVI$_s$	DVI$_s$	0.01	**6.59**
	Init.	0.67	
	Total	11.62	

Forest Covertype ($l = 37877, n = 54$)			Speedup
Solver	Total	1675.46	-
Solver+SSNSV	SSNSV	2.73	7.60
	Init.	35.52	
	Total	220.58	
Solver+ESSNSV	ESSNSV	2.89	10.72
	Init.	36.13	
	Total	156.23	
Solver+DVI$_s$	DVI$_s$	1.27	**79.18**
	Init.	12.57	
	Total	21.16	

7.2. DVI for LAD

We evaluate the performance of DVI$_s$ for LAD on three real data sets: (a) Magic Gamma Telescope data set (Bache & Lichman, 2013); (b) Computer data set (Rasmussen et al.); (c) Houses data set (Pace & Barry, 1997). Fig. 3 shows the rejection ratio of DVI$_s$ rules for the three data sets. We can observe that the rejection ratio of DVI$_s$ on Magic Gamma Telescope data set is about 90%, leading to a 10 times

Table 3. Running time (in seconds) for solving the LAD problems with the given 100 parameter values on three real data sets. In "Solver+DVI$_s$", "Init." reports the running time for solving LAD at the smallest parameter value which is required to run DVI$_s$. Init. is included in the total running time of Solver+DVI$_s$.

Magic Gamma Telescope ($l = 19020, n = 10$)			Speedup
Solver	Total	122.34	-
Solver+DVI$_s$	DVI$_s$	0.28	**9.86**
	Init.	0.12	
	Total	12.41	

Computer ($l = 8192, n = 12$)			Speedup
Solver	Total	5.38	-
Solver+DVI$_s$	DVI$_s$	0.08	**19.21**
	Init.	0.05	
	Total	0.28	

Houses ($l = 20640, n = 8$)			Speedup
Solver	Total	21.43	-
Solver+DVI$_s$	DVI$_s$	0.06	**114.91**
	Init.	0.10	
	Total	0.19	

speedup as indicated in Table 3. For the Computer and Houses data sets, the rejection rates are very close to 100%, i.e., almost all of the data instances' membership can be determined by DVI$_s$. As expected, Table 3 shows that the resulting speedup are about 20 and 115 times, respectively. Notice that, the speedup for the Houses data set is more than two orders of magnitude. These results demonstrate the effectiveness of the proposed DVI rules.

8. Conclusion

In this paper, we develop new screening rules for a class of supervised learning problems by studying their dual formulation with the variational inequalities. Our framework includes two well known models, i.e., SVM and LAD, as special cases. The proposed DVI rules are very effective in identifying the non-support vectors for both SVM and LAD, and thus result in substantial savings in the computational cost and memory usage. Extensive experiments on both synthetic and real data sets demonstrate the effectiveness of the proposed DVI rules. We plan to extend the framework of DVI to other supervised learning problems, e.g., weighted SVM (Yang et al., 2005), RWLS (robust weighted least squres) (Chatterjee & Mächler, 1997), robust PCA (Ding et al., 2006), robust matrix factorization (Ke & Kanade, 2005).

Acknowledgments

References

Achlioptas, D., Mcsherry, F., and Schölkopf, B. Sampling techniques for kernel methods. In *NIPS*, 2002.

Bache, K. and Lichman, M. UCI machine learning repository, 2013. URL http://archive.ics.uci.edu/ml.

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

Buchinsky, M. Recent advances in quantile regression models. *The Journal of Human Resources*, 33:88–126, 1998.

Cao, D. and Boley, D. On approximate solutions to support vector machines. In *SDM*, 2006.

Chatterjee, S. and Mächler, M. Robust regression: a weighted least squares approach. *Communications in Statistics: Theory and Methods*, 26:1381–1394, 1997.

Chen, K., Ying, Z., Zhang, H., and Zhao, L. Analysis of least absolute deviation. *Biometrika*, 95:107–122, 2008.

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., and Reis, J. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47: 547–553, 2009.

Ding, C., Zhou, D., He, X., and Zha, H. R_1-PCA: Rotational invariant L_1-norm principle component analysis for robust subspace factorization. In *ICML*, 2006.

El Ghaoui, L., Viallon, V., and Rabbani, T. Safe feature elimination in sparse supervised learning. *Pacific Journal of Optimization*, 8:667–698, 2012.

Fan, R., Chang, K., Hsieh, C., Wang, X., and Lin, C. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

Güler, O. *Foundations of optimization*. Springer, 2010.

Hastie, T., Rosset, S., Tibshirani, R., and Zhu, J. The entire regularization path for the support vector machine. *Journal of Machine Learning Research*, 5:1391–1415, 2004.

Hettich, S. and Bay, S. UCI KDD Archive. 1999.

Hiriart-Urruty, J.-B. and Lemaréchal, C. *Convex analysis and minimization algorithms volumn I, II*. Springer-Verlag, 1993.

Hsieh, C. J., Chang, K. W., and Lin, C. J. A dual coordinate descent method for large-scale linear SVM. In *ICML*, 2008.

Jin, Z., Ying, Z., and Wei, L. A simple resampling method by perturbing the minimand. *Biometrika*, 88:381–390, 2001.

Joachims, T. Training linear SVMs in linear time. In *ACM KDD*, 2006.

Ke, Q. and Kanade, T. Robust L_1 norm factorization in the presence of outliers and missing data by alternative convex programming. In *CVPR*, 2005.

Liu, J., Zhao, Z., Wang, J., and Ye, J. Safe screening with variational inequalities and its application to lasso. In *ICML*, 2014.

Ogawa, K., Suzuki, Y., and Takeuchi, I. Safe screening of non-support vectors in pathwise SVM computation. In *ICML*, 2013.

Pace, R. and Barry, R. Sparse spatial autoregressions. *Statistics and Probability Letters*, 33:291–297, 1997.

Powell, J. Least absolute deviations estimation for the censored regression model. *Journal of Econometrics*, 25: 303–325, 1984.

Prokhorov, D. Slide presentation in IJCNN01. In *IJCNN 2001 neural network competition*, 2001.

Rao, C. R., Toutenburg, H., Shalabh, and Heumann, C. *Linear Models and Generalizations: Least Squares and Alternatives*. Springer, 2008.

Rasmussen, C., Neal, R., Hinton, G., Camp, D., Revow, M., Ghahramani, Z., Kustra, R., and Tibshirani, R. Data for evaluating learning in valid experiments.

Ruszczyński, A. *Nonlinear Optimization*. Princeton University Press, 2006.

Shalev-Shwartz, S., Singer, Y., and Srebro, N. Pegasos: primal estimated sub-gradient solver for SVM. In *ICML*, 2007.

Tibshirani, R., Bien, J., Friedman, J., Hastie, T., Simon, N., Taylor, J., and Tibshirani, R. Strong rules for discarding predictors in lasso-type problems. *Journal of the Royal Statistical Society Series B*, 74:245–266, 2012.

Wang, J., Zhou, J., Wonka, P., and Ye, J. Lasso screening rules via dual polytope projection. In *NIPS*, 2013.

Wang, L., Gordon, M., and Zhu, J. Regularized least absolute deviations regression and an efficient algorithm for parameter tuning. In *ICDM*, 2006.

Xiang, Z. J., Xu, H., and Ramadge, P. J. Learning sparse representation of high dimensional data on large scale dictionaries. In *NIPS*, 2011.

Yang, X., Song, Q., and Cao, A. Weighted support vector machine for data classification. In *IJCNN*, 2005.

Yu, H., Yang, J., and Han, J. Classifying large data sets using SVMs with hiearchical clusters. In *KDD*, 2003.

Scaling SVM and Least Absolute Deviations via Exact Data Reduction
Supplement Material

In this supplement, we present all of the details we mentioned in the main text.

A. Proof of Lemma 1

Before we prove Lemma 1, let us cite the following technical lemma.

Lemma 15. *(Hiriart-Urruty & Lemaréchal, 1993) The function f is equal to its biconjugate f^{**} if and only if $f \in \Gamma_0(\Re^n)$.*

We are now ready to derive a simple proof of Lemma 1 based on Lemma 15.

Proof. In order to show $\varphi^{**} = \varphi$, it is enough to show $\varphi \in \Gamma_0(\Re)$ according to Lemma 15. Therefore we only to check the following three conditions:

1). Properness: because $\varphi : \Re \to \Re_+$, i.e., there exists $t \in \Re$ such that $\varphi(t)$ is finite, φ is proper.

2). Lower semi-continuality: φ is lower semicontinuous because it is continuous.

3). Convexity: the convexity of φ is due to the its sublinearity, see Definition 1.

Thus, we have $\varphi \in \Gamma_0(\Re)$, which completes the proof. $\qquad\square$

B. Proof of Lemma 2

To prove Lemma 2, we need to following results.

Lemma 16. *(Ruszczyński, 2006) Let $Z \subseteq \Re^n$ be a convex and closed set. Let us define the support function of Z as*

$$\sigma_Z(s) := \sup_{\mathbf{x} \in Z} \mathbf{s}^T \mathbf{x}, \tag{29}$$

and the indicator function ι_Z as

$$\iota_Z(\mathbf{x}) = \begin{cases} 0, & \text{if } \mathbf{x} \in Z, \\ \infty, & \text{otherwise.} \end{cases} \tag{30}$$

Then

$$\sigma_Z^* = \iota_Z, \quad \text{amd } \iota_Z^* = \sigma_Z. \tag{31}$$

Theorem 17. *(Hiriart-Urruty & Lemaréchal, 1993) Let $\sigma \in \Gamma_0(\Re^n)$ be a sublinear function, then σ is the support function of the nonempty closed convex set*

$$S_\sigma := \{\mathbf{s} \in \Re^n : \mathbf{s}^T \mathbf{d} \leq \sigma(\mathbf{d}), \ \forall \mathbf{d} \in \Re^n\}. \tag{32}$$

We are now ready to prove Lemma 2.

Proof. Due to Lemma 16 and Theorem 17, we can see that, there is a nonempty closed convex set $Z \subseteq \Re$ such that

$$\varphi(t) = \sup_{s \in Z} st, \ \forall t \in \Re, \tag{33}$$

where

$$Z := \{s : st \leq \varphi(t), \ \forall t \in \Re\}. \tag{34}$$

Let $t = 1$ and -1 respectively, Eq. (34) implies that

$$\sup_{s \in Z} s \leq \varphi(1) \quad \text{and} \quad \inf_{s \in Z} s \geq \varphi(-1). \tag{35}$$

Therefore, Z is a closed and bounded interval, i.e., $Z = [\alpha, \beta]$ with $\alpha, \beta \in \Re$.

Next, let us show that $\alpha \neq \beta$. In fact, in view of the nonnegativity of φ and Eq. (34), it is easy to see that $0 \in Z$. Therefore, if $\alpha = \beta$, we must have $Z = \{0\}$. Thus, Lemma 16 implies that

$$\varphi = \iota_Z^* \equiv 0, \tag{36}$$

which contradicts the fact that φ is a nonconstant function. Hence, we can conclude that $\alpha < \beta$, which completes the proof. $\square$

C. Derivation of the KKT Condition in Eq. (14)

The problem in (12) can be written as follows:

$$\min_{\theta} \frac{C}{2} \|\mathbf{Z}^T \theta\|^2 - \langle \bar{\mathbf{y}}, \theta \rangle, \tag{37}$$
$$\text{s.t. } \theta_i \in [\alpha, \beta], \; i = 1, \ldots, l.$$

Therefore, we can see that the Lagrangian is

$$L(\theta, \mu, \nu) = \frac{C}{2} \|\mathbf{Z}^T \theta\|^2 - \langle \bar{\mathbf{y}}, \theta \rangle \tag{38}$$
$$+ \sum_{i=1}^{l} \mu_i(\alpha - \theta_i) + \sum_{i=1}^{l} \nu_i(\theta_i - \beta),$$

where $\mu = (\mu_1, \ldots, \mu_l)^T$, $\nu = (\nu_1, \ldots, \nu_l)^T$, and $\mu_i \geq 0$, $\nu_i \geq 0$ for all $i = 1, \ldots, l$. μ and ν are in fact the vector of Lagrangian multipliers.

For simplicity, let us denote $\theta^*(C)$ by θ^*. Then the KKT conditions (Boyd & Vandenberghe, 2004) are

$$\frac{\partial L(\theta, \mu, \nu)}{\partial \theta}\Big|_{\theta^*} = 0 \Rightarrow C\mathbf{Z}\mathbf{Z}^T \theta^* - \bar{\mathbf{y}} - \mu + \nu = 0, \tag{39}$$

$$\begin{aligned} \mu_i(\alpha - \theta_i^*) &= 0, \\ \nu_i(\theta_i^* - \beta) &= 0, \end{aligned} \; i = 1, \ldots, l. \tag{40}$$

Eq. (40) is known as the complementary slackness condition. The equation in (39) actually involves l equations. We can write down the ith equation as follows:

$$C\langle \mathbf{Z}^T \theta^*, a_i \mathbf{x}_i \rangle - \mu_i + \nu_i = b_i y_i. \tag{41}$$

Recall that the ith column of $\mathbf{Z}$ is $a_i \mathbf{x}_i$. In view of Eq. (40) and Eq. (41), we can see that:

1. if $\theta_i^* = \alpha$, then $\nu_i = 0$ and Eq. (41) results in

$$C\langle \mathbf{Z}^T \theta^*, a_i \mathbf{x}_i \rangle \geq b_i y_i; \tag{42}$$

2. if $\theta_i^* \in (\alpha, \beta)$, then $\mu_i = \nu_i = 0$ and Eq. (41) results in

$$C\langle \mathbf{Z}^T \theta^*, a_i \mathbf{x}_i \rangle = b_i y_i; \tag{43}$$

3. if $\theta_i^* = \beta$, then $\mu_i = 0$ and Eq. (41) results in

$$C\langle \mathbf{Z}^T \theta^*, a_i \mathbf{x}_i \rangle \leq b_i y_i. \tag{44}$$

Then, in view of the inequalities in (42), (43) and (44), and Eq. (13), it is straightforward to derive the KKT condition in (14).

D. Proof of Lemma 3

Proof. The first part of the statement is trivial by the definition of $\hat{\mathcal{R}}$ and $\hat{\mathcal{L}}$. Therefore, we only consider the second part of the statement.

Let $\mathbf{G} = \mathbf{ZZ}^T$. By permuting the columns and rows of $\mathbf{G}$, we have

$$\hat{\mathbf{G}} = \begin{pmatrix} \hat{\mathbf{G}}_{11} & \hat{\mathbf{G}}_{12} \\ \hat{\mathbf{G}}_{21} & \hat{\mathbf{G}}_{22} \end{pmatrix} = \begin{pmatrix} [\mathbf{X}^T]_{\hat{\mathcal{S}}^c}^T [\mathbf{X}^T]_{\hat{\mathcal{S}}^c} & [\mathbf{X}^T]_{\hat{\mathcal{S}}^c}^T [\mathbf{X}^T]_{\hat{\mathcal{S}}} \\ [\mathbf{X}^T]_{\hat{\mathcal{S}}}^T [\mathbf{X}^T]_{\hat{\mathcal{S}}^c} & [\mathbf{X}^T]_{\hat{\mathcal{S}}}^T [\mathbf{X}^T]_{\hat{\mathcal{S}}} \end{pmatrix}.$$

As a result, the objective function of problem (12) can be rewritten as

$$\frac{C}{2}[\theta]_{\hat{\mathcal{S}}^c}^T \hat{\mathbf{G}}_{11}[\theta]_{\hat{\mathcal{S}}^c} - \hat{\mathbf{y}}^T[\theta]_{\hat{\mathcal{S}}^c} + R([\theta]_{\hat{\mathcal{S}}}) \tag{45}$$

where

$$\hat{\mathbf{y}} = \mathbf{y}_{\hat{\mathcal{S}}^c} - C\hat{\mathbf{G}}_{12}[\theta]_{\hat{\mathcal{S}}}, \tag{46}$$

$$R([\theta]_{\hat{\mathcal{S}}}) = \frac{C}{2}[\theta]_{\hat{\mathcal{S}}}^T \hat{\mathbf{G}}_{22}[\theta]_{\hat{\mathcal{S}}} - \mathbf{y}_{\hat{\mathcal{S}}}^T[\theta]_{\hat{\mathcal{S}}} \tag{47}$$

Due to the assumption that $[\theta^*(C)]_{\hat{\mathcal{S}}}$ is known, $\hat{\mathbf{y}}$ and $R([\theta]_{\hat{\mathcal{S}}})$ can be treated as constants, and thus problem (12) reduces to problem (15). $\square$

E. Review of the Dual Coordinate Descent Method

Problem (15) can be efficiently solved by the dual coordinate descent method (Hsieh et al., 2008).

More precisely, the optimization procedure starts from an initial point $\hat{\theta}^0 \in \Re^{|\hat{\mathcal{S}}^c|}$ and generates a sequence of points $\{\hat{\theta}^k\}_{k=0}^{\infty}$. The process from $\hat{\theta}^k$ to $\hat{\theta}^{k+1}$ is referred to as an outer iteration. In each outer iteration, we update the components of $\hat{\theta}^k$ one at a time and thus get a sequence of points $\hat{\theta}^{k,i} \in \Re^{|\hat{\mathcal{S}}^c|}$, $i = 1, \ldots, |\hat{\mathcal{S}}^c|$. Suppose we are at the kth outer iteration. To get $\hat{\theta}^{k,i}$ from $\hat{\theta}^{k,i-1}$, we need to solve the following optimization problem:

$$\min_t \quad \frac{C}{2}(\hat{\theta}^{k,i-1} + t\mathbf{e}_i)^T \hat{\mathbf{G}}_{11}(\hat{\theta}^{k,i-1} + t\mathbf{e}_i) \tag{48}$$
$$- \hat{\mathbf{y}}^T(\hat{\theta}^{k,i-1} + t\mathbf{e}_i)$$
$$\text{s.t.} \quad [\hat{\theta}^{k,i-1}]_i + t \in [\alpha, \beta], \quad i = 1, \ldots, l,$$

where $\mathbf{e}_i = (0, \ldots, 1, \ldots, 0)^T$. Clearly, problem (48) is equivalent to the following 1D optimization problem:

$$\min_t \quad \frac{C}{2}[\hat{\mathbf{G}}_{11}]_{i,i} t^2 + (C\mathbf{e}_i^T \hat{\mathbf{G}}_{11}\hat{\theta}^{k,i-1} - [\hat{\mathbf{y}}]_i)t \tag{49}$$
$$\text{s.t.} \quad [\hat{\theta}^{k,i-1}]_i + t \in [\alpha, \beta],$$

which admits a closed form solution t^*. Once t^* is available, we can set $\hat{\theta}^{k,i} = \hat{\theta}^{k,i-1} + t^*\mathbf{e}_i$. For more details, please refer to Hsieh et al. (2008).

F. Proof of Theorem 6

Proof. We will prove the first half of the statement. The second half can be proved analogously. To show $[\theta^*(C)]_i = \alpha$, i.e., $i \in \mathcal{R}$, (R1) implies that we only need to show $C\langle \mathbf{Z}^T \theta^*(C), a_i \mathbf{x}_i \rangle > b_i y_i$. Thus, we can see that

$$
\begin{aligned}
C\langle \mathbf{Z}^T \theta^*(C), a_i \mathbf{x}_i \rangle &= C \left\langle \mathbf{Z}^T \theta^*(C) - \frac{C_0 + C}{2C} \mathbf{Z}^T \theta^*(C_0), a_i \mathbf{x}_i \right\rangle + C \left\langle \frac{C_0 + C}{2C} \mathbf{Z}^T \theta^*(C_0), a_i \mathbf{x}_i \right\rangle \\
&\geq \frac{C_0 + C}{2} \langle \mathbf{Z}^T \theta^*(C_0), a_i \mathbf{x}_i \rangle - C \left\| \mathbf{Z}^T \theta^*(C) - \frac{C_0 + C}{2C} \mathbf{X}^T \theta^*(C_0) \right\| \| a_i \mathbf{x}_i \| \\
&\geq \frac{C_0 + C}{2} \langle \mathbf{Z}^T \theta^*(C_0), a_i \mathbf{x}_i \rangle - \frac{C - C_0}{2} \| \mathbf{Z}^T \theta^*(C_0) \| \| a_i \mathbf{x} \| \\
&> b_i y_i.
\end{aligned}
$$

Note that, the second inequality is due to Theorem 5, and the last line is due to the statement. This completes the proof. $\square$

G. Improving SSNSV via VI

In this section, we describe how to strictly improve SSNSV by using the same technique used in DVI rules in a detailed manner.

Estimation of $\mathbf{w}^*$ via VI

We show that $\Omega_{[s_b, s_a]}$ in Eq. (25) can be strictly improved by the variational inequalities. Consider $\mathcal{F}_{s_a}$. Because $s_a > s_b$, we can see that $\mathbf{w}^*(s_b) \in \mathcal{F}_{s_a}$. Therefore, by Theorem 4, we have

$$
\langle \mathbf{w}^*(s_a), \mathbf{w}^*(s) - \mathbf{w}^*(s_a) \rangle \geq 0, \tag{50}
$$

which is the first constraint in (25). Similarly, consider $\mathcal{F}_{s_b}$. Since $\hat{\mathbf{w}}(s_b) \in \mathcal{F}_{s_b}$, Theorem 4 implies that

$$
\langle \mathbf{w}^*(s), \hat{\mathbf{w}}(s_b) - \mathbf{w}^*(s) \rangle \geq 0,
$$

which is equivalent to

$$
\| \mathbf{w}^*(s) - \tfrac{1}{2} \hat{\mathbf{w}}(s_b) \| \leq \tfrac{1}{2} \| \hat{\mathbf{w}}(s_b) \|. \tag{51}
$$

Clearly, the radius determined by the inequality (51) is only a half of the radius determined by the second constraint in (25). In view of the inequalities in (50) and (51), we can see that $\mathbf{w}^*(s)$ can be bounded inside the following region:

$$
\Omega'_{[s_b, s_a]} := \left\{ \mathbf{w} : \begin{array}{l} \langle \mathbf{w}^*(s_a), \mathbf{w} - \mathbf{w}^*(s_a) \rangle \geq 0, \\ \| \mathbf{w} - \tfrac{1}{2} \hat{\mathbf{w}}(s_b) \| \leq \tfrac{1}{2} \| \hat{\mathbf{w}}(s_b) \| \end{array} \right\}
$$

It is easy to see that $\Omega'_{[s_b, s_a]} \subset \Omega_{[s_b, s_a]}$. As a result, the bounds in (R1$'$) and (R2$'$) with $\Omega'_{[s_b, s_a]}$ are tighter than that of $\Omega_{[s_b, s_a]}$. Thus, SSNSV (Ogawa et al., 2013) can be strictly improved by the estimation in (26). In fact, we have the following theorem:

Theorem 18. *Suppose we are given two parameters $s_a > s_b > 0$, and let $\mathbf{w}^*(s_a)$ and $\hat{\mathbf{w}}(s_b)$ be the optimal solution at $s = s_a$ and a feasible solution at $s = s_b$, respectively. Moreover, let us define*

$$
\rho = -\| \mathbf{w}^*(s_a) \|^2 + \tfrac{1}{2} \langle \mathbf{w}^*(s_a), \hat{\mathbf{w}}(s_b) \rangle
$$

$$
\mathbf{v}^\perp = \mathbf{v} - \frac{\mathbf{v}^T \mathbf{w}^*(s_a)}{\| \mathbf{w}^*(s_a) \|^2} \mathbf{w}^*(s_a), \forall \mathbf{v} \in \Re^n.
$$

Then, for all $s \in [s_b, s_a]$,

$$
\langle \mathbf{w}^*(s_a), \bar{x}_i \rangle > \frac{2 \| \bar{x}_i \|}{\| \hat{\mathbf{w}}(s_b) \|} \rho \text{ and } \ell_i > 1 \Rightarrow i \in \mathcal{R} \Leftrightarrow \alpha_i = 0, \tag{52}
$$

where

$$
\ell_i = -\frac{\langle \mathbf{w}^*(s_a), \bar{\mathbf{x}}_i \rangle}{\| \mathbf{w}^*(s_a) \|^2} \rho + \tfrac{1}{2} \langle \hat{\mathbf{w}}(s_b), \bar{\mathbf{x}}_i \rangle \tag{53}
$$

$$
- \| \bar{\mathbf{x}}_i^\perp \| \sqrt{\frac{1}{4} \| \hat{\mathbf{w}}(s_b) \|^2 - \frac{\rho^2}{\| \mathbf{w}^*(s_a) \|^2}}.
$$

Similarly,

$$u_i < 1 \Rightarrow i \in \mathcal{L} \Leftrightarrow \alpha_i = c, \tag{54}$$

where

$$u_i = \begin{cases} \frac{1}{2}\left(\langle \hat{\mathbf{w}}(s_b), \bar{\mathbf{x}}_i \rangle + \|\hat{\mathbf{w}}(s_b)\|\|\bar{\mathbf{x}}_i\|\right), \\ \qquad \text{if } \langle \mathbf{w}^*(s_a), \bar{x}_i \rangle \geq -\frac{2\|\bar{\mathbf{x}}_i\|}{\|\hat{\mathbf{w}}(s_b)\|}\rho \\ -\frac{\langle \mathbf{w}^*(s_a), \bar{\mathbf{x}}_i \rangle}{\|\mathbf{w}^*(s_a)\|^2}\rho + \frac{1}{2}\langle \hat{\mathbf{w}}(s_b), \bar{\mathbf{x}}_i \rangle \\ \qquad +\|\bar{\mathbf{x}}_i^\perp\|\sqrt{\frac{1}{4}\|\hat{\mathbf{w}}(s_b)\|^2 - \frac{\rho^2}{\|\mathbf{w}^*(s_a)\|^2}}, \\ \qquad \text{otherwise.} \end{cases} \tag{55}$$

For convenience, we call the screening rule presented in Theorem 18 as the "enhanced" SSNSV (ESSNSV).

To prove Theorem 18, we first establish the following technical lemma.

Lemma 19. *Consider the problem as follows:*

$$\min_{\mathbf{w}} f(\mathbf{w}) = \mathbf{v}^T\mathbf{w}, \text{ s.t. } \mathbf{u}^T\mathbf{w} \leq d, \|\mathbf{w} - \mathbf{o}\| \leq r, \tag{56}$$

where $r > 0$. Let $d' = d - \mathbf{u}^T\mathbf{o}$ and the optimal solution of problem (56) be f^*. Then we have

1. *If $\mathbf{v}^T\mathbf{u} + \frac{\|\mathbf{v}\|d'}{r} \geq 0$, then*

$$f^* = \mathbf{v}^T\mathbf{o} - r\|\mathbf{v}\|.$$

2. *Otherwise,*

$$f^* = \mathbf{v}^T\mathbf{o} - \|\mathbf{v}^\perp\|\sqrt{r^2 - \frac{(d')^2}{\|\mathbf{u}\|^2}} + \frac{\mathbf{v}^T\mathbf{u}d'}{\|\mathbf{u}\|^2},$$

where $\mathbf{v}^\perp = \mathbf{v} - \frac{\mathbf{v}^T\mathbf{u}}{\|\mathbf{u}\|^2}\mathbf{u}$.

Notice that, we assume problem (56) is feasible, i.e., $\frac{|\mathbf{u}^T\mathbf{o} - d|}{\|\mathbf{u}\|} \leq r$.

Proof. Let $\mathbf{z} = \mathbf{w} - \mathbf{o}$, problem (56) can be rewritten as:

$$\min_{\mathbf{z}} \mathbf{v}^T\mathbf{z} + \mathbf{v}^T\mathbf{o}, \text{ s.t. } \mathbf{u}^T\mathbf{z} \leq d - \mathbf{u}^T\mathbf{o}, \|\mathbf{z}\| \leq r. \tag{57}$$

Problem (57) reduces to

$$\min_{\mathbf{z}} \mathbf{v}^T\mathbf{z}, \text{ s.t. } \mathbf{u}^T\mathbf{z} \leq d', \|\mathbf{z}\| \leq r. \tag{58}$$

To solve problem (58), we make use of the Lagrangian multiplier method. For notational convenience, let $\mathcal{F} := \{\mathbf{z} : \mathbf{u}^T\mathbf{z} \leq d', \|\mathbf{z}\| \leq r\}$.

$$\begin{aligned} \min_{\mathbf{z}\in\mathcal{F}} \mathbf{v}^T\mathbf{z} &= \min_{\mathbf{z}} \max_{\substack{\mu \geq 0, \\ \nu \geq 0}} \mathbf{v}^T\mathbf{z} + \nu(\mathbf{u}^T\mathbf{z} - d') + \frac{\mu}{2}(\|\mathbf{z}\|^2 - r^2) \\ &= \max_{\substack{\mu \geq 0, \\ \nu \geq 0}} \min_{\mathbf{z}} \mathbf{v}^T\mathbf{z} + \nu(\mathbf{u}^T\mathbf{z} - d') + \frac{\mu}{2}(\|\mathbf{z}\|^2 - r^2) \\ &= \max_{\substack{\mu \geq 0, \\ \nu \geq 0}} -\frac{1}{2\mu}\|\mathbf{v} + \nu\mathbf{u}\|^2 - \nu d' - \frac{\mu r^2}{2}. \end{aligned} \tag{59}$$

Notice that, in Eq. (59), we make the assumption that $\mu > 0$. However, we can not simply exclude this possibility. In fact, if $\mu = 0$, we must have

$$\mathbf{v} + \nu \mathbf{u} = 0, \tag{60}$$

since otherwise the function value of

$$\mathbf{v}^T \mathbf{z} + \nu (\mathbf{u}^T \mathbf{z} - d')$$

in the second line of Eq. (59) can be made arbitrarily small. As a result, we will have

$$g(\mu, \nu) = -\infty,$$

which contradicts the strong duality of problem (56) (Boyd & Vandenberghe, 2004). [Problem (56) is clearly lower bounded since the feasible set is compact.] Therefore, in view of Eq. (60), we can conclude that $\mu = 0$ only if $\mathbf{v}$ point in the opposite direction of $\mathbf{u}$.

Let us first consider the general case, i.e., $\mathbf{v}$ does not point in the opposite direction of $\mathbf{u}$. In view of Eq. (59), let $g(\mu, \nu) = -\frac{1}{2\mu} \|\mathbf{v} + \nu \mathbf{u}\|^2 - \nu d' - \frac{\mu r^2}{2}$. It is easy to see that

$$\frac{\partial g(\mu, \nu)}{\partial \nu} = 0 \Leftrightarrow \nu = -\frac{\mathbf{v}^T \mathbf{u} + \mu d'}{\|\mathbf{u}\|^2}. \tag{61}$$

Since ν has to be nonnegative, we have

$$\nu = \max \left\{ 0, -\frac{\mathbf{v}^T \mathbf{u} + \mu d'}{\|\mathbf{u}\|^2} \right\}. \tag{62}$$

Case 1. If $-\frac{\mathbf{v}^T \mathbf{u} + \mu d'}{\|\mathbf{u}\|^2} \leq 0$, then $\nu = 0$ and thus

$$\frac{\partial g(\mu, \nu)}{\partial \mu} = 0 \Leftrightarrow \mu = \frac{\|\mathbf{v}\|}{r}. \tag{63}$$

Then $g(\mu, \nu) = -r\|\mathbf{v}\|$ and the optimal value of problem (56) is given by

$$\mathbf{v}^T \mathbf{o} - r\|\mathbf{v}\|. \tag{64}$$

Case 2. If $-\frac{\mathbf{v}^T \mathbf{u} + \mu d'}{\|\mathbf{u}\|^2} > 0$, then $\nu = -\frac{\mathbf{v}^T \mathbf{u} + \mu d'}{\|\mathbf{u}\|^2}$ and

$$g(\mu, \nu) = -\frac{1}{2\mu} \|\mathbf{v}^\perp\|^2 - \frac{\mu}{2} \left(r^2 - \frac{(d')^2}{\|\mathbf{u}\|^2} \right) + \frac{\mathbf{v}^T \mathbf{u} d'}{\|\mathbf{u}\|^2}, \tag{65}$$

Thus,

$$\frac{\partial g(\mu, \nu)}{\partial \mu} = 0 \Leftrightarrow \mu = \frac{\|\mathbf{v}^\perp\|}{\sqrt{r^2 - \frac{(d')^2}{\|\mathbf{u}\|^2}}} \tag{66}$$

Then $g(\mu, \nu) = -\|\mathbf{v}^\perp\| \sqrt{r^2 - \frac{(d')^2}{\|\mathbf{u}\|^2}} + \frac{\mathbf{v}^T \mathbf{u} d'}{\|\mathbf{u}\|^2}$ and the optimal value of problem (56) is given by

$$\mathbf{v}^T \mathbf{o} - \|\mathbf{v}^\perp\| \sqrt{r^2 - \frac{(d')^2}{\|\mathbf{u}\|^2}} + \frac{\mathbf{v}^T \mathbf{u} d'}{\|\mathbf{u}\|^2}. \tag{67}$$

Now let us consider the case with $\mathbf{v}$ pointing in the opposite direction of $\mathbf{u}$. We can see that there exists $\gamma = -\frac{\mathbf{u}^T \mathbf{v}}{\|\mathbf{u}\|^2} > 0$ such that $\mathbf{v} = -\gamma \mathbf{u}$. By plugging $\mathbf{v} = -\gamma \mathbf{u}$ in problem (56) and following an analogous argument as before, we can see that the statement in Lemma 19 is also applicable to this case.

Therefore, the proof of the statement is completed. □

We are now ready to prove Theorem 18.

Proof. To prove the statements in (52) and (53), we only need to set

$$
\begin{aligned}
\mathbf{v} &:= \bar{\mathbf{x}}_i, \quad \mathbf{u} : -\mathbf{w}^*(s_a), \quad d := -\|\mathbf{w}^*(s_a)\|^2, \\
\mathbf{o} &:= \tfrac{1}{2}\hat{\mathbf{w}}, \quad r := \tfrac{1}{2}\|\hat{\mathbf{w}}\|, \\
d' &:= \rho = -\|\mathbf{w}^*(s_a)\|^2 + \frac{1}{2}\langle \mathbf{w}^*(s_a), \hat{\mathbf{w}} \rangle,
\end{aligned}
$$

and then apply Lemma 19. Notice that, for case 1, the optimal value

$$
f^* = \frac{1}{2}\left(\langle \hat{\mathbf{w}}(s_b), \bar{\mathbf{x}}_i \rangle - \|\hat{\mathbf{w}}(s_b)\|\|\bar{\mathbf{x}}_i\| \right) \le 0,
$$

and thus none of the non-support vectors can be identified [recall that, according to (R1′), f^* has to be larger than 1 such that $\bar{\mathbf{x}}_i$ can be detected as a non-support vector]. As a result, we only need to consider case 2.

The statement in (54) and (55) follows with an analogous argument by noting that

$$
\max_{\mathbf{w}\in\Theta'_{[s_b,s_a]}} \langle \mathbf{w}, \bar{\mathbf{x}}_i \rangle = - \min_{\mathbf{w}\in\Theta'_{[s_b,s_a]}} -\langle \mathbf{w}, \bar{\mathbf{x}}_i \rangle.
$$

$\square$