Screening Rules for Lasso with Non-Convex Sparse Regularizers

Joseph Salmon

http://josephsalmon.eu

Université de Montpellier

Joint work with A. Rakotomamonjy and G. Gasso

Motivation and objective

Lasso and screening

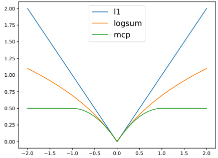
- Learning sparse regression models : $\mathbf{X} \in \mathbb{R}^{n\times d}$, $\mathbf{y} \in \mathbb{R}^n$

$$\min_{\mathbf{w}=(w_1,\dots,w_d)^\top \in \mathbb{R}^d} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 + \lambda \sum_{j=1}^{d} |w_j|$$

- **Safe screening rules** (1), (2) : identify vanishing coordinates of a/the solution by exploiting sparsity, convexity and duality

Extension to non-convex regularizers :

- non-convex regularizers lead to statistically better models but ...
- how to perform screening when the regularizer is non-convex ?



(1). L. El Ghaoui, V. Viallon et T. Rabbani. "Safe feature elimination in sparse supervised learning". In : *Journal of Pacific Optimization* (8 2012), p. 667-698.

(2). Antoine Bonnefoy et al. "Dynamic screening : Accelerating first-order algorithms for the lasso and group-lasso". In : *IEEE Trans. Signal Process.* 63.19 (2015), p. 5121-5132.

Non-convex sparse regression

Non convex regularization : $r_\lambda(\cdot)$ smooth & concave on $[0, \infty[$

$$\min_{\mathbf{w} \in \mathbb{R}^d} \frac{1}{2} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 + \sum_{j=1}^{d} r_\lambda(|w_j|)$$

Examples :

- *Log Sum penalty* (LSP) [(3)]
- *Smoothly Clipped Absolute Deviation* (SCAD) [(4)]
- *capped-ℓ_1 penalty* [(5)]
- *Minimax Concave Penalty* (MCP) [(6)]

Rem: for pros & cons of such formulations *cf.* Soubies *et al.* [(7)]

---

(3). Emmanuel J Candès, Michael B Wakin et Stephen P Boyd. "Enhancing Sparsity by Reweighted l_1 Minimization". In : *J. Fourier Anal. Applicat.* 14.5-6 (2008), p. 877-905.

(4). Jianqing Fan et Runze Li. "Variable selection via nonconcave penalized likelihood and its oracle properties". In : *J. Amer. Statist. Assoc.* 96.456 (2001), p. 1348-1360.

(5). Tong Zhang. "Analysis of multi-stage convex relaxation for sparse regularization". In : *Journal of Machine Learning Research* 11.Mar (2010), p. 1081-1107.

(6). Cun-Hui Zhang. "Nearly unbiased variable selection under minimax concave penalty". In : *Ann. Statist.* 38.2 (2010), p. 894-942.

(7). E. Soubies, L. Blanc-Féraud et G. Aubert. "A Unified View of Exact Continuous Penalties for ℓ_2-ℓ_0 Minimization". In : *SIAM J. Optim.* 27.3 (2017), p. 2034-2060.

Majorization-Minimization

Algorithm: Maximization minimization

---

input : max. iterations $k_{\max}$, stopping criterion ϵ, α, $\mathbf{w}^0 (= 0)$

for $k = 0, \ldots, k_{\max} - 1$ **do**

  Break if stopping criterion smaller than ϵ

  $\lambda_j^k \leftarrow r'_\lambda(|w_j^k|)$ // Majorization

  $$\mathbf{w}^k \leftarrow \arg\min_{\mathbf{w} \in \mathbb{R}^d} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 + \frac{1}{2\alpha}\|\mathbf{w} - \mathbf{w}^k\|^2 + \sum_{j=1}^d \lambda_j^k |w_j|$$ // Minimization

return $\mathbf{w}^k$

Majorization : $r_\lambda(|w_j|) \leq r_\lambda(|w_j^k|) + r'_\lambda(|w_j^k|)(|w_j| - |w_j^k|)$

Minimization : weighted-Lasso formulation

Rem : $\frac{1}{2\alpha}\|\mathbf{w} - \mathbf{w}^k\|^2$ acts as a regularization for MM [(8)] (other majorization alternatives possible, *e.g.*, with gradient information)

---

(8). Yangyang Kang, Zhihua Zhang et Wu-Jun Li. "On the global convergence of majorization minimization algorithms for nonconvex optimization problems". In : *arXiv preprint arXiv :1504.07791* (2015).

Safe Screening / Two-level screening

Safe Screening : for Lasso problems, vanishing coefficients at optimality can be certified without knowing the solution

- prior computation starting from a similar set of tuning parameter (sequential [9] / dual-warm start)
- along the optimization algorithm (dynamic [10])

State-of-the-art safe screening rules : rely on duality gap [11]

Two-level screening for non-convex cases :

- Inner level screening : within each (weighted) Lasso
- Outer level screening : propagate information between Lassos

(9). L. El Ghaoui, V. Viallon et T. Rabbani. "Safe feature elimination in sparse supervised learning". In : *Journal of Pacific Optimization* (8 2012), p. 667-698.

(10). Antoine Bonnefoy et al. "Dynamic screening : Accelerating first-order algorithms for the lasso and group-lasso". In : *IEEE Trans. Signal Process.* 63.19 (2015), p. 5121-5132.

(11). E. Ndiaye et al. "Gap Safe screening rules for sparsity enforcing penalties". In : *Journal of Machine Learning Research* 18.128 (2017), p. 1-33.

Safe Screening / Two-level screening

Safe Screening : for Lasso problems, vanishing coefficients at optimality can be certified without knowing the solution

- prior computation starting from a similar set of tuning parameter (sequential [9] / dual-warm start)
- along the optimization algorithm (dynamic [10])

State-of-the-art safe screening rules : rely on duality gap [11]

Two-level screening for non-convex cases :

- Inner level screening : within each (weighted) Lasso
- Outer level screening : propagate information between Lassos

(9). L. El Ghaoui, V. Viallon et T. Rabbani. "Safe feature elimination in sparse supervised learning". In : *Journal of Pacific Optimization* (8 2012), p. 667-698.

(10). Antoine Bonnefoy et al. "Dynamic screening : Accelerating first-order algorithms for the lasso and group-lasso". In : *IEEE Trans. Signal Process.* 63.19 (2015), p. 5121-5132.

(11). E. Ndiaye et al. "Gap Safe screening rules for sparsity enforcing penalties". In : *Journal of Machine Learning Research* 18.128 (2017), p. 1-33.

Notation

Notation : $X = [\mathbf{x}_1, \ldots, \mathbf{x}_d]$, $\Lambda = (\lambda_1, \ldots, \lambda_d)^\top$

Inner (convex) problems :

(Primal) $$P_\Lambda(\mathbf{w}) \triangleq \tfrac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 + \tfrac{1}{2\alpha}\|\mathbf{w} - \mathbf{w}^k\|^2 + \sum_{j=1}^{d} \lambda_j |w_j|$$

Notation

Notation : $X = [\mathbf{x}_1, \ldots, \mathbf{x}_d]$, $\Lambda = (\lambda_1, \ldots, \lambda_d)^\top$, $\mathbf{s} \in \mathbb{R}^n$, $\mathbf{v} \in \mathbb{R}^d$

Inner (convex) problems :

(Primal) $$P_\Lambda(\mathbf{w}) \triangleq \tfrac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 + \tfrac{1}{2\alpha}\|\mathbf{w} - \mathbf{w}^k\|^2 + \sum_{j=1}^{d} \lambda_j |w_j|$$

(Dual) $$D_\Lambda(\mathbf{s}, \mathbf{v}) \triangleq -\frac{1}{2}\|\mathbf{s}\|^2 - \frac{\alpha}{2}\|\mathbf{v}\|^2 + \mathbf{s}^\top\mathbf{y} - \mathbf{v}^\top\mathbf{w}^k \quad \text{s.t.} \quad |\mathbf{X}^\top\mathbf{s} - \mathbf{v}| \preccurlyeq \Lambda$$

Notation

Notation : $X = [\mathbf{x}_1, \ldots, \mathbf{x}_d]$, $\Lambda = (\lambda_1, \ldots, \lambda_d)^\top$, $\mathbf{s} \in \mathbb{R}^n$, $\mathbf{v} \in \mathbb{R}^d$

Inner (convex) problems :

(Primal) $$P_\Lambda(\mathbf{w}) \triangleq \tfrac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 + \tfrac{1}{2\alpha}\|\mathbf{w} - \mathbf{w}^k\|^2 + \sum_{j=1}^{d} \lambda_j |w_j|$$

(Dual) $$D_\Lambda(\mathbf{s}, \mathbf{v}) \triangleq -\frac{1}{2}\|\mathbf{s}\|^2 - \frac{\alpha}{2}\|\mathbf{v}\|^2 + \mathbf{s}^\top \mathbf{y} - \mathbf{v}^\top \mathbf{w}^k \quad \text{s.t.} \quad |\mathbf{X}^\top \mathbf{s} - \mathbf{v}| \preccurlyeq \Lambda$$

(Dual-Gap) $G_\Lambda(\mathbf{w}, \mathbf{s}, \mathbf{v}) \triangleq P_\Lambda(\mathbf{w}) - D(\mathbf{s}, \mathbf{v})$

Screening weighted Lasso

- Primal optimization problem $P_\Lambda(\mathbf{w})$:

$$\tilde{\mathbf{w}} \leftarrow \underset{\mathbf{w}\in\mathbb{R}^d}{\arg\min} \frac{1}{2}\|\mathbf{y}-\mathbf{X}\mathbf{w}\|^2 + \frac{1}{2\alpha}\|\mathbf{w}-\mathbf{w}^k\|^2 + \sum_{j=1}^{d} \lambda_j |w_j|$$

Screening test : $\boxed{|\mathbf{x}_j^\top \tilde{\mathbf{s}} - \tilde{v}_j| < \lambda_j \implies \tilde{w}_j = 0}$ (impractical)

with $\tilde{\mathbf{s}} \triangleq \frac{\mathbf{y}-\mathbf{X}\tilde{\mathbf{w}}}{\rho(\Lambda)}$, $\tilde{\mathbf{v}} \triangleq \frac{\tilde{\mathbf{w}}-\mathbf{w}^k}{\alpha\rho(\Lambda)}$ (for a scalar $\rho(\Lambda)$ well chosen)

- (Practical) Dynamic Gap safe screening test [12], [13] :

$$\underbrace{|\mathbf{x}_j^\top \mathbf{s} - v_j| + \sqrt{2G_\Lambda(\mathbf{w},\mathbf{s},\mathbf{v})}\left(\|\mathbf{x}_j\| + \frac{1}{\alpha}\right)}_{T_j^{(\Lambda)}(\mathbf{w},\mathbf{s},\mathbf{v})} < \lambda_j$$

given a primal-dual approximate solution triplet $(\mathbf{w},\mathbf{s},\mathbf{v})$

(12). O. Fercoq, A. Gramfort et J. Salmon. “Mind the duality gap : safer rules for the lasso”. In : *ICML*. 2015, p. 333-342.

(13). E. Ndiaye et al. “Gap Safe screening rules for sparsity enforcing penalties”. In : *Journal of Machine Learning Research* 18.128 (2017), p. 1-33.

Screening weighted Lasso

- Primal optimization problem $P_\Lambda(\mathbf{w})$:

$$\tilde{\mathbf{w}} \leftarrow \underset{\mathbf{w}\in\mathbb{R}^d}{\arg\min} \frac{1}{2}\|\mathbf{y}-\mathbf{X}\mathbf{w}\|^2 + \frac{1}{2\alpha}\|\mathbf{w}-\mathbf{w}^k\|^2 + \sum_{j=1}^{d} \lambda_j |w_j|$$

Screening test : $\boxed{|\mathbf{x}_j^\top \tilde{\mathbf{s}} - \tilde{v}_j| < \lambda_j \implies \tilde{w}_j = 0}$ (impractical)

with $\tilde{\mathbf{s}} \triangleq \frac{\mathbf{y}-\mathbf{X}\tilde{\mathbf{w}}}{\rho(\Lambda)}$, $\tilde{\mathbf{v}} \triangleq \frac{\tilde{\mathbf{w}}-\mathbf{w}^k}{\alpha\rho(\Lambda)}$ (for a scalar $\rho(\Lambda)$ well chosen)

- (Practical) Dynamic Gap safe screening test [(12), (13)] :

$$\underbrace{|\mathbf{x}_j^\top \mathbf{s} - v_j| + \sqrt{2G_\Lambda(\mathbf{w},\mathbf{s},\mathbf{v})}\left(\|\mathbf{x}_j\| + \frac{1}{\alpha}\right)}_{T_j^{(\Lambda)}(\mathbf{w},\mathbf{s},\mathbf{v})} < \lambda_j$$

given a primal-dual approximate solution triplet $(\mathbf{w},\mathbf{s},\mathbf{v})$

(12). O. Fercoq, A. Gramfort et J. Salmon. "Mind the duality gap : safer rules for the lasso". In : *ICML*. 2015, p. 333-342.

(13). E. Ndiaye et al. "Gap Safe screening rules for sparsity enforcing penalties". In : *Journal of Machine Learning Research* 18.128 (2017), p. 1-33.

Inner level screening and speed-ups

- After iteration k, one receives approximate solutions $\mathbf{w}^k$, $\mathbf{s}^k$ and $\mathbf{v}^k$ for weighted Lasso with weights Λ^k

  Set of screened variables :

$$\mathcal{S} \triangleq \left\{ j \in [\![1, d]\!] : T_j^{(\Lambda^k)}(\mathbf{w}^k, \mathbf{s}^k, \mathbf{v}^k) < \lambda_j^k \right\}$$

- Speed-ups : reduced weighted Lasso problem size substituting

$$X \leftarrow X_{\mathcal{S}^c}$$

  Rem : most beneficiary with coordinate descent type solvers

Outer screening level / screening propagation

Before iteration $k+1$

- change of weights $\Lambda^{k+1} = \{\lambda_j^{k+1}\}_{j=1,\dots,d}$
- update $(\mathbf{w}^{k+1}, \mathbf{s}^{k+1}, \mathbf{v}^{k+1}) \leftarrow \left(\mathbf{w}^k, \frac{\mathbf{y}-\mathbf{X}\mathbf{w}^k}{\rho(\Lambda^{k+1})}, \frac{\mathbf{w}^{k+1}-\mathbf{w}^k}{\rho(\Lambda^{k+1})}\right)$

Screening propagation test

$$T_j^{(\Lambda^k)}(\hat{\mathbf{w}}, \hat{\mathbf{s}}, \hat{\mathbf{v}}) + \|\mathbf{x}_j\|(a + \sqrt{2b}) + c + \frac{1}{\alpha}\sqrt{2b} < \lambda_j^{k+1}$$

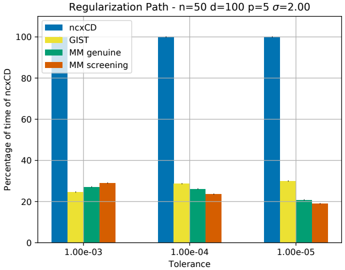
with

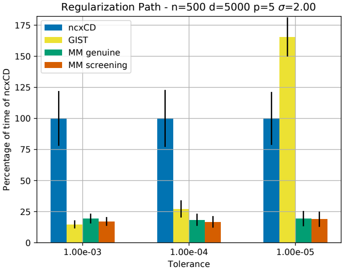
$$\begin{aligned} \|\mathbf{s}^{k+1} - \mathbf{s}^k\| &\leq a \\ |G_\Lambda(\mathbf{w}^k, \mathbf{s}^k, \mathbf{v}^k) - G_{\Lambda^{k+1}}(\mathbf{w}^{k+1}, \mathbf{s}^{k+1}, \mathbf{v}^{k+1})| &\leq b \\ |v_j^{k+1} - v_j^k| &\leq c \end{aligned}$$

Rem: same flavor as sequential screening [14]

(14). L. El Ghaoui, V. Viallon et T. Rabbani. “Safe feature elimination in sparse supervised learning”. In : *Journal of Pacific Optimization* (8 2012), p. 667-698.

Experiments (log-sum penalty)





- ncxCD : coordinate descent
- GIST : majorization + iterative-soft thresholding
- MM-genuine : screening inside proximal weighted Lasso steps
- MM-screening : adding screening propagation to the later

Conclusion

- First approach for screening with non-convex regularizers
- Convexification and propagation
- Limits (they exist !) : $\lambda_j > 0$ (cannot handle MCP easily)

- Variants : active set extension [15] following Massias *et al.* [16]
- More technical details [17] and code online
  https://github.com/arakotom/screening_ncvx_penalty

(15). A. Rakotomamonjy et al. *Provably Convergent Working Set Algorithm for Non-Convex Regularized Regression*. Rapp. tech. 2020.

(16). M. Massias, A. Gramfort et J. Salmon. “Celer : a Fast Solver for the Lasso with Dual Extrapolation”. In : *ICML*. 2018.

(17). A. Rakotomamonjy, G. Gasso et J. Salmon. “Screening Rules for Lasso with Non-Convex Sparse Regularizers”. In : *ICML*. T. 97. 2019, p. 5341-5350.

BenchOpt : https://benchopt.github.io/

BenchOpt : package to simplify, make more transparent and more reproducible [18] the comparisons of optimization algorithms

(18). J. B. Buckheit et D. L. Donoho. "Wavelab and reproducible research". In : *Wavelets and statistics*. Springer, 1995, p. 55-81.

BenchOpt : https://benchopt.github.io/

BenchOpt : package to simplify, make more transparent and more reproducible [18] the comparisons of optimization algorithms

Languages available : Python (default), R, Julia, C/C++

(18). J. B. Buckheit et D. L. Donoho. "Wavelab and reproducible research". In : *Wavelets and statistics*. Springer, 1995, p. 55-81.

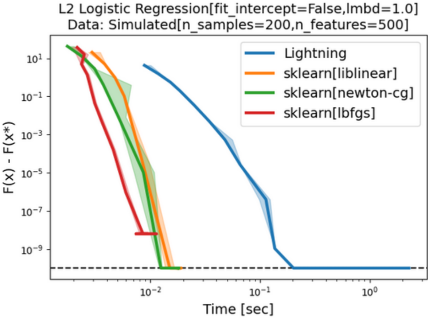
BenchOpt : https://benchopt.github.io/

BenchOpt : package to simplify, make more transparent and more reproducible [18] the comparisons of optimization algorithms

Languages available : Python (default), R, Julia, C/C++

```
$ git clone https://github.com/benchopt/benchmark_logreg_l2
$ benchopt run ./benchmark_logreg_l2
```

Running these commands will fetch the benchmark files and give you a benchmark plot on l2-regularized logistic regression:



(18). J. B. Buckheit et D. L. Donoho. "Wavelab and reproducible research". In : *Wavelets and statistics*. Springer, 1995, p. 55-81.

Disclaimer on `BenchOpt`

Use-cases : research, review, fast speed check on a machine

"*For now we handle convex batch methods, but we can do much more with your help (stochastic, non-convex, etc.)*" T. Moreau

"*We are family ! Come work with us :)*" A. Gramfort

Give it a try : https://benchopt.github.io/

Papers and code

Contact:

Joseph Salmon

joseph.salmon@umontpellier.fr

http://josephsalmon.eu

Github: @josephsalmon

Twitter: @salmonjsph

References I

- Bonnefoy, Antoine et al. "Dynamic screening : Accelerating first-order algorithms for the lasso and group-lasso". In : *IEEE Trans. Signal Process.* 63.19 (2015), p. 5121-5132.
- Buckheit, J. B. et D. L. Donoho. "Wavelab and reproducible research". In : *Wavelets and statistics*. Springer, 1995, p. 55-81.
- Candès, Emmanuel J, Michael B Wakin et Stephen P Boyd. "Enhancing Sparsity by Reweighted l_1 Minimization". In : *J. Fourier Anal. Applicat.* 14.5-6 (2008), p. 877-905.
- El Ghaoui, L., V. Viallon et T. Rabbani. "Safe feature elimination in sparse supervised learning". In : *Journal of Pacific Optimization* (8 2012), p. 667-698.
- Fan, Jianqing et Runze Li. "Variable selection via nonconcave penalized likelihood and its oracle properties". In : *J. Amer. Statist. Assoc.* 96.456 (2001), p. 1348-1360.
- Fercoq, O., A. Gramfort et J. Salmon. "Mind the duality gap : safer rules for the lasso". In : *ICML*. 2015, p. 333-342.

References II

- Kang, Yangyang, Zhihua Zhang et Wu-Jun Li. "On the global convergence of majorization minimization algorithms for nonconvex optimization problems". In : *arXiv preprint arXiv :1504.07791* (2015).
- Massias, M., A. Gramfort et J. Salmon. "Celer : a Fast Solver for the Lasso with Dual Extrapolation". In : *ICML*. 2018.
- Ndiaye, E. et al. "Gap Safe screening rules for sparsity enforcing penalties". In : *Journal of Machine Learning Research* 18.128 (2017), p. 1-33.
- Rakotomamonjy, A., G. Gasso et J. Salmon. "Screening Rules for Lasso with Non-Convex Sparse Regularizers". In : *ICML*. T. 97. 2019, p. 5341-5350.
- Rakotomamonjy, A. et al. *Provably Convergent Working Set Algorithm for Non-Convex Regularized Regression*. Rapp. tech. 2020.

References III

- Soubies, E., L. Blanc-Féraud et G. Aubert. "A Unified View of Exact Continuous Penalties for ℓ_2-ℓ_0 Minimization". In : *SIAM J. Optim.* 27.3 (2017), p. 2034-2060.
- Zhang, Cun-Hui. "Nearly unbiased variable selection under minimax concave penalty". In : *Ann. Statist.* 38.2 (2010), p. 894-942.
- Zhang, Tong. "Analysis of multi-stage convex relaxation for sparse regularization". In : *Journal of Machine Learning Research* 11.Mar (2010), p. 1081-1107.

Appendix

Computation of ρ, needed for feasibility :

$$j^\dagger = \underset{j:\lambda_j>0}{\arg\max} \underbrace{\frac{1}{\lambda_j}\left|\mathbf{x}_j^\top(\mathbf{y}-\mathbf{X}\mathbf{w}) - \frac{1}{\alpha}(\hat{w}_j - \hat{w}_j)\right|}_{\rho^\Lambda(j)} \quad . \tag{1}$$

with $\mathbf{w}^k$ coming from the previous problem,*i.e.*, solving :

$$P_\Lambda(\mathbf{w}) \triangleq \frac{1}{2}\|\mathbf{y}-\mathbf{X}\mathbf{w}\|^2 + \frac{1}{2\alpha}\|\mathbf{w}-\mathbf{w}^k\|^2 + \sum_{j=1}^{d}\lambda_j|w_j|$$