Exploiting regularity in sparse Generalized Linear Models

M. Massias♠, S. Vaiter†, A. Gramfort♠, J. Salmon⋆
♠ INRIA, Université Paris-Saclay, France. † CNRS & Institut de Mathématiques de Bourgogne, France.
⋆ IMAG, Univ Montpellier, CNRS, France.

Sparse GLM

- $y \in \mathbb{R}^n$: observations
- $X = [X_1 | \dots | X_p] \in \mathbb{R}^{n \times p}$: design matrix, line: $\mathbf{x}_i$
- $\lambda > 0$: trade-off parameter between data-fit and regularization

$$\hat{\beta} \in \arg\min_{\beta \in \mathbb{R}^p} \underbrace{\sum_{i=1}^n f_i(\beta^\top \mathbf{x}_i) + \lambda \|\beta\|_1}_{\mathcal{P}(\beta)} \qquad \textbf{(Primal)}$$

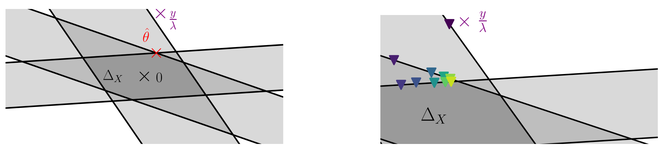
$$\hat{\theta} = \arg\max_{\theta \in \Delta_X} \underbrace{\left(- \sum_{i=1}^n f_i^*(-\lambda \theta_i) \right)}_{\mathcal{D}(\theta)} \qquad \textbf{(Dual)}$$

$$\hat{\theta} = -\nabla F(X\hat{\beta})/\lambda \quad \text{with} \quad F(u) \overset{\text{def.}}{=} \sum_{i=1}^n f_i(u_i)$$

$$\mathcal{D}(\theta) \leq \mathcal{D}(\hat{\theta}) = \mathcal{P}(\hat{\beta}) \leq \mathcal{P}(\beta) \qquad \textbf{(Primal-dual relations)}$$

$$E \overset{\text{def.}}{=} \{j \in [p] : |X_j^\top \hat{\theta}| = 1\} = \left\{ j \in [p] \,:\, |X_j \nabla F(X\hat{\beta})| = \lambda \right\} \qquad \textbf{(Equicorrelation set)}$$

$\Delta_X = \left\{ \theta \in \mathbb{R}^n \,:\, \forall j \in [p],\ |X_j^\top \theta| \leq 1 \right\}$: dual feasible set
$(n = 2, p = 3$ example$)$



Our focus: solve the primal iteratively with cyclic coordinate descent (CD), *i.e.,* minimize $\mathcal{P}(\beta)$ *w.r.t.* β_1, then β_2, etc.

Stopping criterion?

$$\forall \beta, (\exists \theta \in \Delta_X,\ \text{gap}(\beta, \theta) \leq \epsilon) \Rightarrow \mathcal{P}(\beta) - \mathcal{P}(\hat{\beta}) \leq \epsilon$$

1 epoch of CD costs $\mathcal{O}(np) \rightarrow$ compute the gap every 10 epochs.
Key property to speed-up solvers:

$$j \notin E \Rightarrow \hat{\beta}_j = 0, \quad aka \quad |X_j^\top \hat{\theta}| < 1 \Rightarrow \hat{\beta}_j = 0$$

Solving the primal restricted to E is easier! But E is unknown.

Duality matters

Safe screening rule proxy: find a region $\mathcal{C} \subset \mathbb{R}^n$ containing $\hat{\theta}$:

$$\sup_{\theta \in \mathcal{C}} |X_j^\top \theta| < 1 \Rightarrow |X_j^\top \hat{\theta}| < 1 \Rightarrow j \notin E \Rightarrow \hat{\beta}_j = 0$$

Gap Safe rule [1]: $\mathcal{C} = \mathcal{B}\left(\theta, \rho = \sqrt{\frac{2}{\gamma \lambda^2} \text{gap}(\beta, \theta)}\right)$

$$\boxed{\forall (\beta, \theta), d_j(\theta) \overset{\text{def.}}{=} \frac{1 - |X_j^\top \theta|}{\|X_j\|} > \rho \Rightarrow \hat{\beta}_j = 0}$$

Alternative approach: **working set** (WS) [2] for a dual point θ, solve the primal restricted to coordinates j with the smallest $d_j(\theta)$, update θ and define a new WS until convergence.

New dual point: extrapolation [3]

Rescaled residuals: at CD epoch t, for β^t, take

$$\boxed{\theta^t_{\text{res}} \overset{\text{def.}}{=} -\nabla F(X\beta^t) / \max(\lambda, \|X^\top \nabla F(X\beta^t)\|_\infty)}$$

pros	cons
$\mathcal{O}(np)$ cost (1 epoch of CD) converges to $\hat{\theta}$	ignores past information workload "imbalanced"

Thm (sign identification): $\exists T,\ t \geq T \implies \text{sign}\, \beta^t = \text{sign}\, \hat{\beta}$

Thm (VAR sequence): After sign identification, $X\beta^t$ is an (asymptotic) VAR sequence:

$$X\beta^{t+1} = AX\beta^t + b$$

$\rightarrow$ can be extrapolated.

Extrapolated residuals (our approach, $r^t = X\beta^t$):

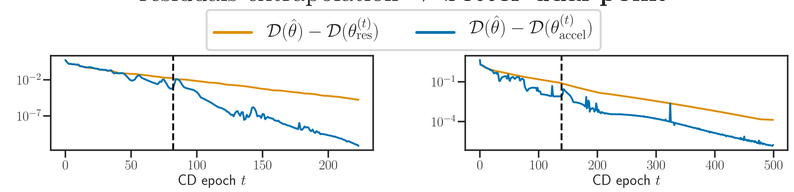
- form $U^t = [r^{t+1-K} - r^{t-K}, \dots, r^t - r^{t-1}] \in \mathbb{R}^{n \times K}$ $(K = 5)$
- solve $(U^t)^\top U^t z = \mathbf{1}_K$, set $c = z / z^\top \mathbf{1}_K$

For $t > K$: $\quad r^t_{\text{accel}} \overset{\text{def.}}{=} \sum_{k=1}^K c_k r^{t+1-k}$

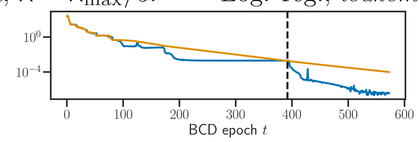
$$\boxed{\theta^t_{\text{acc}} \overset{\text{def.}}{=} -\nabla F(r^t_{\text{accel}}) / \max(\lambda, \|X^\top \nabla F(r^t_{\text{accel}})\|_\infty)}$$

Performance

residuals extrapolation $\Rightarrow$ **better dual point**

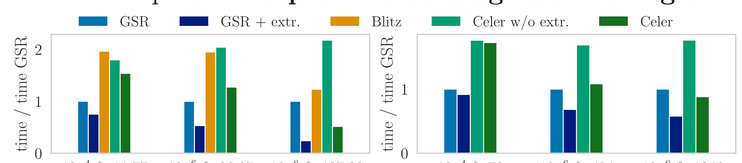


Lasso, *leukemia*, $\lambda = \lambda_{\max}/5.$ Log. reg., *leukemia*, $\lambda = \lambda_{\max}/10.$

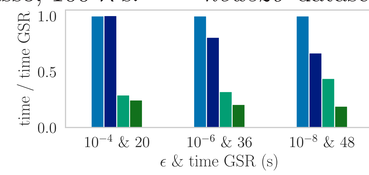


Multitask Lasso, M/EEG data, $\lambda = \lambda_{\max}/20.$

better dual point $\Rightarrow$ **improves screening and working sets**



rcv1 dataset, Lasso, 100 λ's. *news20* dataset, Logreg, 100 λ's.



MEG data, Multitask Lasso, 10 λ's.

Code: drop-in sklearn replacement

https://github.com/mathurinm/celer

$\rightarrow$ support for Lasso, LogisticRegression, GroupLasso, MultitaskLasso

- pip-installable: `pip install celer`
- documentation with examples
- continuous integration & bug tracker

Lasso article: Celer: a fast solver for the Lasso with dual extrapolation, M. Massias, A. Gramfort and J. Salmon, *ICML* 2018.

This work was funded by ERC Starting Grant SLAB ERC-YStG-676943 and by the chair Machine Learning for Big Data of Télécom ParisTech.

[1] Ndiaye, E., Fercoq, O., Gramfort, A., and Salmon, J. Gap safe screening rules for sparsity enforcing penalties. *J. Mach. Learn. Res.*, 18(128):1–33, 2017.
[2] T. B. Johnson and C. Guestrin. Blitz: A principled meta-algorithm for scaling sparse optimization. In *ICML*, pages 1171–1179, 2015.
[3] Scieur, D., d'Aspremont, A., and Bach, F. Regularized nonlinear acceleration. In *NIPS*, pp. 712–720, 2016.