# Coordinate Descent for SLOPE

Johan Larsson[1] Quentin Klopfenstein[2,3] Mathurin Massias[3,4]
Jonas Wallin[1]

[1]Department of Statistics, Lund University [2]Luxembourg Centre for Systems Biomedicine
[3]University of Luxembourg, Luxembourg [4]University of Lyon, Inria, CNRS, ENS de Lyon [5]UCB
Lyon 1, LIP UMR 5668, F-69342

March 31, 2023

  

# Synopsis

**The Problem**

SLOPE is a sparsity-inducing model with appealing properties, but the best algorithms (up til now) for solving SLOPE are slow.

**Our Contribution**

A hybrid algorithm based on coordinate descent and proximal gradient descent.

# Sorted L-One Penalized Estimation (SLOPE)

For a design matrix $X \in \mathbb{R}^{n \times p}$ and response vector $y \in \mathbb{R}^n$, the solution to SLOPE is

$$\beta^* \in \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} \left\{ P(\beta) = \frac{1}{2} \|y - X\beta\|^2 + J(\beta) \right\}$$

where

$$J(\beta) = \sum_{j=1}^{p} \lambda_j |\beta_{(j)}|$$

is the sorted $\ell_1$ norm, defined through

$$|\beta_{(1)}| \geq |\beta_{(2)}| \geq \cdots \geq |\beta_{(p)}|, \tag{1}$$

with $\lambda$ being a fixed non-increasing and non-negative sequence.

# Sorted L-One Penalized Estimation (SLOPE)

For a design matrix $X \in \mathbb{R}^{n\times p}$ and response vector $y \in \mathbb{R}^n$, the solution to SLOPE is

$$\beta^* \in \operatorname*{arg\,min}_{\beta\in\mathbb{R}^p} \left\{ P(\beta) = \frac{1}{2}\|y - X\beta\|^2 + J(\beta) \right\}$$

where

$$J(\beta) = \sum_{j=1}^{p} \lambda_j |\beta_{(j)}|$$

is the sorted $\ell_1$ norm, defined through

$$|\beta_{(1)}| \geq |\beta_{(2)}| \geq \cdots \geq |\beta_{(p)}|, \tag{1}$$

with $\lambda$ being a fixed non-increasing and non-negative sequence.

### Generalizations

- $\lambda_1 = \cdots = \lambda_p \to \ell_1$ (the lasso penalty)
- $\lambda_1 > \lambda_2 = \cdots = \lambda_p = 0 \to \ell_\infty$

# Properties

- Clustering (Bogdan, Dupuis, et al. 2022; Schneider and Tardivel 2020; Figueiredo and Nowak 2016)
- Control of false discovery rate (Bogdan, Berg, Su, et al. 2013; Bogdan, Berg, Sabatti, et al. 2015)
- Recovery of sparsity and ordering patterns (Bogdan, Dupuis, et al. 2022)
- Convexity

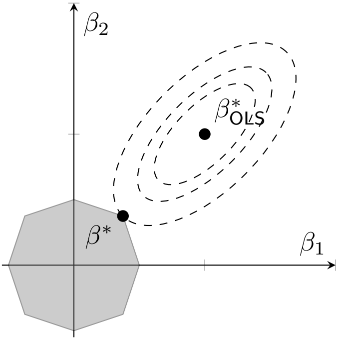


**Figure 1:** The SLOPE solution seen as a constrained problem.

# Why Does Not Everyone Use SLOPE?

- The lasso is much more popular than SLOPE.

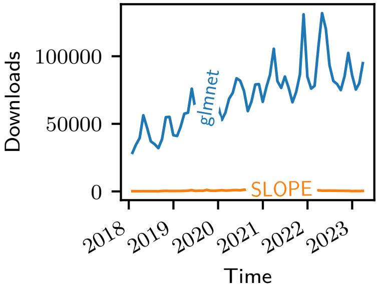


**Figure 2:** CRAN download statistics for the SLOPE and glmnet (lasso) packages.

# Why Does Not Everyone Use SLOPE?

- The lasso is much more popular than SLOPE.
- One reason is that current state-of-the-art algorithms for fitting the lasso are faster.

  **Example**: Fitting the `bcTCGA` data set with the R-package SLOPE takes 43 seconds versus 0.14 seconds for glmnet (lasso).

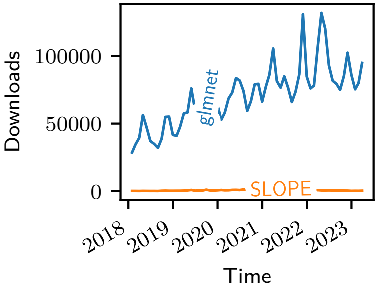


**Figure 2:** CRAN download statistics for the SLOPE and glmnet (lasso) packages.

- Coordinate descent (CD) works great for the lasso (Friedman, Hastie, and Tibshirani 2010).

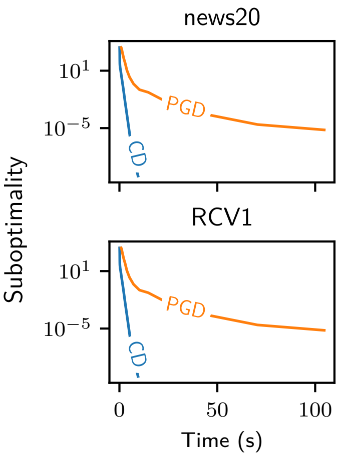


**Figure 3:** Coordinate descent versus proximal gradient descent for the lasso.

# Coordinate Descent

- Coordinate descent (CD) works great for the lasso (Friedman, Hastie, and Tibshirani 2010).
- Unfortunately, we cannot directly use CD for SLOPE since the sorted $\ell_1$ norm is not separable:

$$J(\beta) = \sum_{j=1}^{p} \lambda_j |\beta_{(j)}|.$$

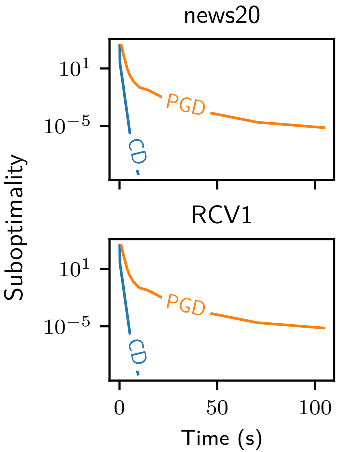


**Figure 3:** Coordinate descent versus proximal gradient descent for the lasso.

# The SLOPE Problem is Not Separable

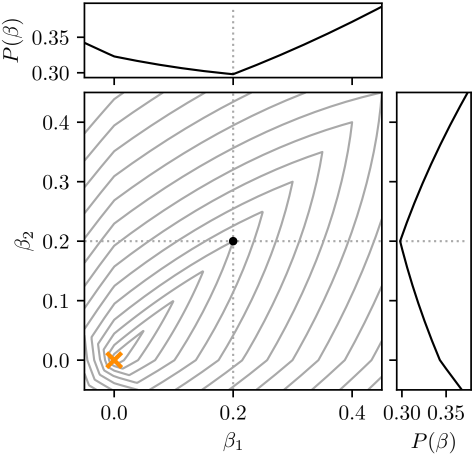


**Figure 4:** A *naive* coordinate descent algorithm cannot advance from the current iterate (●) to reach the optimum (✖).

# Clusters Are Not Known In Advance

If the clusters were known, the problem would become separable,

$$\min_{z \in \mathbb{R}^{m^*}} \left( \frac{1}{2} \Big\| y - X \sum_{i=1}^{m^*} \sum_{j \in \mathcal{C}_i^*} z_i \operatorname{sign}(\beta_j^*) e_j \Big\|^2 + \sum_{i=1}^{m^*} |z_i| \sum_{j \in \mathcal{C}_i^*} \lambda_j \right),$$

and we could solve it using CD.

# Clusters Are Not Known In Advance

If the clusters were known, the problem would become separable,

$$\min_{z \in \mathbb{R}^{m^*}} \left( \frac{1}{2} \Big\| y - X \sum_{i=1}^{m^*} \sum_{j \in \mathcal{C}_i^*} z_i \operatorname{sign}(\beta_j^*) e_j \Big\|^2 + \sum_{i=1}^{m^*} |z_i| \sum_{j \in \mathcal{C}_i^*} \lambda_j \right),$$

and we could solve it using CD.

**Idea**

Alternate between gradient descent steps that identify the clusters (via partial smoothness) and coordinate descent steps on the clusters, which enable fast convergence.

# Hybrid Algorithm

- Every $v$th iteration, take a full proximal gradient step. This allows clusters to split (or merge).
- At all other iterations, take coordinate descent steps on the clusters.

# Hybrid Algorithm

- Every $v$th iteration, take a full proximal gradient step. This allows clusters to split (or merge).
- At all other iterations, take coordinate descent steps on the clusters.

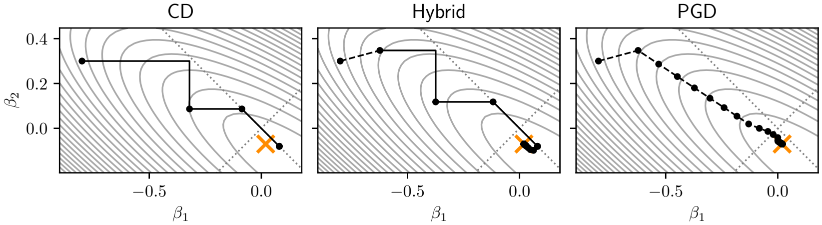


**Figure 5:** Our algorithm (hybrid) is a combination of CD and PGD.

# Coordinate Descent Steps

When updating the $k$th cluster, we let

$$\beta_i(z) = \begin{cases} \operatorname{sign}(\beta_i) z, & \text{if } i \in \mathcal{C}_k, \\ \beta_i, & \text{otherwise.} \end{cases}$$

# Coordinate Descent Steps

When updating the $k$th cluster, we let

$$\beta_i(z) = \begin{cases} \mathrm{sign}(\beta_i)z, & \text{if } i \in \mathcal{C}_k, \\ \beta_i, & \text{otherwise.} \end{cases}$$

Minimizing the objective in this direction amounts to solving the following one-dimensional problem:

$$\min_{z\in\mathbb{R}} \Big( G(z) = P(\beta(z)) = \frac{1}{2}\|y - X\beta(z)\|^2 + H(z) \Big),$$

where

$$H(z) = |z| \sum_{j\in\mathcal{C}_k} \lambda_{(j)^-_z} + \sum_{j\notin\mathcal{C}_k} |\beta_j|\lambda_{(j)^-_z}$$

is the *partial sorted $\ell_1$ norm* with respect to the $k$-th cluster and where we write $\lambda_{(j)^-_z}$ to indicate that the inverse sorting permutation $(j)^-_z$ is defined with respect to $\beta(z)$.

# The Partial Sorted $\ell_1$ Norm

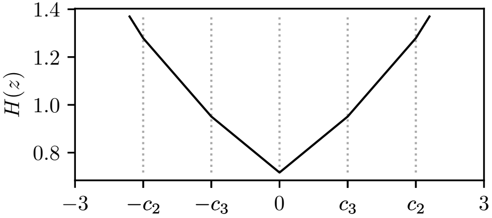


**Figure 6:** The partial sorted $\ell_1$ norm with $\beta = [-3, 1, 3, 2]^T$, $k = 1$, and so $c_1, c_2, c_3 = (3, 2, 1)$.

# How Do We Minimize Over One Cluster?

The optimality condition, using the directional derivative, is

$$\forall \delta \in \{-1, 1\}, \quad G'(z; \delta) \geq 0,$$

with

$$G'(z;\delta) = \delta \sum_{j \in \mathcal{C}_k} X_{:j}^\top (X\beta(z) - y) + H'(z;\delta).$$

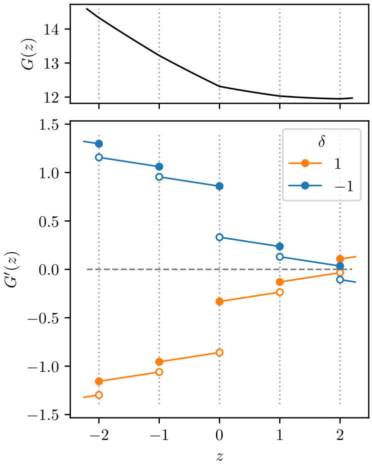


**Figure 7:** $G$ and its directional derivative $G'(\cdot;\delta)$ for an example with $\beta = [-3, 1, 3, 2]^T$, $k = 1$, and consequently $c^{\setminus k} = \{1, 2\}$.

# The SLOPE Thresholding Operator

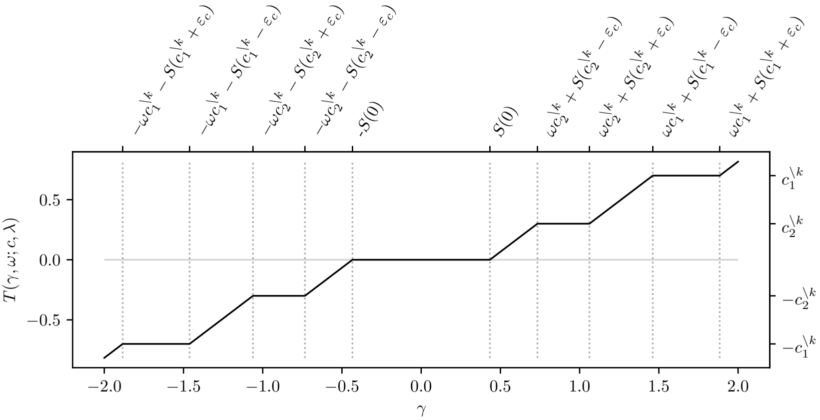


**Figure 8:** The SLOPE Thresholding Operator

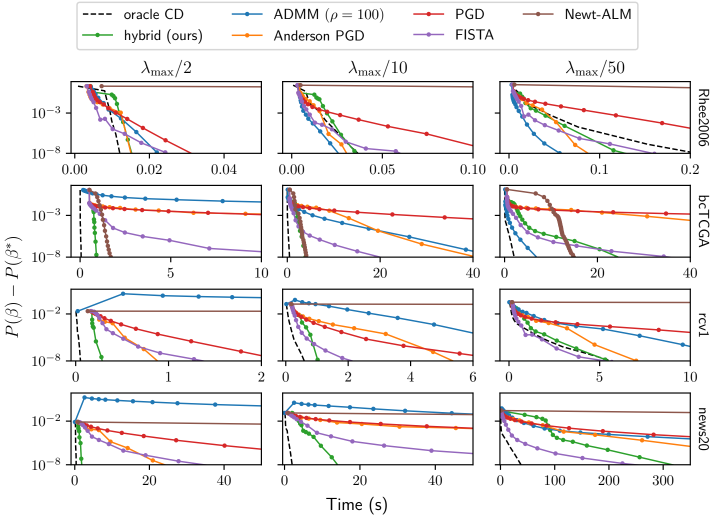


**Figure 9:** Benchmarks on real data

# Experiments

## Simulated Data

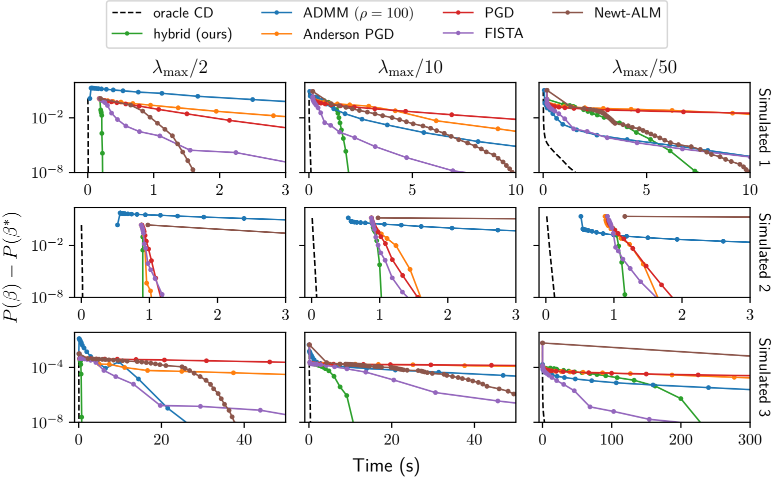


**Figure 10:** Benchmarks on simulated data. Scenario 1: $n = 200$ and $p = 20\,000$, $X$. Scenario 2: $n = 20\,000$ and $p = 200$. Scenario 3: $n = 200$, $p = 200\,000$, and sparse $X$.

# Wrap Up

- Experiments were set up using Benchopt (benchopt.github.io)
- Code is available at github.com/jolars/slopecd
- Add your own solver for SLOPE at github.com/benchopt/benchmark_SLOPE

# References I


[1] Małgorzata Bogdan, Ewout van den Berg, Chiara Sabatti, et al. "SLOPE – Adaptive Variable Selection via Convex Optimization". In: *The annals of applied statistics* 9.3 (Sept. 2015), pp. 1103–1140. ISSN: 1932-6157. DOI: 10.1214/15-AOAS842. pmid: 26709357. URL: https://projecteuclid.org/euclid.aoas/1446488733 (visited on 12/17/2018).

[2] Małgorzata Bogdan, Ewout van den Berg, Weijie Su, et al. "Statistical Estimation and Testing via the Sorted L1 Norm". Oct. 29, 2013. arXiv: 1310.1969 [math, stat]. URL: http://arxiv.org/abs/1310.1969 (visited on 04/16/2020).

[3] Małgorzata Bogdan, Xavier Dupuis, et al. "Pattern Recovery by SLOPE". May 17, 2022. DOI: 10.48550/arXiv.2203.12086. arXiv: 2203.12086 [math, stat]. URL: http://arxiv.org/abs/2203.12086 (visited on 06/03/2022).

# References II


[4] Mario Figueiredo and Robert Nowak. "Ordered Weighted L1 Regularized Regression with Strongly Correlated Covariates: Theoretical Aspects". In: *Artificial Intelligence and Statistics*. Artificial Intelligence and Statistics. May 2, 2016, pp. 930–938. URL: http://proceedings.mlr.press/v51/figueiredo16.html (visited on 11/05/2019).

[5] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. "Regularization Paths for Generalized Linear Models via Coordinate Descent". In: *Journal of Statistical Software* 33.1 (Jan. 2010), pp. 1–22. DOI: 10.18637/jss.v033.i01. URL: http://www.jstatsoft.org/v33/i01/.

[6] Ulrike Schneider and Patrick Tardivel. "The Geometry of Uniqueness, Sparsity and Clustering in Penalized Estimation". Aug. 18, 2020. DOI: 10.48550/arXiv.2004.09106. arXiv: 2004.09106 [math, stat]. URL: http://arxiv.org/abs/2004.09106 (visited on 06/03/2022).