Supplemental Material: 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]

[1]State Key Lab of CAD&CG, Zhejiang University, China
[2] Tencent AI Lab, Shenzhen, China, [3] University of Michigan, USA

In this supplement, we first present the detailed proofs of all the theorems in the main text and then report the rest experiment results which are omitted in the experiment section due to the space limitation.

A. Proof for Theorem 1

Proof. of Theorem 1:
(i) : Let $\bar{\mathbf{X}} = (\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, ..., \bar{\mathbf{x}}_n)$ and $\mathbf{z} = \mathbf{1} - \bar{\mathbf{X}}^T\mathbf{w}$, the primal problem (P*) then is equivalent to

$$\min_{\mathbf{w}\in\mathbb{R}^p, \mathbf{z}\in\mathbb{R}^n} \frac{\alpha}{2}||\mathbf{w}||^2 + \beta||\mathbf{w}||_1 + \frac{1}{n}\sum_{i=1}^n \ell([\mathbf{z}]_i),$$
$$\text{s.t. } \mathbf{z} = \mathbf{1} - \bar{\mathbf{X}}^T\mathbf{w}.$$

The Lagrangian then becomes

$$L(\mathbf{w}, \mathbf{z}, \theta) = \frac{\alpha}{2}||\mathbf{w}||^2 + \beta||\mathbf{w}||_1 + \frac{1}{n}\sum_{i=1}^n \ell([\mathbf{z}]_i) + \frac{1}{n}\langle \mathbf{1} - \bar{\mathbf{X}}^T\mathbf{w} - \mathbf{z}, \theta\rangle \tag{17}$$

$$= \underbrace{\frac{\alpha}{2}||\mathbf{w}||^2 + \beta||\mathbf{w}||_1 - \frac{1}{n}\langle\bar{\mathbf{X}}\theta, \mathbf{w}\rangle}_{:=f_1(\mathbf{w})} + \underbrace{\frac{1}{n}\sum_{i=1}^n \ell([\mathbf{z}]_i) - \frac{1}{n}\langle \mathbf{z}, \theta\rangle}_{:=f_2(\mathbf{z})} + \frac{1}{n}\langle \mathbf{1}, \theta\rangle \tag{18}$$

We first consider the subproblem $\min_{\mathbf{w}} L(\mathbf{w}, \mathbf{z}, \theta)$:

$$0 \in \partial_{\mathbf{w}} L(\mathbf{w}, \mathbf{z}, \theta) = \partial_{\mathbf{w}} f_1(\mathbf{w}) = \alpha\mathbf{w} - \frac{1}{n}\bar{\mathbf{X}}\theta + \beta\partial||\mathbf{w}||_1 \Leftrightarrow$$
$$\frac{1}{n}\bar{\mathbf{X}}\theta \in \alpha\mathbf{w} + \beta\partial||\mathbf{w}||_1 \Rightarrow \mathbf{w} = \frac{1}{\alpha}\mathcal{S}_\beta(\frac{1}{n}\bar{\mathbf{X}}\theta) \tag{19}$$

By substituting (19) into $f_1(\mathbf{w})$, we get

$$f_1(\mathbf{w}) = \frac{\alpha}{2}||\mathbf{w}||^2 + \beta||\mathbf{w}||_1 - \langle\alpha\mathbf{w} + \beta\partial||\mathbf{w}||_1, \mathbf{w}\rangle = -\frac{\alpha}{2}||\mathbf{w}||^2 = -\frac{1}{2\alpha}||\mathcal{S}_\beta(\frac{1}{n}\bar{\mathbf{X}}\theta)||^2. \tag{20}$$

Then, we consider the problem $\min_{\mathbf{z}} L(\mathbf{w}, \mathbf{z}, \theta)$:

$$0 = \nabla_{[\mathbf{z}]_i} L(\mathbf{w}, \mathbf{z}, \theta) = \nabla_{[\mathbf{z}]_i} f_2(\mathbf{z}) = \begin{cases} -\frac{1}{n}[\theta]_i, & \text{if } [\mathbf{z}]_i < 0, \\ \frac{1}{\gamma n}[\mathbf{z}]_i - \frac{1}{n}[\theta]_i, & \text{if } 0 \le [\mathbf{z}]_i \le \gamma, \\ \frac{1}{n} - \frac{1}{n}[\theta]_i, & \text{if } [\mathbf{z}]_i > \gamma. \end{cases}$$
$$\Rightarrow [\theta]_i = \begin{cases} 0, & \text{if } [\mathbf{z}]_i < 0, \\ \frac{1}{\gamma}[\mathbf{z}]_i, & \text{if } 0 \le [\mathbf{z}]_i \le \gamma, \\ 1, & \text{if } [\mathbf{z}]_i > \gamma. \end{cases} \tag{21}$$

Thus, we have

$$f_2(\mathbf{z}) = \begin{cases} -\frac{\gamma}{2n}||\theta||^2, & \text{if } [\theta]_i \in [0,1], \forall i \in [n], \\ -\infty, & \text{otherwise .} \end{cases} \tag{22}$$

Combining Eq. (17), Eq. (20) and Eq. (22), we obtain the dual problem:

$$\min_{\theta \in [0,1]^n} \frac{1}{2\alpha}||\mathcal{S}_\beta(\frac{1}{n}\bar{\mathbf{X}}\theta)||^2 + \frac{\gamma}{2n}||\theta||^2 - \frac{1}{n}\langle \mathbf{1}, \theta \rangle \tag{23}$$

(ii) : From Eq. (19) and Eq. (21), we get the KKT conditions:

$$\mathbf{w}^*(\alpha,\beta) = \frac{1}{\alpha}\mathcal{S}_\beta(\frac{1}{n}\bar{\mathbf{X}}^T\theta^*(\alpha,\beta))$$

$$[\theta^*(\alpha,\beta)]_i = \begin{cases} 0, & \text{if } 1 - \langle \bar{\mathbf{x}}_i, \mathbf{w}^*(\alpha,\beta)\rangle < 0, \\ \frac{1}{\gamma}(1 - \langle \bar{\mathbf{x}}_i, \mathbf{w}^*(\alpha,\beta)\rangle), & \text{if } 0 \le 1 - \langle \bar{\mathbf{x}}_i, \mathbf{w}^*(\alpha,\beta)\rangle \le \gamma, \\ 1, & \text{if } 1 - \langle \bar{\mathbf{x}}_i, \mathbf{w}^*(\alpha,\beta)\rangle > \gamma. \end{cases} \quad i = 1, ..., n.$$

The proof is complete. □

B. Proof for Lemma 1

Proof. of Theorem 1:
1) It is the conclusion of the analysis above.
2) After feature screening, the primal problem (P*) is scaled into:

$$\min_{\tilde{\mathbf{w}} \in \mathbb{R}^{|\hat{\mathcal{F}}^c|}} \frac{\alpha}{2}||\tilde{\mathbf{w}}||^2 + \beta||\tilde{\mathbf{w}}||_1 + \frac{1}{n}\sum_{i=1}^{n} \ell(1 - \langle [\bar{\mathbf{x}}_i]_{\hat{\mathcal{F}}^c}, \tilde{\mathbf{w}} \rangle), \tag{scaled-P^*-1}$$

Thus, we can easily derive out the dual problem of (scaled-P^*-1):

$$\min_{\tilde{\theta} \in [0,\alpha]^n} \tilde{D}(\tilde{\theta}; \alpha, \beta) = \frac{1}{2\alpha}||\mathcal{S}_\beta(\frac{1}{n}_{\hat{\mathcal{F}}^c}[\bar{\mathbf{X}}]\tilde{\theta})||^2 + \frac{\gamma}{2n}||\tilde{\theta}||^2 - \frac{1}{n}\langle \mathbf{1}, \tilde{\theta} \rangle. \tag{scaled-D^*-1}$$

and also the KKT conditions:

$$\tilde{\mathbf{w}}^*(\alpha,\beta) = \frac{1}{\alpha}\mathcal{S}_\beta(\frac{1}{n}_{\hat{\mathcal{F}}^c}[\bar{\mathbf{X}}]\tilde{\theta}^*(\alpha,\beta)) \tag{scaled-KKT-1}$$

$$[\tilde{\theta}^*(\alpha,\beta)]_i = \begin{cases} 0, & \text{if } 1 - \langle [\bar{\mathbf{x}}_i]_{\hat{\mathcal{F}}^c}, \tilde{\mathbf{w}}^*(\alpha,\beta)\rangle < 0, \\ \frac{1}{\gamma}(1 - \langle [\bar{\mathbf{x}}_i]_{\hat{\mathcal{F}}^c}, \tilde{\mathbf{w}}^*(\alpha,\beta)\rangle), & \text{if } 0 \le 1 - \langle [\bar{\mathbf{x}}_i]_{\hat{\mathcal{F}}^c}, \tilde{\mathbf{w}}^*(\alpha,\beta) \le \gamma, \\ 1, & \text{if } 1 - \langle [\bar{\mathbf{x}}_i]_{\hat{\mathcal{F}}^c}, \tilde{\mathbf{w}}^*(\alpha,\beta) > \gamma, \end{cases} \tag{scaled-KKT-2}$$

Then, it is obvious that $\tilde{\mathbf{w}}^*(\alpha,\beta) = [\mathbf{w}^*(\alpha,\beta)]_{\hat{\mathcal{F}}^c}$, since essentially, problem (scaled-P^*-1) can be derived by substituting 0 to the weights for the eliminated features in problem (P*) and optimize over the rest weights.
Since the solutions $\mathbf{w}^*(\alpha,\beta)$ and $\theta^*(\alpha,\beta)$ satisfy the conditions KKT-1 and KKT-2 and $\langle [\bar{\mathbf{x}}_i]_{\hat{\mathcal{F}}^c}, \tilde{\mathbf{w}}^*(\alpha,\beta)\rangle = \langle \bar{\mathbf{x}}_i, \mathbf{w}^*(\alpha,\beta)\rangle$ for all i , we know $\tilde{\mathbf{w}}^*(\alpha,\beta)$ and $\theta^*(\alpha,\beta)$ satisfy the conditions scaled-KKT-1 and scaled-KKT-2. So they are the solutions of problems (scaled-P^*-1) and (scaled-D^*-1). Thus, due to the uniqueness of the solution of problem (scaled-D^*-1), we have

$$\theta^*(\alpha,\beta) = \tilde{\theta}^*(\alpha,\beta) \tag{24}$$

From 1) we have, $[\tilde{\theta}^*(\alpha,\beta)]_{\hat{\mathcal{R}}^c} = 0$ and $[\tilde{\theta}^*(\alpha,\beta)]_{\hat{\mathcal{L}}^c} = 1$. Therefore, from the dual problem (scaled-D*), we can see that $[\tilde{\theta}^*(C,\alpha)]_{\hat{\mathcal{D}}^c}$ can be recovered from the following problem:

$$\min_{\hat{\theta} \in [0,1]^{|\hat{\mathcal{D}}^c|}} \frac{1}{2\alpha}||\mathcal{S}_\beta(\frac{1}{n}\hat{\mathbf{G}}_1\hat{\theta} + \frac{1}{n}\hat{\mathbf{G}}_2\mathbf{1})||^2 + \frac{\gamma}{2n}||\hat{\theta}||^2 - \frac{1}{n}\langle \mathbf{1}, \hat{\theta} \rangle,$$

Since $[\tilde{\theta}^*(\alpha,\beta)]_{\hat{\mathcal{D}}^c} = [\theta^*(\alpha,\beta)]_{\hat{\mathcal{D}}^c}$, the proof is therefore completed. □

C. Proof for Lemma 2

Proof. Due to the α-strong convexity of the objective $P(\mathbf{w};\alpha,\beta)$, we have

$$P(\mathbf{w}^*(\alpha_0,\beta_0);\alpha,\beta_0) \geq P(\mathbf{w}^*(\alpha,\beta_0);\alpha,\beta_0) + \frac{\alpha}{2}||\mathbf{w}^*(\alpha_0,\beta_0) - \mathbf{w}^*(\alpha,\beta_0)||^2$$

$$P(\mathbf{w}^*(\alpha,\beta_0);\alpha_0,\beta_0) \geq P(\mathbf{w}^*(\alpha_0,\beta_0);\alpha_0,\beta_0) + \frac{\alpha_0}{2}||\mathbf{w}^*(\alpha_0,\beta_0) - \mathbf{w}^*(\alpha,\beta_0)||^2$$

which are equivalent to

$$\begin{aligned}
&\frac{\alpha}{2}||\mathbf{w}^*(\alpha_0,\beta_0)||^2 + \beta_0||\mathbf{w}^*(\alpha_0,\beta_0)||_1 + \frac{1}{n}\sum_{i=1}^n \ell(1-\langle\bar{\mathbf{x}}_i,\mathbf{w}^*(\alpha_0,\beta_0)\rangle)\\
\geq&\frac{\alpha}{2}||\mathbf{w}^*(\alpha,\beta_0)||^2 + \beta_0||\mathbf{w}^*(\alpha,\beta_0)||_1 + \frac{1}{n}\sum_{i=1}^n \ell(1-\langle\bar{\mathbf{x}}_i,\mathbf{w}^*(\alpha,\beta_0)\rangle)\\
&+\frac{\alpha}{2}||\mathbf{w}^*(\alpha_0,\beta_0) - \mathbf{w}^*(\alpha,\beta_0)||^2\\
&\frac{\alpha_0}{2}||\mathbf{w}^*(\alpha,\beta_0)||^2 + \beta_0||\mathbf{w}^*(\alpha,\beta_0)||_1 + \frac{1}{n}\sum_{i=1}^n \ell(1-\langle\bar{\mathbf{x}}_i,\mathbf{w}^*(\alpha,\beta_0)\rangle)\\
\geq&\frac{\alpha_0}{2}||\mathbf{w}^*(\alpha_0,\beta_0)||^2 + \beta_0||\mathbf{w}^*(\alpha_0,\beta_0)||_1 + \frac{1}{n}\sum_{i=1}^n \ell(1-\langle\bar{\mathbf{x}}_i,\mathbf{w}^*(\alpha_0,\beta_0)\rangle)\\
&+\frac{\alpha_0}{2}||\mathbf{w}^*(\alpha_0,\beta_0) - \mathbf{w}^*(\alpha,\beta_0)||^2
\end{aligned}$$

Adding the above two inequalities together, we get

$$\begin{aligned}
&\frac{\alpha-\alpha_0}{2}||\mathbf{w}^*(\alpha_0,\beta_0)||^2 \geq \frac{\alpha-\alpha_0}{2}||\mathbf{w}^*(\alpha,\beta_0)||^2 + \frac{\alpha_0+\alpha}{2}||\mathbf{w}^*(\alpha_0,\beta_0) - \mathbf{w}^*(\alpha,\beta_0)||^2\\
\Rightarrow&||\mathbf{w}^*(\alpha,\beta_0) - \frac{\alpha_0+\alpha}{2\alpha}\mathbf{w}^*(\alpha_0,\beta_0)||^2 \leq \frac{(\alpha-\alpha_0)^2}{4\alpha^2}||\mathbf{w}^*(\alpha_0,\beta_0)||^2
\end{aligned} \tag{25}$$

Substitute the prior that $[\mathbf{w}^*(\alpha,\beta_0)]_{\hat{\mathcal{F}}} = 0$ into (25), we get

$$\begin{aligned}
&||[\mathbf{w}^*(\alpha,\beta_0)]_{\hat{\mathcal{F}}^c} - \frac{\alpha_0+\alpha}{2\alpha}[\mathbf{w}^*(\alpha_0,\beta_0)]_{\hat{\mathcal{F}}^c}||^2\\
\leq&\frac{(\alpha-\alpha_0)^2}{4\alpha^2}||\mathbf{w}^*(\alpha_0,\beta_0)||^2 - \frac{(\alpha_0+\alpha)^2}{4\alpha^2}||[\mathbf{w}^*(\alpha_0,\beta_0)]_{\hat{\mathcal{F}}}||^2.
\end{aligned}$$

The proof is complete. □

D. Proof for Lemma 3

Proof. Firstly, we need to extend the definition of $D(\theta;\alpha,\beta)$ to $\mathbb{R}^n$:

$$\tilde{D}(\theta;\alpha,\beta) = \begin{cases} D(\theta;\alpha,\beta), & \text{if } \theta \in [0,1]^n,\\ +\infty, & \text{otherwise} \end{cases} \tag{26}$$

Due to the strong convexity of objective $\tilde{D}(\theta;\alpha,\beta)$, we have

$$\begin{aligned}
\tilde{D}(\theta^*(\alpha_0,\beta_0),\alpha,\beta_0) &\geq \tilde{D}(\theta^*(\alpha,\beta_0),\alpha,\beta_0) + \frac{\gamma}{2n}||\theta^*(\alpha_0,\beta_0) - \theta^*(\alpha,\beta_0)||^2,\\
\tilde{D}(\theta^*(\alpha,\beta_0),\alpha_0,\beta_0) &\geq \tilde{D}(\theta^*(\alpha_0,\beta_0),\alpha_0,\beta_0) + \frac{\gamma}{2n}||\theta^*(\alpha_0,\beta_0) - \theta^*(\alpha,\beta_0)||^2.
\end{aligned}$$

Since $\theta^*(\alpha_0,\beta_0), \theta^*(\alpha,\beta_0) \in [0,1]^n$, the above inequalities are equivalent to

$$
\begin{aligned}
&\frac{1}{2\alpha}||\mathcal{S}_{\beta_0}(\frac{1}{n}\bar{\mathbf{X}}^T\theta^*(\alpha_0,\beta_0))||^2+\frac{\gamma}{2n}||\theta^*(\alpha_0,\beta_0)||^2-\frac{1}{n}\langle\mathbf{1},\theta^*(\alpha_0,\beta_0)\rangle\\
\geq&\frac{1}{2\alpha}||\mathcal{S}_{\beta_0}(\frac{1}{n}\bar{\mathbf{X}}^T\theta^*(\alpha,\beta_0))||^2+\frac{\gamma}{2n}||\theta^*(\alpha,\beta_0)||^2-\frac{1}{n}\langle\mathbf{1},\theta^*(\alpha,\beta_0)\rangle\\
&+\frac{\gamma}{2n}||\theta^*(\alpha_0,\beta_0)-\theta^*(\alpha,\beta_0)||^2,\\
&\frac{1}{2\alpha_0}||\mathcal{S}_{\beta_0}(\frac{1}{n}\bar{\mathbf{X}}^T\theta^*(\alpha,\beta_0))||^2+\frac{\gamma}{2n}||\theta^*(\alpha,\beta_0)||^2-\frac{1}{n}\langle\mathbf{1},\theta^*(\alpha,\beta_0)\rangle\\
\geq&\frac{1}{2\alpha_0}||\mathcal{S}_{\beta_0}(\frac{1}{n}\bar{\mathbf{X}}^T\theta^*(\alpha_0,\beta_0))||^2+\frac{\gamma}{2n}||\theta^*(\alpha_0,\beta_0)||^2-\frac{1}{n}\langle\mathbf{1},\theta^*(\alpha_0,\beta_0)\rangle+\frac{\gamma}{2n}||\theta^*(\alpha_0,\beta_0)-\theta^*(\alpha,\beta_0)||^2.
\end{aligned}
$$

Adding the above two inequalities, we get

$$
\begin{aligned}
&\frac{\gamma(\alpha-\alpha_0)}{2n}||\theta^*(\alpha_0,\beta_0)||^2-\frac{\alpha-\alpha_0}{n}\langle\mathbf{1},\theta^*(\alpha_0,\beta_0)\\
\geq&\frac{\gamma(\alpha-\alpha_0)}{2n}||\theta^*(\alpha,\beta_0)||^2-\frac{\alpha-\alpha_0}{n}\langle\mathbf{1},\theta^*(\alpha,\beta_0)\rangle+\frac{\gamma(\alpha_0+\alpha)}{2n}||\theta^*(\alpha_0,\beta_0)-\theta^*(\alpha,\beta_0)||^2
\end{aligned}
$$

That is equivalent to

$$
\begin{aligned}
&||\theta^*(\alpha,\beta_0)||^2-\langle\frac{\alpha-\alpha_0}{\gamma\alpha}\mathbf{1}+\frac{\alpha_0+\alpha}{\alpha}\theta^*(\alpha_0,\beta_0),\theta^*(\alpha,\beta_0)\rangle\\
\leq&-\frac{\alpha_0}{\alpha}||\theta^*(\alpha_0,\beta_0)||^2-\frac{\alpha-\alpha_0}{\gamma\alpha}\langle\mathbf{1},\theta^*(\alpha_0,\beta_0)\rangle
\end{aligned}
\tag{27}
$$

That is

$$
||\theta^*(\alpha,\beta_0)-(\frac{\alpha-\alpha_0}{2\gamma\alpha}\mathbf{1}+\frac{\alpha_0+\alpha}{2\alpha}\theta^*(\alpha_0,\beta_0))||^2\leq(\frac{\alpha-\alpha_0}{2\alpha})^2||\theta^*(\alpha_0,\beta_0)-\frac{1}{\gamma}\mathbf{1}||^2 \tag{28}
$$

Substitute the priors that $[\theta^*(\alpha,\beta_0)]_{\hat{\mathcal{R}}}=0$ and $[\theta^*(\alpha,\beta_0)]_{\hat{\mathcal{L}}}=1$ into (28), we have

$$
\begin{aligned}
&||[\theta^*(\alpha,\beta_0)]_{\hat{\mathcal{D}}^c}-(\frac{\alpha-\alpha_0}{2\gamma\alpha}\mathbf{1}+\frac{\alpha_0+\alpha}{2\alpha}[\theta^*(\alpha_0,\beta_0)]_{\hat{\mathcal{D}}^c})||^2\\
\leq&(\frac{\alpha-\alpha_0}{2\alpha})^2||\theta^*(\alpha_0,\beta_0)-\frac{1}{\gamma}\mathbf{1}||^2-||\frac{(2\gamma-1)\alpha+\alpha_0}{2\gamma\alpha}\mathbf{1}-\frac{\alpha_0+\alpha}{2\alpha}[\theta^*(\alpha_0,\beta_0)]_{\hat{\mathcal{L}}}||^2\\
&-||\frac{\alpha-\alpha_0}{2\gamma\alpha}\mathbf{1}+\frac{\alpha_0+\alpha}{2\alpha}[\theta^*(\alpha_0,\beta_0)]_{\hat{\mathcal{R}}}||^2.
\end{aligned}
$$

The proof is complete. □

E. Proof for Lemma 4

Before the proof of Lemma 4, we should prove that the optimization problem in (1) is equivalent to

$$
s^i(\alpha,\beta_0)=\max_{\theta\in\Theta}\left\{\frac{1}{n}|\langle[\bar{\mathbf{x}}^i]_{\hat{\mathcal{D}}^c},\theta\rangle+\langle[\bar{\mathbf{x}}^i]_{\hat{\mathcal{L}}},\mathbf{1}\rangle|\right\},i\in\hat{\mathcal{F}}^c. \tag{29}
$$

To avoid notational confusion, we denote the feasible region Θ in (1) as $\tilde{\Theta}$. Then,

$$
\begin{aligned}
&\max_{\theta\in\tilde{\Theta}}\left\{\left|\frac{1}{n}[\bar{\mathbf{X}}\theta]_i\right|\right\}=\max_{\theta\in\Theta}\left\{\frac{1}{n}\left|\bar{\mathbf{x}}^i\theta\right|\right\}\\
=&\max_{\theta\in\tilde{\Theta}}\left\{\frac{1}{n}\left|[\bar{\mathbf{x}}^i]_{\hat{\mathcal{D}}^c}[\theta]_{\hat{\mathcal{D}}^c}+[\bar{\mathbf{x}}^i]_{\hat{\mathcal{L}}}[\theta]_{\hat{\mathcal{L}}}+[\bar{\mathbf{x}}^i]_{\hat{\mathcal{R}}}[\theta]_{\hat{\mathcal{R}}}\right|\right\}\\
=&\max_{\theta\in\Theta}\left\{\frac{1}{n}|\langle[\bar{\mathbf{x}}^i]_{\hat{\mathcal{D}}^c},[\theta]_{\hat{\mathcal{D}}^c}\rangle+\langle[\bar{\mathbf{x}}^i]_{\hat{\mathcal{L}}},\mathbf{1}\rangle|\right\}=s^i(\alpha,\beta_0).
\end{aligned}
$$

The last equation holds since $[\theta]_{\hat{\mathcal{L}}} = \mathbf{1}$, $[\theta]_{\hat{\mathcal{R}}} = 0$ and $[\theta_{\hat{\mathcal{D}}^c}] \in \Theta$.

Proof. of Lemma 4:

$$s^i(\alpha, \beta_0) = \max_{\theta \in B(\mathbf{c}, r)} \{\frac{1}{n} |\langle [\bar{\mathbf{x}}^i]_{\hat{\mathcal{D}}^c}, \theta \rangle + \langle [\bar{\mathbf{x}}^i]_{\hat{\mathcal{L}}}, \mathbf{1} \rangle|\}.$$

$$\begin{aligned} &= \max_{\eta \in B(\mathbf{0}, r)} \{\frac{1}{n} |\langle [\bar{\mathbf{x}}^i]_{\hat{\mathcal{D}}^c}, \mathbf{c} \rangle + \langle [\bar{\mathbf{x}}^i]_{\hat{\mathcal{L}}}, \mathbf{1} \rangle + \langle [\bar{\mathbf{x}}^i]_{\hat{\mathcal{D}}^c}, \eta \rangle|\} \\ &= \frac{1}{n} \big(|\langle [\bar{\mathbf{x}}^i]_{\hat{\mathcal{D}}^c}, \mathbf{c} \rangle + \langle [\bar{\mathbf{x}}^i]_{\hat{\mathcal{L}}}, \mathbf{1}| + \|[\bar{\mathbf{x}}^i]_{\hat{\mathcal{D}}^c}\| r \big) \end{aligned}$$

The last equality holds since $-\|[\bar{\mathbf{x}}^i]_{\hat{\mathcal{D}}^c}\| r \leq \langle [\bar{\mathbf{x}}^i]_{\hat{\mathcal{D}}^c}, \eta \rangle| \leq \|[\bar{\mathbf{x}}^i]_{\hat{\mathcal{D}}^c}\| r$. The proof is complete. □

F. Proof for Theorem 4

Proof. (1) It can be obtained from the the rule (R1).
(2) It is from the definition of $\hat{\mathcal{F}}$.

□

G. Proof for Lemma 5

Firstly, we need to point out that the optimization problems in (2) and (3) are equivalent to the problems:

$$u_i(\alpha, \beta_0) = \max_{\mathbf{w} \in \mathcal{W}} \{1 - \langle [\bar{\mathbf{x}}_i]_{\hat{\mathcal{F}}^c}, \mathbf{w} \rangle\}, i \in \hat{\mathcal{D}}^c, \tag{30}$$

$$l_i(\alpha, \beta_0) = \min_{\mathbf{w} \in \mathcal{W}} \{1 - \langle [\bar{\mathbf{x}}_i]_{\hat{\mathcal{F}}^c}, \mathbf{w} \rangle\}, i \in \hat{\mathcal{D}}^c \tag{31}$$

They follow from the fact that $[\mathbf{w}]_{\hat{\mathcal{F}}^c} \in \mathcal{W}$ and

$$\begin{aligned} &\{1 - \langle \mathbf{w}, \bar{\mathbf{x}}_i \rangle\} \\ =&\{1 - \langle [\mathbf{w}]_{\hat{\mathcal{F}}^c}, [\bar{\mathbf{x}}_i]_{\hat{\mathcal{F}}^c} \rangle - \langle [\mathbf{w}]_{\hat{\mathcal{F}}}, [\bar{\mathbf{x}}_i]_{\hat{\mathcal{F}}} \rangle\} \\ =&\{1 - \langle [\mathbf{w}]_{\hat{\mathcal{F}}^c}, [\bar{\mathbf{x}}_i]_{\hat{\mathcal{F}}^c} \rangle\} \text{ (since } [\mathbf{w}]_{\hat{\mathcal{F}}} = 0). \end{aligned}$$

Proof. of Lemma 5:

$$\begin{aligned} u_i(\alpha, \beta_0) &= \max_{\mathbf{w} \in B(\mathbf{c}, r)} \{1 - \langle [\bar{\mathbf{x}}_i]_{\hat{\mathcal{F}}^c}, \mathbf{w} \rangle\} \\ &= \max_{\eta \in B(\mathbf{0}, r)} \{1 - \langle [\bar{\mathbf{x}}_i]_{\hat{\mathcal{F}}^c}, \mathbf{c} \rangle - \langle [\bar{\mathbf{x}}_i]_{\hat{\mathcal{F}}^c}, \eta \rangle\} \\ &= 1 - \langle [\bar{\mathbf{x}}_i]_{\hat{\mathcal{F}}^c}, \mathbf{c} \rangle + \max_{\eta \in B(\mathbf{0}, r)} \{-\langle [\bar{\mathbf{x}}_i]_{\hat{\mathcal{F}}^c}, \eta \rangle\} \\ &= 1 - \langle [\bar{\mathbf{x}}_i]_{\hat{\mathcal{F}}^c}, \mathbf{c} \rangle + \|[\bar{\mathbf{x}}_i]_{\hat{\mathcal{F}}^c}\| r \end{aligned}$$

$$\begin{aligned} l_i(\alpha, \beta_0) &= \min_{\mathbf{w} \in B(\mathbf{c}, r)} \{1 - \langle [\bar{\mathbf{x}}_i]_{\hat{\mathcal{F}}^c}, \mathbf{w} \rangle\} \\ &= \min_{\eta \in B(\mathbf{0}, r)} \{1 - \langle [\bar{\mathbf{x}}_i]_{\hat{\mathcal{F}}^c}, \mathbf{c} \rangle - \langle [\bar{\mathbf{x}}_i]_{\hat{\mathcal{F}}^c}, \eta \rangle\} \\ &= 1 - \langle [\bar{\mathbf{x}}_i]_{\hat{\mathcal{F}}^c}, \mathbf{c} \rangle + \min_{\eta \in B(\mathbf{0}, r)} \{-\langle [\bar{\mathbf{x}}_i]_{\hat{\mathcal{F}}^c}, \eta \rangle\} \\ &= 1 - \langle [\bar{\mathbf{x}}_i]_{\hat{\mathcal{F}}^c}, \mathbf{c} \rangle - \|[\bar{\mathbf{x}}_i]_{\hat{\mathcal{F}}^c}\| r \end{aligned}$$

The proof is complete. □

H. Proof for Theorem 5

Proof. (1) It can be obtained from the the rule (R2).

(2) It is from the definitions of $\hat{\mathcal{R}}$ and $\hat{\mathcal{L}}$. □

I. Proof for Theorem 2

Proof. of Theorem 2:
We prove this theorem by verifying that the solutions $\mathbf{w}^*(\alpha,\beta)=\mathbf{0}$ and $\theta^*(\alpha,\beta)=\mathbf{1}$ satisfy the conditions KKT-1 and KKT-2.
Firstly, since $\beta\geq\beta_{\max}=||\frac{1}{n}\bar{\mathbf{X}}\mathbf{1}||_\infty$, we have $\mathcal{S}_\beta(\frac{1}{n}\bar{\mathbf{X}}\mathbf{1})=0$. Thus $\mathbf{w}^*(\alpha,\beta)=\mathbf{0}$ and $\theta^*(\alpha,\beta)=\mathbf{1}$ satisfy the condition KKT-1.
Then, for all $i\in[n]$, we have

$$1-\langle\bar{\mathbf{x}}_i,\mathbf{w}^*(\alpha,\beta)\rangle=1-0>\gamma.$$

Thus $\mathbf{w}^*(\alpha,\beta)=\mathbf{0}$ and $\theta^*(\alpha,\beta)=\mathbf{1}$ satisfy the condition KKT-2. Hence, they are the solutions for the primal problem (P*) and the dual problem (D*), respectively. □

J. Proof for Theorem 3

Proof. of Theorem 3:
Similar with the proof of Theorem 2, we prove this theorem by verifying that the solutions $\mathbf{w}^*(\alpha,\beta)=\frac{1}{\alpha}\mathcal{S}_\beta(\frac{1}{n}\bar{\mathbf{X}}\theta^*(\alpha,\beta))$ and $\theta^*(\alpha,\beta)=\mathbf{1}$ satisfy the conditions KKT-1 and KKT-2.

1. **Case 1:** $\alpha_{\max}(\beta)\leq 0$. Then for all $\alpha>0$, we have

$$\begin{aligned}
&\min_{i\in[n]}\{1-\langle\bar{\mathbf{x}}_i,\mathbf{w}^*(\alpha,\beta)\rangle\}\\
=&\min_{i\in[n]}\{1-\frac{1}{\alpha}\langle\bar{\mathbf{x}}_i,\mathcal{S}_\beta(\frac{1}{n}\bar{\mathbf{X}}\theta^*(\alpha,\beta))\rangle\}=\min_{i\in[n]}\{1-\frac{1}{\alpha}\langle\bar{\mathbf{x}}_i,\mathcal{S}_\beta(\frac{1}{n}\bar{\mathbf{X}}\mathbf{1})\rangle\}\\
=&1-\frac{1}{\alpha}\max_{i\in[n]}\langle\bar{\mathbf{x}}_i,\mathcal{S}_\beta(\frac{1}{n}\bar{\mathbf{X}}\mathbf{1})\rangle=1-(1-\gamma)\frac{1}{\alpha}\alpha_{\max}(\beta)\\
\geq&1>\gamma
\end{aligned}$$

   Then, $\mathcal{L}=[n]$ and $\mathbf{w}^*(\alpha,\beta)=\frac{1}{\alpha}\mathcal{S}_\beta(\frac{1}{n}\bar{\mathbf{X}}\theta^*(\alpha,\beta))$ and $\theta^*(\alpha,\beta)=\mathbf{1}$ satisfy the conditions KKT-1 and KKT-2. Hence, they are the optimal solution for the primal and dual problems (P*) and (D*).

2. **Case 2:** $\alpha_{\max}(\beta)>0$. Then for any $\alpha\geq\alpha_{\max}(\beta)$, we have

$$\begin{aligned}
&\min_{i\in[n]}\{1-\langle\bar{\mathbf{x}}_i,\mathbf{w}^*(\alpha,\beta)\rangle\}\\
=&\min_{i\in[n]}\{1-\frac{1}{\alpha}\langle\bar{\mathbf{x}}_i,\mathcal{S}_\beta(\frac{1}{n}\bar{\mathbf{X}}\theta^*(\alpha,\beta))\rangle\}=\min_{i\in[n]}\{1-\frac{1}{\alpha}\langle\bar{\mathbf{x}}_i,\mathcal{S}_\beta(\frac{1}{n}\bar{\mathbf{X}}\mathbf{1})\rangle\}\\
=&1-\frac{1}{\alpha}\max_{i\in[n]}\langle\bar{\mathbf{x}}_i,\mathcal{S}_\beta(\frac{1}{n}\bar{\mathbf{X}}\mathbf{1})\rangle=1-(1-\gamma)\frac{1}{\alpha}\alpha_{\max}(\beta)\geq 1-(1-\gamma)=\gamma.
\end{aligned}$$

   Thus, $\mathcal{E}\cup\mathcal{L}=[n]$ and $\mathbf{w}^*(\alpha,\beta)=\frac{1}{\alpha}\mathcal{S}_\beta(\frac{1}{n}\bar{\mathbf{X}}\theta^*(\alpha,\beta))$ and $\theta^*(\alpha,\beta)=\mathbf{1}$ satisfy the conditions KKT-1 and KKT-2. Hence, they are the optimal solution for the primal and dual problems (P*) and (D*).

The proof is complete. □

K. Proof for Theorem 6

Proof. of Theorem 6:
(1) Given the reference solutions pair $\mathbf{w}^*(\alpha_{i-1,j},\beta_j)$ and $\theta^*(\alpha_{i-1,j},\beta_j)$, if we do ISS first in SIFS and apply ISS and IFS for infinite times. If after p times of triggering, no new inactive features or samples are identified, then we can denote the sequence of $\hat{\mathcal{F}},\hat{\mathcal{R}}$ and $\hat{\mathcal{L}}$ as:

$$\hat{\mathcal{F}}_0^A=\hat{\mathcal{R}}_0^A=\hat{\mathcal{L}}_0^A=\emptyset\overset{ISS}{\longrightarrow}\hat{\mathcal{F}}_1^A,\hat{\mathcal{R}}_1^A,\hat{\mathcal{L}}_1^A\overset{IFS}{\longrightarrow}\hat{\mathcal{F}}_2^A,\hat{\mathcal{R}}_2^A,\hat{\mathcal{L}}_2^A\overset{ISS}{\longrightarrow}...\hat{\mathcal{F}}_p^A,\hat{\mathcal{R}}_p^A,\hat{\mathcal{L}}_p^A\overset{IFS/ISS}{\longrightarrow}... \tag{32}$$

$$\text{with } \hat{\mathcal{F}}_p^A = \hat{\mathcal{F}}_{p+1}^A = \hat{\mathcal{F}}_{p+2}^A = ..., \hat{\mathcal{R}}_p^A = \hat{\mathcal{R}}_{p+1}^A = \hat{\mathcal{R}}_{p+2}^A = ..., \text{ and } \hat{\mathcal{L}}_p^A = \hat{\mathcal{L}}_{p+1}^A = \hat{\mathcal{L}}_{p+2}^A = ... \tag{33}$$

In the same way, if we do IFS first in SIFS and no new inactive feature or samples are identified after q times of triggering of ISS and IFS, then the sequence can be denoted as:

$$\hat{\mathcal{F}}_0^B = \hat{\mathcal{R}}_0^B = \hat{\mathcal{L}}_0^B = \emptyset \xrightarrow{IFS} \hat{\mathcal{F}}_1^B, \hat{\mathcal{R}}_1^B, \hat{\mathcal{L}}_1^B \xrightarrow{ISS} \hat{\mathcal{F}}_2^B, \hat{\mathcal{R}}_2^B, \hat{\mathcal{L}}_2^B \xrightarrow{IFS} ...\hat{\mathcal{F}}_q^B, \hat{\mathcal{R}}_q^B, \hat{\mathcal{L}}_q^B \xrightarrow{IFS/ISS} ... \tag{34}$$

$$\text{with } \hat{\mathcal{F}}_q^B = \hat{\mathcal{F}}_{q+1}^B = \hat{\mathcal{F}}_{q+2}^B = ..., \hat{\mathcal{R}}_q^B = \hat{\mathcal{R}}_{q+1}^B = \hat{\mathcal{R}}_{q+2}^B = ..., \text{ and } \hat{\mathcal{L}}_q^B = \hat{\mathcal{L}}_{q+1}^B = \hat{\mathcal{L}}_{q+2}^B = ... \tag{35}$$

We first prove that $\hat{\mathcal{F}}_k^B \subseteq \hat{\mathcal{F}}_{k+1}^A, \hat{\mathcal{R}}_k^B \subseteq \hat{\mathcal{R}}_{k+1}^A$ and $\hat{\mathcal{L}}_k^B \subseteq \hat{\mathcal{L}}_{k+1}^A$ hold for all $k \geq 0$ by induction.

1) When $k = 0$, the equalities $\hat{\mathcal{F}}_0^B \subseteq \hat{\mathcal{F}}_1^A, \hat{\mathcal{R}}_0^B \subseteq \hat{\mathcal{R}}_1^A$ and $\hat{\mathcal{L}}_0^B \subseteq \hat{\mathcal{L}}_1^A$ hold since $\hat{\mathcal{F}}_0^B = \hat{\mathcal{R}}_0^B = \hat{\mathcal{L}}_0^B = \emptyset$.
2) If $\hat{\mathcal{F}}_k^B \subseteq \hat{\mathcal{F}}_{k+1}^A, \hat{\mathcal{R}}_k^B \subseteq \hat{\mathcal{R}}_{k+1}^A$ and $\hat{\mathcal{L}}_k^B \subseteq \hat{\mathcal{L}}_{k+1}^A$ hold, by the synergy effect of ISS and IFS, we have $\hat{\mathcal{F}}_{k+1}^B \subseteq \hat{\mathcal{F}}_{k+2}^A, \hat{\mathcal{R}}_{k+1}^B \subseteq \hat{\mathcal{R}}_{k+2}^A$ and $\hat{\mathcal{L}}_{k+1}^B \subseteq \hat{\mathcal{L}}_{k+2}^A$ hold.
Thus, $\hat{\mathcal{F}}_k^B \subseteq \hat{\mathcal{F}}_{k+1}^A, \hat{\mathcal{R}}_k^B \subseteq \hat{\mathcal{R}}_{k+1}^A$ and $\hat{\mathcal{L}}_k^B \subseteq \hat{\mathcal{L}}_{k+1}^A$ hold for all $k \geq 0$.
Similar with the analysis in (1), we can also prove that $\hat{\mathcal{F}}_k^A \subseteq \hat{\mathcal{F}}_{k+1}^B, \hat{\mathcal{R}}_k^A \subseteq \hat{\mathcal{R}}_{k+1}^B$ and $\hat{\mathcal{L}}_k^A \subseteq \hat{\mathcal{L}}_{k+1}^B$ hold for all $k \geq 0$.
Combine (1) and (2), we can get

$$\hat{\mathcal{F}}_0^B \subseteq \hat{\mathcal{F}}_1^A \subseteq \hat{\mathcal{F}}_2^B \subseteq \hat{\mathcal{F}}_3^A .... \tag{36}$$
$$\hat{\mathcal{F}}_0^A \subseteq \hat{\mathcal{F}}_1^B \subseteq \hat{\mathcal{F}}_2^A \subseteq \hat{\mathcal{F}}_3^B .... \tag{37}$$
$$\hat{\mathcal{R}}_0^B \subseteq \hat{\mathcal{R}}_1^A \subseteq \hat{\mathcal{R}}_2^B \subseteq \hat{\mathcal{R}}_3^A .... \tag{38}$$
$$\hat{\mathcal{R}}_0^A \subseteq \hat{\mathcal{R}}_1^B \subseteq \hat{\mathcal{R}}_2^A \subseteq \hat{\mathcal{R}}_3^A .... \tag{39}$$
$$\hat{\mathcal{L}}_0^B \subseteq \hat{\mathcal{L}}_1^A \subseteq \hat{\mathcal{L}}_2^B \subseteq \hat{\mathcal{L}}_3^A .... \tag{40}$$
$$\hat{\mathcal{L}}_0^A \subseteq \hat{\mathcal{L}}_1^B \subseteq \hat{\mathcal{L}}_2^A \subseteq \hat{\mathcal{L}}_3^B .... \tag{41}$$

by the first equality of (33), (36) and (37), we can get $\hat{\mathcal{F}}_p^A = \hat{\mathcal{F}}_q^B$. Similarly, we can get $\hat{\mathcal{R}}_p^A = \hat{\mathcal{R}}_q^B$ and $\hat{\mathcal{L}}_p^A = \hat{\mathcal{L}}_q^B$.
(2) If p is odd, then by (36), (38 and (40), we have $\hat{\mathcal{F}}_p^A \subseteq \hat{\mathcal{F}}_{p+1}^B, \hat{\mathcal{R}}_p^A \subseteq \hat{\mathcal{R}}_{p+1}^B$ and $\hat{\mathcal{L}}_p^A \subseteq \hat{\mathcal{L}}_{p+1}^B$. Thus $q \leq p + 1$.
Else if p is even, then by (37), (39) and (41), we have $\hat{\mathcal{F}}_p^A \subseteq \hat{\mathcal{F}}_{p+1}^B, \hat{\mathcal{R}}_p^A \subseteq \hat{\mathcal{R}}_{p+1}^B$ and $\hat{\mathcal{L}}_p^A \subseteq \hat{\mathcal{L}}_{p+1}^B$. Thus $q \leq p + 1$.
Do the same analysis for q, we can get $p \leq q + 1$.
Hence, $|p - q| \leq 1$.
The proof is complete.

□

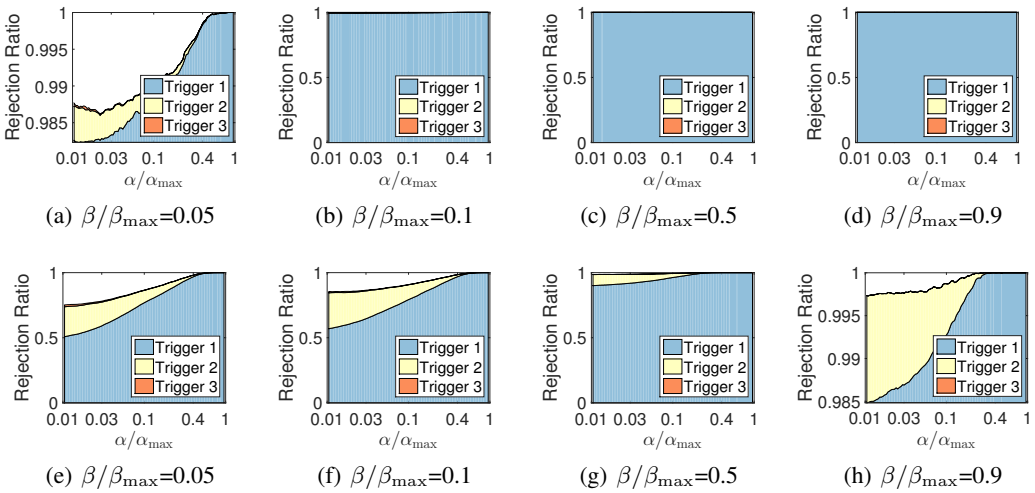
L. Experiment Result

L.1. Verification of the Synergy Effect

Here, we verify the synergy effect between ISS and IFS in SIFS from the experiment results on the dataset real-sim. In Fig. 4, SIFS performs ISS (sample screening) first, while in Fig. 5, it performs IFS (feature screening) first. All the rejection ratios (Fig. 4(a)-(d)) of the 1st triggering of IFS when SIFS performs ISS first are much higher than (at least equal to) those (Fig. 5(a)-(d)) when SIFS performs IFS first. In turn, all the rejection ratios (Fig. 5(e)-(h)) of the 1st triggering of ISS when SIFS performs IFS first are also much higher than those (Fig. 4(e)-(h)) when SIFS performs ISS first. This demonstrates that the screening result of ISS can reinforce the capability of IFS and vice versa, which is the so called synergy effect. At last, in Fig. 5 and Fig. 4, we can see that the overall rejection ratios at the end of SIFS are the same, so no matter which (ISS or IFS) we perform first in SIFS, SIFS has the same screening performances in the end. This is consistent with Theorem 6.

L.2. The Rest Experiment Result

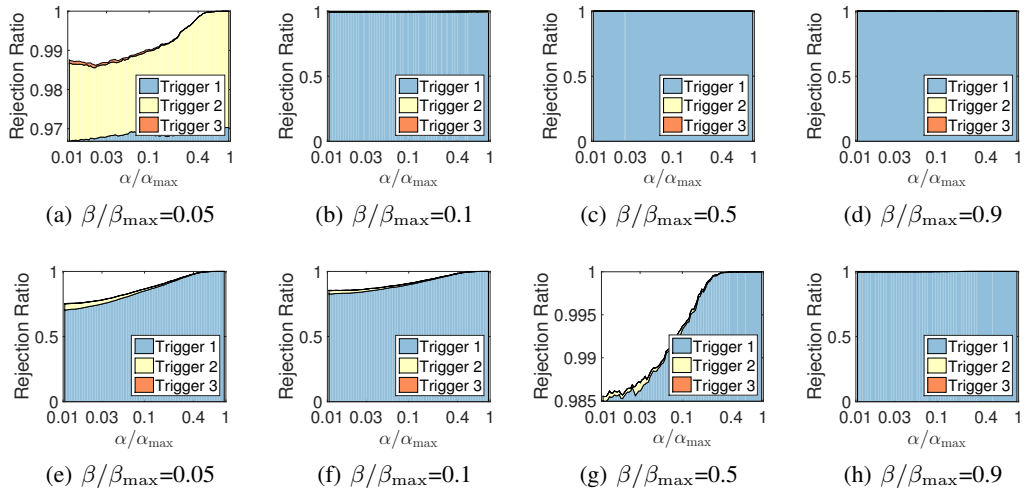
Below, we report the rejection ratios of SIFS on syn1 (Fig. 6), syn3 (Fig. 7), rcv1-train (Fig. 8), rcv1-test(Fig. 9), url (Fig. 10) and kddb (Fig. 11), which are omitted in the main text due to the space limitation.



(a) $\beta/\beta_{\max}$=0.05 (b) $\beta/\beta_{\max}$=0.1 (c) $\beta/\beta_{\max}$=0.5 (d) $\beta/\beta_{\max}$=0.9

(e) $\beta/\beta_{\max}$=0.05 (f) $\beta/\beta_{\max}$=0.1 (g) $\beta/\beta_{\max}$=0.5 (h) $\beta/\beta_{\max}$=0.9

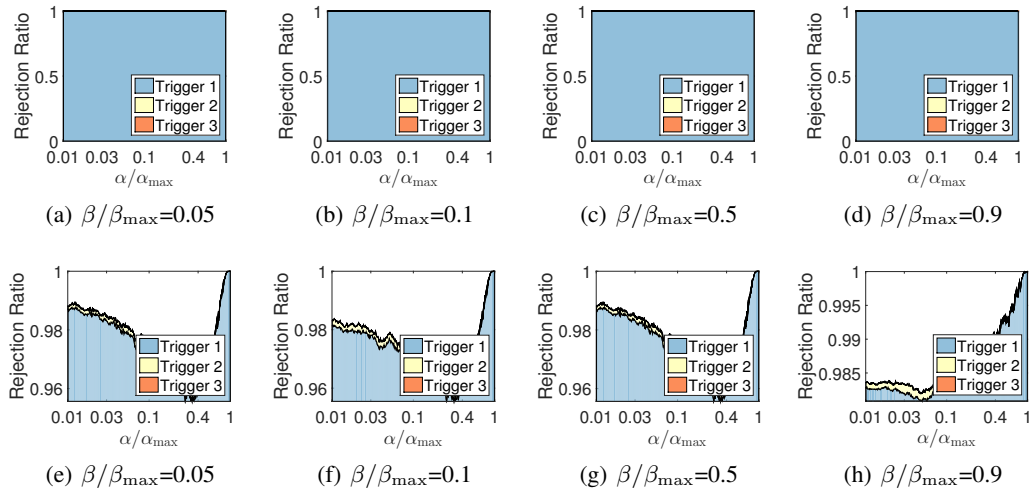
Figure 4. Rejection ratios of SIFS on real-sim when it performs **ISS first** (first row: Feature Screening, second row: Sample Screening).



(a) $\beta/\beta_{\max}$=0.05 (b) $\beta/\beta_{\max}$=0.1 (c) $\beta/\beta_{\max}$=0.5 (d) $\beta/\beta_{\max}$=0.9

(e) $\beta/\beta_{\max}$=0.05 (f) $\beta/\beta_{\max}$=0.1 (g) $\beta/\beta_{\max}$=0.5 (h) $\beta/\beta_{\max}$=0.9

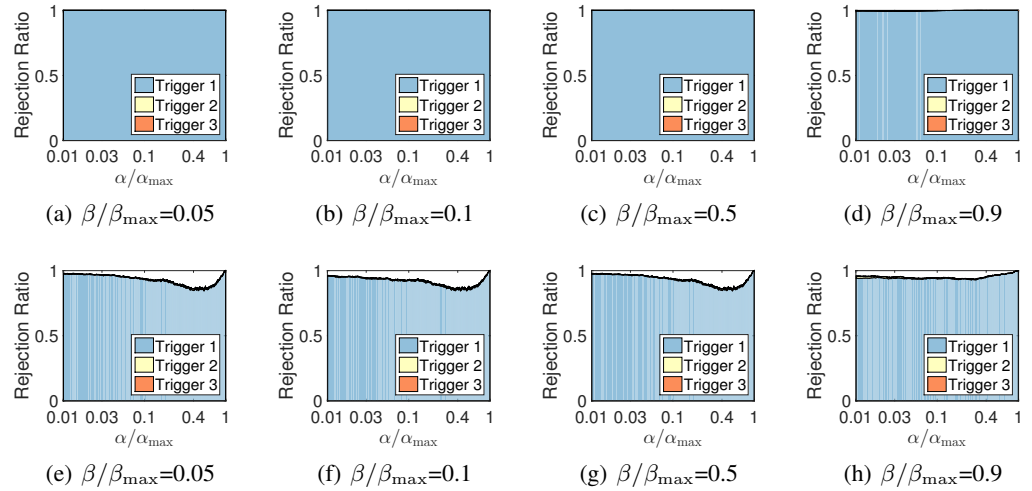
Figure 5. Rejection ratios of SIFS on real-sim when it performs **IFS first**(first row: Feature Screening, second row: Sample Screening).



(a) $\beta/\beta_{\max}$=0.05 (b) $\beta/\beta_{\max}$=0.1 (c) $\beta/\beta_{\max}$=0.5 (d) $\beta/\beta_{\max}$=0.9

(e) $\beta/\beta_{\max}$=0.05 (f) $\beta/\beta_{\max}$=0.1 (g) $\beta/\beta_{\max}$=0.5 (h) $\beta/\beta_{\max}$=0.9

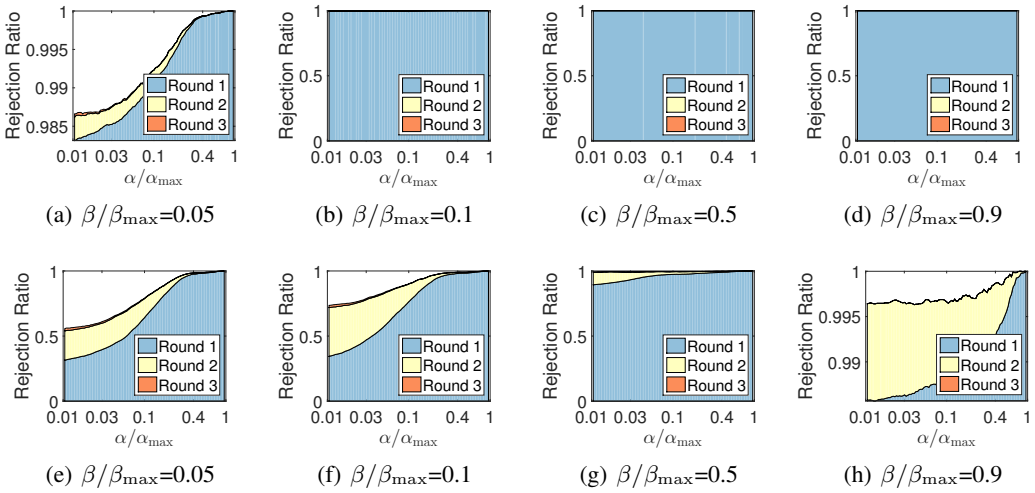
Figure 6. Rejection ratios of SIFS on syn1 (first row: Feature Screening, second row: Sample Screening).



(a) $\beta/\beta_{\max}$=0.05 (b) $\beta/\beta_{\max}$=0.1 (c) $\beta/\beta_{\max}$=0.5 (d) $\beta/\beta_{\max}$=0.9

(e) $\beta/\beta_{\max}$=0.05 (f) $\beta/\beta_{\max}$=0.1 (g) $\beta/\beta_{\max}$=0.5 (h) $\beta/\beta_{\max}$=0.9

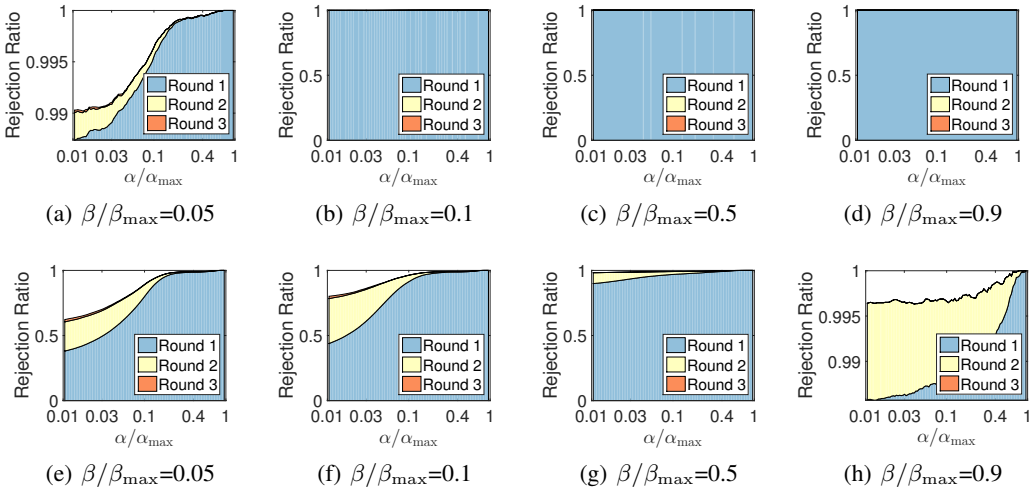
Figure 7. Rejection ratios of SIFS on syn3 (first row: Feature Screening, second row: Sample Screening).



(a) $\beta/\beta_{\max}$=0.05 (b) $\beta/\beta_{\max}$=0.1 (c) $\beta/\beta_{\max}$=0.5 (d) $\beta/\beta_{\max}$=0.9

(e) $\beta/\beta_{\max}$=0.05 (f) $\beta/\beta_{\max}$=0.1 (g) $\beta/\beta_{\max}$=0.5 (h) $\beta/\beta_{\max}$=0.9

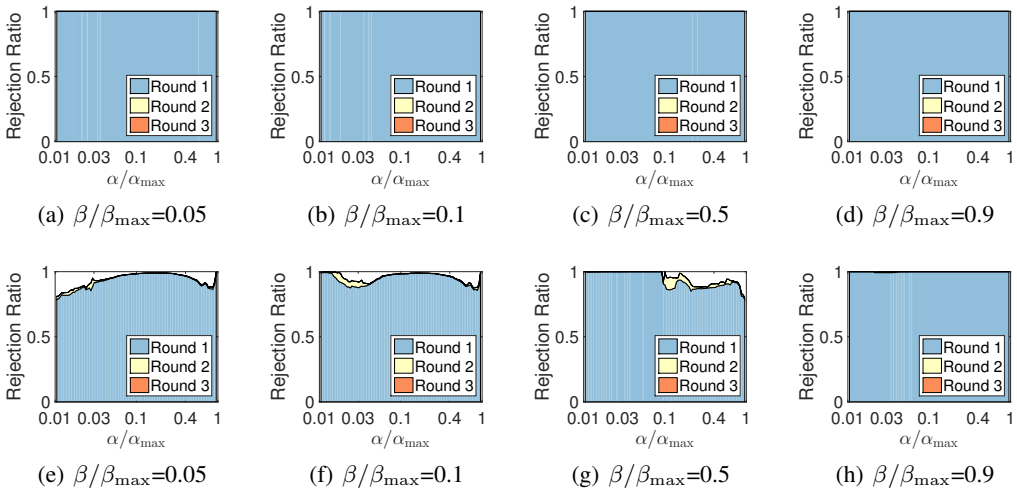
Figure 8. Rejection ratios of SIFS on rcv1-train dataset (first row: Feature Screening, second row: Sample Screening).



(a) $\beta/\beta_{\max}$=0.05 (b) $\beta/\beta_{\max}$=0.1 (c) $\beta/\beta_{\max}$=0.5 (d) $\beta/\beta_{\max}$=0.9

(e) $\beta/\beta_{\max}$=0.05 (f) $\beta/\beta_{\max}$=0.1 (g) $\beta/\beta_{\max}$=0.5 (h) $\beta/\beta_{\max}$=0.9

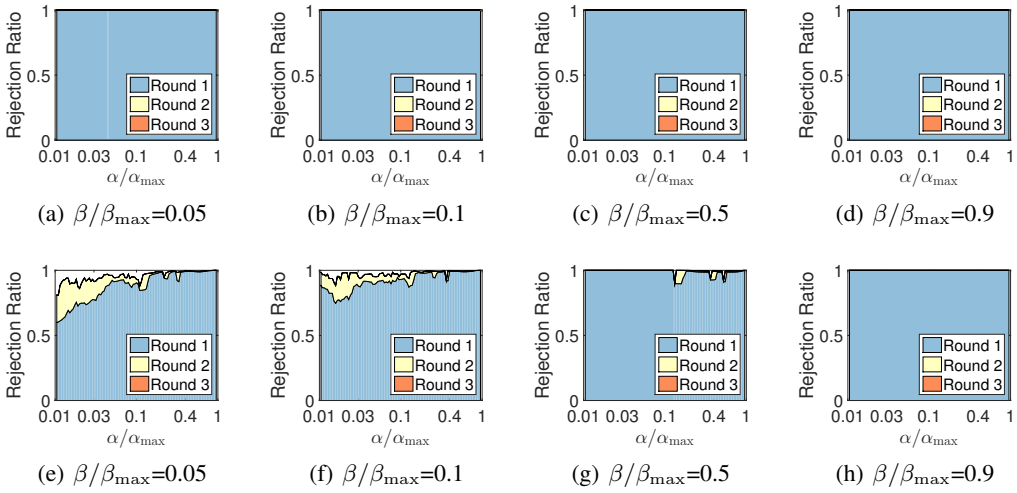
Figure 9. Rejection ratios of SIFS on rcv1-test dataset (first row: Feature Screening, second row: Sample Screening).



(a) $\beta/\beta_{\max}$=0.05 (b) $\beta/\beta_{\max}$=0.1 (c) $\beta/\beta_{\max}$=0.5 (d) $\beta/\beta_{\max}$=0.9

(e) $\beta/\beta_{\max}$=0.05 (f) $\beta/\beta_{\max}$=0.1 (g) $\beta/\beta_{\max}$=0.5 (h) $\beta/\beta_{\max}$=0.9

Figure 10. Rejection ratios of SIFS on url dataset (first row: Feature Screening, second row: Sample Screening).



(a) $\beta/\beta_{\max}$=0.05 (b) $\beta/\beta_{\max}$=0.1 (c) $\beta/\beta_{\max}$=0.5 (d) $\beta/\beta_{\max}$=0.9

(e) $\beta/\beta_{\max}$=0.05 (f) $\beta/\beta_{\max}$=0.1 (g) $\beta/\beta_{\max}$=0.5 (h) $\beta/\beta_{\max}$=0.9

Figure 11. Rejection ratios of SIFS on kddb dataset (first row: Feature Screening, second row: Sample Screening).