Screening Rules for Lasso with Non-Convex Sparse Regularizers

G. Gasso
Joint work with A. Rakotomamonjy, J. Salmon

May 23, 2019

LITIS EA 4108, INSA Rouen Normandie

Table of contents

At a glance

- Sparse high dimensional problems
  - Signal denoising
  - Compressive sensing
  - Bioinformatics ...

$$\min_w \quad \frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 \quad + \quad \lambda\,\Omega(\|\mathbf{w}\|)$$

Contribution

- Screening rules
  - Safely set $w_j = 0$ with few computation burden
- Speeding up Lasso solvers with non convex regularization

Context

Sparse Least squares

$$\min_{w \in \mathbb{R}^d} \quad \frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 \quad + \quad \lambda\ \|\mathbf{w}\|_0$$

- $\mathbf{y} \in \mathbb{R}^n$: observations
- $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_d] \in \mathbb{R}^{n \times d}$: design matrix, d features
- $\lambda > 0$: trade-off parameter between data-fit and regularization

Sparsity by the counting pseudo-norm

1. $\Omega(\mathbf{w}) = \sum_{j=1}^{d} \mathbb{I}_{w_j \neq 0}$
2. Number of non-zeros components of $\mathbf{w}$

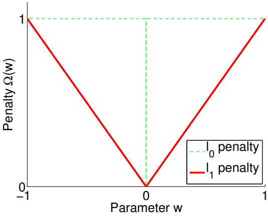
Sparse Least squares

$$\min_{w \in \mathbb{R}^d} \quad \frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 \quad + \quad \lambda \, \|\mathbf{w}\|_0$$

- $\mathbf{y} \in \mathbb{R}^n$: observations
- $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_d] \in \mathbb{R}^{n \times d}$: design matrix, d features
- $\lambda > 0$: trade-off parameter between data-fit and regularization

Convex relaxation

- ℓ_1-norm $\Omega(\mathbf{w}) = \|\mathbf{w}\|_1$
- Leading to Lasso problem (convex problem)



Lasso basics

Solving the Lasso : Cyclic Coordinate Descent

$$\min_{\mathbf{w}\in\mathbb{R}^d} P(\mathbf{w}) \triangleq \frac{1}{2}\|\mathbf{y} - \sum_{j=1}^{d} \mathbf{x}_j \mathbf{w}_j\|_2^2 + \lambda \sum_{j=1}^{d} |w_j|$$

Algorithm 1: Cyclic CD

Initialization: $\mathbf{w}^0 = \mathbf{0}$;

for $t = 1, \cdots, T$ **do**

$w_1^t \leftarrow \text{argmin}_{w_1\in\mathbb{R}}\ P(w_1,\ w_2^{t-1}, \cdots, w_{d-1}^{t-1},\ w_d^{t-1})$;

$w_2^t \leftarrow \text{argmin}_{w_2\in\mathbb{R}}\ P(w_1^t,\ w_2, \cdots, w_{d-1}^{t-1},\ w_d^{t-1})$;

$\vdots$;

$w_d^t \leftarrow \text{argmin}_{w_d\in\mathbb{R}}\ P(w_1^t,\ w_2^t, \cdots, w_{d-1}^{t-1},\ w_d)$;

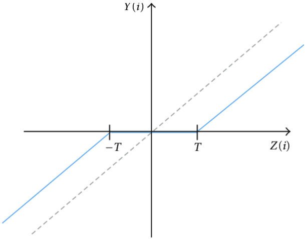
end

Coordinate-wise update

Soft thresholding

$$\mathbf{w}_j \leftarrow \mathrm{ST}\left(\frac{\lambda}{\|\mathbf{x}_j\|^2}, \mathbf{w}_j + \frac{\mathbf{x}_j^\top(\mathbf{y} - \mathbf{X}\mathbf{w})}{\|\mathbf{x}_j\|^2}\right)$$

$$\mathrm{ST}(\tau, z) = \max(0, 1 - \tau/|z|)\, z$$



- Easy computation
- $\mathcal{O}(n)$ operations for an update

Benefit of Lasso

$$\mathbf{w}^\star = \mathrm{argmin}_{\mathbf{w}\in\mathbb{R}^d}\ \frac{1}{2}\|\mathbf{y} - \sum_{j=1}^{d} \mathbf{x}_j \mathbf{w}_j\|_2^2 + \lambda \sum_{j=1}^{d} |w_j|$$

Key property

- Sparse solution $\mathbf{w}^\star$ is expected
- Let $\mathcal{S}_{\mathbf{w}^\star} = \left\{ j = 1, \cdots, d \quad w_j^\star \neq 0 \right\}$ the (small) support of $\mathbf{w}^\star$
- For large λ: $|\mathcal{S}_{\mathbf{w}^\star}| = p \ll d$

Holy grail

- Identify beforehand $\mathcal{S}_{\mathbf{w}^\star}$
- Leverage on it to solve a reduced problem

$$\mathbf{w}^\star_{\mathcal{S}_{\mathbf{w}^\star}} = \mathrm{argmin}_{\omega\in\mathbb{R}^p}\ \frac{1}{2}\|\mathbf{y} - \mathbf{X}_{\mathcal{S}_{\mathbf{w}^\star}}\omega\|_2^2 + \lambda\|\omega\|_1$$

Accelerate Lasso solving

Approaches

- Screening: remove parameter j whenever it is certified that $j \notin \mathcal{S}_{\mathbf{w}^\star}$

- Active set: identify parameters j likely in $\mathcal{S}_{\mathbf{w}^\star}$

Issue

How to identify $\mathcal{S}_{\mathbf{w}^\star}$ or subset of it ?

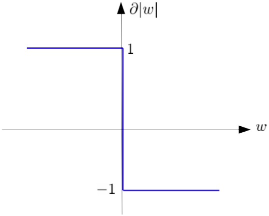
$\Longrightarrow$ Exploit the dual of Lasso and its optimality condition

Optimality condition

$$\mathbf{w}^\star = \mathrm{argmin}_{\mathbf{w}\in\mathbb{R}^d} \ \frac{1}{2}\|\mathbf{y} - \sum_{j=1}^{d} \mathbf{x}_j w_j\|_2^2 + \lambda \sum_{j=1}^{d} |w_j|$$

Subgradient of $|w|$

$$\partial_w |w| = \begin{cases} [-1, 1] & \text{if} \quad w = 0 \\ \{\mathrm{sign}(w)\} & \text{if} \quad w \neq 0 \end{cases}$$



Necessary and sufficient optimality condition

$$\forall j \ \ \exists g_j \in \partial_w |w_j|, \quad \mathbf{x}_j^\top (\mathbf{y} - \mathbf{X}\mathbf{w}) - \lambda g_j = 0$$

- Screening condition: owing to definition of g_j

$$|\mathbf{x}_j^\top (\mathbf{y} - \mathbf{X}\mathbf{w}^\star)| < \lambda \quad \Rightarrow \quad w_j^\star = 0$$

Dual of Lasso

The dual

$$\max_{\boldsymbol{\theta}\in\mathbb{R}^n} \quad D(\boldsymbol{\theta}) \triangleq \frac{1}{2}\|\mathbf{y}\|_2^2 - \frac{1}{2}\|\mathbf{y}/\lambda - \boldsymbol{\theta}\|_2^2$$
$$\text{s.t.} \quad |\mathbf{x}_j^\top \boldsymbol{\theta}| \leq 1 \quad \forall j = 1, \cdots, d$$

$\boldsymbol{\theta}^\star$ = solution of the projection of $\mathbf{y}/\lambda$ onto a polyhedral

Primal dual link

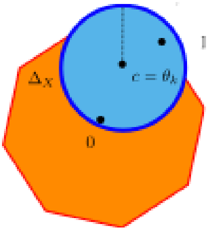
$$\boldsymbol{\theta}^\star = (\mathbf{y} - \mathbf{X}\mathbf{w}^\star)/\lambda$$

$\boldsymbol{\theta}$ is the scaled residual

Screening rule

$$|\mathbf{x}_j^\top \boldsymbol{\theta}^\star| < \lambda \quad \Rightarrow \quad w_j = 0$$

Useless rule as we still not know $\mathbf{w}^\star$



A proxy to the screening rule

- Find a region $\mathcal{C} \in \mathbb{R}^n$ containing $\boldsymbol{\theta}^\star$
- If $\sup_{\boldsymbol{\theta}\in\mathcal{C}} |\mathbf{x}_j^\top \boldsymbol{\theta}| < 1 \;\Rightarrow\; |\mathbf{x}_j^\top \boldsymbol{\theta}^\star| < 1 \;\Rightarrow\; w_j^\star = 0$

Choice of $\mathcal{C}$

- $\mathcal{C}$ is a ball of center $\mathbf{c} \in \mathbb{R}^n$ and radius $\rho > 0$
- Simple solution: $\sup_{\boldsymbol{\theta}\in\mathcal{C}} |\mathbf{x}_j^\top \boldsymbol{\theta}| = |\mathbf{x}_j^\top \mathbf{c}| + \rho\|\mathbf{x}_j\|_2$

Safe screening test

$$\text{if} \quad |\mathbf{x}_j^\top \mathbf{c}| + \rho\|\mathbf{x}_j\|_2 < \lambda \quad \Rightarrow \quad w_j^\star = 0$$

Computation requirement: $\mathcal{O}(n)$

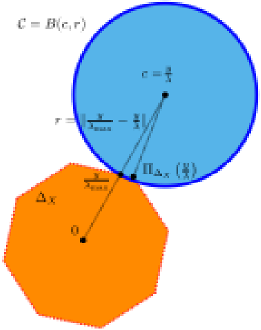
Review of Lasso screening rules

- To get a practical and useful rule
  - choose c close to $\boldsymbol{\theta}^\star$
  - choose the radius ρ as small as possible
- Leading to different screening rule

Static rule El Ghaoui et al. (2012)

$$\mathbf{c} = \mathbf{y}/\lambda, \quad \rho = \|\mathbf{y}/\lambda - \mathbf{y}/\lambda_{\max}\|$$

$\lambda_{\max} = \|\mathbf{X}^\top \mathbf{y}\|_\infty$ is the maximal correlation

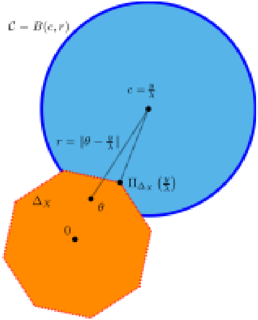


- To get a practical and useful rule
  - choose c close to $\boldsymbol{\theta}^\star$
  - choose the radius ρ as small as possible
- Leading to different screening rule

Dynamic rule Bonnefoy et al. (2014)

$$\mathbf{c} = \mathbf{y}/\lambda, \quad \rho = \|\boldsymbol{\theta}^k - \mathbf{y}/\lambda_{\max}\|$$

$\boldsymbol{\theta}^k = (\mathbf{y} - \mathbf{w}^k)/\alpha^k$ is a feasible scaled residual



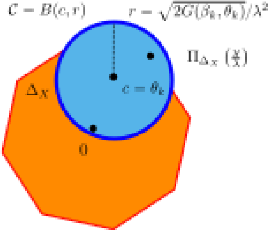
Review of Lasso screening rules

- To get a practical and useful rule
  - choose c close to $\boldsymbol{\theta}^\star$
  - choose the radius ρ as small as possible
- Leading to different screening rule

Duality Gap safe rule Fercocq et al. (2015)

$$\mathbf{c} = \boldsymbol{\theta}^k, \quad \rho = \sqrt{2\mathrm{Gap}(P(\mathbf{w}^k) - D(\boldsymbol{\theta}^k))}$$

$\boldsymbol{\theta}^k = (\mathbf{y} - \mathbf{w}^k)/\alpha^k$ is a feasible scaled residual, $\mathrm{Gap}(P(\mathbf{w}) - D(\boldsymbol{\theta}))$ is the duality gap (stopping criterion of a lasso solver)



Lasso with Gap screening rule

Algorithm 2: Cyclic CD with screening

Initialization: $\mathbf{w}^0 = \mathbf{0}$, t=0 ;

repeat

    if *t mod F = 0* **then**

        Design feasible residual $\boldsymbol{\theta}^t$;

        Set $\rho = \sqrt{2\text{Gap}(P(\mathbf{w}^t) - D(\boldsymbol{\theta}^t)}$;

        Screen safely parameters $w_j = 0$;

    end

    foreach $\ell \in \mathcal{S}_{\hat{\mathbf{w}}}$ *(not screeneed out)* **do**

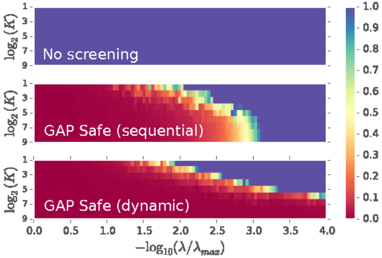
        $w_\ell^t \leftarrow \text{argmin}_{\mathbf{w}_\ell \in \mathbb{R}} P(w_1^t, \cdots, w_\ell \quad , \cdots, w_{d-1}^{t-1}, w_d^{t-1})$;

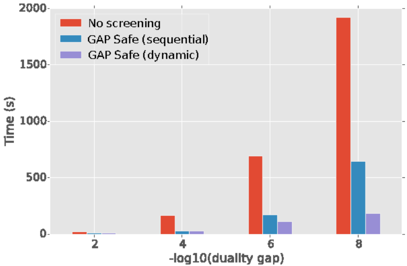
    end

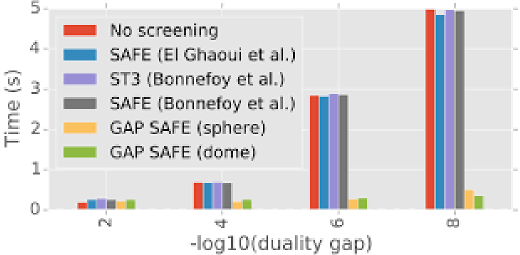
    $t = t + 1$;

until *$Gap(P(\mathbf{w}^t) - D(\boldsymbol{\theta}^t) < \epsilon$*;

Lasso screening rule in play







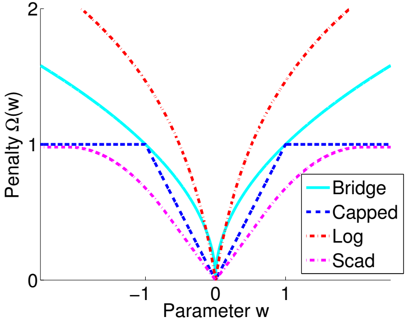
Leukemia dataset

$$\frac{d}{n} = \frac{7129}{72}$$

Non-convex Lasso

Why non-convex Lasso?

- Lasso tends to select larger support $\mathcal{S}_{\hat{\mathbf{w}}}$ (more parameters than needed)
- Remedy: use non-convex approximation of the sparsity $\|\mathbf{w}\|_0$



Log (**?**): $\Omega(\mathbf{w}) = \sum_{j=1}^{d} \log(|w_j| + \eta)$,

SCAD (**?**)

$$\mathbf{w}^\star = \mathrm{argmin}_{\mathbf{w}\in\mathbb{R}^d}\ \frac{1}{2}\|\mathbf{y} - \sum_{j=1}^{d} \mathbf{x}_j w_j\|_2^2 + \lambda \sum_{j=1}^{d} \Omega(|w_j|)$$

Issues

- Non-convex relaxations promote better sparsity. . .
- but their optimization is more challenging
- How to design screening rules as in the convex Lasso case?

Adopted optimization approach

$$\mathbf{w}^\star = \text{argmin}_{\mathbf{w}\in\mathbb{R}^d} \ \tfrac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda \sum_{j=1}^{d} \Omega(|w_j|)$$

Assumption

- Ω is concave, lower semi-continuous, differentiable on $[0, \ \infty)$
- Leading to a convex surrogate

$$\Omega(|w_j|) \leq \Omega(|w_j'|) + \Omega'(|w_j'|)\left(|w_j| - |w_j'|\right)$$

Majorization-Minimization Kang et al. (2015)

- At iteration t we assume knowing a $\mathbf{w}^t$
- Next iterate is obtained by

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \frac{1}{2\alpha}\|\mathbf{w} - \mathbf{w}^t\|_2^2 + \lambda \sum_{j=1}^{d} \Omega_\lambda'(|w_j^t|)|w_j| \ ,$$

Our MM algorithm

Algorithm 3: MM algorithm

Initialization: $\mathbf{w}^0 = \mathbf{0}$, t=0, set $\alpha > 0$;

repeat

  for $j = 1, \cdots, d$ **do**

    compute $\lambda_j = \lambda\ \Omega'_\lambda(|w_j^t|)$;

  end

  Solve the Proximal Weighted Lasso problem ;

  $\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \frac{1}{2\alpha}\|\mathbf{w} - \mathbf{w}^t\|_2^2 + \sum_{j=1}^d \lambda_j |w_j|$;

  $t = t + 1$;

until *convergence*;

Speeding up this solver

- Design screening rule for the Weighted Lasso
- Ensure screened out parameters remain at zero across MM iterations

Screen Proximal Weighted Lasso

Dual problem

$$\max_{\substack{\boldsymbol{\theta}\in\mathbb{R}^n \\ \boldsymbol{\beta}\in\mathbb{R}^d}} D(\boldsymbol{\theta},\boldsymbol{\beta}) \triangleq -\frac{1}{2}\|\boldsymbol{\theta}\|_2^2 - \frac{\alpha}{2}\|\boldsymbol{\beta}\|_2^2 + \boldsymbol{\theta}^\top \mathbf{y} - \boldsymbol{\beta}^\top \mathbf{w}^t$$

$$\text{s.t.} \quad |\mathbf{x}_j^\top \boldsymbol{\theta} - \boldsymbol{\beta}_j| \leq \lambda_j \quad \forall j$$

- Primal-dual link: $\mathbf{y} - \mathbf{X}\mathbf{w} = \boldsymbol{\theta} \quad \mathbf{w} - \mathbf{w}^t = \boldsymbol{\beta}$

Screening from optimality condition

$$|\mathbf{x}_j^\top \boldsymbol{\theta}^\star - \beta_j^\star| < \lambda_j \implies \mathbf{w}_j^\star = 0$$

- Unhelpful rule as it requires the optimal solution
- Effective screening rule: find an upper bound γ_j such that

$$|\mathbf{x}_j^\top \boldsymbol{\theta}^\star - \beta_j^\star| < \gamma_j < \lambda_j \implies w_j^\star = 0$$

Machinery of our screening rule

- Let $(\hat{\mathbf{w}}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}})$ with $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\beta}}$ being dual feasible, a primal-dual solution
- We can get a first upper bound

$$\begin{aligned} |\mathbf{x}_j^\top \boldsymbol{\theta}^\star - \beta_j^\star| &= |\mathbf{x}_j^\top \hat{\boldsymbol{\theta}} - \hat{\beta}_j + \mathbf{x}_j^\top (\boldsymbol{\theta}^\star - \hat{\boldsymbol{\theta}}) - (\beta_j^\star - \hat{\beta}_j)| \\ &\leq |\mathbf{x}_j^\top \hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\beta}}_j| + \|\mathbf{x}_j\| \, \|\boldsymbol{\theta}^\star - \hat{\boldsymbol{\theta}}\| + |\boldsymbol{\beta}_j^\star - \hat{\boldsymbol{\beta}}_j| \end{aligned}$$

- Get rid of the optimal solution $\boldsymbol{\theta}^\star$ and $\boldsymbol{\beta}^\star$: use duality gap

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\|_2^2 + \alpha \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_2^2 \leq 2(P(\hat{\mathbf{w}}) - D(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}}))$$

- All together

$$\underbrace{|\mathbf{x}_j^\top \hat{\boldsymbol{\theta}} - \hat{\beta}_j| + \sqrt{2\mathrm{gap}(\hat{\mathbf{w}}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}})}\left(\|\mathbf{x}_j\| + \frac{1}{\alpha}\right)}_{T_j^{(\lambda_j)}(\hat{\mathbf{w}}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}})} < \lambda_j \implies w_j = 0$$

Proximal Weighted Lasso with Gap screening rule

Algorithm 4: Cyclic PWL with screening

Inputs : $\mathbf{X}, \mathbf{w}^t, \{\lambda_j\}, \mathbf{w}^0, \alpha$;
Initialization: $k = 0$;
repeat
    if *k mod F = 0* **then**
        Design feasible dual variables $\boldsymbol{\theta}^k$ and $\boldsymbol{\theta}^k$;
        Compute the duality gap ;
        Screen safely parameters $w_j = 0$;
    end
    foreach $\ell \in \mathcal{S}^t_{\hat{\mathbf{w}}}$ *(not screeneed out)* **do**
        update w_j coordinate-wisely ;
    end
    $k = k + 1$;
until *convergence*;

Propagation of screened set

Algorithm 5: MM algorithm with screening

Initialization: $\mathbf{w}^0 = \mathbf{0}$, t=0, set $\alpha > 0$;
for $t = 1, \cdots, T$ **do**
  for $j = 1, \cdots, d$ **do**
    compute $\lambda_j^t = \lambda \, \Omega'_\lambda(|w_j^t|)$
  end
  Solve the Proximal Weighted Lasso problem ;
  $\mathbf{w}^{t+1}, \mathcal{S}_{\mathbf{w}}^{t+1} \leftarrow \text{ScreeningCyclicPWL}\left(\mathbf{X}, \mathbf{y}, \left\{\lambda_j^t\right\}, \mathbf{w}^t, \alpha\right)$
end

Remarks

- $\Lambda^t = \left\{\lambda_j^t\right\}$ changes at each MM iteration! $\implies$ $\mathcal{S}_{\mathbf{w}}^{t+1}$ changes across iterations
- Can we guarantee that some parameters w_j screened out at iteration t remained screened at $t+1$?

Propagation of screened set

Algorithm 6: MM algorithm with screening

Initialization: $\mathbf{w}^0 = \mathbf{0}$, t=0, set $\alpha > 0$;

for $t = 1, \cdots, T$ **do**

  for $j = 1, \cdots, d$ **do**

    compute $\lambda_j^t = \lambda\ \Omega'_\lambda(|w_j^t|)$

  end

  if *t mod K* **then**

    Propagate screened set

  end

  Solve the Proximal Weighted Lasso problem ;

  $\mathbf{w}^{t+1}, \mathcal{S}_{\mathbf{w}}^{t+1} \leftarrow \text{ScreeningCyclicPWL}\left(\mathbf{X}, \mathbf{y}, \left\{\lambda_j^t\right\}, \mathbf{w}^t, \alpha\right)$

end

Alleluia (maybe a hype?)

- We can propagate screened variables by checking

$$T_j^{(\lambda_j^t)}(\hat{\mathbf{w}}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}}) + c_1\|\mathbf{x}_j\| + c_2 \leq \lambda_j^{t+1} \implies w_j = 0$$

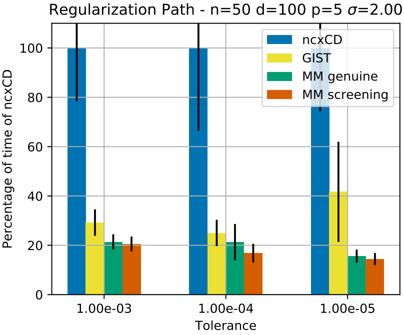
Evaluation

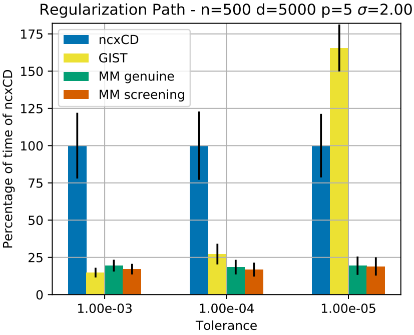
Empirical evaluation

Synthetic problem

$$\mathbf{y} = \mathbf{X}\,\mathbf{w} + \varepsilon$$

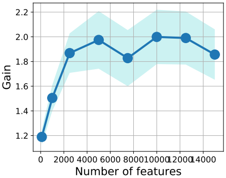
Regularization : $\Omega(\mathbf{w}) = \sum_{j=1}^{d} \log(|w_j| + \eta)$
Comparing running time for regularization path computation

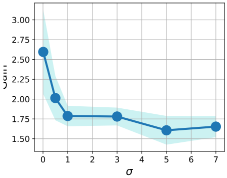


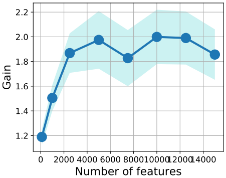


Empirical evaluation (continued)

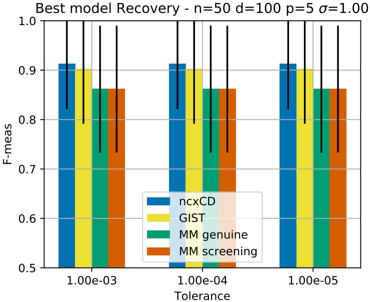
Benefit of screening set propagation

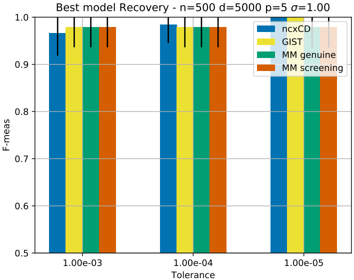






Best model recovery (ability to retrieve the true support $\mathcal{S}_{\mathbf{w}}$)



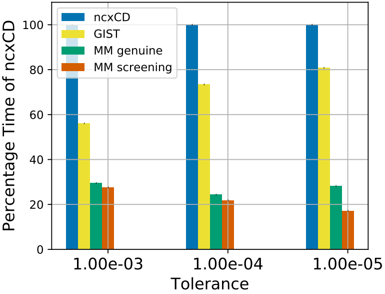


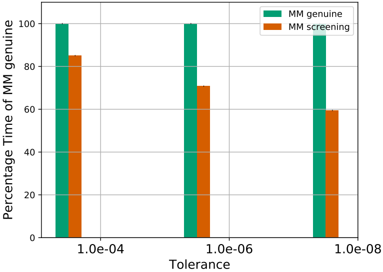
Empirical evaluation (end)

Real world datasets

(Left) Leukemia with $n = 50$, $d = 7129$,

(Right) Newsgroup with $n = 961$ and d = 21319





To sum up

- We address Lasso problem with non-convex regularization
- Design efficient screening rules
  - Screening rule for inner Majorization-Minimization Lasso
  - Propagation of the screening conditions
- Benefit: speed up of non-convex Lasso solver
- Future work
  - Under which condition the propagation rule is efficient?
  - Extension to other learning problem