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

June, 2016

Motivation: To reduce the cost for training of sparse model

The training process occurs in two steps:

1. Identifying active features/samples at optimal solution
2. Optimizing the model over active features/samples

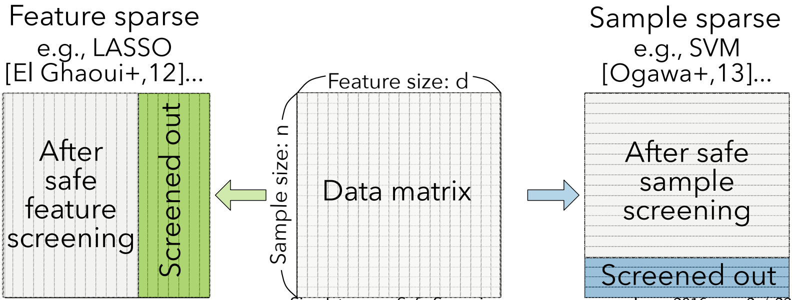
Motivation: To reduce the cost for training of sparse model

The training process occurs in two steps:

1. Identifying active features/samples at optimal solution
2. Optimizing the model over active features/samples

Safe screening:

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



Contribution: Simultaneous safe screening

Safe screening has been individually studied either

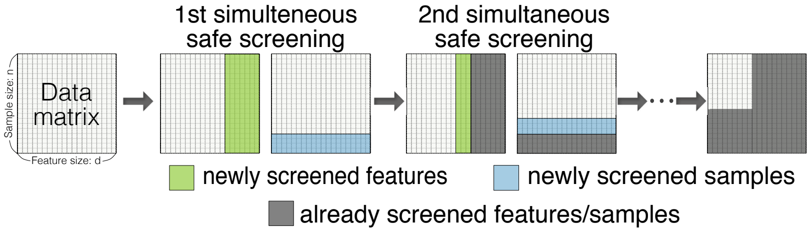
- for feature screening
- for sample screening

We consider doubly sparse modelings:

induce both of feature sparsity and sample sparsity

- e.g., L_1-penalized SVMs

Advantage: synergy effect



Target problems: Doubly Sparse Modeling

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

Regularized Empirical Risk Minimization (Primal problem):

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

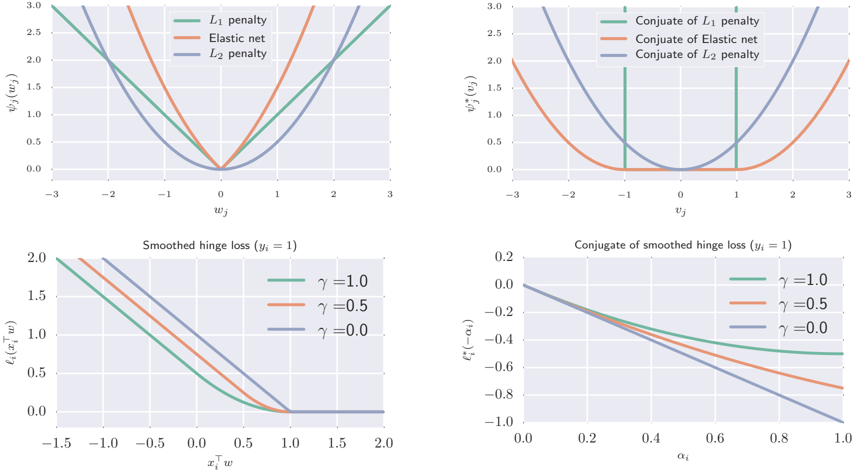
e.g, L_1-penalized smoothed SVMs:

- Penalty function: Elastic Net $\psi(w) = \|w\|_1 + \frac{1}{2}\|w\|_2^2$
- Loss function (classification): Smoothed hinge loss

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

Regularized Empirical Risk Minimization (Dual problem):

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



Safe fature screening (for Elastic net):

allows us to identify a part of non-active features $w_j^* = 0$

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

$$\Downarrow$$

Safe fature screening rule: $UB(|X_{:j}^\top \alpha^*|) \leq \lambda n \Rightarrow w_j^* = 0$

Procedure:

- Construct the region of the optimal solution Θ_{α^*}
- For $j \in [d]$:
  1. Compute $UB(|X_{:j}^\top \alpha^*|) := \max X_{:j}^\top \alpha \;\text{ s.t. }\; \alpha \in \Theta_{\alpha^*}$
  2. Check safe feature screening rule

Safe sample screening (for smoothed hinge loss ($y_i = 1$)):

allows us to identify a part of non-active samples $\alpha_i^* = \{0, 1\}$

$$\text{KKT condtion: } x_i^\top w^* \in \begin{cases} [1, \infty) & (\alpha_i^* = 0) \\ (-\infty, 1-\gamma] & (\alpha_i^* = 1) \\ -\gamma\alpha_i^* + 1 & (\alpha_i^* \in (0,1)) \end{cases}$$

Safe sample screening rules: $\Downarrow$

$$LB(x_i^\top w^*) \geq 1 \Rightarrow \alpha_i^* = 0, \quad UB(x_i^\top w^*) \leq 1 - \gamma \Rightarrow \alpha_i^* = 1$$

Procedure:

- Construct the region of the optimal solution Θ_{w^*}
- For $i \in [n]$:
  1. Compute $LB(x_i^\top w^*) := \min x_i^\top \alpha \;\; \text{s.t.} \;\; w \in \Theta_{w^*}$,
     $UB(x_i^\top w^*) := \max x_i^\top \alpha \;\; \text{s.t.} \;\; w \in \Theta_{w^*}$
  2. Check safe sample screening rules

Construct the region of the optimal solution $\Theta_{\alpha^*}, \Theta_{w^*}$

Theorem 3 in [Ndiaye+, 15]

If D_λ is γ/n-strongly concave then

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

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

- ℓ_i is γ-smooth $\Rightarrow$ D_λ is γ/n-strongly concave

Θ_{w^*} as well as Θ_{α^*}

If P_λ is λ-strongly convex then

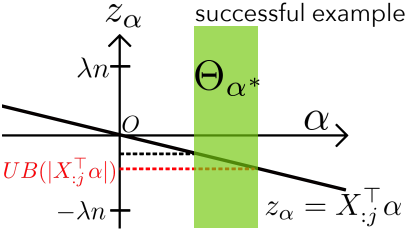
$$w^* \in \Theta_{w^*} := \{ w \mid \|\hat{w} - w\|_2 \leq \sqrt{2(P_\lambda(\hat{w}) - D_\lambda(\hat{\alpha}))/\lambda} \},$$

- ψ is Elastic net $\Rightarrow$ P_λ is λ-strongly convex

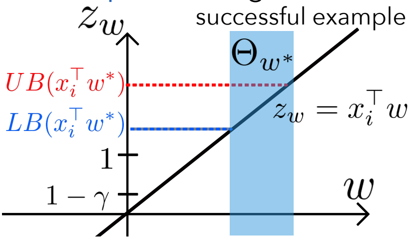
Schematic illustration of safe screening

If $\Theta_{\alpha^*}, \Theta_{w^*}$ are sphere then we have closed form solutions of $UB(|X_{:j}^\top \alpha^*|), LB(x_i^\top w^*), UB(x_i^\top w^*)$.

safe feature screening



safe sample screening



Compute $UB(|X_{:j}^\top \alpha^*|), LB(x_i^\top w^*), UB(x_i^\top w^*)$

Closed form solutions (since $\Theta_{\alpha^*}, \Theta_{w^*}$ are sphere)

$$
\begin{aligned}
X_{:j}^\top \alpha^* \le UB(|X_{:j}^\top \alpha^*|) &:= \max_{\alpha} X_{:j}^\top \alpha \ \text{ s.t. } \ \alpha \in \Theta_{\alpha^*} \\
&= X_{:j}^\top \hat{\alpha} + \|X_{:j}\|_2 \sqrt{2nG_\lambda(\hat{w}, \hat{\alpha})/\gamma}, \\
x_i^\top w^* \ge LB(x_i^\top w^*) &:= \min_{w} x_i^\top w \ \text{ s.t. } \ w \in \Theta_{w^*} \\
&= x_i^\top \hat{w} - \|x_i\|_2 \sqrt{2G_\lambda(\hat{w}, \hat{\alpha})/\lambda}, \\
x_i^\top w^* \le UB(x_i^\top w^*) &:= \max_{w} x_i^\top w \ \text{ s.t. } \ w \in \Theta_{w^*} \\
&= x_i^\top \hat{w} + \|x_i\|_2 \sqrt{2G_\lambda(\hat{w}, \hat{\alpha})/\lambda}
\end{aligned}
$$

, where $G_\lambda(\hat{w}, \hat{\alpha}) := P_\lambda(\hat{w}) - D_\lambda(\hat{\alpha})$

Optimization with safe screening

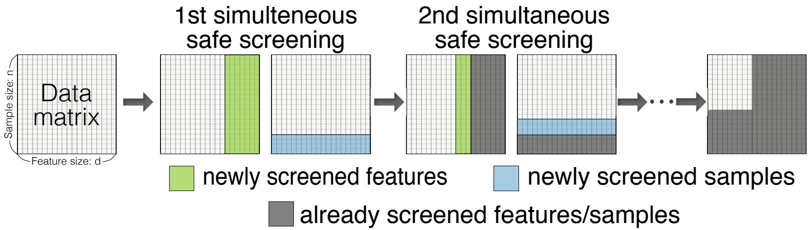
We need good accurate solution $\hat{w}$ and $\hat{\alpha}$!

Dynamic screening [Bonnefoy+, 14]

- Input: $\hat{w}_0, \hat{\alpha}_0, t \leftarrow 0$
- 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$(\hat{w}_t, \hat{\alpha}_t)$
  3. $t \leftarrow t + 1$

Synergy effect by simultaneous screeing

- Results of safe feature/sample screening can improve a safe sample/feature screening performance
- By alternately iterating feature and sample screening: more and more features and samples could be screened out



safe sample screening $\rightarrow$ safe feature screening

individual safe feature screening:

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

simultaneous safe feature screening:

We know $\alpha_i^* = \{0, \pm 1\}$ for $i \in \mathcal{S}$, $\bar{\mathcal{S}} := [n] \setminus \mathcal{S}$

$$\tilde{UB}(|X_{:j}^\top \alpha^*|)$$

$$:= \max_{\alpha} |X_{:j}^\top \alpha| \ \text{ s.t. } \ \|\hat{\alpha} - \alpha\|_2 \leq \sqrt{2G_\lambda(\hat{w}, \hat{\alpha})/\lambda}, \alpha_i = \alpha_i^* \ \forall i \in \mathcal{S}$$

$$= |X_{\mathcal{S},j}^\top \alpha_{\mathcal{S}}^*| + |X_{\bar{\mathcal{S}},j}^\top \alpha_{\bar{\mathcal{S}}}| - \|X_{\bar{\mathcal{S}},j}\|_2 \sqrt{2nG_\lambda(\hat{w}, \hat{\alpha})/\gamma - \|\hat{\alpha}_{\mathcal{S}} - \alpha_{\mathcal{S}}^*\|_2^2}$$

$$\leq UB(|X_{:j}^\top \alpha^*|)$$

safe feature screening $\rightarrow$ safe sample screening

individual safe sample screening:

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

simultaneous safe sample screening:

We know $w_j^* = 0$ for $j \in \mathcal{F}$, $\bar{\mathcal{F}} := [d] \setminus \mathcal{F}$.

$$\begin{aligned} &\tilde{LB}(x_i^\top w^*) \\ &= \min_w x_i^\top w \;\; \text{s.t.} \;\; \|\hat{w} - w\|_2 \le \sqrt{2G_\lambda(\hat{w}, \hat{\alpha})/\lambda}, w_j = 0 \; \forall j \in \mathcal{F} \\ &= x_{i\bar{\mathcal{F}}}^\top \hat{w}_{\bar{\mathcal{F}}} - \|x_{i\bar{\mathcal{F}}}\|_2 \sqrt{2G_\lambda(\hat{w}, \hat{\alpha})/\lambda - \|\hat{w}_{\mathcal{F}}\|_2^2} \\ &\ge LB(x_i^\top w^*) \end{aligned}$$

$\tilde{UB}(x_i^\top w^*)$ also

Safe keeping

allows us to identify a part of active features/samples

Safe feature keeping

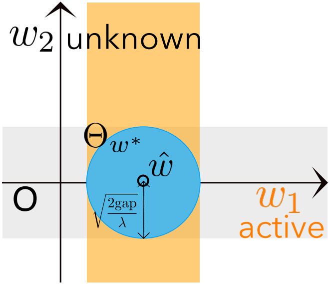
If P_λ is λ-strongly convex then

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

Safe sample keeping

If D_λ is γ/n-strongly convex then

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



Advantages of safe keeping

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

$$\Downarrow$$

we can calculate

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

$$\Downarrow$$

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

Summarize:

- Θ_{α^*} can safe feature screening and sample keeping
- Θ_{w^*} can safe sample screening and feature keeping

Optimization with simultaneous safe screening and keeping

Algorithm (Dynamic screening)

- Input: $\hat{w}_0, \hat{\alpha}_0, t \leftarrow 0$
- While convergence do;
  1. Safe keeping using $(\hat{w}_t, \hat{\alpha}_t)$
  2. While convergence do; (simultaneous safe screening)
     - Safe feature screening using the result of sample screening
     - Safe sample screening using the result of feature screening
  3. $(\hat{w}_{t+1}, \hat{\alpha}_{t+1}) \leftarrow$ Optimization update$(\hat{w}_t, \hat{\alpha}_t)$
  4. $t \leftarrow t + 1$

Experiments: Computation time saving

Setups:

- L_1-penalized smoothed hinge SVC
- $\lambda_{\max} := \|\mathrm{diag}(y)X^\top \mathbf{1}\|_\infty$
- train at 100 λs evenly allocated in $[10^{-4}\lambda_{\max}, \lambda_{\max}]$ in the logarithmic scale

Solvers:

- Stochastic Primal-Dual Coordinate (SPDC)
- Stochastic Dual Coordinate Ascent (SDCA)

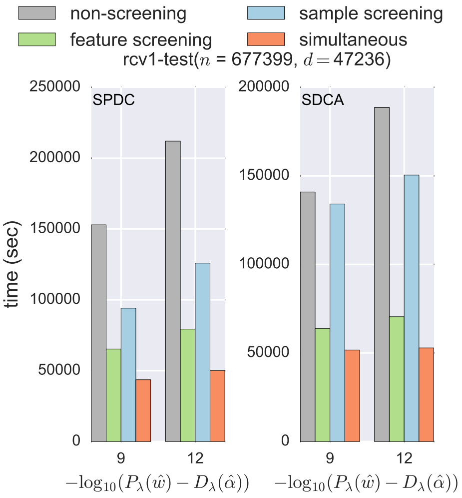
Experiments: Computation time saving

Setups:

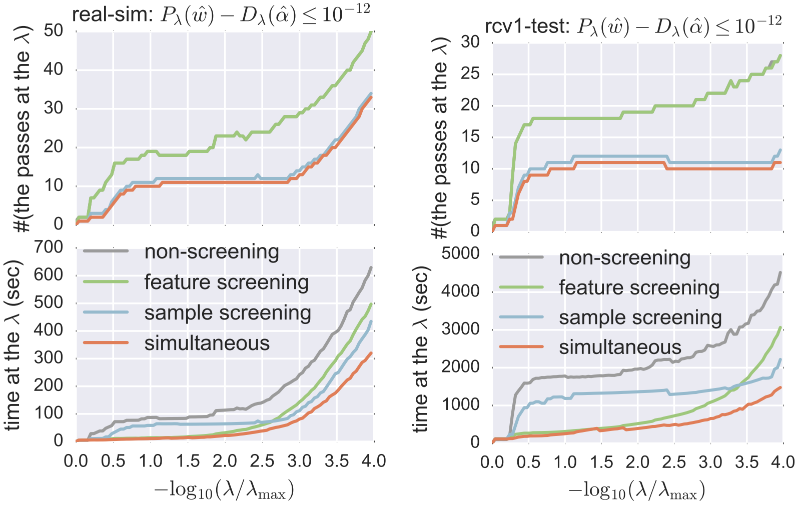
- L_1-penalized smoothed hinge SVC
- $\lambda_{\max} := \|\mathrm{diag}(y)X^\top \mathbf{1}\|_\infty$
- train at 100 λs evenly allocated in $[10^{-4}\lambda_{\max}, \lambda_{\max}]$ in the logarithmic scale

Solvers:

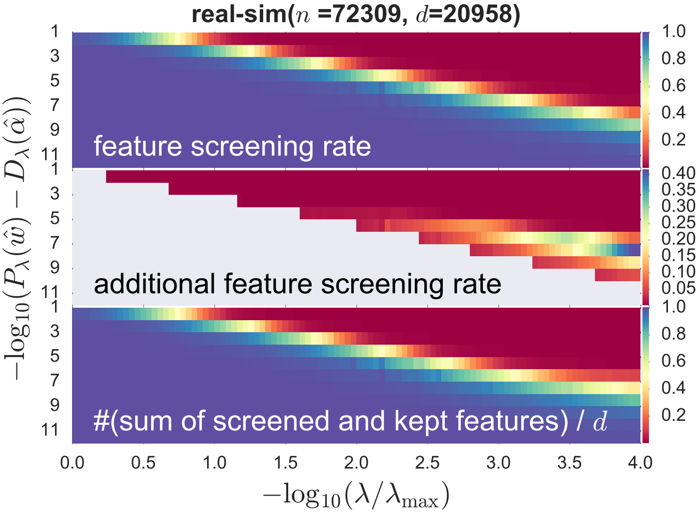
- Stochastic Primal-Dual Coordinate (SPDC)
- Stochastic Dual Coordinate Ascent (SDCA)



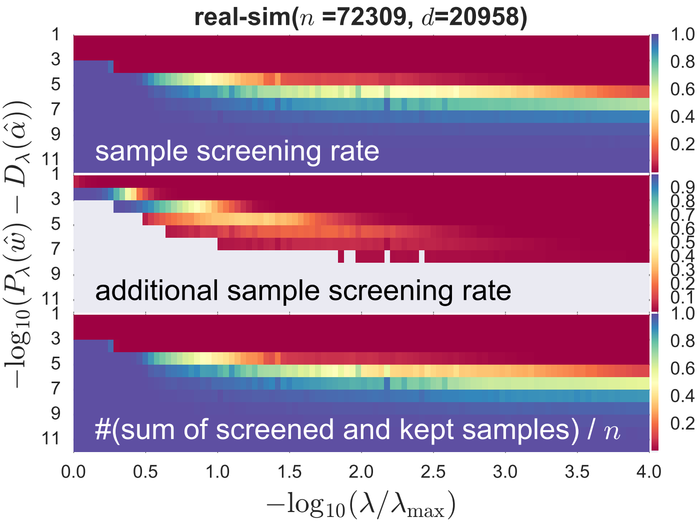
Computation time and iteration at each λ



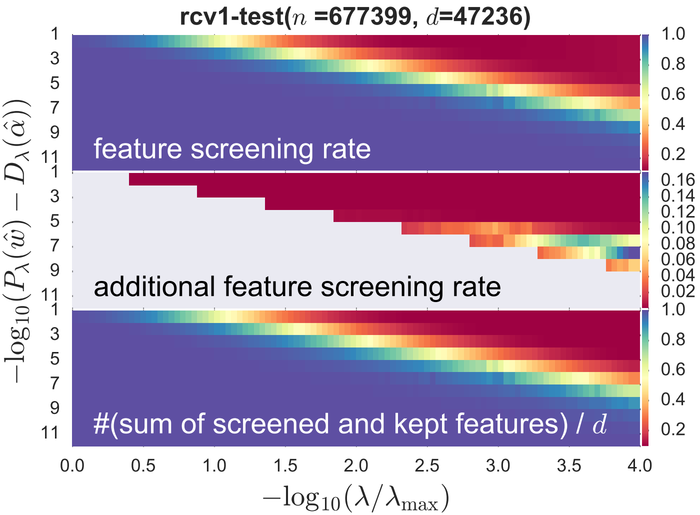
Safe feature screening and keeping rates



Safe sample screening and keeping rates



Safe feature screening and keeping rates



Safe sample screening and keeping rates