Simultaneous Safe Screening of Features and Samples in Doubly Sparse Modeling

Atsushi Shibagaki†, Masayuki Karasuyama†, Kohei Hatano‡, and Ichiro Takeuchi† († Nagoya Institute of Technology, ‡ Kyushu University (Japan))

Introduction

We consider regularized empirical risk minimizations induce feature/sample sparsity

- **Motivation:** To reduce computational cost of optimization
- **Approch:** Identifying non-active features/samples at the optimal solution

Previous works:

- **Safe feature screening**[1]**:** Identifying non-active features for feature sparse models
- **Safe sample screening**[2]**:** Identifying non-active samples for sample sparse models

Safe screening has been individually studied either for feature or sample screening

- **Main contribution (simultaneous safe screening of features and samples) :**

Safely screening features and samples simultaneously by alternatively iterating feature and sample screening steps for feature and sample (doubly) sparse models

Preliminaries

Safe feature screening (for Elastic net penalty)

KKT condtion: Safe fature screening rule :

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

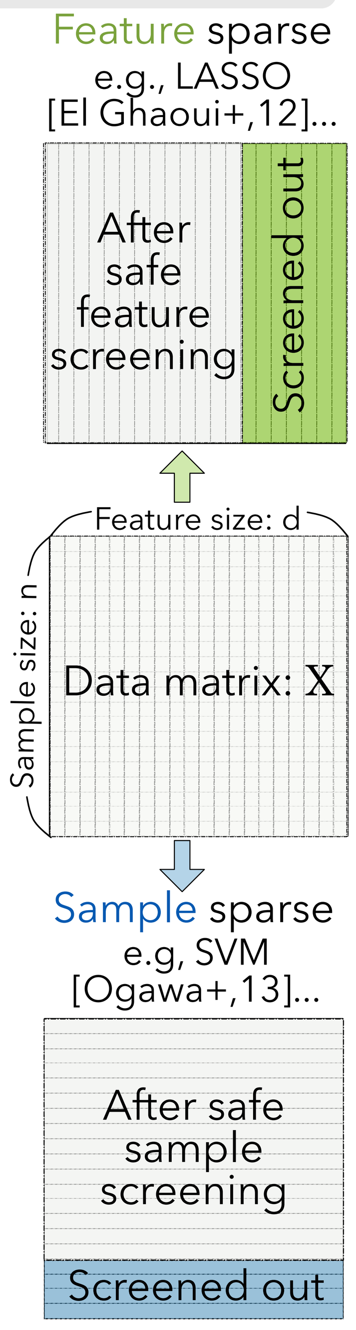
$$|X_{:j}^\top\alpha^*| \le UB(|X_{:j}^\top\alpha^*|) := \max_{\alpha} |X_{:j}^\top\alpha| \ \text{ s.t. } \ \alpha \in \Theta_{\alpha^*}$$
$$= |X_{:j}^\top\hat{\alpha}| + \|X_{:j}\|_2\sqrt{2n(P_\lambda(\hat{w})-D_\lambda(\hat{\alpha}))/\gamma}$$

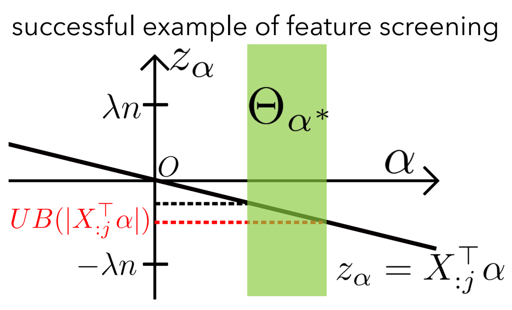
Θ_{α^*}: Region of dual optimal solution

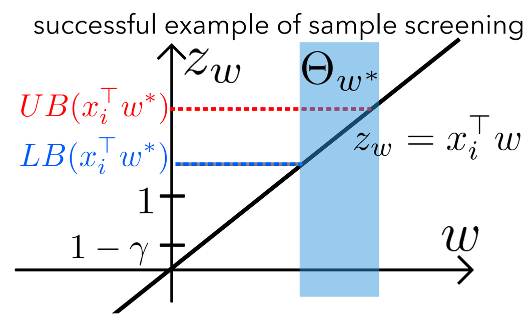
[Ndiaye+, 15] If the D_λ is γ/n-strongly concave then

$$\alpha^* \in \Theta_{\alpha^*} := \{\, \alpha \mid \|\hat{\alpha}-\alpha\|_2 \le \sqrt{2n(P_\lambda(\hat{w})-D_\lambda(\hat{\alpha}))/\gamma}\,\},$$

for any $\hat{w} \in \mathrm{dom}P_\lambda, \hat{\alpha} \in \mathrm{dom}D_\lambda$







Safe sample screening (for smoothed hinge loss)

KKT condtion: Safe sample screening rules:

$$x_i^\top w^* \in \begin{cases} [1,\infty) & (\alpha_i^*=0) \implies LB(x_i^\top w^*) \ge 1 \Rightarrow \alpha_i^*=0, \\ (-\infty, 1-\gamma] & (\alpha_i^*=1) \implies UB(x_i^\top w^*) \le 1-\gamma \Rightarrow \alpha_i^*=1 \\ -\gamma\alpha_i^*+1 & (\alpha_i^* \in (0,1)) \end{cases}$$

$$x_i^\top w^* \ge LB(x_i^\top w^*) := \min_{w\in\Theta_{w^*}} x_i^\top w = x_i^\top\hat{w} - \|x_i\|_2\sqrt{2(P_\lambda(\hat{w})-D_\lambda(\hat{\alpha}))/\lambda}$$
$$x_i^\top w^* \le UB(x_i^\top w^*) := \max_{w\in\Theta_{w^*}} x_i^\top w = x_i^\top\hat{w} + \|x_i\|_2\sqrt{2(P_\lambda(\hat{w})-D_\lambda(\hat{\alpha}))/\lambda}$$

Θ_{w^*}: Region of primal optimal solution

P_λ is λ-strongly convex $\Rightarrow w^* \in \Theta_{w^*} := \{w \mid \|\hat{w}-w\|_2 \le \sqrt{2(P_\lambda(\hat{w})-D_\lambda(\hat{\alpha}))/\lambda}\}$

Dynamic screening [Bonnefoy+, 14]

We need good accurate solution $\hat{w}$ and $\hat{\alpha}$ for good safe screening performances !

While convergence do;

1. Safe screening using $(\hat{w}_t, \hat{\alpha}_t)$
2. $(\hat{w}_{t+1}, \hat{\alpha}_{t+1}) \leftarrow$ Optimization update$(\hat{w}_t, \hat{\alpha}_t)$

Summary

Primal space Θ_{w^*}**:** safe sample screening and safe feature keeping

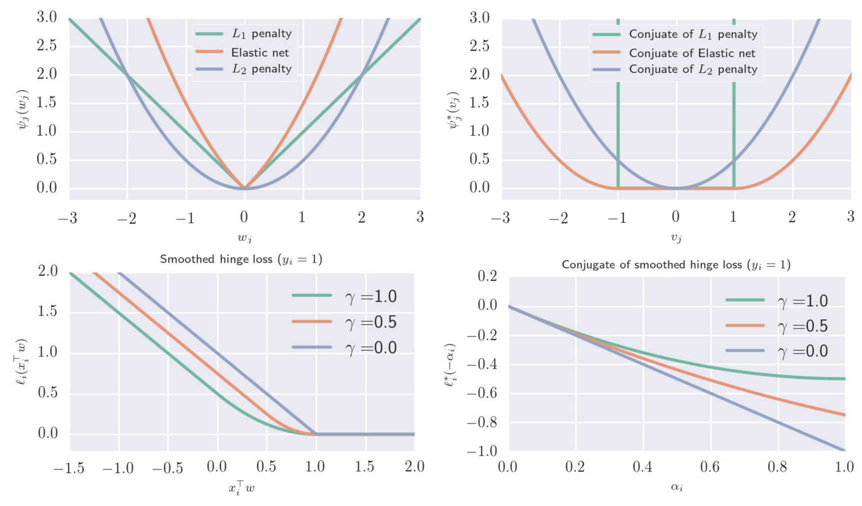
Dual space Θ_{α^*}**:** safe feature screening and safe sample keeping

Formulations

- Data: $\{(x_i, y_i)\}_{i\in[n]}$, Data matrix $(n \times d)$: X

$$-\text{(Primal)} \quad w^* = \arg\min_{w\in\mathbb{R}^d} P_\lambda(w) := \lambda\psi(w) + \frac{1}{n}\sum_{i\in[n]}\ell_i(x_i^\top w)$$

$$-\text{(Dual)} \quad \alpha^* = \arg\max_{\alpha\in\mathrm{dom}D_\lambda} D_\lambda(\alpha) := -\lambda\psi^*(\frac{1}{\lambda n}X^\top\alpha) - \frac{1}{n}\sum_{i\in[n]}\ell_i^*(-\alpha_i),$$



- Elastic net: $\psi(w) := \|w\|_1 + \frac{1}{2}\|w\|_2^2$
- Smoothed hinge loss:

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

Simultaneous Safe Screening

- Results of safe sample screening can improve a safe feature screening (and vice-versa)

safe feature screening using the result of sample screening

- We know $\alpha_i^* = \{0, \pm 1\}$ for $i \in \mathcal{S}$ by safe sample screening ($\bar{\mathcal{S}} := [n] \setminus \mathcal{S}$)
- We can get the tighter upper bound of $|X_{:j}^\top\alpha^*|$:

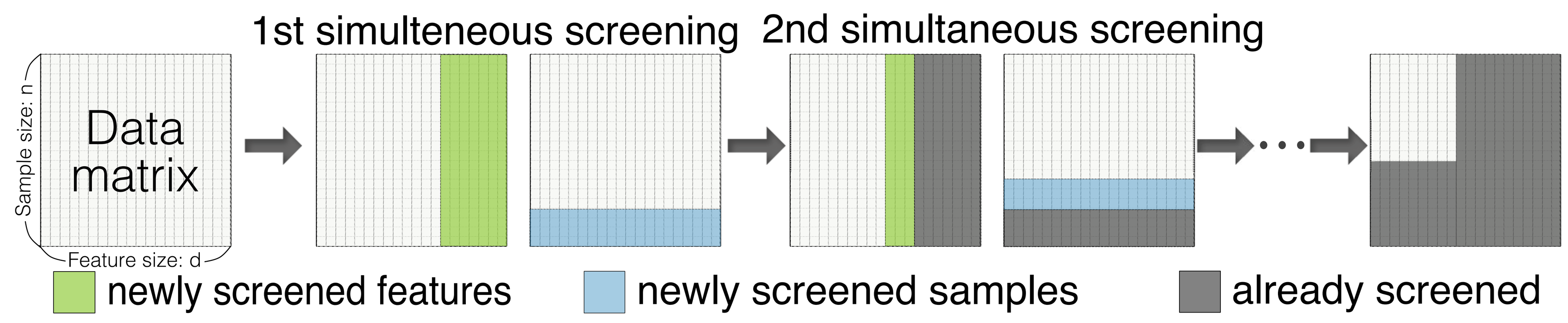
$$\tilde{UB}(|X_{:j}^\top\alpha^*|) := \max_{\alpha} |X_{:j}^\top\alpha| \ \text{ s.t. } \ \alpha \in \Theta_{\alpha^*}, \alpha_i = \alpha_i^* \ \forall i \in \mathcal{S}$$
$$= |X_{\mathcal{S},j}^\top\alpha_{\mathcal{S}}^*| + |X_{\bar{\mathcal{S}},j}^\top\hat{\alpha}_{\bar{\mathcal{S}}}| - \|X_{\bar{\mathcal{S}},j}\|_2\sqrt{2n(P_\lambda(\hat{w})-D_\lambda(\hat{\alpha}))/\gamma - \|\hat{\alpha}_{\mathcal{S}} - \alpha_{\mathcal{S}}^*\|_2^2}$$

safe sample screening using the result of feature screening

- We know $w_j^* = 0$ for $j \in \mathcal{F}$ by safe feature screening ($\bar{\mathcal{F}} := [d] \setminus \mathcal{F}$)
- We can get the tighter bounds of $x_i^\top w^*$:

$$\tilde{LB}(x_i^\top w^*) = \min_{w} x_i^\top w \ \text{ s.t. } \ w \in \Theta_{w^*}, w_j = w_j^* \ \forall j \in \mathcal{F}$$
$$= x_{i\bar{\mathcal{F}}}^\top\hat{w}_{\bar{\mathcal{F}}} - \|x_{i\bar{\mathcal{F}}}\|_2\sqrt{2(P_\lambda(\hat{w})-D_\lambda(\hat{\alpha}))/\lambda - \|\hat{w}_{\mathcal{F}}\|_2^2}$$

- $\tilde{UB}(x_i^\top w^*)$ also
- More and more features and samples could be screened out by alternately iterating feature and sample screening



Safe keeping

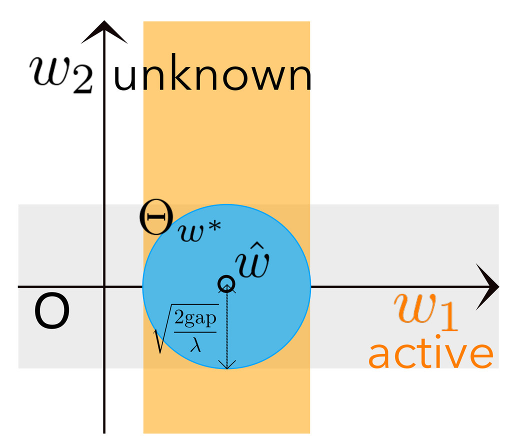
- allows us to identify a part of active features/samples

Safe feature keeping: If P_λ is λ-strongly convex then

$$|\hat{w}_j| - \sqrt{2(P_\lambda(\hat{w})-D_\lambda(\hat{\alpha}))/\lambda} > 0 \ \Rightarrow \ w_j^* \neq 0$$

Safe sample keeping: If D_λ is γ/n-strongly convex then

$|\hat{\alpha}_i| - \sqrt{2n(P_\lambda(\hat{w})-D_\lambda(\hat{\alpha}))/\gamma} > 0$ and
$|\hat{\alpha}_i| + \sqrt{2n(P_\lambda(\hat{w})-D_\lambda(\hat{\alpha}))/\gamma} < 1 \Rightarrow \alpha_i^* \notin \{0, \pm 1\}$



Advantages:

- We do not have to waste the screening rule evaluation costs for active features/samples
- By combining safe screening and safe keeping:

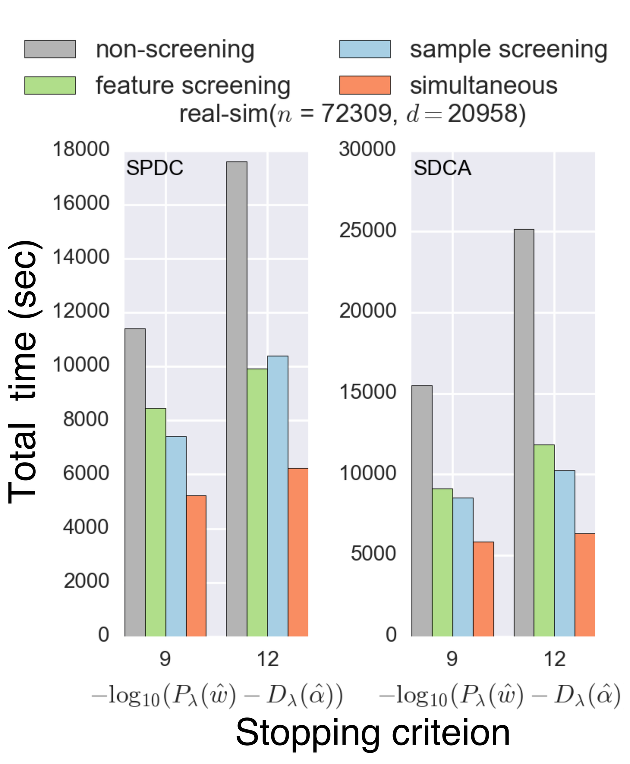
#(features/samples aren't determined to be active or non-active)

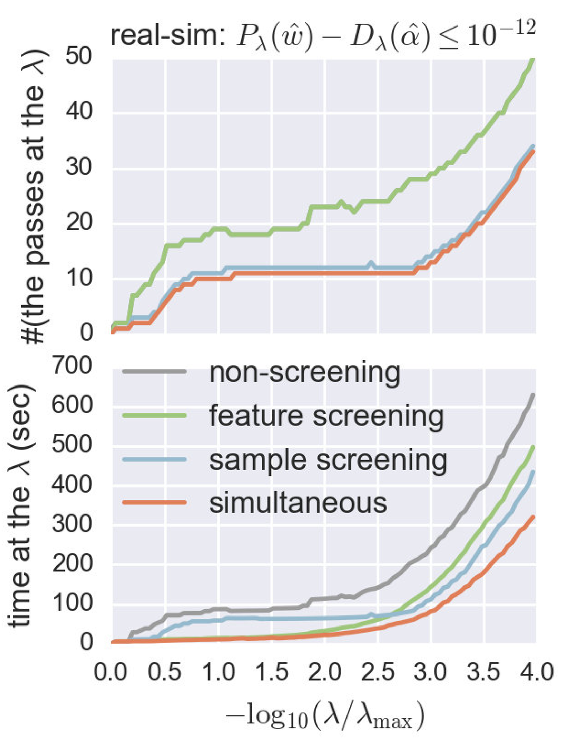
can be also used as a stopping criteria of dynamic screening and simultaneous screening

Experiments

Elastic net + smoothed hinge (P_λ: λ-strongly convex, D_λ: γ/n-strongly concave)

- Computation time savings

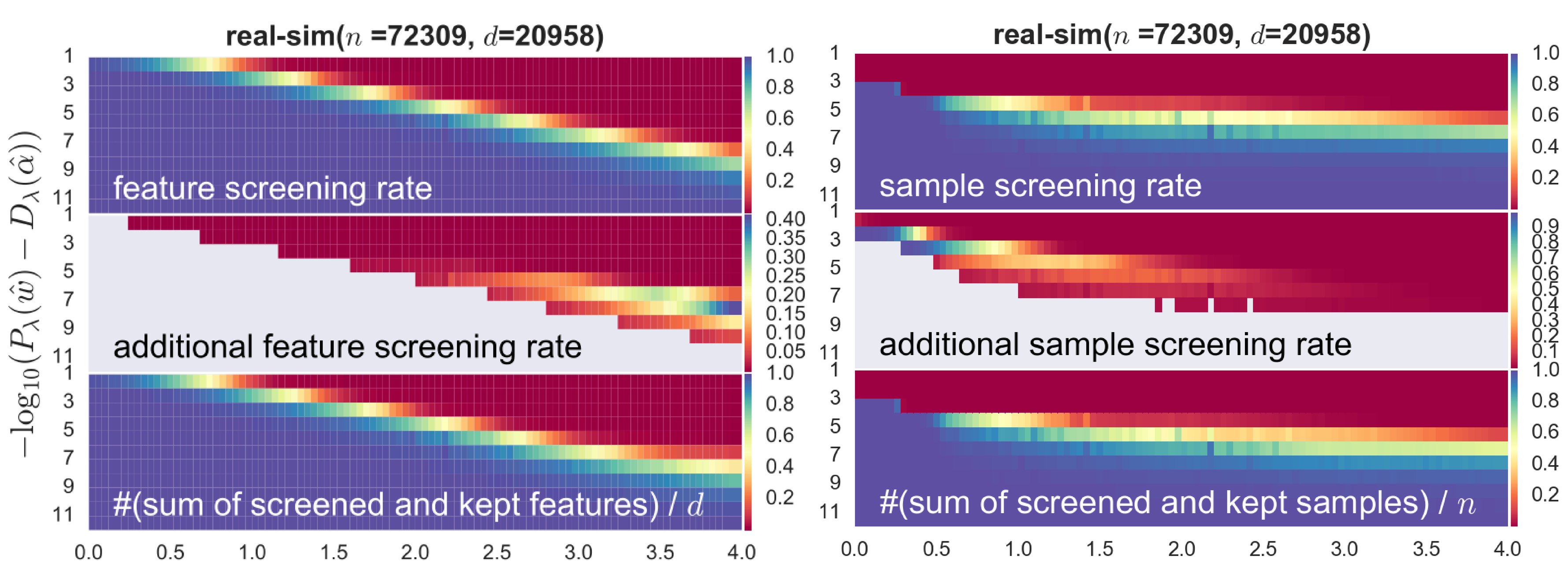




- $\gamma = 0.5$
- $\lambda_{\max} := \|\mathrm{diag}(y)X^\top\mathbf{1}\|_\infty$
- train at 100 different λ evenly allocated in $[10^{-4}\lambda_{\max}, \lambda_{\max}]$ in the logarithmic scale
- we used warm-start
- Solvers:
  - Stochastic Primal-Dual Coordinate
  - Stochastic Dual Coordinate Ascent

- Screening and keeping rates

- screening rate := #(screened features or samples) / #($w_j^* = 0$ or $\alpha_i^* = \{0, \pm 1\}$)
- additional screening rate by simulatenous screening: In gray area, the individual safe screening performances are good enough (screening rate > 0.95) and additional screening is unnecessary



References: [1] L. El Ghaoui, V. Viallon, and T. Rabbani. Safe feature elimination for the lasso and sparse supervised learning problems. Pacific Journal of Optimization, 2012. [2] K. Ogawa, Y. Suzuki, and I. Takeuchi. Safe screening of non-support vectors in pathwise svm computation. ICML, 2013.
[3] E. Ndiaye, O. Fercoq, A. Gramfort, and J. Salmon. Gap safe screening rules for sparse multi-task and multi-class models. NIPS, 2015. [4] Bonnefoy, Antoine, Emiya, Valentin, Ralaivola, Liva, and Gribonval, Remi. A dynamic screening principle for the lasso. EUSIPCO, 2014.