Post-Doc Interview Presentation

Johan Larsson

Department of Statistics, Lund University

January 16, 2024

Outline

Coordinate Descent for SLOPE (Latest Work)

Coordinate Descent for SLOPE

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 (CD) and proximal gradient descent.

Coordinate Descent for SLOPE

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 (CD) and proximal gradient descent.

A collaboration with Quentin Klopfenstein, Mathurin Massias, and Jonas Wallin.

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 ℓ_1 norm, defined through

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

with λ 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 ℓ_1 norm, defined through

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

with λ being a fixed non-increasing and non-negative sequence.

Special Cases

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

Properties

SLOPE has many appealing 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

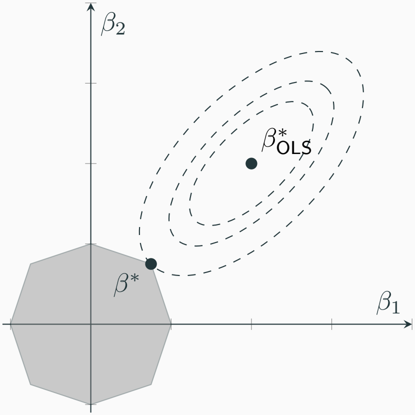

Figure 1: The SLOPE solution seen as a constrained problem.

Properties

SLOPE has many appealing 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

So why isn't SLOPE more popular?

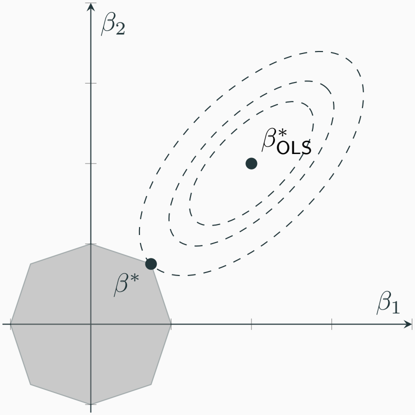

Figure 1: The SLOPE solution seen as a constrained problem.

Coordinate Descent

- Partly because the best solvers for the lasso use coordinate descent.

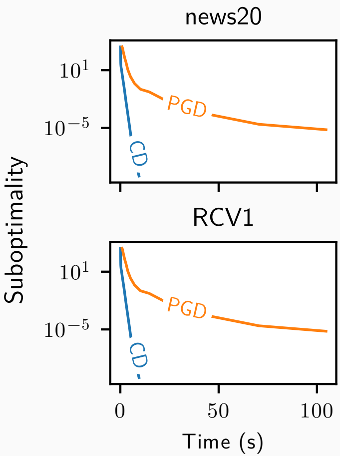

Figure 2: Coordinate descent versus proximal gradient descent for the lasso.

Coordinate Descent

- Partly because the best solvers for the lasso use coordinate descent.
- Simple optimization method: at each iteration, update a single coordinate (coefficient).

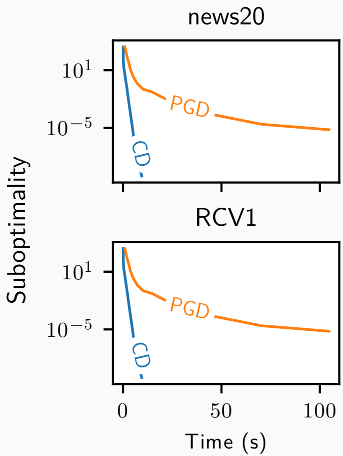

Figure 2: Coordinate descent versus proximal gradient descent for the lasso.

Coordinate Descent and Inseparability

- Unfortunately, we cannot use basic coordinate descent for SLOPE since the sorted ℓ_1 norm is inseparable:

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

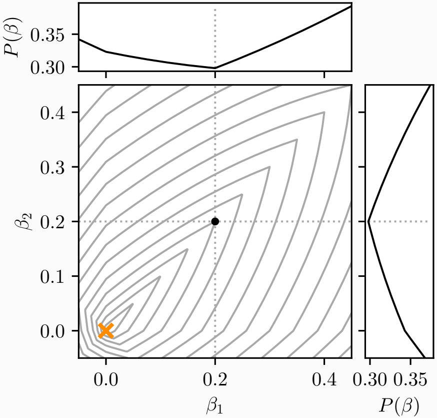

Figure 3: A *naive* coordinate descent algorithm cannot advance from the current iterate (●) to reach the optimum (✖).

Coordinate Descent and Inseparability

- Unfortunately, we cannot use basic coordinate descent for SLOPE since the sorted ℓ_1 norm is inseparable:

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

- But if we fix the clusters, we have separability and can solve SLOPE using coordinate descent.

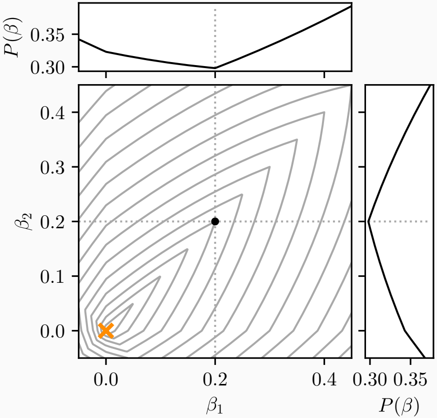

Figure 3: A *naive* coordinate descent algorithm cannot advance from the current iterate (●) to reach the optimum (✖).

Coordinate Descent and Inseparability

- Unfortunately, we cannot use basic coordinate descent for SLOPE since the sorted ℓ_1 norm is inseparable:

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

- But if we fix the clusters, we have separability and can solve SLOPE using coordinate descent.
- **Idea:** Alternate between gradient descent steps (identify the clusters) and coordinate descent steps on the clusters (converge quickly).

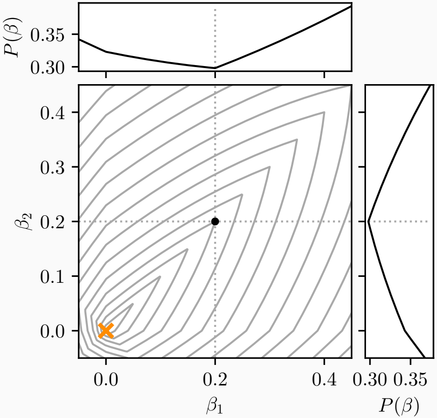

Figure 3: A *naive* coordinate descent algorithm cannot advance from the current iterate (●) to reach the optimum (✖).

Hybrid Algorithm

- Every vth 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 vth 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.

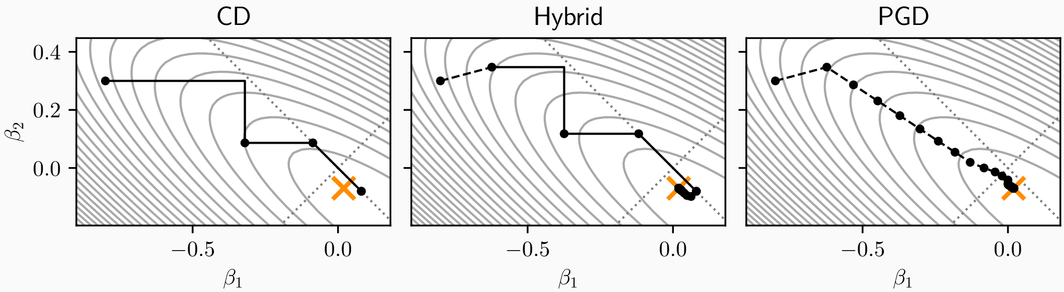

Figure 4: Our algorithm (hybrid) is a combination of CD and PGD.

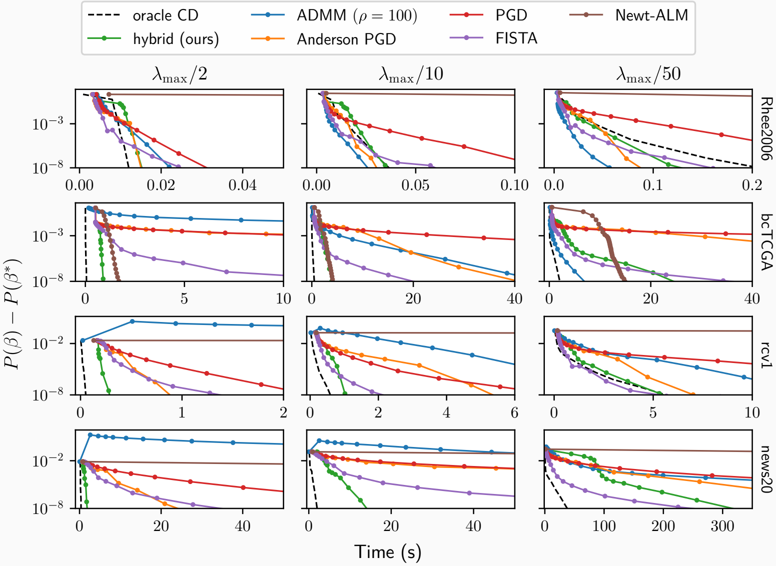

Figure 5: Benchmarks on real data

Previous Work

The Strong Screening Rule for SLOPE

Basic idea:

- When $p \gg n$, SLOPE and lasso solutions have small support.
- If we can estimate the support (before fitting the model), we save a lot of time.
- If the screening method is cheap, we have a net gain.

The Strong Screening Rule for SLOPE

Basic idea:

- When $p \gg n$, SLOPE and lasso solutions have small support.
- If we can estimate the support (before fitting the model), we save a lot of time.
- If the screening method is cheap, we have a net gain.

A game-changer for the lasso. But for SLOPE, there were no screening rules before our work (Larsson, Bogdan, and Wallin 2020).

The Strong Screening Rule for SLOPE

Basic idea:

- When $p \gg n$, SLOPE and lasso solutions have small support.
- If we can estimate the support (before fitting the model), we save a lot of time.
- If the screening method is cheap, we have a net gain.

A game-changer for the lasso. But for SLOPE, there were no screening rules before our work (Larsson, Bogdan, and Wallin 2020).

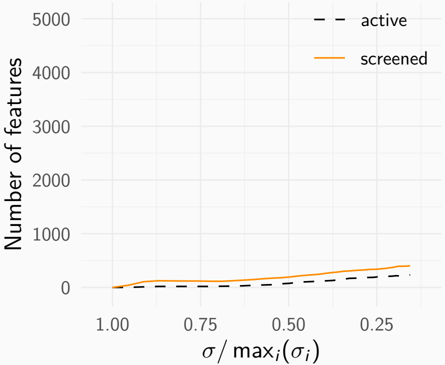

Figure 6: Number of features screened along the SLOPE path for for a data set with 200 observations and 5000 features.

The Hessian Screening Rule

In this paper (Larsson and Wallin 2022) we continued our work on screening rules, but for the lasso instead.

Our contribution: a new rule that uses second-order information to better predict the support along the regularization path.

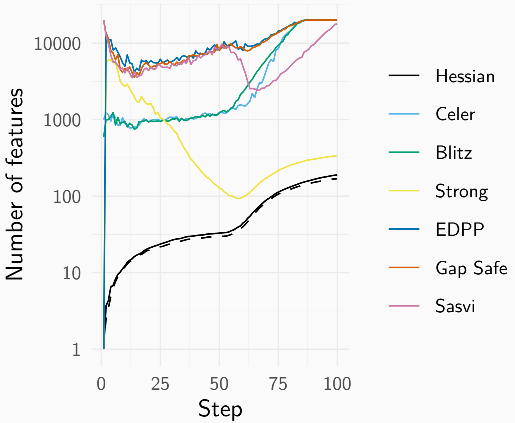

Figure 7: Number of features (predictors) screened along the SLOPE path for designs with varying correlation (ρ).

Benchopt

Benchopt (Moreau et al. 2022) strives to make benchmarking easy, transparent, and reproducible.

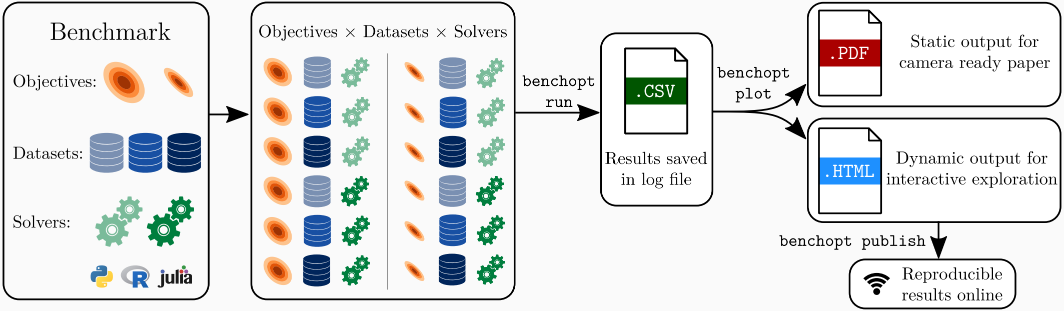

Figure 8: How Benchopt works.

Ongoing Work

Regularization And Normalization

- Normalization is essential for regularized methods, but there is almost no work on the topic.
- What effects do different types of normalization have on the solutions of regularized methods?

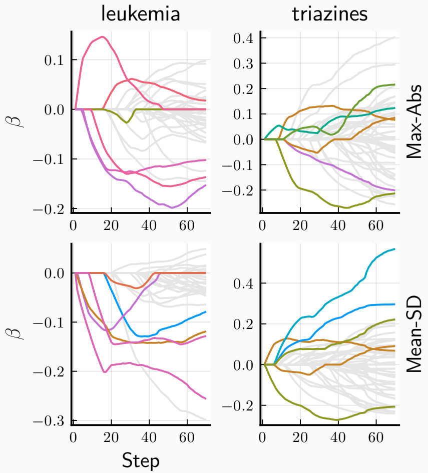

Figure 9: Lasso paths for two types of normalization.

Thank you!

[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).

[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] Johan Larsson, Małgorzata Bogdan, and Jonas Wallin. **"The Strong Screening Rule for SLOPE".** In: *Advances in Neural Information Processing Systems 33*. 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Ed. by Hugo Larochelle et al. Vol. 33. Virtual: Curran Associates, Inc., Dec. 6–12, 2020, pp. 14592–14603. ISBN: 978-1-71382-954-6. URL: https://papers.nips.cc/paper%5C%5Ffiles/paper/2020/hash/a7d8ae4569120b5bec12e7b6e9648b86-Abstract.html.

[6] Johan Larsson and Jonas Wallin. **"The Hessian Screening Rule".** In: *Advances in Neural Information Processing Systems 35*. 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Ed. by S. Koyejo et al. Vol. 35. New Orleans, USA: Curran Associates, Inc., Nov. 28–Dec. 9, 2022, pp. 15823–15835. ISBN: 978-1-71387-108-8. URL: https://papers.nips.cc/paper%5C%5Ffiles/paper/2022/hash/65a925049647eab0aa06a9faf1cd470b-Abstract-Conference.html.

[7] Thomas Moreau et al. **"Benchopt: Reproducible, Efficient and Collaborative Optimization Benchmarks".** In: *Advances in Neural Information Processing Systems 35*. 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Ed. by S. Koyejo et al. Vol. 35. New Orleans, USA: Curran Associates, Inc., Nov. 28–Dec. 9, 2022, pp. 25404–25421. ISBN: 978-1-71387-108-8. URL: https://proceedings.neurips.cc/paper%5C%5Ffiles/paper/2022/hash/a30769d9b62c9b94b72e21e0ca73f338-Abstract-Conference.html.

[8] 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).

Coordinate Descent Steps

When updating the kth 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 kth cluster, we let

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

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 ℓ_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 ℓ_1 Norm

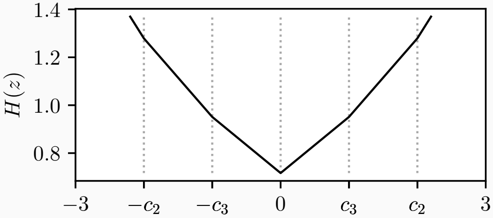

Figure 10: The partial sorted ℓ_1 norm with $\beta = [-3, 1, 3, 2]^T$, $k = 1$, and so $c_1, c_2, c_3 = (3, 2, 1)$.

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

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

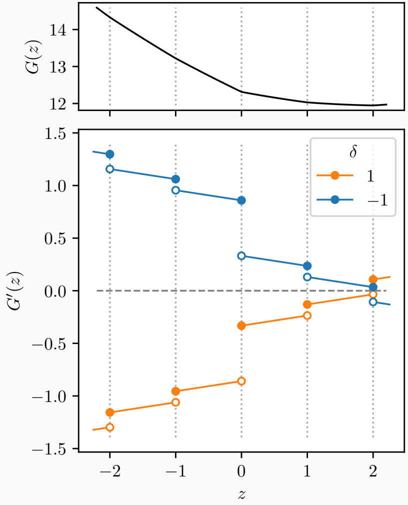

Figure 11: G and its directional derivative $G'(\cdot;\delta)$.

The SLOPE Thresholding Operator

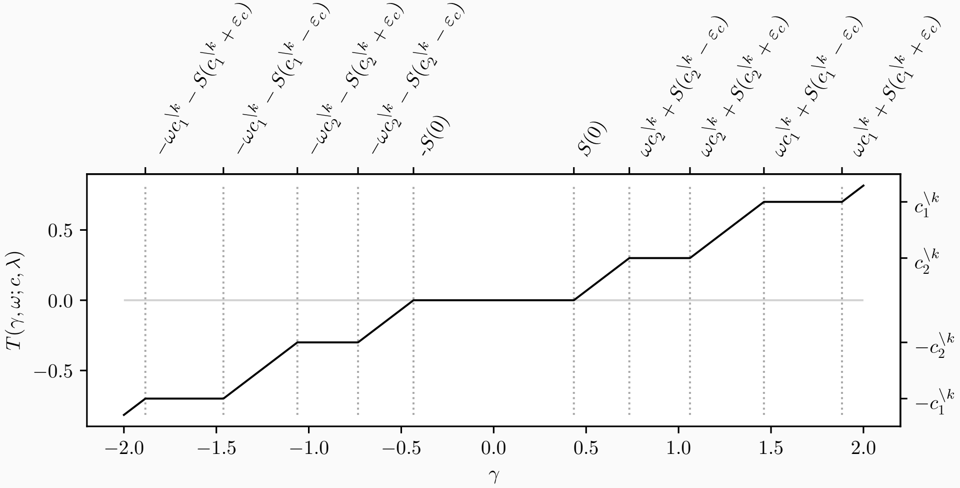

Figure 12: The SLOPE Thresholding Operator