Scaling Up Sparse Support Vector Machines
by Simultaneous Feature and Sample Reduction

Weizhong Zhang* 1 2 Bin Hong* 1 3 Wei Liu 2 Jieping Ye 3 Deng Cai 1 Xiaofei He 1 Jie Wang 3

Abstract
Sparse support vector machine (SVM) is a popular classification technique that can simultaneously learn a small set of the most interpretable features and identify the support vectors. It has achieved great successes in many real-world applications. However, for large-scale problems involving a huge number of samples and extremely high-dimensional features, solving sparse SVMs remains challenging. By noting that sparse SVMs induce sparsities in both feature and sample spaces, we propose a novel approach, which is based on accurate estimations of the primal and dual optima of sparse SVMs, to simultaneously identify the features and samples that are guaranteed to be irrelevant to the outputs. Thus, we can remove the identified inactive samples and features from the training phase, leading to substantial savings in both the memory usage and computational cost without sacrificing accuracy. To the best of our knowledge, the proposed method is the first static feature and sample reduction method for sparse SVM. Experiments on both synthetic and real datasets (e.g., the kddb dataset with about 20 million samples and 30 million features) demonstrate that our approach significantly outperforms state-of-the-art methods and the speedup gained by our approach can be orders of magnitude.

1. Introduction
Sparse support vector machine (SVM) (Bi et al., 2003; Wang et al., 2006) is a powerful technique that can simultaneously perform classification by margin maximization and variable selection by \( \ell_1 \)-norm penalty. The last few years have witnessed many successful applications of sparse SVMs, such as text mining (Joachims, 1998; Yoshikawa et al., 2014), bioinformatics (Narasimhan & Aggarwal, 2013) and image processing (Mohr & Obermayer, 2004; Kotsia & Pitas, 2007). Many algorithms (Hastie et al., 2004; Fan et al., 2008; Catanzaro et al., 2008; Hsieh et al., 2008; Shalev-Shwartz et al., 2011) have been proposed to efficiently solve sparse SVM problems. However, the applications of sparse SVMs to large-scale learning problems, which involve a huge number of samples and extremely high-dimensional features, remain challenging.

An emerging technique, called screening (El Ghaoui et al., 2012), has been shown to be promising in accelerating large-scale sparse learning. The essential idea of screening is to quickly identify the zero coefficients in the sparse solutions without solving any optimization problems such that the corresponding features or samples—that are called inactive features or samples—can be removed from the training phase. Then, we only need to perform optimization on the reduced datasets instead of the full datasets, leading to substantial savings in the computational cost and memory usage. Here, we need to emphasize that screening differs greatly from feature selection methods, although they look similar at the first glance. To be precise, screening is devoted to accelerating the training of many sparse models including Lasso, Sparse SVM, etc., while feature selection is the goal of these models. In the past few years, many screening methods are proposed for a large set of sparse learning techniques, such as Lasso (Tibshirani et al., 2012; Xiang & Ramadge, 2012; Wang et al., 2013), group Lasso (Ndiaye et al., 2016), \( \ell_1 \)-regularized logistic regression (Wang et al., 2014), and SVM (Ogawa et al., 2013). Empirical studies indicate that screening methods can lead to orders of magnitude of speedup in computation time.

However, most existing screening methods study either feature screening or sample screening individually (Shibagaki et al., 2016) and their applications have very different scenarios. Specifically, to achieve better performance (say, in terms of speedup), we favor feature screening methods when the number of features \( p \) is much larger than the number of samples \( n \), while sample screening methods are...
preferable when $n \gg p$. Note that there is another class of sparse learning techniques, like sparse SVMs, which induce sparsities in both feature and sample spaces. All these screening methods are helpless in accelerating the training of these models with large $n$ and $p$. We also cannot address this problem by simply combining the existing feature and sample screening methods. The reason is that they could mistakenly discard relevant data as they are specifically designed for different sparse models. Recently, Shibagaki et al. (Shibagaki et al., 2016) consider this problem and propose a method to simultaneously identify the inactive features and samples in a dynamic manner (Bonnefoy et al., 2014); that is, during the optimization process, they trigger their testing rule when there is a sufficient decrease in the duality gap. Thus, the method in (Shibagaki et al., 2016) can discard more inactive features and samples as the optimization proceeds and one has small-scale problems to solve in the late stage of the optimization. Nevertheless, the overall speedup can be limited as the problems’ size can be large in the early stage of the optimization. To be specific, the method in (Shibagaki et al., 2016) depends heavily on the duality gap during the optimization process. The duality gap in the early stage can always be large, which makes the dual and primal estimations inaccurate and finally results in ineffective screening rules. Hence, it is essentially solving a large problem in the early stage.

In this paper, to address the limitations in the dynamic screening method, we propose a novel screening method that can simultaneously identify Inactive Features and Samples (SIFS) for sparse SVMs in a static manner, that is, we only need to perform SIFS once before (instead of during) optimization. Thus, we only need to run the optimization algorithm on small-scale problems. The major technical challenge in developing SIFS is that we need to accurately estimate the primal and dual optima. The more accurate the estimations are, the more effective SIFS is in detecting inactive features and samples. Thus, our major technical contribution is a novel framework, which is based on the strong convexity of the primal and dual problems of sparse SVMs [see problems (P$^*$) and (D$^*$) in Section 2] for deriving accurate estimations of the primal and dual optima (see Section 3). Another appealing feature of SIFS is the so-called synergy effect (Shibagaki et al., 2016). Specifically, the proposed SIFS consists of two parts, i.e., Inactive Feature Screening (IFS) and Inactive Samples Screening (ISS). We show that discarding inactive features (samples) identified by IFS (ISS) leads to a more accurate estimation of the primal (dual) optimum, which in turn dramatically enhances the capability of ISS (IFS) in detecting inactive samples (features). Thus, SIFS applies IFS and ISS in an alternating manner until no more inactive features and samples can be identified, leading to much better performance in scaling up large-scale problems than the application of ISS or IFS individually. Moreover, SIFS (see Section 4) is safe in the sense that the detected features and samples are guaranteed to be absent from the sparse representations. To the best of our knowledge, SIFS is the first static screening rule for sparse SVM that is able to simultaneously detect inactive features and samples. Experiments (see Section 5) on both synthetic and real datasets demonstrate that SIFS significantly outperforms the state-of-the-art (Shibagaki et al., 2016) in improving the efficiency of sparse SVMs and the speedup can be orders of magnitude. Detailed proofs of theoretical results in the main text are in the supplementary supplements.

Notations: Let $\| \cdot \|_1$, $\| \cdot \|_2$, and $\| \cdot \|_{\infty}$ be the $\ell_1$, $\ell_2$, and $\ell_{\infty}$ norms, respectively. We denote the inner product of vectors $x$ and $y$ by $\langle x, y \rangle$, and the $i$-th component of $x$ by $[x]_i$. Let $[p] = \{1, 2, \ldots, p\}$ for a positive integer $p$. Given a subset $\mathcal{J} := \{j_1, \ldots, j_k\}$ of $[p]$, let $|\mathcal{J}|$ be the cardinality of $\mathcal{J}$. For a vector $x$, let $[x]_{\mathcal{J}} = ([x]_{j_1}, \ldots, [x]_{j_k})^T$. For a matrix $X$, let $[X]_{\mathcal{J}} = ([x_{1j}], \ldots, [x_{pj}])^T$ and $[X] = ([x^1]^T, \ldots, [x^p]^T)^T$, where $x^i$ and $x_j$ are the $i$th row and $j$th column of $X$, respectively. For a scalar $t$, we denote $\max\{0, t\}$ by $[t]_+$. 

## 2. Basics and Motivations

In this section, we briefly review some basics of sparse SVMs and then motivate SIFS via the KKT conditions. Specifically, we focus on the $\ell_1$-regularized SVM with a smoothed hinged loss that has strong theoretical guarantees (Shalev-Shwartz & Zhang, 2016), which takes the form of

$$
\min_{w \in \mathbb{R}^p} P(w; \alpha, \beta) = \frac{1}{n} \sum_{i=1}^{n} \ell(1 - \langle \bar{x}_i, w \rangle) + \frac{\alpha}{2} \| w \|^2 + \beta \| w \|_1,
$$

(P$^*$)

where $w$ is the parameter vector to be estimated, $\{\bar{x}_i, y_i\}_{i=1}^n$ is the training set, $\bar{x}_i \in \mathbb{R}^p$, $y_i \in \{-1, +1\}$, $\alpha$ and $\beta$ are positive parameters, and the loss function $\ell(\cdot) : \mathbb{R} \rightarrow \mathbb{R}$ is

$$
\ell(t) = \begin{cases} 
0, & \text{if } t < 0, \\
\frac{t^2}{2}, & \text{if } 0 \leq t \leq \gamma, \\
\gamma - \frac{t}{2}, & \text{if } t > \gamma,
\end{cases}
$$

where $\gamma \in (0, 1)$. We present the Lagrangian dual problem of problem (P$^*$) and the KKT conditions in the following theorem, which plays a fundamentally important role in developing our screening rule.

**Theorem 1.** Let $\bar{X} = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n)$ and $S_3(\cdot)$ be the soft-thresholding operator (Hastie et al., 2015), i.e., $[S_3(u)]_i = \text{sign}([u]_i) |[u]_i| - \beta)_+$. Then, for problem (P$^*$), the followings hold:
(i) The dual problem of \((P^*)\) is
\[
\min_{\theta \in [0,1]^n} \mathcal{D}(\theta; \alpha, \beta) = \frac{1}{2\alpha} \left\| \mathbf{S}_\beta \left( \frac{1}{n} \mathbf{X}_\theta \right) \right\|^2 + \frac{\gamma}{2n} \|\theta\|^2
\]
\[= \frac{1}{n} (1, \theta), \tag{D*} \]
where \(1 \in \mathbb{R}^n\) is a vector with all components equal to 1.

(ii) Denote the optima of \((P^*)\) and \((D^*)\) by \(\mathbf{w}^*(\alpha, \beta)\) and \(\theta^*(\alpha, \beta)\), respectively. Then,
\[
\mathbf{w}^*(\alpha, \beta) = \frac{1}{\alpha} \mathbf{S}_\beta \left( \frac{1}{n} \mathbf{X} \mathbf{w}^*(\alpha, \beta) \right) \tag{KKT-1},
\]
\[
[\theta^*(\alpha, \beta)]_i = \begin{cases} 
0, & \text{if } 1 - \langle \mathbf{x}_i, \mathbf{w}^*(\alpha, \beta) \rangle < 0; \\
1, & \text{if } 1 - \langle \mathbf{x}_i, \mathbf{w}^*(\alpha, \beta) \rangle > \gamma; \\
\frac{1}{n} (1 - \langle \mathbf{x}_i, \mathbf{w}^*(\alpha, \beta) \rangle), & \text{otherwise}. \tag{KKT-2}
\end{cases}
\]

According to KKT-1 and KKT-2, we define 4 index sets:
\[
\mathcal{F} = \left\{ j \in [p] : \frac{1}{n} \mathbf{X} \mathbf{w}^*(\alpha, \beta)_j \leq 0 \right\},
\]
\[
\mathcal{R} = \left\{ i \in [n] : 1 - \langle \mathbf{w}^*(\alpha, \beta), \mathbf{x}_i \rangle < 0 \right\},
\]
\[
\mathcal{E} = \left\{ i \in [n] : 1 - \langle \mathbf{w}^*(\alpha, \beta), \mathbf{x}_i \rangle \in [0, \gamma] \right\},
\]
\[
\mathcal{L} = \left\{ i \in [n] : 1 - \langle \mathbf{w}^*(\alpha, \beta), \mathbf{x}_i \rangle > \gamma \right\},
\]

which imply that
\[
(i) : i \in \mathcal{F} \Rightarrow [\mathbf{w}^*(\alpha, \beta)]_i = 0,
\]
\[
(ii) : \begin{cases} 
i \in \mathcal{R} \Rightarrow [\theta^*(\alpha, \beta)]_i = 0; \\
i \in \mathcal{L} \Rightarrow [\theta^*(\alpha, \beta)]_i = 1. \tag{R} \end{cases}
\]

Thus, we call the \(j^{th}\) feature inactive if \(j \in \mathcal{F}\). The samples in \(\mathcal{L}\) are the so-called support vectors and we call the samples in \(\mathcal{R}\) and \(\mathcal{L}\) inactive samples.

Suppose that we are given subsets of \(\mathcal{F}\), \(\mathcal{R}\), and \(\mathcal{L}\), then by (R), we can see that many coefficients of \(\mathbf{w}^*(\alpha, \beta)\) and \(\theta^*(\alpha, \beta)\) are known. Thus, we may have much less unknowns to solve and the problem size can be dramatically reduced. We formalize this idea in Lemma 1.

**Lemma 1.** Given index sets \(\bar{\mathcal{F}} \subseteq \mathcal{F}\), \(\bar{\mathcal{R}} \subseteq \mathcal{R}\), and \(\bar{\mathcal{L}} \subseteq \mathcal{L}\), the followings hold
\[
(i) : [\mathbf{w}^*(\alpha, \beta)]_{\bar{\mathcal{F}}} = 0, [\theta^*(\alpha, \beta)]_{\bar{\mathcal{R}}} = 0, [\theta^*(\alpha, \beta)]_{\bar{\mathcal{L}}} = 1.
\]
\[
(ii) : \text{Let } \bar{\mathcal{D}} = \mathcal{D} \cup \bar{\mathcal{L}}, \bar{\mathcal{G}}_1 = \bar{\mathcal{F}} \bar{\mathcal{F}} + \bar{\mathcal{L}} \bar{\mathcal{L}}, \text{and } \bar{\mathcal{G}}_2 = \bar{\mathcal{F}} \bar{\mathcal{L}}, \text{where } \bar{\mathcal{D}} = [p] \setminus \bar{\mathcal{F}}, \bar{\mathcal{D}} = [n] \setminus \bar{\mathcal{D}}, \text{and } \bar{\mathcal{D}} = [n] \setminus \bar{\mathcal{L}}. \text{Then,}\]
\[
[\theta^*(\alpha, \beta)]_{\bar{\mathcal{D}}} \text{ solves the following scaled dual problem:}
\]
\[
\min_{\theta \in [0,1]^{\bar{\mathcal{D}}}} \left\{ \frac{1}{2\alpha} \left\| \mathbf{S}_\beta \left( \frac{1}{n} \bar{\mathcal{G}}_1 \hat{\theta} + \frac{1}{n} \bar{\mathcal{G}}_2 \right) \right\|^2 + \frac{\gamma}{2n} \|\hat{\theta}\|^2 \right. \]
\[= \frac{1}{n} (1, \hat{\theta}), \text{ (scaled-D*)} \]

Lemma 1 indicates that, if we can identify index sets \(\bar{\mathcal{F}}\) and \(\bar{\mathcal{D}}\) and the cardinalities of \(\bar{\mathcal{F}}\) and \(\bar{\mathcal{D}}\) are much smaller than the feature dimension \(p\) and the dataset size \(n\), we only need to solve a problem (scaled-D*) that may be much smaller than problem (D*) to exactly recover the optima \(\mathbf{w}^*(\alpha, \beta)\) and \(\theta^*(\alpha, \beta)\) without sacrificing any accuracy.

However, we cannot directly apply the rules in (R) to identify subsets of \(\mathcal{F}\), \(\mathcal{R}\), and \(\mathcal{L}\), as they require the knowledge of \(\mathbf{w}^*(\alpha, \beta)\) and \(\theta^*(\alpha, \beta)\) that are usually unavailable.

Inspired by the idea in (El Ghaoui et al., 2012), we can first estimate regions \(\mathcal{W}\) and \(\Theta\) that contain \(\mathbf{w}^*(\alpha, \beta)\) and \(\theta^*(\alpha, \beta)\), respectively. Then, by denoting
\[
\hat{\mathcal{F}} := \left\{ j \in [p] : \max_{\theta \in \Theta} \left\{ \frac{1}{n} \mathbf{X} \mathbf{w} \right\}_j \leq \beta \right\}, \tag{1}
\]
\[
\hat{\mathcal{R}} := \left\{ i \in [n] : \max_{\mathbf{w} \in \mathcal{W}} (1 - \mathbf{x}_i \theta) < 0 \right\}; \tag{2}
\]
\[
\hat{\mathcal{L}} := \left\{ i \in [n] : \min_{\mathbf{w} \in \mathcal{W}} (1 - \mathbf{x}_i \theta) > \gamma \right\}, \tag{3}
\]

since it is easy to know that \(\hat{\mathcal{F}} \subseteq \mathcal{F}, \hat{\mathcal{R}} \subseteq \mathcal{R}, \text{ and } \hat{\mathcal{L}} \subseteq \mathcal{L}\), the rules in (R) can be relaxed as follows:
\[
(i) : j \in \hat{\mathcal{F}} \Rightarrow [\mathbf{w}^*(\alpha, \beta)]_j = 0, \tag{R1}
\]
\[
(ii) : \begin{cases} 
i \in \hat{\mathcal{R}} \Rightarrow [\theta^*(\alpha, \beta)]_i = 0; \\
i \in \hat{\mathcal{L}} \Rightarrow [\theta^*(\alpha, \beta)]_i = 1. \tag{R2} \end{cases}
\]

In view of R1 and R2, we sketch the development of SIFS as follows.

**Step 1:** Derive estimations \(\mathcal{W}\) and \(\Theta\) such that \(\mathbf{w}^*(\alpha, \beta) \in \mathcal{W}\) and \(\theta^*(\alpha, \beta) \in \Theta\), respectively.

**Step 2:** Develop SIFS by deriving the relaxed screening rules R1 and R2, i.e., by solving the optimization problems in Eq. (1), Eq. (2) and Eq. (3).

### 3. Estimate the Primal and Dual Optima

In this section, we first show that the primal and dual optima admit closed form solutions for specific values of \(\alpha\) and \(\beta\) (see Section 3.1). Then, in Sections 3.2 and 3.3, we present accurate estimates of the primal and dual optima, respectively.

#### 3.1. Effective Intervals of the Parameters \(\alpha\) and \(\beta\)

We first show that, if the value of \(\beta\) is sufficiently large, no matter what \(\alpha\) is, the primal solution is 0.
Theorem 2. Let $\beta_{\text{max}} = \|\frac{1}{n}X1\|_{\infty}$. Then, for $\alpha > 0$ and $\beta \geq \beta_{\text{max}}$, we have

$$w^*(\alpha, \beta) = 0, \quad \theta^*(\alpha, \beta) = 1.$$ 

For any $\beta$, the next result shows that, if $\alpha$ is large enough, the primal and dual optima admit closed form solutions.

Theorem 3. If we denote

$$\alpha_{\text{max}}(\beta) = \frac{1}{1 - \gamma} \max_{i \in [n]} \{\langle x_i, S_\beta(\frac{1}{n}X1) \rangle\},$$

then for all $\alpha \in [\alpha_{\text{max}}(\beta), 0), \infty) \cap (0, \infty)$, we have

$$w^*(\alpha, \beta) = \frac{1}{\alpha} S_\beta(\frac{1}{n}X1), \quad \theta^*(\alpha, \beta) = 1.$$ 

By Theorems 2 and 3, we only need to consider the cases with $\beta \in (0, \beta_{\text{max}}]$ and $\alpha \in (0, \alpha_{\text{max}}(\beta)]$.

3.2. Primal Optimum Estimation

In Section 1, we mention that the proposed SIFS consists of IFS and ISS, and an alternating application of IFS and ISS can improve the estimation of the primal and dual optima, which can in turn make ISS and IFS more effective in identifying inactive samples and features, respectively. Lemma 2 shows that discarding inactive features by IFS leads to a more accurate estimation of the primal optimum.

Lemma 2. Suppose that the reference solution $w^*(\alpha_0, \beta_0)$ with $\beta_0 \in (0, \beta_{\text{max}}]$ and $\alpha_0 \in (0, \alpha_{\text{max}}(\beta_0)]$ is known. Consider problem $\mathcal{P}^*$ with parameters $\alpha > 0$ and $\beta_0$. Let $\bar{\mathcal{F}}$ be the index set of the inactive features identified by the previous IFS steps, i.e., $\{w^*(\alpha_0, \beta_0)\}_{\bar{\mathcal{F}}} = 0$. We define

$$c = \frac{\alpha_0 + \alpha}{2\alpha} w^*(\alpha_0, \beta_0), \quad r^2 = \frac{(\alpha_0 - \alpha)^2}{4\alpha^2} \|w^*(\alpha_0, \beta_0)\|^2$$

Then, the following holds:

$$\{w^*(\alpha, \beta_0)\}_{\bar{\mathcal{F}}} \in \mathcal{W} := \{w : \|w - c\| \leq r\}.$$ 

As $\bar{\mathcal{F}}$ is the index set of identified inactive features, we have $\{w^*(\alpha, \beta_0)\}_{\bar{\mathcal{F}}} = 0$. Hence, we only need to find an accurate estimation of $\{w^*(\alpha, \beta_0)\}_{\bar{\mathcal{F}}}$. Lemma 2 shows that $\{w^*(\alpha, \beta_0)\}_{\bar{\mathcal{F}}}$ lies in a ball of radius $r$ centered at $c$. Note that, before we perform ISS, the set $\bar{\mathcal{F}}$ is empty and thus the second term on the right hand side (RHS) of Eq. (6) is 0. If we apply ISS multiple times (alternating with ISS), the set $\bar{\mathcal{F}}$ will be monotonically increasing. Thus, Eq. (6) implies that the radius will be monotonically decreasing, leading to a more accurate primal optimum estimation.

3.3. Dual Optimum Estimation

Similar to Lemma 2, the next result shows that ISS can improve the estimation of the dual optimum.

Lemma 3. Suppose that the reference solution $\theta^*(\alpha_0, \beta_0)$ with $\beta_0 \in (0, \beta_{\text{max}}]$ and $\alpha_0 \in (0, \alpha_{\text{max}}(\beta_0)]$ is known. Consider problem $\mathcal{P}^*$ with parameters $\alpha > 0$ and $\beta_0$. Let $\mathcal{R}$ and $\mathcal{L}$ be the index sets of inactive samples identified by the previous ISS steps, i.e., $\{\theta^*(\alpha_0, \beta_0)\}_{\mathcal{R}} = 0$, $\{\theta^*(\alpha_0, \beta_0)\}_{\mathcal{L}} = 1$, and $\mathcal{D} = \mathcal{R} \cup \mathcal{L}$. We define

$$c = \frac{\alpha - \alpha_0}{2\alpha \gamma} 1 + \frac{\alpha_0 + \alpha}{2\alpha} \|\theta^*(\alpha_0, \beta_0)\|_{\mathcal{P}^*}, \quad r^2 = \frac{(\alpha_0 - \alpha)^2}{4\alpha^2} \|\theta^*(\alpha_0, \beta_0) - \frac{1}{\gamma} 1\|^2$$

Then, the following holds:

$$\{\theta^*(\alpha, \beta_0)\}_{\mathcal{P}^*} \in \Theta := \{\theta : \|\theta - c\| \leq r\}.$$ 

4. The Proposed SIFS Screening Rule

We first present the IFS and ISS rules in Sections 4.1 and 4.2, respectively. Then, in Section 4.3, we develop the SIFS screening rule by an alternating application of IFS and ISS.

4.1. Inactive Feature Screening (IFS)

Suppose that $w^*(\alpha_0, \beta_0)$ and $\theta^*(\alpha_0, \beta_0)$ are known, we derive IFS to identify inactive features for problem $\mathcal{P}^*$ at $(\alpha, \beta_0)$ by solving the optimization problem in Eq. (1) (see Section E in the supplementary material):

$$s^i(\alpha, \beta_0) = \max_{\theta \in \Theta} \left\{ \frac{1}{n} \langle \hat{x}_i, \theta \rangle + \langle |x_i|_{\mathcal{P}^*}, 1 \rangle \right\}, i \in \mathcal{F}^c,$$

(11)
where $\Theta$ is given by Eq. (10) and $\hat{F}$ and $\hat{D} = \hat{R} \cup \hat{L}$ are the index sets of inactive features and samples that have been identified in previous screening processes, respectively. The next result shows the closed form solution of problem (11).

**Lemma 4.** Consider problem (11). Let $c$ and $r$ be given by Eq. (8) and Eq. (9). Then, for all $i \in \hat{F}^c$, we have

$$s^i(\alpha, \beta_0) = \frac{1}{n} (||\langle [x^i]_D, c \rangle + \langle [x^i]_L, I \rangle| + ||[x^i]_D||_2 r).$$

We are now ready to present the IFS rule.

**Theorem 4.** Consider problem $(P^*)$. We suppose that $w^*(\alpha_0, \beta_0)$ and $\theta^*(\alpha_0, \beta_0)$ are known. Then,

(1): The feature screening rule IFS takes the form of

$$s^i(\alpha, \beta_0) \leq \beta_0 \Rightarrow [w^*(\alpha, \beta_0)]_i = 0, \forall i \in \hat{F}^c \quad (\text{IFS})$$

(2): We update the index set $\hat{F}$ by

$$\hat{F} \leftarrow \hat{F} \cup \{ i : s^i \leq \beta_0, i \in \hat{F}^c \}. \quad (12)$$

Recall that (Lemma 3), previous sample screening results give us a more tighter dual estimation, i.e., a smaller feasible region $\Theta$ for problem (11), which results in a smaller $s^i(\alpha, \beta_0)$. It finally leads us to a more powerful feature screening rule IFS. This is the so called synergy effect.

### 4.2. Inactive Sample Screening (ISS)

Similar to IFS, we derive ISS to identify inactive samples by solving the optimization problems in Eq. (2) and Eq. (3) (see Section G in the supplementary material for details):

$$u_i(\alpha, \beta_0) = \max_{w \in \mathcal{W}} \{ 1 - \langle [x^i]_{\bar{x}^c}, w \rangle \}, i \in \hat{D}^c, \quad (13)$$

$$l_i(\alpha, \beta_0) = \min_{w \in w^*} \{ 1 - \langle [x^i]_{\bar{x}^c}, w \rangle \}, i \in \hat{D}^c, \quad (14)$$

where $\mathcal{W}$ is given by Eq. (7) and $\hat{F}$ and $\hat{D} = \hat{R} \cup \hat{L}$ are the index sets of inactive features and samples that have been identified in previous screening processes. We show that problems (13) and (14) admit closed form solutions.

**Lemma 5.** Consider problems (13) and (14). Let $c$ and $r$ be given by Eq. (5) and Eq. (6). Then,

$$u_i(\alpha, \beta_0) = 1 - \langle [x^i]_{\bar{x}^c}, c \rangle + ||[x^i]_{\bar{x}^c}||_2 r, i \in \hat{D}^c,$$

$$l_i(\alpha, \beta_0) = 1 - \langle [x^i]_{\bar{x}^c}, c \rangle - ||[x^i]_{\bar{x}^c}||_2 r, i \in \hat{D}^c.$$

We are now ready to present the ISS rule.

**Theorem 5.** Consider problem $(D^*)$. We suppose that $w^*(\alpha_0, \beta_0)$ and $\theta^*(\alpha_0, \beta_0)$ are known. Then,

(1): The sampling screening rule ISS takes the form of

$$u_i(\alpha, \beta_0) < 0 \Rightarrow [\theta^*(\alpha, \beta_0)]_i = 0,$$

$$l_i(\alpha, \beta_0) > \gamma \Rightarrow [\theta^*(\alpha, \beta_0)]_i = 1, \forall i \in \hat{D}^c \quad (\text{ISS})$$

(2): We update the the index sets $\hat{R}$ and $\hat{L}$ by

$$\hat{R} \leftarrow \hat{R} \cup \{ i : u_i(\alpha, \beta_0) < 0, i \in \hat{D}^c \},$$

$$\hat{L} \leftarrow \hat{L} \cup \{ i : l_i(\alpha, \beta_0) > \gamma, i \in \hat{D}^c \}. \quad (15), (16)$$

The synergy effect also exists here. Recall that (Lemma 2), previous feature screening results lead a smaller feasible region $\mathcal{W}$ for the problems (13) and (14), which results in smaller $u_i(\alpha, \beta_0)$ and bigger $l_i(\alpha, \beta_0)$. It finally leads us to a more accurate sample screening rule ISS.

### 4.3. The Proposed SIFS Rule by An Alternating Application of IFS and ISS

In real applications, the optimal parameter values of $\alpha$ and $\beta$ are usually unknown. To determine appropriate parameter values, common approaches, like cross validation and stability selection, need to solve the model over a grid of parameter values $\{(\alpha_{ij}, \beta_j) : i \in [M], j \in [N]\}$ with $\beta_{max} > \beta > ... > \beta_0 > 0$ and $\alpha_{max}(\beta_j) > \alpha_{i,j} > ... > \alpha_{M,j} > 0$. This can be very time-consuming. Inspired by Strong Rule (Tibshirani et al., 2012) and SAFE (El Ghaoui et al., 2012), we develop a sequential version of SIFS in Algorithm 1. Specifically, given the primal and dual opti...
at \((\alpha_{ij}, \beta_j)\). We repeat this process until we solve problem \((P^*)\) at all pairs of parameter values.

Note that we insert \(\alpha_{ij}\) into every sequence \(\{\alpha_{ij} : i \in [M]\}\) (see line 1 in Algorithm 1) to obtain a closed-form solution as the first reference solution. In this way, we can avoid solving problem at \((\alpha_{ij}, \beta_j)\), \(j \in [N]\) directly (without screening), which is time consuming. At last, we would like to point out that the values \(\{(\alpha_{ij}, \beta_j) : i \in [M], j \in [N]\}\) in SIFS can be specified by users arbitrarily.

SIFS applies ISS and IFS in an alternating manner to re-inforce their capability in identifying inactive samples and features. In Algorithm 1, we apply ISS first. Of course, we can also apply IFS first. The theorem below demonstrates that the orders have no impact on the performance of SIFS.

**Theorem 6.** Given the optimal solutions \(w^*(\alpha_{i-1,j}, \beta_j)\) and \(\theta^*(\alpha_{i-1,j}, \beta_j)\) at \((\alpha_{i-1,j}, \beta_j)\) for SIFS, we assume SIFS with ISS first stops after applying ISS and IFS for \(p\) times and denote the identified inactive features and samples as \(F^A_p, R^A_p\) and \(L^A_p\). Similarly, when we apply IFS first, the results are denoted as \(F^B_q, R^B_q\) and \(L^B_q\). Then, the followings hold:

1. With different orders of applying ISS and IFS, the difference of the times of ISS and IFS we need to apply in SIFS can never be larger than 1, that is, \(|p - q| \leq 1.0\).

2. Remark 2. From Remark 1, we can see that our SIFS can also be applied to solve a single problem, due to the existence of the free reference solution pair.

**5. Experiments**

We evaluate SIFS on both synthetic and real datasets in terms of three measurements. The first one is the scaling ratio: \(1 - \frac{(n - \tilde{n})(p - \tilde{p})}{np}\), where \(\tilde{n}\), \(\tilde{p}\), \(n\), and \(p\) are the numbers of inactive samples and features identified by SIFS, sample size, and feature dimension of the datasets. The second measure is rejection ratios of each triggering of ISS and IFS in SIFS: \(\frac{n_i}{n_0}\) and \(\frac{p_i}{p_0}\), where \(n_i\) and \(p_i\) are the numbers of inactive samples and features identified in \(i\)-th triggering of ISS and IFS in SIFS. \(n_0\) and \(p_0\) are the numbers of inactive samples and features in the solution. The third measure is speedup, i.e., the ratio of the running time of the solver without screening to that with screening.

Recall that, we can integrate SIFS with any solvers for problem \((P^*)\). In this experiment, we use Accelerated Proximal Stochastic Dual Coordinate Ascent (Accelerated-Prox-SDCA) (Shalev-Shwartz & Zhang, 2016), as it is one of the state-of-the-arts. As we mentioned in the introduction section that screening differs greatly from features selection methods, it is not appropriate to make comparisons with feature selection methods. To this end, we only choose the state-of-art screening method for Sparse SVMs in (Shibagaki et al., 2016) as a baseline in the experiments.

For each dataset, we solve problem \((P^*)\) at a grid of turning parameter values. Specifically, we first compute \(\beta_{\max}\) by Theorem 2 and then select 10 values of \(\beta\) that are equally spaced on the logarithmic scale of \(\beta/\beta_{\max}\) from 1 to 0.05. Then, for each value of \(\beta\), we first compute \(\alpha_{\max}(\beta)\) by Theorem 3 and then select 100 values of \(\alpha\) that are equally spaced on the logarithmic scale of \(\alpha/\alpha_{\max}(\beta)\) from 1 to 0.01. Thus, for each dataset, we solve problem \((P^*)\) at 1000 pairs of parameter values in total. We write the code in C++ along with Eigen library for some numerical computations. We perform all the computations on a single core of Intel(R) Core(TM) i7-5930K 3.50GHz, 128GB MEM.

**5.1. Simulation Studies**

We evaluate SIFS on 3 synthetic datasets named syn1, syn2 and syn3 with sample and feature size \((n, p) \in \{(1000, 1000), (10000, 10000), (1000, 10000)\}\). We present each data point as \(x = [x_1; x_2]\) with \(x_1 \in \mathbb{R}^{0.02p}\) and \(x_2 \in \mathbb{R}^{0.98p}\). We use Gaussian distributions \(G_1 = N(u, 0.75I), G_2 = N(-u, 0.75I)\) and \(G_3 = N(0, 1)\) to generate the data points, where \(u = 1.5I\) and \(I \in \mathbb{R}^{0.02p \times 0.02p}\) is the identity matrix. To be precise, \(x_1\) for positive and negative points are sampled from \(G_1\) and \(G_2\), respectively. For each entry in \(x_2\), it has chance \(\eta = 0.02\) to be sampled from \(G_3\) and chance \(1 - \eta\) to be 0.

![Figure 1](image-url) Scaling ratios of ISS, IFS, and SIFS (from left to right).
Table 1 reports the running time (in seconds) for solving problem (P\(^*\)) at 1000 pairs of parameter values on three synthetic datasets.

Recall that, SIFS detects the inactive features and samples in a static manner, i.e., we perform SIFS only once before the optimization and thus the size of the problem we need to perform optimization on is fixed. However, the method in (Shibagaki et al., 2016) detects inactive features and samples in a dynamic manner (Bonnefoy et al., 2014), i.e., they perform their method along with the optimization and thus the size of the problem would keep decreasing during the iterative process. Thus, comparing SIFS with the method in (Shibagaki et al., 2016) in terms of rejection ratios is inapplicable. We compare the performance of SIFS with the method in (Shibagaki et al., 2016) in terms of speedup. Specifically, we compare the speedup gained by SIFS and the method in (Shibagaki et al., 2016) for solving problem (P\(^*\)) at 1000 pairs of parameter values. The code of the method in (Shibagaki et al., 2016) is obtained from (https://github.com/husk214/s3fs).
Fig. 3 shows the rejection ratios of SIFS on the real-sim dataset (other results are in the supplementary material). In Fig. 3, we can see that some inactive features and samples are identified in the 2nd and 3rd triggering of ISS and IFS, which verifies the necessity of the alternating application of ISS and IFS. SIFS is efficient since it always stops in 3 times of triggering. In addition, most of (> 98%) the inactive features can be identified in the 1st triggering of IFS while identifying inactive samples needs to apply ISS two or more times. It may result from two reasons: 1) We run ISS first, which reinforces the capability of IFS due to the synergy effect (see Sections 4.1 and 4.2), see Section L.1 in the supplementary material for further verification; 2) Feature screening here may be easier than sample screening.

Table 3 reports the running time of solver without and with the method in (Shibagaki et al., 2016) and SIFS for solving problem (P*) at 1000 pairs of parameter values on five real datasets. The speedup gained by SIFS is up to 300 times faster than the solver with the method in (Shibagaki et al., 2016) on both datasets url and kddb. Take the dataset kddb as an example. The solver with SIFS takes about 13 hours to solve problem (P*) for all 1000 pairs of parameter values, while the solver with the method in (Shibagaki et al., 2016) needs 11 days to finish the same task.

6. Conclusion

In this paper, we develop a novel data reduction method SIFS to simultaneously identify inactive features and samples for sparse SVM. Our major contribution is a novel framework for an accurate estimation of the primal and dual optima based on strong convexity. To the best of our knowledge, the proposed SIFS is the first static screening method that is able to simultaneously identify inactive features and samples for sparse SVMs. An appealing feature of SIFS is that all detected features and samples are guaranteed to be irrelevant to the outputs. Thus, the model learned on the reduced data is identical to the one learned on the full data. Experiments on both synthetic and real datasets demonstrate that SIFS can dramatically reduce the problem size and the resulting speedup can be orders of magnitude. We plan to generalize SIFS to more complicated models, e.g., SVM with a structured sparsity-inducing penalty.
Acknowledgements

This work was supported by the National Basic Research Program of China (973 Program) under Grant 2013CB336500, National Natural Science Foundation of China under Grant 61233011 and National Youth Top-notch Talent Support Program.

References


