# Simultaneous Safe Screening of Features and Samples in Doubly Sparse Modeling

**Atsushi Shibagaki** [†] SHIBAGAKI.A.MLLAB.NIT@GMAIL.COM
**Masayuki Karasuyama** [†] KARASUYAMA@NITECH.AC.JP
**Kohei Hatano** [‡] HATANO@INF.KYUSHU-U.AC.JP
**Ichiro Takeuchi** [†] TAKEUCHI.ICHIRO@NITECH.AC.JP
[†] Nagoya Institute of Technology, Nagoya, 466-8555, Japan
[‡] Kyushu University, Fukuoka, 819-0395, Japan

## Abstract

The problem of learning a sparse model is conceptually interpreted as the process of identifying *active* features/samples and then optimizing the model over them. Recently introduced *safe screening* allows us to identify a part of non-active features/samples. So far, safe screening has been individually studied either for feature screening or for sample screening. In this paper, we introduce a new approach for safely screening features and samples *simultaneously* by alternatively iterating feature and sample screening steps. A significant advantage of considering them simultaneously rather than individually is that they have a *synergy* effect in the sense that the results of the previous safe feature screening can be exploited for improving the next safe sample screening performances, and vice-versa. We first theoretically investigate the synergy effect, and then illustrate the practical advantage through intensive numerical experiments for problems with large numbers of features and samples.

## 1. Introduction

In many areas of science and industry, large-scale datasets with many features and samples are collected and analyzed for data-driven scientific discovery and decision making. One promising approach to handle large numbers of features and samples is introducing *sparsity* constraints in statistical models (Hastie et al., 2015). The most common approach for inducing *feature sparsity*, e.g., in a linear least-square model fitting, is using sparsity-inducing penalties such as $L_1$-norm of the coefficients (Tibshirani, 1996). A feature sparse model depends only on a subset of features (called *active* features), and the rest of the features (called *non-active* features) are irrelevant. On the other hand, the most popular machine learning algorithm that induces *sample sparsity* would be the support vector machine (SVM) (Boser et al., 1992). In the SVM, the large margin principle enhances sample sparsity in the sense that it depends only on a subset of samples (called *support vectors (SVs)* or *active* samples) and does not depend on the rest of the samples (called *non-SVs* or *non-active* samples).

The problem of learning a sparse model is conceptually interpreted as the process of identifying active features or samples and then optimizing the model over them. Recently, a new approach called *safe screening* has been studied by several authors. Safe screening enables to identify a subset of non-active features or samples before or during the model training process. A nice thing about safe screening is that it is guaranteed to have no false negatives, i.e., safe screening never identify active features or samples as non-active. It means that, if we train the model by only using the remaining set of features or samples after safe screening, the solution is guaranteed to be optimal. The basic technical idea behind safe screening is to bound the solution of the problem within a region, and show that some features or samples cannot be active wherever the optimal solution is located within the region.

After the seminal work by (El Ghaoui et al., 2012), *safe feature screening* (safely screening a part of non-active features in sparse feature models such as LASSO) has been intensively studied in the literature (Xiang et al., 2011; Wang et al., 2013; Bonnefoy et al., 2014; Liu et al., 2014; Wang et al., 2014b; Xiang et al., 2014; Fercoq et al., 2015; Ndiaye et al., 2015). Safe feature screening exploits the fact that the sparseness of a feature is characterized by a property of the dual solution, i.e., if the dual solution satis-

*Proceedings of the $33^{rd}$ International Conference on Machine Learning*, New York, NY, USA, 2016. JMLR: W&CP volume 48. Copyright 2016 by the author(s).

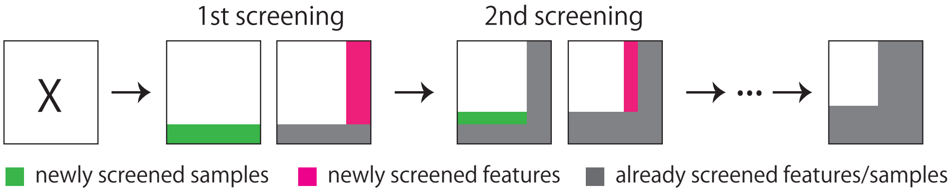

*Figure 1.* Schematic illustration of the proposed approach. By iterating safe feature screening and safe sample screening, irrelevant features and samples are safely removed out from the training set. A significant advantage of considering them simultaneously rather than individually is that they have a *synergy* effect in the sense that the screening performances (the number of features/samples that can be safely removed out) are gradually improved by exploiting the results in earlier steps.

fies a certain condition in the dual space, the feature is guaranteed to be non-active. Safe feature screening is beneficial especially when the number of features is large. There are also several studies on *safe sample screening* (safely screening a part of non-active samples in sparse sample models such as the SVM) (Ogawa et al., 2013; Wang et al., 2014a; Zimmert et al., 2015). The basic idea behind safe sample screening is that the sparseness of a sample is characterized by a property of the primal solution. If the primal solution satisfies a certain condition in the primal space, the sample is guaranteed to be non-active. Safe sample screening is useful when the number of samples is large.

In this paper, we consider problems where the numbers of features and samples are both large. In these problems, we consider a class of learning algorithms that induce both of feature sparsity and sample sparsity, which we call *doubly sparse modeling*. Our main contribution in this paper is to develop a safe screening method that can identify both of non-active features and non-active samples simultaneously in doubly sparse modeling. Specifically, we propose a novel method for simultaneously constructing two regions, one in the dual space and the other in the primal space. The former is used for safe feature screening, while the latter is used for safe sample screening.

A significant advantage of considering safe feature screening and safe sample screening simultaneously rather than individually is that they have a *synergy* effect. Specifically, we show that, after we know that a part of features are non-active based on safe feature screening, we can potentially improve the performance of safe sample screening. Our basic idea behind this property is that, by fixing a part of the primal variables, the region in the primal space, which is used for safe sample screening, can be made smaller (we can also show the converse similarly). These findings suggest that, safe screening performances of features and samples can be both improved by alternatively repeating them. Figure 1 is a schematic illustration of the simultaneous safe screening approach.

Another interesting finding we first introduce in this paper is that, by simultaneously considering regions both in the dual and the primal spaces, we can also identify features and samples that are guaranteed to be active. We call this technique as *safe keeping*. While safe screening assures no false negative findings, safe keeping guarantees no false positive findings of active features/samples. By combining these two techniques, we can better identify active features/samples. A practical advantage of safe keeping is that we do not have to consider the safe screening rules anymore for features and samples which are identified as active by safe keeping. This is helpful for reducing the computational costs of safe screening rule evaluations especially in the context of dynamic safe screening (Bonnefoy et al., 2014).

**Notation:** For any natural number $n$, we define $[n] := \{1, \ldots, n\}$. For an $n \times d$ matrix $M$, its $i$-th row and $j$-th column are denoted as $M_{i:}$ and $M_{:j}$, respectively, for $i \in [n]$ and $j \in [d]$. The $L_1$ norm, the $L_2$ norm and the $L_\infty$ norm of a vector $v \in \mathbb{R}^m$ are respectively written as $\|v\|_1 := \sum_{k\in[m]} |v_k|$, $\|v\|_2 := \sqrt{\sum_{k\in[m]} |v_k|^2}$ and $\|v\|_\infty := \max_{k\in[m]} |v_k|$. For a scalar $z$, we define $[z]_+ := \max\{0, z\}$. We write the subdifferential operator as $\partial$, and remind that the subdifferential of $L_1$ norm is given as $\partial\|v\|_1 = \{g \mid \|g\|_\infty \leq 1, g^\top v = \|v\|_1\}$. For a function $f$, we denote its domain as $\mathrm{dom} f$.

## 2. Preliminaries

In this section, we first describe the problem formulation in §2.1. Then, we briefly summarize the basic concepts of safe feature screening and safe sample screening in §2.2 and §2.3, respectively.

### 2.1. Problem formulation

Consider classification and regression problems with the number of samples $n$ and the number of features $d$. The training set is written as $\{(x_i, y_i)\}_{i\in[n]}$ where $x_i \in \mathbb{R}^d$, and $y_i \in \{-1, 1\}$ for classification and $y_i \in \mathbb{R}$ for regression. The $n \times d$ input data matrix (design matrix) is denoted as $X := [x_1, \ldots, x_n]^\top$.

We consider a linear classification and regression function in the form of $f(x) = x^\top w$, and study the problem of estimating the parameter $w \in \mathbb{R}^d$ by solving a class of regularized empirical risk minimization problems:

$$\min_{w \in \mathbb{R}^d} P_\lambda(w) := \lambda \psi(w) + \frac{1}{n} \sum_{i \in [n]} \ell_i(x_i^\top w), \tag{1}$$

where $\psi$ is a penalty function, $\ell_i$ is a loss function for the $i$-th sample[1], and $\lambda > 0$ is a trade-off parameter for controlling the balance between the penalty and the loss.

In this paper, we study doubly sparse modeling, i.e., a pair of penalty and loss function that induces sparsities both in features and samples. As specific working examples, we consider $L_1$-penalized smoothed hinge SV classification and $L_1$-penalized smoothed $\varepsilon$-insensitive SV regression. The use of these smoothed loss functions are known to produce almost same solutions as the original hinge or $\varepsilon$-insensitive loss functions (see e.g., Shalev-Shwartz & Zhang, 2015). We will discuss other doubly sparse modeling problem in §5.

Remembering that the original SVMs are trained with $L_2$ penalty, by combining an additional $L_1$-penalty, our penalty function $\psi$ is written as *elastic net penalty*:

$$\psi(w) := \|w\|_1 + \frac{\beta}{2} \|w\|_2^2, \tag{2}$$

where $\beta > 0$ is a balancing parameter which we omit hereafter by substituting $\beta = 1$ for notational simplicity. The smoothed hinge loss and the smoothed $\varepsilon$-insensitive loss are respectively written as

$$\ell_i(x_i^\top w) := \begin{cases} 0 & (y_i x_i^\top w > 1), \\ 1 - y_i x_i^\top w - \frac{\gamma}{2} & (y_i x_i^\top w < 1 - \gamma), \\ \frac{1}{2\gamma}(1 - y_i x_i^\top w)^2 & (\text{otherwise}), \end{cases} \tag{3}$$

$$\ell_i(x_i^\top w) := \begin{cases} 0 & (|x_i^\top w - y_i| < \varepsilon), \\ |x_i^\top w - y_i| - \varepsilon - \frac{\gamma}{2} & (|x_i^\top w - y_i| > \varepsilon + \gamma), \\ \frac{1}{2\gamma}(|x_i^\top w - y_i| - \varepsilon)^2 & (\text{otherwise}), \end{cases} \tag{4}$$

where $\gamma > 0$ is a tuning parameter.

Using Fenchel's duality theorem (see, e.g., Corollary 31.2.1 in (Rockafellar, 1970)), the dual problem of (1) is written as

$$\max_{\alpha} D_\lambda(\alpha) := -\lambda \psi^* \left( \frac{1}{\lambda n} X^\top \alpha \right) - \frac{1}{n} \sum_{i \in [n]} \ell_i^*(-\alpha_i), \tag{5}$$

where $\psi^*$ and $\ell^*$ is convex conjugate function of $\psi$ and $\ell$, respectively. The convex conjugate function of the penalty term in (2) is given as $\psi^*(v) = \frac{1}{2} \sum_{j=1}^d ([|v_j| - 1]_+)^2$. The convex conjugate functions of the smoothed hinge loss and the smoothed $\varepsilon$-insensitive loss are respectively written as $\ell_i^*(\alpha_i) = \frac{\gamma}{2}\alpha_i^2 + y_i \alpha_i$ for $y_i \alpha_i \in [-1, 0]$ and $\infty$ otherwise, and $\ell_i^*(\alpha_i) = \frac{\gamma}{2}\alpha_i^2 + y_i \alpha_i + \varepsilon |\alpha_i|$ for $\alpha_i \in [-1, 1]$ and $\infty$ otherwise. We call the problems in (1) and (5) as *primal problem* and *dual problem*, respectively, and denote the primal optimal solution as $w^* \in \mathbb{R}^d$ and the dual optimal solution as $\alpha^* \in \mathbb{R}^n$.

### 2.2. Safe feature screening

The goal of safe feature screening is to identify a part of non-active features $\{j \in [d] \mid w_j^* = 0\}$ before or during the optimization process. Safe feature screening is built on the following KKT optimality condition (see Theorem 31.3 in (Rockafellar, 1970))

$$\frac{1}{\lambda n} X^\top \alpha^* \in \partial \psi(w^*). \tag{6}$$

In the case of our specific regularization term (2), the optimality condition (6) is written as

$$\frac{1}{\lambda n} X_{:j}^\top \alpha^* \in \begin{cases} \dfrac{w_j^*}{|w_j^*|} + w_j^* & (w_j^* \neq 0), \\ [-1, 1] & (w_j^* = 0). \end{cases} \tag{7}$$

The optimality condition (7) indicates that $|X_{:j}^\top \alpha^*| \leq \lambda n \Rightarrow w_j^* = 0$. The basic idea behind safe feature screening is to construct a region $\Theta_{\alpha^*} \subset \mathbb{R}^n$ in the dual space so that $\alpha^* \in \Theta_{\alpha^*}$, and then compute an upper bound $UB(|X_{:j}^\top \alpha^*|) := \max_{\alpha \in \Theta_{\alpha^*}} |X_{:j}^\top \alpha|$. Using this upper bound, we can construct a safe feature screening rule in the form of $UB(|X_{:j}^\top \alpha^*|) \leq \lambda n \Rightarrow w_j^* = 0$.

After the seminal work (El Ghaoui et al., 2012), many different approaches for constructing a region $\Theta_{\alpha^*}$ have been developed (see §1). Among those, we use an approach in (Ndiaye et al., 2015). Noting the fact that the convex conjugate function $\ell^*$ is $\gamma$-strongly convex, and henceforth the dual objective function $D_\lambda(\alpha)$ is $\gamma/n$-strongly concave, we can define a region

$$\Theta_{\alpha^*} := \{\alpha \mid \|\hat{\alpha} - \alpha\|_2 \leq \sqrt{2nG_\lambda(\hat{w}, \hat{\alpha})/\gamma}\}, \tag{8}$$

where $G_\lambda(\hat{w}, \hat{\alpha}) := P_\lambda(\hat{w}) - D_\lambda(\hat{\alpha})$ is the duality gap defined by an arbitrary pair of primal feasible solution $\hat{w} \in \mathrm{dom} P_\lambda$ and dual feasible solution $\hat{\alpha} \in \mathrm{dom} D_\lambda$. Since the region $\Theta_{\alpha^*}$ is a sphere, we can explicitly write the upper bound as

$$UB(|X_{:j}^\top \alpha^*|) = |X_{:j}^\top \hat{\alpha}| + \|X_{:j}\|_2 \sqrt{2nG_\lambda(\hat{w}, \hat{\alpha})/\gamma}. \tag{9}$$

### 2.3. Safe sample screening

The goal of safe sample screening is to identify a part of non-active samples $\{i \in [n] \mid \alpha_i^* = 0, \pm 1\}$ before or during the optimization process. Here, we slightly abuse the

[1] Here, we use individual loss function $\ell_i$ for $i \in [n]$ because it implicitly depends on $y_i$ (see, e.g., (3) and (4)).

word "non-active" in the sense that we call a sample to be non-active not only when the corresponding $\alpha_i^*$ is 0, but also when it is $\pm 1$. Although the $i$-th sample can be removed out only when $\alpha_i^* = 0$, we have similar computational advantages when we can guarantee that $\alpha_i^* = \pm 1$ because the size of the optimization problem can be reduced.

Safe sample screening is also built on the KKT optimality condition

$$x_i^\top w^* \in \partial \ell_i^*(-\alpha_i^*). \tag{10}$$

In the case of smoothed hinge loss, the KKT condition (10) is written when $y_i = 1$ as

$$x_i^\top w^* \in \begin{cases} [1, \infty) & (\alpha_i^* = 0) \\ (-\infty, 1-\gamma] & (\alpha_i^* = 1) \\ -\gamma\alpha_i^* + 1 & (\alpha_i^* \in (0,1)). \end{cases} \tag{11}$$

We construct a region $\Theta_{w^*} \subset \mathbb{R}^d$ in the primal space so that $w^* \in \Theta_{w^*}$, and then compute a lower bound $LB(x_i^\top w^*) := \min_{w \in \Theta_{w^*}} x_i^\top w$ and an upper bound $UB(x_i^\top w^*) := \max_{w \in \Theta_{w^*}} x_i^\top w$. The optimality condition (11) suggests that $LB(x_i^\top w^*) \geq 1 \Rightarrow \alpha_i^* = 0, UB(x_i^\top w^*) \leq 1-\gamma \Rightarrow \alpha_i^* = 1$. Similarly, for $y_i = -1$, the optimality condition is written as

$$x_i^\top w^* \in \begin{cases} (-\infty, -1] & (\alpha_i^* = 0) \\ [\gamma - 1, \infty) & (\alpha_i^* = -1) \\ -\gamma\alpha_i^* - 1 & (\alpha_i^* \in (-1,0)), \end{cases} \tag{12}$$

suggesting that $UB(y_i x_i^\top w^*) \leq -1 \Rightarrow \alpha_i^* = 0, LB(y_i x_i^\top w^*) \geq \gamma - 1 \Rightarrow \alpha_i^* = -1$. In the case of smoothed $\varepsilon$-insensitive loss, the optimality condition (10) is written as

$$x_i^\top w^* \in \begin{cases} [y_i - \varepsilon, y_i + \varepsilon] & (\alpha_i^* = 0), \\ [\gamma + y_i + \varepsilon, \infty) & (\alpha_i^* = -1), \\ (-\infty, -\gamma + y_i - \varepsilon] & (\alpha_i^* = 1), \\ -\gamma\alpha_i^* + y_i + \varepsilon & (\alpha_i^* \in (-1,0)), \\ -\gamma\alpha_i^* + y_i - \varepsilon & (\alpha_i^* \in (0,1)). \end{cases} \tag{13}$$

It indicates that $LB(x_i^\top w^*) \geq y_i - \varepsilon$ and $UB(x_i^\top w^*) \leq y_i + \varepsilon \Rightarrow \alpha_i^* = 0, LB(x_i^\top w^*) \geq \gamma + y_i + \varepsilon \Rightarrow \alpha_i^* = -1, UB(x_i^\top w^*) \leq -\gamma + y_i - \varepsilon \Rightarrow \alpha_i^* = 1$.

In order to develop a sphere region $\Theta_{w^*}$ in the primal space, we extend the duality GAP-based safe feature screening approach proposed in (Ndiaye et al., 2015) into safe sample screening context. The result is summarized in the following lemma.

**Lemma 1.** *For any $\hat{w} \in \mathrm{dom}P_\lambda$ and $\hat{\alpha} \in \mathrm{dom}D_\lambda$,*

$$w^* \in \Theta_{w^*} = \{w \mid \|\hat{w} - w\|_2 \leq \sqrt{2G_\lambda(\hat{w}, \hat{\alpha})/\lambda}\}, \tag{14}$$

*Furthermore, using the sphere form region $\Theta_{w^*}$ in (14), for any $\hat{w} \in \mathrm{dom}P_\lambda$ and $\hat{\alpha} \in \mathrm{dom}D_\lambda$, a pair of lower and upper bounds of $x_i^\top w^*$ are given as*

$$LB(x_i^\top w^*) = x_i^\top \hat{w} - \|x_i\|_2 \sqrt{2G_\lambda(\hat{w}, \hat{\alpha})/\lambda}, \tag{15a}$$

$$UB(x_i^\top w^*) = x_i^\top \hat{w} + \|x_i\|_2 \sqrt{2G_\lambda(\hat{w}, \hat{\alpha})/\lambda}. \tag{15b}$$

The proof of Lemma 1 is presented in Appendix.

## 3. Simultaneous safe screening

We have shown that, for doubly sparse modeling problems, safe screening rules for both of features and samples can be constructed respectively. In this paper we further develop the framework in which safe feature screening and safe sample screening are alternately iterated. A significant additional benefit of this framework is that the result of the previous safe feature screening can be exploited for making the primal region $\Theta_{w^*}$ smaller, meaning that the performance of the next safe sample screening can be improved, and vice-versa.

The following two theorems formally state these ideas. First, Theorem 2 states that we can obtain tighter upper bound for feature screening by exploiting the result of the previous safe sample screening.

**Theorem 2.** *Consider a safe feature screening problem given arbitrary pair of primal and dual feasible solution $\hat{w} \in \mathrm{dom}P_\lambda$ and $\hat{\alpha} \in \mathrm{dom}D_\lambda$. Furthermore, suppose that the result of the previous safe sample screening step assures the optimal values $\alpha_i^*$ for a subset of the samples $i \in \mathcal{S} \subset [n]$. Let $\mathcal{U}_s := [n] \setminus \mathcal{S}$, $r_D := \sqrt{2nG_\lambda(\hat{w}, \hat{\alpha})/\gamma}$, and $\tilde{\alpha}$ be an $n$-dimensional vector whose element is defined as $\tilde{\alpha}_i = \hat{\alpha}_i$ for $i \in \mathcal{U}_s$ and $\tilde{\alpha}_i = \alpha_i^*$ for $i \in \mathcal{S}$. Then, $|X_{:j}^\top \alpha^*|$ is bounded from above by the following upper bound:*

$$\tilde{UB}(|X_{:j}^\top \alpha^*|) := |X_{:j}^\top \tilde{\alpha}| + \|X_{\mathcal{U}_s j}\|_2 \sqrt{r_D^2 - \|\hat{\alpha}_{\mathcal{S}} - \alpha_{\mathcal{S}}^*\|_2^2}, \tag{16}$$

*and the upper bound in (16) is tighter than or equal to that in (9), i.e., $\tilde{UB}(|X_{:j}^\top \alpha^*|) \leq UB(|X_{:j}^\top \alpha^*|)$.*

The proof of Theorem 2 is presented in Appendix. By replacing the upper bounds in (9) with that in (16) in the safe feature screening step, there are more chance for screening out non-active features.

Next, Theorem 3 states that we can obtain tighter lower and upper bounds for sample screening by exploiting the result of the previous safe feature screening.

**Theorem 3.** *Consider a safe sample screening problem given arbitrary pair of primal and dual feasible solutions $\hat{w} \in \mathrm{dom}P_\lambda$ and $\hat{\alpha} \in \mathrm{dom}D_\lambda$. Furthermore, suppose that the result of the previous safe feature screening step assures that $w_j^* = 0$ for a subset of the features $j \in \mathcal{F} \subset [d]$.*

*Let $\mathcal{U}_f := [d] \setminus \mathcal{F}$, $r_P := \sqrt{2G_\lambda(\hat{w}, \hat{\alpha})/\lambda}$, and $\tilde{w}$ be a $d$-dimensional vector whose element is defined as $\tilde{w}_j = \hat{w}_j$ for $j \in \mathcal{U}_f$ and $\tilde{w}_j = 0$ for $j \in \mathcal{F}$. Then, $x_i^\top w^*$ is bounded from below and above respectively by the following lower and upper bounds:*

$$\tilde{LB}(x_i^\top w^*) := x_i^\top \tilde{w} - \|x_{i\mathcal{U}_f}\|_2 \sqrt{r_P^2 - \|\hat{w}_{\mathcal{F}}\|_2^2} \tag{17a}$$

$$\tilde{UB}(x_i^\top w^*) := x_i^\top \tilde{w} + \|x_{i\mathcal{U}_f}\|_2 \sqrt{r_P^2 - \|\hat{w}_{\mathcal{F}}\|_2^2} \tag{17b}$$

*and these bounds in (17) are tighter than or equal to those in (15), i.e., $\tilde{LB}(x_i^\top w^*) \geq LB(x_i^\top w^*)$ and $\tilde{UB}(x_i^\top w^*) \leq UB(x_i^\top w^*)$.*

The proof of Theorem 3 is presented in Appendix. By replacing the lower and the upper bounds in (15) with those in (17) in the safe sample screening step, there are more chance to be able to screen out non-active samples.

Theorems 2 and 3 suggests that, by alternately iterating feature screening and sample screening, more and more features and samples could be screened out. This iteration process can be terminated when there are few chances to be able to screen out additional features and samples. Such a termination condition can be developed by using the results in the next section.

## 4. Safe keeping of active features and samples

Safe screening studies initiated by the seminal work by (El Ghaoui et al., 2012) enabled us to identify a part of non-active features/samples before actually solving the optimization problem. In other words, safe screening is interpreted as an active set prediction method without *false negative error* (an error that truly active features/samples are predicted as non-active). In this section, we show that, by exploiting the two regions in the dual and the primal spaces, we can develop an active set prediction method without *false positive error* (an error that truly non-active features/samples are predicted as active). We call the latter approach as *safe feature/sample keeping*.

Safe feature keeping rule can be constructed by using the region in the primal space. Using $\Theta_{w^*}$ in (14), we can get the lower bound of $|w_j^*|$ for $j \in [d]$ as $LB(|w_j^*|) := |\hat{w}_j| - \sqrt{2G_\lambda(\hat{w}, \hat{\alpha})\lambda}$. Using this lower bound, safe feature keeping rule is simply formulated as the following theorem.

**Theorem 4.** *For an arbitrary pair of primal feasible solution $\hat{w} \in \mathrm{dom}P_\lambda$ and dual feasible solution $\hat{\alpha} \in \mathrm{dom}D_\lambda$,*

$$|\hat{w}_j| - \sqrt{2G_\lambda(\hat{w}, \hat{\alpha})/\lambda} > 0 \;\Rightarrow\; w_j^* \neq 0 \textit{ for } j \in [d].$$

Similarly, safe sample keeping rule can be constructed by using a region in the dual space. The condition for $\alpha_i^*$ being active is written as $\alpha_i^* \neq -1, 0, 1$. This can be guaranteed when the condition $\alpha_i^* \in (0, 1)$ or $\alpha_i^* \in (-1, 0)$ holds for the $i$-th element of $\forall \alpha \in \Theta_{\alpha^*}$. Since $\Theta_{\alpha^*}$ is a sphere, safe sample keeping rule can be simply derived by (8).

**Theorem 5.** *For an arbitrary pair of primal feasible solution $\hat{w} \in \mathrm{dom}P_\lambda$ and dual feasible solution $\hat{\alpha} \in \mathrm{dom}D_\lambda$,*

$$|\hat{\alpha}_i| - \sqrt{2nG_\lambda(\hat{w}, \hat{\alpha})/\gamma} > 0 \textit{ and } |\hat{\alpha}_i| + \sqrt{2nG_\lambda(\hat{w}, \hat{\alpha})/\gamma} < 1$$
$$\Rightarrow \alpha_i^* \neq 0, \pm 1 \textit{ for } i \in [n].$$

When we use safe screening approaches, there is a trade-off between the computational costs of evaluating safe screening rules and the computational time saving by screening out some features/samples. If we know in advance that some of the features/samples cannot be non-active by using safe keeping approaches, we do not have to waste the rule evaluation costs for those features/samples. Safe screening rule evaluation costs would be more significant in dynamic screening and our simultaneous screening scenarios because rules are repeatedly evaluated. By combining safe screening and safe keeping approaches, we can get an information about how many features/samples are not yet determined to be active or non-active. This information can be also used as a stopping criteria of dynamic screening and our simultaneous screening. We can stop evaluating safe screening rules when there only remain few features/samples that have not been determined to be active or non-active.

We finally note that, in our particular working problem of $L_1$ smooth SVC and $L_1$ smooth SVR, safe keeping is also possible by using the KKT optimality conditions in (7) for features, and (11) - (13) for samples. We describe the details in the Appendix.

## 5. LP-based simultaneous safe screening

In this section, we consider another empirical risk minimization problem that induces sparsities both in features and samples. Specifically, we study a problem with $L_1$-penalty $\psi(w) = \|w\|_1$ and vanilla hinge loss $\ell_i(x_i^\top w) = [1 - y_i x_i^\top w]_+$, which we call *LP-based SVM* because it is casted into a linear program (LP). LP-based SVM has been studied in (Bradley & Mangasarian, 1998; Zhu et al., 2004), and also in boosting context. LPBoost (Demiriz et al., 2002) solves LP-based SVM via the column generation approach of linear programming. Sparse LPBoost (Hatano & Takimoto, 2009) is similar to simultaneous screening in that it iteratively solves LP subproblems for features and samples. LP-based SVM induces feature sparsity due to $L_1$-penalty and sample sparsity due to the property of hinge loss.b

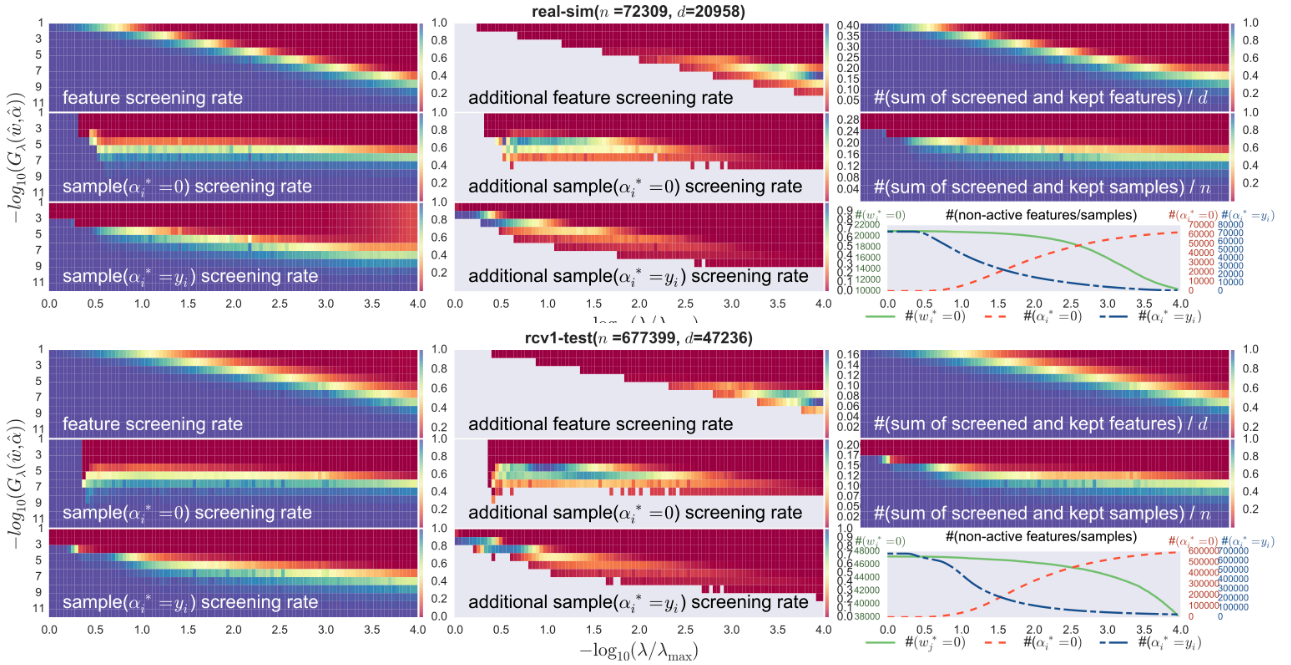

*Figure 2.* Safe screening and keeping rates for classification problems (for `real-sim` and `rcv1-test` datasets). The three plots in the left show the individual safe feature/sample screening rates (the middle and the bottom ones are for $\alpha_i^* = 0$ and $\alpha_i^* = \pm 1$, respectively). The three plots in the center show the additional safe screening rates by simultaneously considering feature and sample screenings. The gray area in these center plots corresponds to the blue area in the corresponding left plot. In these gray area, the individual safe screening performances are good enough (screening rate $> 0.95$) and additional screening is unnecessary. The top right and middle right plots show the safe keeping rates for feature and samples, respectively. The bottom right plot shows the numbers of non-active features and samples for various values of $\lambda$.

### 5.1. Safe feature screening for LP-based SVM

Feature safe screening for LP-based SVM was studied in the seminal safe feature screening paper by El Ghaoui et al. (2012). However, the method presented in their paper requires a precise optimal solution of an LP-based SVM with a different penalty parameter $\lambda$. This requirement is impractical because precise optimal solutions are often difficult to get numerically. Here, we present a novel safe feature screening method for LP-based SVM that only requires an arbitrary pair of a primal feasible solution $\hat{w} \in \mathrm{dom} P_\lambda$ and a dual feasible solution $\hat{\alpha} \in \mathrm{dom} D_\lambda$. The proposed safe feature screening method for LP-based SVM is summarized in the following theorem.

**Theorem 6.** *Consider safe feature screening problem given an arbitrary pair of a primal feasible solution $\hat{w} \in \mathrm{dom} P_\lambda$ and a dual feasible solution $\hat{\alpha} \in \mathrm{dom} D_\lambda$. Let $\ell_q := \lfloor y^\top \hat{\alpha} \rfloor$, $u_q = \lfloor P_\lambda(\hat{w}) \rfloor$, $Z := [y_1 x_1, \ldots, y_n x_n]^\top \in \mathbb{R}^{n \times d}$, and $Z'_{:j} \in \mathbb{R}^n$, $j \in [d]$, be the vector obtained by sorting $Z_{:j}$ in increasing order. Furthermore, let $n_{Z'_{:j}}$ and $p_{Z'_{:j}}$ represent the numbers of negative and positive elements of $Z'_{:j}$, respectively. Then, $LB(X_{:j}^\top \alpha^*) < -\lambda n$ and $UB(X_{:j}^\top \alpha^*) > \lambda n \Rightarrow w_j^* = 0$, where*

$$LB(X_{:j}^\top \alpha^*) := \begin{cases} \sum_{i=1}^{l_q} Z'_{ij} + (y^\top \hat{\alpha} - l_q) Z'_{(l_q+1)j} & (n_{Z'_{:j}} < l_q + 1), \\ \sum_{i=1}^{u_q} Z'_{ij} + (P_\lambda(\hat{w}) - u_q), Z'_{u_q j} & (n_{Z'_{:j}} > u_q) \\ \sum_{i=1}^{n} \min\{0, Z'_{ij}\} & (\text{otherwise}), \end{cases}$$

$$UB(X_{:j}^\top \alpha^*) := \begin{cases} \sum_{i=n-l_q}^{n} Z'_{ij} + (y^\top \hat{\alpha} - l_q) Z'_{(n-l_q-1)j} & (p_{Z'_{:j}} < l_q + 1), \\ \sum_{i=n-u_q}^{n} Z'_{ij} + (P_\lambda(\hat{w}) - u_q) Z'_{(n-u_q-1)j} & (p_{Z'_{:j}} > u_q), \\ \sum_{i=1}^{n} \max\{0, Z'_{ij}\} & (\text{otherwise}). \end{cases}$$

The proof of Theorem 6 is presented in Appendix.

### 5.2. Safe sample screening for LP-based SVM

Here, we develop a novel safe sample screening method for LP-based SVM as summarized in the following theorem.

**Theorem 7.** *Consider safe sample screening given an arbitrary primal feasible solution $\hat{w} \in \mathrm{dom} P_\lambda$. Let $g_{\ell_i}(w)$ be a subgradient of vanilla hinge loss $\ell_i(x_i^\top w)$ for $w \in \mathrm{dom} P_\lambda$,*

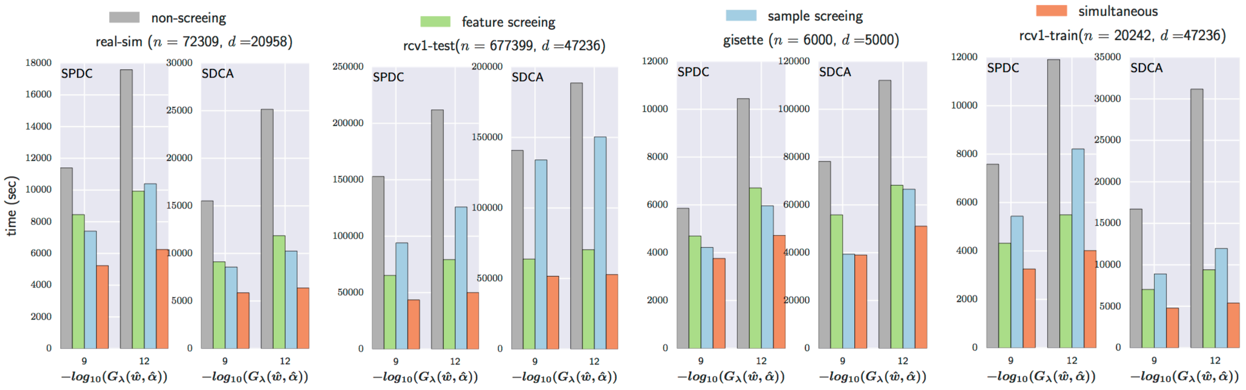

*Figure 3.* Total computation time for training 100 solutions for various values of $\lambda$ in classification problems.

*and define* $k := \lambda\|\hat{w}\|_1 + \frac{1}{n}\sum_{i\in[n]} g_{\ell_i(\hat{w})}^\top \hat{w}$. *Then,*

$$\begin{aligned}
y_i = +1 \text{ and } LB(x_i^\top w^*) > +1 &\Rightarrow \alpha_i^* = +1,\\
y_i = +1 \text{ and } UB(x_i^\top w^*) < +1 &\Rightarrow \alpha_i^* = 0,\\
y_i = -1 \text{ and } LB(x_i^\top w^*) > -1 &\Rightarrow \alpha_i^* = 0,\\
y_i = -1 \text{ and } UB(x_i^\top w^*) < -1 &\Rightarrow \alpha_i^* = -1,
\end{aligned}$$

*where*

$$LB(x_i^\top w^*) := \max_{\mu>0}\{\mu k\} \text{ s.t. } \left\| -\frac{1}{\lambda}x_i - \frac{\mu}{\lambda n}\sum_{i=1}^n g_{\ell_i}(\hat{w}) \right\|_\infty \le \mu,$$

$$UB(x_i^\top w^*) := \max_{\mu>0}\{\mu k\} \text{ s.t. } \left\| \frac{1}{\lambda}x_i - \frac{\mu}{\lambda n}\sum_{i=1}^n g_{\ell_i}(\hat{w}) \right\|_\infty \le \mu.$$

The proof of Theorem 7 is presented in Appendix. Although the lower and the upper bounds are not explicitly presented, these optimization problems can be easily solved because they are just linear programs with one variable $\mu > 0$.

As we discussed in §3, by alternatively iterating safe feature screening in §5.1 and safe sample screening in §5.2, we can make the regions in the dual and the primal regions step by step, indicating that the chance of screening out more features and samples increases [2].

## 6. Numerical experiments

We demonstrate the advantage of simultaneous safe screening through numerical experiments. After we describe the experimental setups in §6.1, we report the results on safe screening and keeping rates, and computation time savings in §6.2 and §6.2, respectively. Due to the space limitation, we only show the results on classification problems in the main text. Other experimental results are presented in Appendix.

### 6.1. Experimental setups

Table 1 summarizes the datasets used in the experiments. We picked up four datasets whose numbers of features and samples are both large from libsvm dataset repository (Chang & Lin, 2011). Here, we report the results on $L_1$-penalized smoothed hinge SV classification. We set $\lambda_{\max} := \|Z^\top \mathbf{1}\|_\infty$, and considered problems with various values of the penalty parameter $\lambda$ between $\lambda_{\max}$ and $10^{-4}\lambda_{\max}$. The parameter in the smoothed hinge loss $\gamma$ is set to be 0.5. The proposed methods can be used with any optimization solvers as long as they provide both primal and dual sequences of solutions that converge to the optimal solution. For the experiments, we used Proximal Stochastic Dual Coordinate Ascent (SDCA) (Shalev-Shwartz & Zhang, 2015) and Stochastic Primal-Dual Coordinate (SPDC) (Zhang & Lin, 2015) because they are state-of-the-art optimization methods for general large-scale regularized empirical risk minimization problems. We wrote the code in C++. The code is available on `https://github.com/takeuchi-lab/s3fs`. All the computations were conducted by using a single core of an Intel Xeon CPU E5-2643 v2 (3.50GHz), 64GB MEM.

*Table 1.* Benchmark datasets used in the experiments.

| dataset name | sample size: $n$ | feature size: $d$ | #(nnz)/$nd$ |
|---|---:|---:|---|
| `real-sim` | 72,309 | 20,958 | 0.002448 |
| `rcv1-test` | 677,399 | 47,236 | 0.025639 |
| `gisette` | 6,000 | 5,000 | 0.991000 |
| `rcv1-train` | 20,242 | 47,236 | 0.001568 |

#(nnz) indicates the number of non-zero elements.

### 6.2. Safe screening and keeping rates

We compared the simultaneous screening rates with individual safe screening rates. Figure 2 shows the results.

[2] The tighter bounds can be obtained by exploiting the previous safe screening results as we discussed in §3 although we do not explicitly present those bounds here due to the space limitation.

In the $3 \times 3$ subplots, the left plots indicate the individual screening rates defined as (#(screened features or samples) / #($w_j^* = 0$ or $\alpha_i^* = 0$ or $\alpha_i^* = \pm 1$)). The center plots indicate the additional screening rates by the synergy effect defined as (#(additionally screened features or samples) / #($w_j^* = 0$ or $\alpha_i^* = 0$ or $\alpha_i^* = \pm 1$)). The top plots represent the results on feature screening, while the middle and the bottom plots show the results on sample screening (for each of $\alpha_i^* = 0$ and $\alpha_i^* = \pm 1$). We investigated the screening rates for various values of $\lambda$ in the horizontal axis and for various quality of the solutions measured in terms of the duality gap in the vertical axis. In all the datasets, we observed that it is valuable to consider both feature and sample screening when the numbers of features and samples are both large. In addition, we confirmed that there are improvements in screening rates by the synergy effect both in feature and sample screenings especially when duality gap $G_\lambda(\hat{w}, \hat{\alpha})$ is large. Note that gray areas in the center plots corresponds to the blue area in the corresponding left plot, where the individual safe screening performances are good enough (screening rate $> 0.95$) and additional screening is unnecessary.

The top left and the middle left plots show the rates of features and samples, respectively, that are determined to be active or non-active by using safe keeping and safe screening approaches, respectively. We see that, by combining safe keeping and safe screening approaches, a large portion of features/samples can be determined to be active or non-active without actually solving the optimization problems.

### 6.3. Computation time savings

We compared the computational costs of simultaneous safe screening and individual safe feature/sample screening with the naive baseline (denoted as “non-screening”). We compared the computation costs in a realistic model building scenario. Specifically, we computed a sequence of solutions at 100 different penalty parameter values evenly allocated in $[10^{-4}\lambda_{\max}, \lambda_{\max}]$ in the logarithmic scale. In all the cases, we used *warm-start* approach, i.e., when we computed a solution at a new $\lambda$, we used the solution at the previous $\lambda$ as the initial starting point of the optimizer. In addition, whenever possible, we used *dynamic safe screening* strategies (Bonnefoy et al., 2014) in which safe screening rules are evaluated every time the duality gap $G_\lambda(\hat{w}, \hat{\alpha})$ was $0.1$ times smaller than before. Here, we exploited the information obtained by safe keeping as well, i.e., we did not evaluate safe screening rules for features and samples which are safely kept as active, and the rate of features/samples that are determined to be active or non-active (see the left top and left middle plots in Figure 2) is used as the stopping criterion for safe feature rule evaluations.

Figures 3 show the entire computation time for training $100$ different solutions. In all the datasets, simultaneous safe screening was significantly faster than individual safe feature/sample screening and non-screening. Figure 4 shows a sequence of computation times for various values of $\lambda$ for the classification problem on `rcv1-test` and `real-sim` datasets with SPDC optimization solver. These plots suggest that the computation time savings by individual safe feature screening was better than individual safe sample screening when $\lambda$ is large because the feature screening rates are high when $\lambda$ is large, while the difference between the two individual screening approaches gets smaller as $\lambda$ gets smaller (see Figure 2). Simultaneous safe screening was consistently faster than individual safe feature/sample screening and non-screening in all the problem setups.

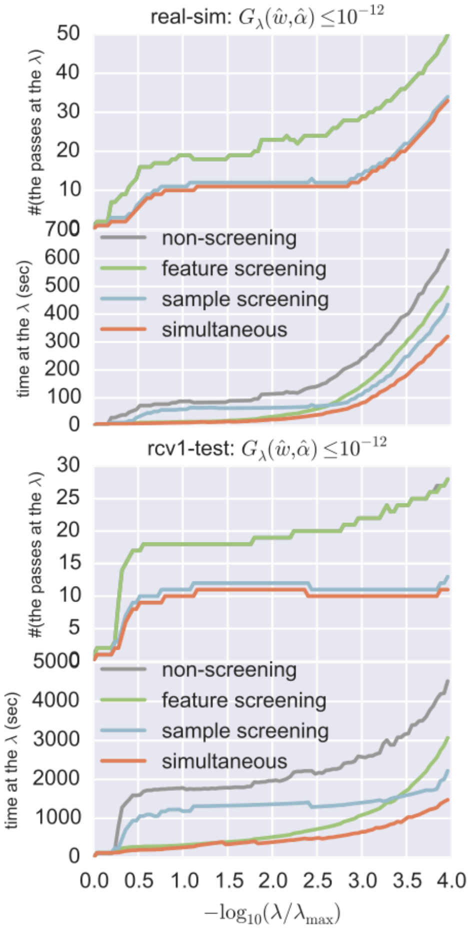

*Figure 4.* Number of optimization steps and computation time (for `real-sim` and `rcv1-test` datasets). Sequences of the number of passes through the entire dataset and computation time to convergence for various values of $\lambda$ for classification problems with SPDC solver are plotted.

## Acknowledgements

For this work, MK was partially supported by JSPS KAKENHI Grant Number 26280083. KH was partially supported by JSPS KAKENHI Grant Number 16K00305 and MEXT KAKENHI Grant Number 24106010 (the ELC project). IT was partially supported by JST CREST 13217726, CREST 15656320, JSPS KAKENHI Grant Number 26280083, MEXT KAKENHI Grant Number 16H00886, and JST support program for starting up innovation-hub on materials research by information integration initiative.

## References

Bertsekas, D. P. *Nonlinear Programming (2nd edition)*. Athena Scientific, 1999.

Bonnefoy, Antoine, Emiya, Valentin, Ralaivola, Liva, and Gribonval, Rémi. A dynamic screening principle for the lasso. In *Signal Processing Conference (EUSIPCO), 2014 Proceedings of the 22nd European*, pp. 6–10. IEEE, 2014.

Boser, Bernhard E, Guyon, Isabelle M, and Vapnik, Vladimir N. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pp. 144–152. ACM, 1992.

Bradley, Paul S and Mangasarian, Olvi L. Feature selection via concave minimization and support vector machines. In *ICML*, volume 98, pp. 82–90, 1998.

Chang, Chih-Chung and Lin, Chih-Jen. Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

Demiriz, Ayhan, Bennett, Kristin P, and Shawe-Taylor, John. Linear programming boosting via column generation. *Machine Learning*, 46(1-3):225–254, 2002.

El Ghaoui, Laurent, Viallon, Vivian, and Rabbani, Tarek. Safe feature elimination for the lasso and sparse supervised learning problems. *Pacific Journal of Optimization*, 8(4):667–698, 2012.

Fercoq, Olivier, Gramfort, Alexandre, and Salmon, Joseph. Mind the duality gap: safer rules for the lasso. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 333–342, 2015.

Hastie, Trevor, Tibshirani, Robert, and Wainwright, Martin. *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC Press, 2015.

Hatano, Kohei and Takimoto, Eiji. Linear Programming Boosting by Column and Row Generation. In *Proceedings of the 12th International Conference on Dicovery Science (DS 2009)*, volume 5808 of *LNCS*, pp. 401–408, 2009.

Johnson, Tyler B. and Guestrin, Carlos. Blitz: A principled meta-algorithm for scaling sparse optimization. In *Proceedings of the 32nd International Conference on Machine Learning*, 2015.

Liu, Jun, Zhao, Zheng, Wang, Jie, and Ye, Jieping. Safe Screening with Variational Inequalities and Its Application to Lasso. In *Proceedings of the 31st International Conference on Machine Learning*, 2014.

Ndiaye, Eugene, Fercoq, Olivier, Gramfort, Alexandre, and Salmon, Joseph. Gap safe screening rules for sparse multi-task and multi-class models. In *Advances in Neural Information Processing Systems*, pp. 811–819, 2015.

Ogawa, Kohei, Suzuki, Yoshiki, and Takeuchi, Ichiro. Safe screening of non-support vectors in pathwise svm computation. In *Proceedings of the 30th International Conference on Machine Learning*, pp. 1382–1390, 2013.

Rockafellar, Ralph Tyrell. *Convex analysis*. Princeton university press, 1970.

Shalev-Shwartz, Shai and Zhang, Tong. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. *Mathematical Programming*, pp. 1–41, 2015.

Tibshirani, Robert. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

Vainsencher, Daniel, Liu, Han, and Zhang, Tong. Local smoothness in variance reduced optimization. In *Advances in Neural Information Processing Systems 28*, pp. 2179–2187. 2015.

Wang, Jie, Zhou, Jiayu, Wonka, Peter, and Ye, Jieping. Lasso screening rules via dual polytope projection. In *Advances in Neural Information Processing Systems*, pp. 1070–1078, 2013.

Wang, Jie, Wonka, Peter, and Ye, Jieping. Scaling svm and least absolute deviations via exact data reduction. *Proceedings of The 31st International Conference on Machine Learning*, 2014a.

Wang, Jie, Zhou, Jiayu, Liu, Jun, Wonka, Peter, and Ye, Jieping. A safe screening rule for sparse logistic regression. In *Advances in Neural Information Processing Systems*, pp. 1053–1061, 2014b.

Xiang, Zhen J, Xu, Hao, and Ramadge, Peter J. Learning sparse representations of high dimensional data on large scale dictionaries. In *Advances in Neural Information Processing Systems*, pp. 900–908, 2011.

Xiang, Zhen James, Wang, Yun, and Ramadge, Peter J. Screening tests for lasso problems. *arXiv preprint arXiv:1405.4897*, 2014.

Zhang, Yuchen and Lin, Xiao. Stochastic primal-dual coordinate method for regularized empirical risk minimization. *Proceedings of The 32nd International Conference on Machine Learning*, pp. 353–361, 2015.

Zhu, Ji, Rosset, Saharon, Hastie, Trevor, and Tibshirani, Rob. 1-norm support vector machines. *Advances in neural information processing systems*, 16(1):49–56, 2004.

Zimmert, Julian, de Witt, Christian Schroeder, Kerg, Giancarlo, and Kloft, Marius. Safe screening for support vector machines. *NIPS 2015 Workshop on Optimization in Machine Learning (OPT)*, 2015.

# A. Proofs

### A.1. Proof of Lemma 1

*Proof.* Since $P_\lambda(w)$ is $\lambda$-strongly convex, $\forall w_1, w_2 \in \mathrm{dom}P_\lambda$,

$$P_\lambda(w_1) \geq P_\lambda(w_2) + g_{P_\lambda}(w_2)^\top (w_1 - w_2) + \frac{\lambda}{2}\|w_1 - w_2\|_2^2,$$

where, $g_{P_\lambda}(w) \in \partial P_\lambda(w)$. On the other hand, $\forall \hat{w} \in \mathrm{dom}P_\lambda$, $g_{P_\lambda}(w^*)^\top(\hat{w} - w^*) \geq 0$ (see Proposition B.24 in (**?**)). Also, from weak duality, $\forall \hat{\alpha} \in \mathrm{dom}D_\lambda$, $D(\hat{\alpha}) \leq P_\lambda(w^*)$. By substituting $w_1 = \hat{w}, w_2 = w^*$,

$$\frac{\lambda}{2}\|\hat{w} - w^*\|_2^2 \leq P_\lambda(\hat{w}) - D_\lambda(\hat{\alpha}).$$

Therefore, $w^*$ is within a region $\Theta_{w^*}$, where

$$\Theta_{w^*} := \left\{ w \mid \|\hat{w} - w\|_2 \leq \sqrt{2G_\lambda(\hat{w}, \hat{\alpha})/\lambda} \right\}.$$

Since $\Theta_{w^*}$ is Sphere, a lower bound of $x_i^\top w^*$ and an upper bound of $x_i^\top w^*$ are given in closed form as follows:

$$LB(x_i^\top w^*) = x_i^\top \hat{w} - \|x_i\|_2\sqrt{2G_\lambda(\hat{w}, \hat{\alpha})/\lambda},$$
$$UB(x_i^\top w^*) = x_i^\top \hat{w} + \|x_i\|_2\sqrt{2G_\lambda(\hat{w}, \hat{\alpha})/\lambda}.$$

■

### A.2. Proof of Theorem 2

*Proof.* Supposing that the result of the previous safe sample screening step assures the optimal values $\alpha_i^*$ for a subset of the samples $i \in \mathcal{S} \subset [n]$, the dual optimal solution region is written as

$$\tilde{\Theta}_{\alpha^*} := \left\{ \alpha \in \Theta_{\alpha^*} \mid \alpha_i = \alpha_i^* \; \forall i \in \mathcal{S} \right\}.$$

Then, $X_{:j}^\top \alpha^*$ is bounded from above by the following upper bound:

$$\begin{aligned}
&\tilde{UB}(X_{:j}^\top \alpha^*) \\
&:= \max_{\alpha \in \tilde{\Theta}_{\alpha^*}} X_{:j}^\top \alpha \\
&= \sum_{i \in \mathcal{S}} \alpha_i^* X_{ij} + \max_{\alpha_{\mathcal{U}_s}} X_{\mathcal{U}_s j}^\top \alpha_{\mathcal{U}_s} \\
&\qquad \text{s.t } \|\hat{\alpha}_{\mathcal{U}_s} - \alpha_{\mathcal{U}_s}\|_2^2 \leq r_D^2 - \|\hat{\alpha}_{\mathcal{S}} - \alpha_{\mathcal{S}}^*\|_2^2 \\
&= \sum_{i \in \mathcal{S}} \alpha_i^* X_{ij} + X_{\mathcal{U}_s j}^\top \hat{\alpha}_{\mathcal{U}_s} + \|X_{\mathcal{U}_s j}\|_2 \sqrt{r_D^2 - \|\hat{\alpha}_{\mathcal{S}} - \alpha_{\mathcal{S}}^*\|_2^2} \\
&= X_{:j}^\top \tilde{\alpha} + \|X_{\mathcal{U}_s j}\|_2 \sqrt{r_D^2 - \|\hat{\alpha}_{\mathcal{S}} - \alpha_{\mathcal{S}}^*\|_2^2}.
\end{aligned}$$

Similarly, $X_{:j}^\top \alpha^*$ is bounded from below by the following lower bound:

$$\tilde{LB}(X_{:j}^\top \alpha) := X_{:j}^\top \tilde{\alpha} - \|X_{\mathcal{U}_s j}\|_2 \sqrt{r_D^2 - \|\hat{\alpha}_{\mathcal{S}} - \alpha_{\mathcal{S}}^*\|_2^2}.$$

Therefore,

$$\tilde{UB}(|X_{:j}^\top \alpha|) = |X_{:j}^\top \tilde{\alpha}| + \|X_{\mathcal{U}_s j}\|_2 \sqrt{r_D^2 - \|\hat{\alpha}_{\mathcal{S}} - \alpha_{\mathcal{S}}^*\|_2^2}.$$

Since $\tilde{\Theta}_{\alpha^*} \subset \Theta_{\alpha^*}$, the upper bound in (16) is tighter than or equal to that in (9), i.e., $\tilde{UB}(|X_{:j}^\top \alpha^*|) \leq UB(|X_{:j}^\top \alpha^*|)$. ■

### A.3. Proof of Theorem 3

*Proof.* Supposing that the result of the previous safe feature screening step assures that $w_j^* = 0$ for a subset of the features $j \in \mathcal{F} \subset [d]$, the primal optimal solution region is written as

$$\tilde{\Theta}_{w^*} := \left\{ w \in \Theta_{w^*} \mid w_j = 0 \, \forall j \in \mathcal{F} \right\}.$$

Then, $x_i^\top w^*$ is bounded from below by the following lower bound:

$$\begin{aligned}
&\tilde{LB}(x_i^\top w) \\
&:= \min_{w \in \tilde{\Theta}_{w^*}} x_i^\top w \\
&= \min_{w} x_i^\top w \;\text{ s.t. } \|\hat{w} - w\|_2^2 \leq r_P^2, \hat{w}_j = 0 \, \forall j \in \mathcal{F} \\
&= \min_{w} x_{i\mathcal{U}_f}^\top w_{\mathcal{U}_f} \;\text{ s.t. } \|\hat{w}_{\mathcal{U}_f} - w_{\mathcal{U}_f}\|_2^2 \leq r_P^2 - \|\hat{w}_{\mathcal{F}}\|_2^2 \\
&= x_{i\mathcal{U}_f}^\top \hat{w}_{\mathcal{U}_f} - \|x_{i\mathcal{U}_f}\|_2 \sqrt{r_P^2 - \|\hat{w}_{\mathcal{F}}\|_2^2}.
\end{aligned}$$

Similarly, $x_i^\top w^*$ is bounded from above by the following upper bound:

$$UB(x_i^\top w^*) = x_{i\mathcal{U}_f}^\top \hat{w}_{\mathcal{U}_f} + \|x_{i\mathcal{U}_f}\|_2 \sqrt{r_P^2 - \|\hat{w}_{\mathcal{F}}\|_2^2}.$$

Since $\tilde{\Theta}_{w^*} \subset \Theta_{w^*}$, these bounds in (17) are tighter than or equal to those in (15), i.e., $\tilde{LB}(x_i^\top w^*) \geq LB(x_i^\top w^*)$ and $\tilde{UB}(x_i^\top w^*) \leq UB(x_i^\top w^*)$. ■

### A.4. Proof of Theorem 6

The convex conjugate functions of $L_1$-penalty and vanilla hinge loss are respectively written as

$$\psi^*(v) := \begin{cases} 0 & (\|v\|_\infty \leq 1), \\ \infty & (\text{otherwise}), \end{cases} \tag{19}$$

$$\ell_i^*(\alpha_i) := \begin{cases} y_i\alpha_i & y_i\alpha_i \in [-1, 0], \\ \infty & (\text{otherwise}), \end{cases} \tag{20}$$

and the dual problem is written as

$$\begin{aligned}
\max_{\alpha} D_\lambda(\alpha) &:= \max_{\alpha} \left\{ y^\top \alpha \right\} \\
&\quad \text{s.t. } \left\| \frac{1}{\lambda n}\alpha_i x_i \right\|_\infty \leq 1, \; y_i\alpha_i \in [0, 1] \;\; \forall i \in [n].
\end{aligned}$$

We first construct the the dual optimal solution region $\tilde{\Theta}_{\alpha^*}$.

**Lemma 8.** *For an arbitrary pair of primal feasible solution $\hat{w} \in \mathrm{dom}P_\lambda$ and dual feasible solution $\hat{\alpha} \in \mathrm{dom}D_\lambda$, the dual optimal solution region is written as*

$$\Theta_{\alpha^*} := \left\{ \forall i \; y_i\alpha_i \in [0,1] \;\middle|\; y^\top \hat{\alpha} \le y^\top \alpha \le P_\lambda(\hat{w}) \right\}.$$

*Proof of Lemma 8* . From the optimality and weak duality $y^\top \hat{\alpha} \le y^\top \alpha^*$ and $y^\top \alpha^* \le P_\lambda(\hat{w})$, respectively. Therefore,

$$\alpha^* \in \hat{\Theta}_{\alpha^*} := \left\{ \alpha \in \mathrm{dom}D_\lambda \;\middle|\; y^\top \hat{\alpha} \le y^\top \alpha \le P_\lambda(\hat{w}) \right\}.$$

Noting that $\hat{\Theta}_{\alpha^*} \subseteq \Theta_{\alpha^*}$, $\alpha^* \in \Theta_{\alpha^*}$. ■

*Proof of Theorem 6.* From Lemma 8,

$$X_{:j}^\top \alpha^* \ge LB(X_{:j}^\top \alpha^*) := \min_{\alpha \in \Theta_{\alpha^*}} X_{:j}^\top \alpha$$

Moreover,

$$LB(X_{:j}^\top \alpha^*) = \min_{\alpha \in \Theta_{\alpha^*}} Z_{:j}^\top \alpha_y,$$

where $\alpha_y := [y_1\alpha_1, \ldots, y_n\alpha_n]^\top$. Let us define three $n$-dimensional vectors $\bar{\alpha}^{(1)}$, $\bar{\alpha}^{(2)}$ and $\bar{\alpha}^{(3)}$ as follows:

$$\bar{\alpha}_i^{(1)} := \begin{cases} y_i & (Z'_{ij} < 0) \\ 0 & (\text{otherwise}), \end{cases}$$

$$\bar{\alpha}_i^{(2)} := \begin{cases} y_i & (Z'_{ij} \le Z'_{l_q j}) \\ y_i(y^\top \hat{\alpha} - l_q) & (Z'_{ij} = Z'_{(l_q+1)j}) \\ 0 & (otherwise), \end{cases}$$

$$\bar{\alpha}_i^{(3)} := \begin{cases} y_i & (Z'_{ij} \le Z'_{u_q j}) \\ y_i(P_\lambda(\hat{w}) - u_q) & (Z'_{ij} = Z'_{(u_q+1)j}) \\ 0 & (otherwise). \end{cases}$$

If $l_q + 1 \le n_{Z'_{:j}} \le u_q$, then $\bar{\alpha}^{(1)}$ is an element of $\Theta_{\alpha^*}$ and minimizes $X_{:j}^\top \alpha$. If $n_{Z'_{:j}} < l_q + 1$ then $\bar{\alpha}^{(1)} \notin \Theta_{\alpha^*}$, $\bar{\alpha}^{(2)}$ is an element of $\Theta_{\alpha^*}$ and minimizes $X_{:j}^\top \alpha$ because $y^\top \bar{\alpha}^{(2)} = y^\top \hat{\alpha}$. If $m_{Z'_{:j}} > u_q$ then $\bar{\alpha}^{(1)} \notin \Theta_{\alpha^*}$, meaning that $\bar{\alpha}^{(3)}$ is an element of $\Theta_{\alpha^*}$ and minimizes $X_{:j}^\top \alpha$ because $y^\top \bar{\alpha}^{(3)} = P_\lambda$. Therefore,

$$LB(X_{:j}^\top \alpha^*) :=$$
$$\begin{cases} \sum_{i=1}^{l_q} Z'_{ij} + (y^\top \hat{\alpha} - l_q) Z'_{(l_q+1)j} & (n_{Z'_{:j}} < l_q + 1), \\ \sum_{i=1}^{u_q} Z'_{ij} + (P_\lambda(\hat{w}) - u_q), Z'_{u_q j} & (n_{Z'_{:j}} > u_q) \\ \sum_{i=1}^{n} \min\{0, Z'_{ij}\} & (\text{otherwise}), \end{cases}$$

Similarly, from Lemma 8,

$$X_{:j}^\top \alpha^* \le UB(X_{:j}^\top \alpha^*) := \max_{\alpha \in \Theta_{\alpha^*}} X_{:j}^\top \alpha$$

Moreover,

$$UB(X_{:j}^\top \alpha^*) = \max_{\alpha \in \Theta_{\alpha^*}} Z_{:j}^\top \alpha_y.$$

Let us define three $n$-dimensional vectors $\bar{\alpha}^{(4)}$, $\bar{\alpha}^{(5)}$ and $\bar{\alpha}^{(6)}$ as follows:

$$\bar{\alpha}_i^{(4)} := \begin{cases} y_i & (Z'_{ij} > 0) \\ 0 & (\text{otherwise}), \end{cases}$$

$$\bar{\alpha}_i^{(5)} := \begin{cases} y_i & (Z'_{ij} \ge Z'_{(n-l_q)j}) \\ y_i(y^\top \hat{\alpha} - l_q) & (Z'_{ij} = Z'_{(n-l_q-1)j}) \\ 0 & (otherwise), \end{cases}$$

$$\bar{\alpha}_i^{(6)} := \begin{cases} y_i & (Z'_{ij} \ge Z'_{(n-u_q)j}) \\ y_i(P_\lambda(\hat{w}) - u_q) & (Z'_{ij} = Z'_{(n-u_q-1)j}) \\ 0 & (otherwise). \end{cases}$$

If $l_q + 1 \le p_{Z'_{:j}} \le u_q$, then $\bar{\alpha}^{(4)}$ is an element of $\Theta_{\alpha^*}$ and maximizes $X_{:j}^\top \alpha$. If $p_{Z'_{:j}} < l_q + 1$ then $\bar{\alpha}^{(4)} \notin \Theta_{\alpha^*}$, $\bar{\alpha}^{(5)}$ is an element of $\Theta_{\alpha^*}$ and maximizes $X_{:j}^\top \alpha$ because $y^\top \bar{\alpha}^{(5)} = y^\top \hat{\alpha}$. If $p_{Z'_{:j}} > u_q$ then $\bar{\alpha}^{(4)} \notin \Theta_{\alpha^*}$, meaning that $\bar{\alpha}^{(6)}$ is an element of $\Theta_{\alpha^*}$ and maximizes $X_{:j}^\top \alpha$ because $y^\top \bar{\alpha}^{(6)} = P_\lambda$. Therefore,

$$UB(X_{:j}^\top \alpha^*) :=$$
$$\begin{cases} \sum_{i=n-l_q}^{n} Z'_{ij} + (y^\top \hat{\alpha} - l_q) Z'_{(n-l_q-1)j} & (p_{Z'_{:j}} < l_q + 1), \\ \sum_{i=n-u_q}^{n} Z'_{ij} + (P_\lambda(\hat{w}) - u_q) Z'_{(n-u_q-1)j} & (p_{Z'_{:j}} > u_q), \\ \sum_{i=1}^{n} \max\{0, Z'_{ij}\} & (\text{otherwise}). \end{cases}$$

On the other hand, from KKT condition(6),

$$\frac{1}{\lambda n} X_{:j}^\top \alpha^* \in \begin{cases} \frac{w_j^*}{|w_j^*|} & (w_j^* \ne 0) \\ [-1, 1] & (\text{otherwise}). \end{cases} \tag{21}$$

Therefore, if $LB(X_{:j}^\top \alpha^*) < -\lambda n$ and $UB(X_{:j}^\top \alpha^*) > \lambda n$ then $w_j^* = 0$. ■

### A.5. Proof of Theorem 7

First, we construct the primal optimal solution region $\Theta_{w^*}$.

**Lemma 9.** *The primal optimal solution region $\Theta_{w^*}$ is given $\forall \hat{w} \in \mathrm{dom}P_\lambda$ as*

$$\Theta_{w^*} = \left\{ w \in \mathrm{dom}P_\lambda \;\middle|\; \lambda\|w\|_1 + g_\ell(\hat{w})^\top w \le k \right\}, \tag{22}$$

*where $g_\ell(w) := \frac{1}{n}\sum_{i \in [n]} g_{\ell_i}(w)$.*

*Proof.* From Proposition B.24 in (**?**),

$$(\lambda g_\psi(w^*) + g_\ell(w^*))^\top (w^* - \hat{w}) \le 0, \forall \hat{w} \in \mathrm{dom}P_\lambda,$$

where $g_\psi(w) \in \partial\psi(w)$. Form the convexity of $\ell_i$ for $i \in [n]$ and the definition of subgradient

$$\ell_i(w^*) \ge \ell_i(\hat{w}) + g_{\ell_i}(\hat{w})(w^* - \hat{w}), \forall \hat{w} \in \mathrm{dom}P_\lambda$$
$$\ell_i(\hat{w}) \ge \ell_i(w^*) + g_{\ell_i}(w^*)(\hat{w} - w^*), \forall \hat{w} \in \mathrm{dom}P_\lambda,$$

and thus, $g_{\ell_i}(w^*)^\top(w^* - \hat{w}) \geq g_{\ell_i}(\hat{w})^\top(w^* - \hat{w}), \forall \hat{w} \in \mathrm{dom}P_\lambda$. Therefore, $\forall \hat{w} \in \mathrm{dom}P_\lambda$,

$$\lambda g_\psi(w^*)^\top w^* + g_\ell(\hat{w})^\top w^* \leq \lambda g_\psi(w^*)^\top \hat{w} + g_\ell(\hat{w})^\top \hat{w}.$$

Since $g_\psi(\hat{w})^\top \hat{w} = \|\hat{w}\|_1 = \max_{s \in [-1,1]^d} s^\top \hat{w}$ and $g_\psi(w^*) \in [-1,1]^d$, we have

$$\lambda g_\psi(w^*)^\top \hat{w} \leq \lambda g_\psi(\hat{w})^\top \hat{w}.$$

By combining these results,

$$\lambda\|w^*\|_1 + g_\ell(\hat{w})^\top w^* \leq k, \ \ \forall \hat{w} \in \mathrm{dom}P_\lambda.$$

■

*Proof of Theorem 7.* From Lemma 9,

$$x_i^\top w^* \geq LB(x_i^\top w^*) := \min_{w \in \Theta_{w^*}} x_i^\top w.$$

Using a Lagrange multiplier $\mu > 0$,

$$\begin{aligned} LB(x_i^\top w^*) &= \min x_i^\top w \ \text{ s.t } \ w \in \Theta_{w^*} \qquad (23)\\ &= \min_{w} \max_{\mu > 0} \{x_i^\top w + \mu(\lambda\|w\|_1 + g_\ell(\hat{w})^\top w - k)\}\\ &= \max_{\mu > 0} \{\mu k + \min(\underbrace{x_i^\top w + \mu\lambda\|w\|_1 + \mu g_\ell(\hat{w})^\top w}_{L(w)})\} \end{aligned}$$

Since $0 \in \partial L$, which is written as $\partial L = x_i + \mu\lambda\partial\psi(w) + \mu g_\ell(\hat{w})$, we have

$$\mu\lambda g_\psi(w) = -x_i - \mu g_\ell(\hat{w}) \tag{24}$$

Substituting $\mu\lambda\|w\| = -x_i^\top w - \mu g_\ell(\hat{w})^\top w$ into (23),

$$\begin{aligned} LB(x_i^\top w^*) &= \max_{\mu > 0} \{\mu k\}\\ \text{s.t. } &\| -\frac{1}{\lambda} x_i^\top w - \frac{\mu}{\lambda} g_\ell(\hat{w})^\top w\|_\infty \leq \mu, \end{aligned}$$

where the constraint comes from (24). Similarly, since

$$x_i^\top w^* \leq UB(x_i^\top w^*) := \max_{w \in \Theta_{w^*}} x_i^\top w = -\min_{w \in \Theta_{w^*}} x_i^\top w,$$

$$\begin{aligned} UB(x_i^\top w^*) &= \max_{\mu > 0} \{\mu k\}\\ \text{s.t. } &\| \frac{1}{\lambda} x_i^\top w - \frac{\mu}{\lambda} g_\ell(\hat{w})^\top w\|_\infty \leq \mu. \end{aligned}$$

■

## B. Safe keeping by using KKT optimality conditions

In this appendix, we describe another type of safe keeping approaches based on KKT optimality conditions.

**Theorem 10.** *For an arbitrary pair of primal feasible solution $\hat{w} \in \mathrm{dom}P_\lambda$ and dual feasible solution $\hat{\alpha} \in \mathrm{dom}D_\lambda$,*

$$LB(X_{:j}^\top \alpha^*) < -\lambda n \text{ and } \lambda n < UB(X_{:j}^\top \alpha^*) \ \Rightarrow \ w_j^* \neq 0$$

*for $j \in [d]$, where*

$$\begin{aligned} LB(X_{:j}^\top \alpha^*) &:= X_{:j}^\top \hat{\alpha} - \|X_{:j}\|_2 \sqrt{2nG_\lambda(\hat{w}, \hat{\alpha})/\gamma},\\ UB(X_{:j}^\top \alpha^*) &:= X_{:j}^\top \hat{\alpha} + \|X_{:j}\|_2 \sqrt{2nG_\lambda(\hat{w}, \hat{\alpha})/\gamma}. \end{aligned}$$

*Proof.* In the case that $D_\lambda$ is $\gamma/n$-strongly concave, $X_{:j}^\top \alpha^*$ is bounded from below and above respectively by the following lower and upper bounds:

$$\begin{aligned} LB(X_{:j}^\top \alpha^*) &:= X_{:j}^\top \hat{\alpha} - \|X_{:j}\|_2 \sqrt{2nG_\lambda(\hat{w}, \hat{\alpha})/\gamma},\\ UB(X_{:j}^\top \alpha^*) &:= X_{:j}^\top \hat{\alpha} + \|X_{:j}\|_2 \sqrt{2nG_\lambda(\hat{w}, \hat{\alpha})/\gamma}. \end{aligned}$$

On the other hand, in the case of our specific regularization term (2), from KKT optimality condition (6), if $-\lambda n < X_{:j}^\top \alpha^* < \lambda n$ then $w_j^* \neq 0$.

Therefore,

$$LB(X_{:j}^\top \alpha^*) < -\lambda n \text{ and } \lambda n < UB(X_{:j}^\top \alpha^*) \ \Rightarrow \ w_j^* \neq 0$$

■

Similarly, we can develop safe sample keeping based on KKT optimality condition.

**Theorem 11.** *For an arbitrary pair of primal feasible solution $\hat{w} \in \mathrm{dom}P_\lambda$ and dual feasible solution $\hat{\alpha} \in \mathrm{dom}D_\lambda$, if $\ell_i$ is smoothed hinge loss then, for $y_i = +1$,*

$$1 - \gamma < LB(x_i^\top w^*) \text{ and } UB(x_i^\top w^*) < 1 \ \Rightarrow \ \alpha_i^* \notin \{0, +1\},$$

*and, for $y_i = -1$,*

$$-1 < LB(x_i^\top w^*) \text{ and } UB(x_i^\top w^*) < \gamma - 1 \ \Rightarrow \ \alpha_i^* \notin \{-1, 0\}.$$

*If $\ell_i$ is smoothed $\varepsilon$-insensitive loss then*

$$\begin{gathered} -\gamma + y_i - \varepsilon < LB(x_i^\top w^*) \text{ and } UB(x_i^\top w^*) < y_i - \varepsilon\\ \text{or}\\ y_i + \varepsilon < LB(x_i^\top w^*) \text{ and } UB(x_i^\top w^*) < \gamma + y_i + \varepsilon\\ \Rightarrow \ \alpha_i^* \notin \{-1, 0, +1\}, \end{gathered}$$

*for $j \in [d]$, where*

$$\begin{aligned} LB(x_i^\top w^*) &= x_i^\top \hat{w} - \|x_i\|_2 \sqrt{2G_\lambda(\hat{w}, \hat{\alpha})/\lambda},\\ UB(x_i^\top w^*) &= x_i^\top \hat{w} + \|x_i\|_2 \sqrt{2G_\lambda(\hat{w}, \hat{\alpha})/\lambda}. \end{aligned}$$

*Proof.* In the case that $P_\lambda$ is $\lambda$-strongly convex, $x_i^\top w^*$ is bounded from below and above respectively by the following lower and upper bounds:

$$\begin{aligned} LB(x_i^\top w^*) &= x_i^\top \hat{w} - \|x_i\|_2 \sqrt{2G_\lambda(\hat{w},\hat{\alpha})/\lambda}, \\ UB(x_i^\top w^*) &= x_i^\top \hat{w} + \|x_i\|_2 \sqrt{2G_\lambda(\hat{w},\hat{\alpha})/\lambda}. \end{aligned}$$

On the other hand, from KKT optimality condition (6), in the case of smoothed hinge loss (3), if $y_i = +1$ and $1 - \gamma < x_i^\top w^* < 1$ then $\alpha_i^* \in \{0, +1\}$, if $y_i = 1$ and $-1 < x_i^\top w^* < \gamma - 1$ then $\alpha_i^* \in \{-1, 0\}$. Therefore,

$$\begin{aligned} & y_i = +1 \text{ and } 1 - \gamma < LB(x_i^\top w^*) \text{ and } UB(x_i^\top w^*) < 1 \\ & \quad \Rightarrow \ \alpha_i^* \notin \{0, +1\}, \\ & y_i = -1 \text{ and } -1 < LB(x_i^\top w^*) \text{ and } UB(x_i^\top w^*) < \gamma - 1 \\ & \quad \Rightarrow \ \alpha_i^* \notin \{-1, 0\}. \end{aligned}$$

Also, in the case of smoothed $\varepsilon$-insensitive (4), if $-\gamma + y_i - \varepsilon < x_i^\top w^* < y_i - \varepsilon$ or $y_i + \varepsilon < x_i^\top w^* < \gamma + y_i + \varepsilon$ then, $\alpha_i^* \notin \{-1, 0, +1\}$ . Therefore,

$$\begin{aligned} & -\gamma + y_i - \varepsilon < LB(x_i^\top w^*) \text{ and } UB(x_i^\top w^*) < y_i - \varepsilon \\ & \qquad\qquad \text{or} \\ & y_i + \varepsilon < LB(x_i^\top w^*) \text{ and } UB(x_i^\top w^*) < \gamma + y_i + \varepsilon \\ & \qquad \Rightarrow \ \alpha_i^* \notin \{-1, 0, +1\}, \end{aligned}$$

■

## C. Other experiments

In this appendix, we show the rest of the experimental results.

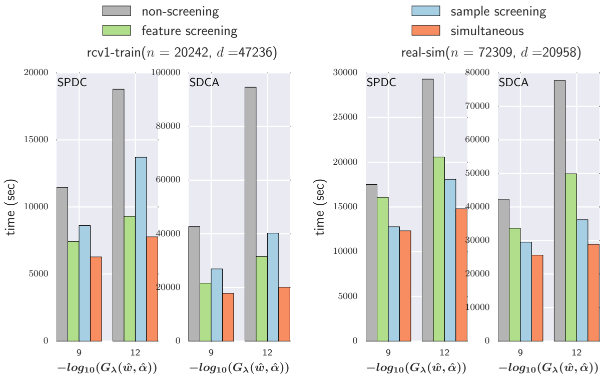

*Figure 5.* Total computation time for training 100 solutions for various values of $\lambda$ in regression problems.

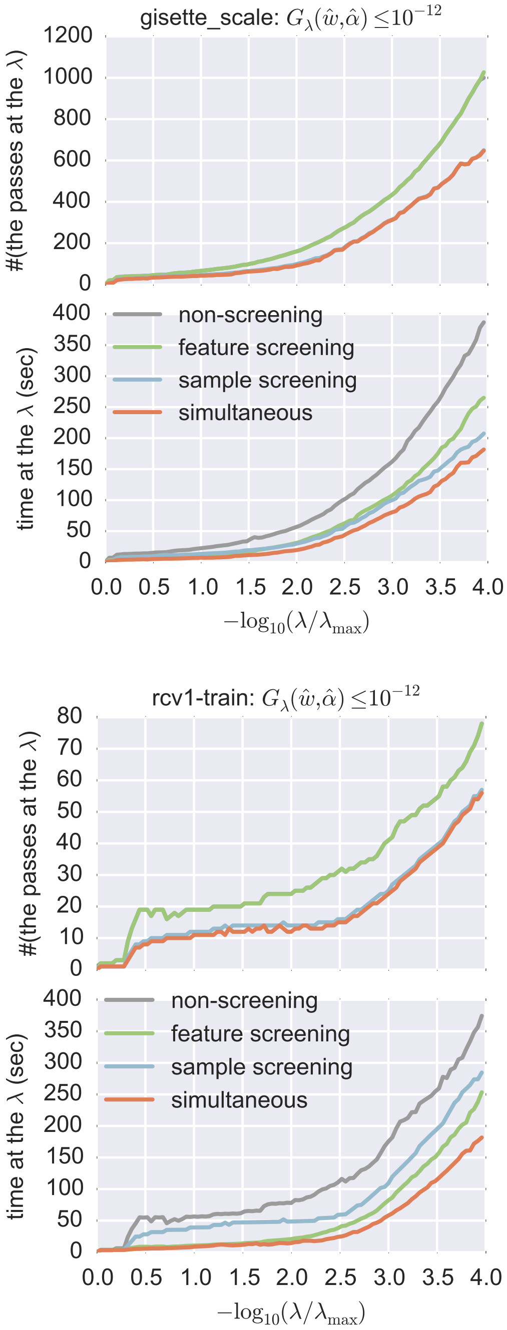

*Figure 6.* Number of optimization steps and computation time in classification problems (`rcv1-train` and `rcv1-train` datasets). See the caption in Figure 4.

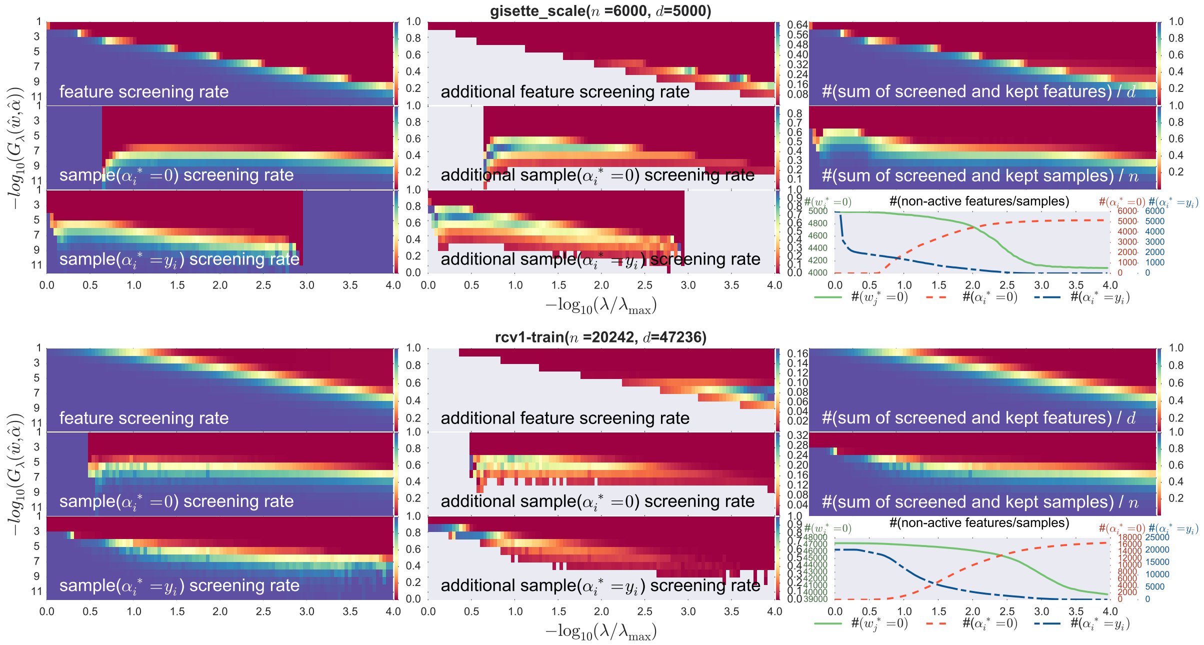

*Figure 7.* Safe screening and keeping rates in classification problems (`rcv1-train` and `rcv1-train` datasets). See the caption in Figure 2.

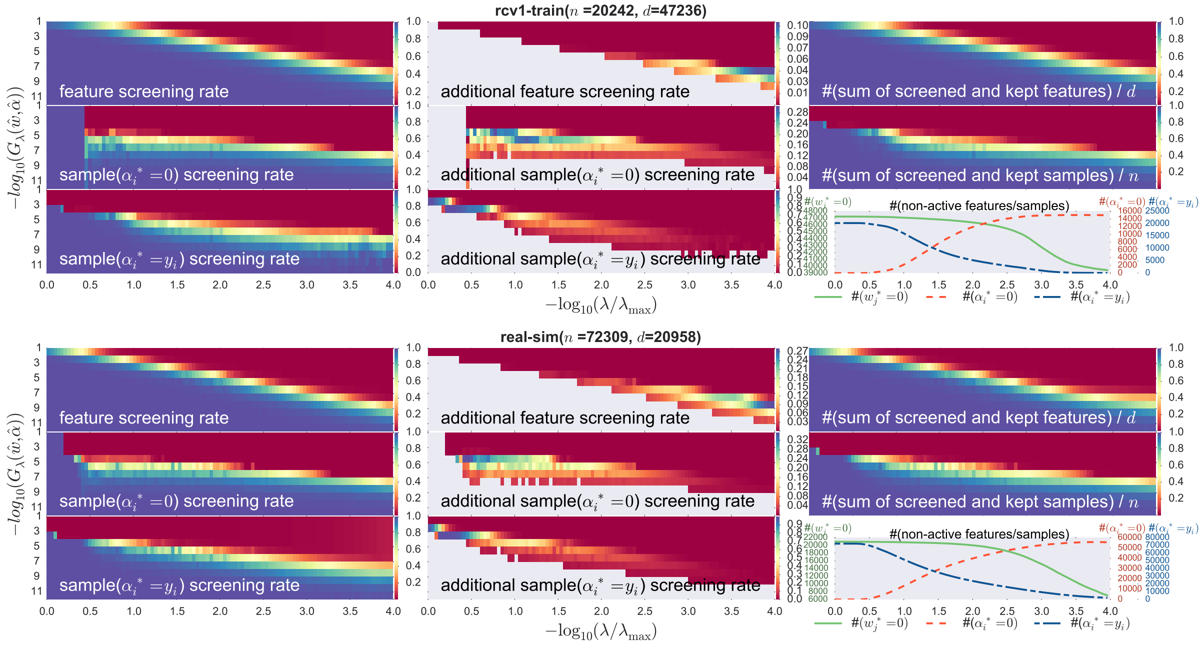

*Figure 8.* Safe screening and keeping rates in regression problems. See the caption in Figure 2.