Department of Industrial and Systems Engineering
University of Southern California

Comparing Solution Paths of Sparse Quadratic Minimization with a Stieltjes Matrix

May 7, 2024

Ziyu He, Shaoning Han, Andrés Gómez, Ying Cui, Jong-Shi Pang

Hyperparameter selection for sparse estimation

Basic Sparse Estimation:

$$\underset{\ell \leq x \leq u}{\overset{\substack{\text{Constraints } \ell \in \mathbb{R}^n_-,\, u \in \mathbb{R}^+ \\ \downarrow}}{\textbf{minimize}}} \quad \underbrace{\frac{1}{2}x^\top Q x + q^\top x}_{\substack{Q \in \mathbb{R}^{n \times n}\text{: Stieltjes matrix} \\ \text{Applications e.g., Markov random field}}} \quad + \quad \overset{\substack{\textbf{Hyperparameter } \gamma \geq 0 \\ \downarrow}}{\gamma} \quad \underbrace{\Phi(x)}_{\substack{\text{Sparsity inducing} \\ \text{regularizer}}} \tag{1}$$

Hyperparameter selection for sparse estimation

Basic Sparse Estimation:

$$\underset{\ell\le x\le u}{\overset{\overset{\text{Constraints } \ell \in \mathbb{R}^n_-,\, u\in\mathbb{R}^+}{\downarrow}}{\textbf{minimize}}} \quad \underbrace{\frac{1}{2}x^\top Qx + q^\top x}_{\substack{Q\in\mathbb{R}^{n\times n}\text{: Stieltjes matrix}\\ \text{Applications e.g., Markov random field}}} \quad + \quad \overset{\overset{\textbf{Hyperparameter } \gamma\ge 0}{\downarrow}}{\gamma}\ \underbrace{\Phi(x)}_{\substack{\text{Sparsity inducing}\\ \text{regularizer}}} \tag{1}$$

Hyperparameter Selection:

- Select the best γ by some criteria for (1)'s solution, e.g., test error.

Hyperparameter selection for sparse estimation

Basic Sparse Estimation:

$$\underset{\ell\leq x\leq u}{\overset{\overset{\text{Constraints } \ell\in\mathbb{R}^n_-,\, u\in\mathbb{R}^+}{\downarrow}}{\textbf{minimize}}} \quad \underbrace{\frac{1}{2}x^\top Qx+q^\top x}_{\substack{Q\in\mathbb{R}^{n\times n}\text{: Stieltjes matrix}\\ \text{Applications e.g., Markov random field}}} \quad + \quad \overset{\overset{\text{Hyperparameter } \gamma\geq 0}{\downarrow}}{\gamma}\ \underbrace{\Phi(x)}_{\substack{\text{Sparsity inducing}\\ \text{regularizer}}} \tag{1}$$

Hyperparameter Selection:

- Select the best γ by some criteria for (1)'s solution, e.g., test error.

Parametric Programming:

- A whole path of (1)'s solution as a function of γ.

Choice of Φ: a dilemma

The ideal choice

- Weighted ℓ_0 : $\sum_{i=1}^{n} p_i |x_i|_0$
- Mixed integer hence can be computationally prohibitive.

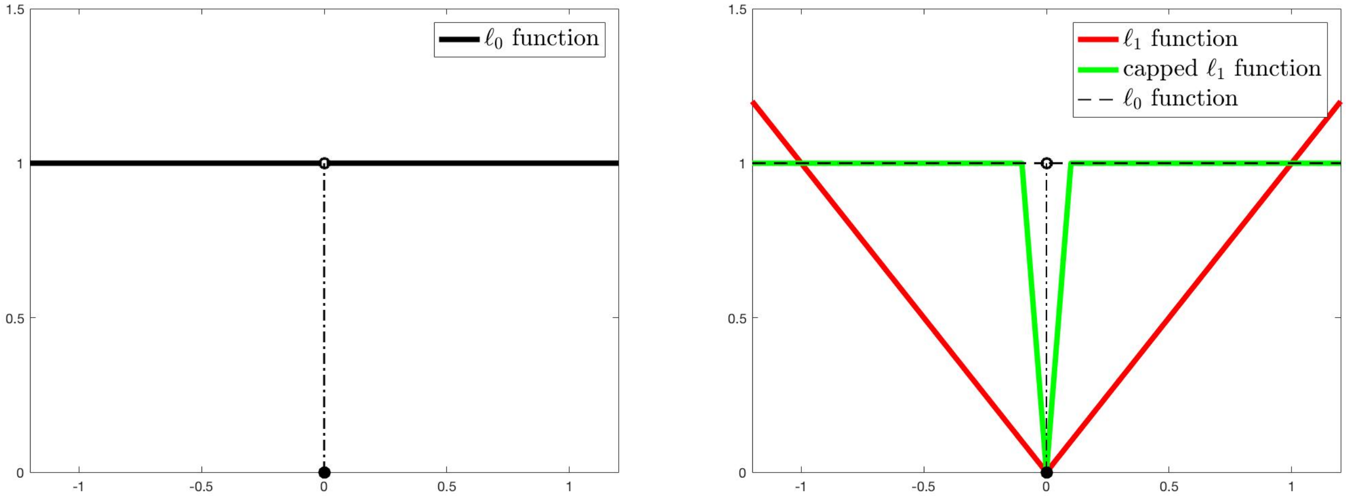
Convex relaxation

- Weighted ℓ_1 : $\sum_{i=1}^{n} p_i |x_i|$
- Easier to compute but can give us undesirable results.

Nonconvex surrogate

- Weighted capped ℓ_1: $\sum_{i=1}^{n} p_i \min(\frac{|x_i|}{\delta}, 1), \delta > 0$

ℓ_0, ℓ_1 and capped ℓ_1



(a) ℓ_0 function

(b) ℓ_0, ℓ_1 and Capped ℓ_1

Figure 1: Capped ℓ_1 is a better approximation to ℓ_0 than ℓ_1

Choice of Φ: a dilemma

The ideal choice

- Weighted ℓ_0 : $\sum_{i=1}^{n} p_i |x_i|_0$
- Mixed integer hence can be computationally prohibitive.

Convex relaxation

- Weighted ℓ_1 : $\sum_{i=1}^{n} p_i |x_i|$
- Easier to compute but can us give undesirable results.

Nonconvex surrogate

- Weighted capped ℓ_1: $\sum_{i=1}^{n} p_i \min(\frac{|x_i|}{\delta}, 1), \delta > 0$
- A better approximation to ℓ_0 but nonconvex and nonsmooth.
- Analytical properties? How to compute?
- D(irectional)-stationary [$\Leftrightarrow$ strongly local] solution path.

The overall goals

- Studying and comparing paths from ℓ_0, ℓ_1 and capped ℓ_1.

- Emphasizing on the d-stationary **(d-stat.)** path of capped ℓ_1
  - Analytical **(e.g., number of pieces)**
  - Computational **(the first rigorous computational study for nonconvex paths).**

- Highlighting the benefit of nonconvex approaches in balancing
  - Computational effort
  - Statistical and optimization performances

Previous studies

ℓ_1-path (Q is PD)

- Continuous piecewise affine [Efron et al., 2004].
- In worst case exponentially many pieces [Mairal and Yu, 2012].
- **Cases guaranteed to have polynomialy many pieces?**

Previous studies

ℓ_1-path (Q is PD)

- Continuous piecewise affine [Efron et al., 2004].
- In worst case exponentially many pieces [Mairal and Yu, 2012].
- **Cases guaranteed to have polynomial pieces?**

ℓ_0-path (equal weights $p_i \equiv 1$)

- Discontinuous piecewise affine ($n + 1$ pieces) [Soussen et al., 2015].
- **Unequal weights?**

Previous studies

ℓ_1-path (Q is PD)

- Continuous piecewise affine [Efron et al., 2004].
- In worst case exponentially many pieces [Mairal and Yu, 2012].
- **Cases guaranteed to have polynomial pieces?**

ℓ_0-path (equal weights $p_i \equiv 1$)

- Discontinuous piecewise affine ($n+1$ pieces) [Soussen et al., 2015].
- **Unequal weights?**

Other nonconvex surrogates (e.g., ℓ_p, MCP)

- Parametric nonlinear systems, e.g., γ in quadratic terms.
- Piecewise smooth paths, cannot be exactly traced in finite time.
- Either heuristic [Yukawa and Amari, 2015].
- Or fails to approximate the exact ℓ_0-path [Zhang, 2010].
- **Capped ℓ_1 doesn't have these issues, but how to compute?**

Our contributions: analytical

Deriving special classes:

- Guaranteed to have polynomial many pieces.
- Worst case exponential complexity.

Our contributions: analytical

Deriving special classes:

- Guaranteed to have polynomial many pieces.
- Worst case exponential complexity.

Regularizer	**Pieces**	**Optimality**	**Class**
ℓ_1	$2n+1$	global	$\ell=0$, Stieltjes Q
ℓ_0	exponential	global	Non-Stieltjes Q
(unequal weights)	$n+1$	global	$\ell=0$, Stieltjes Q
Capped ℓ_1	$2n^2+3n+1$	global	$\ell=0$, Stieltjes Q
	$n+1$	d-stat.	$\ell=0$, Stieltjes Q

Table 1: Summary of some analytical results

Our contributions: analytical

Deriving special classes:

- Guaranteed to have polynomial many pieces.
- Worst case exponential complexity.

Regularizer	**Pieces**	**Optimality**	**Class**
ℓ_1	$2n+1$	global	$\ell = 0$, Stieltjes Q
ℓ_0	exponential	global	Non-Stieltjes Q
(unequal weights)	$n+1$	global	$\ell = 0$, Stieltjes Q
Capped ℓ_1	$2n^2+3n+1$	global	$\ell = 0$, Stieltjes Q
	$n+1$	d-stat.	$\ell = 0$, Stieltjes Q

Table 1: Summary of some analytical results

- **They are all piecewise affine.**
- **Stieltjes structure is the key for polynomial complexity.**

Our contributions: computational

- A rigorous method to compute d-stat. paths for capped ℓ_1.
  - **Can be discontinuous!**

Our contributions: computational

- A rigorous method to compute d-stat. paths for capped ℓ_1.
  - **Can be discontinuous!**
- Efficient algorithm (GHP) to restore discontinuity.
  - **Complexity is strongly polynomial (Stieltjes Q).**

Our contributions: computational

- A rigorous method to compute d-stat. paths for capped ℓ_1.
  - **Can be discontinuous!**
- Efficient algorithm (GHP) to restore discontinuity.
  - **Complexity is strongly polynomial (Stieltjes Q).**
- Benefits of capped ℓ_1 d-stat. path supported by numerical results:
  - **Way faster than computing ℓ_0.**
  - **Superior optimization and statistical performance than ℓ_1.**

Pivoting method: high level ideas

$$\underset{\ell \leq x \leq u}{\textbf{minimize}} \quad \frac{1}{2} x^\top Q x + q^\top x + \gamma \sum_{i=1}^{n} p_i \min\left\{ \frac{|x_i|}{\delta}, 1 \right\} \tag{2}$$

- D-stat. path of (2) is piecewise affine in γ.

Pivoting method: high level ideas

$$\operatorname*{\mathbf{minimize}}_{\ell \le x \le u} \quad \frac{1}{2} x^\top Q x + q^\top x + \gamma \sum_{i=1}^{n} p_i \min\left\{ \frac{|x_i|}{\delta}, 1 \right\} \tag{2}$$

- D-stat. path of (2) is piecewise affine in γ.
- To trace the path is to compute these pieces one by one.
  - **In the direction of $\gamma \downarrow 0$**

Pivoting method: high level ideas

$$\underset{\ell \leq x \leq u}{\textbf{minimize}} \quad \frac{1}{2} x^\top Q x + q^\top x + \gamma \sum_{i=1}^{n} p_i \min\left\{ \frac{|x_i|}{\delta}, 1 \right\} \tag{2}$$

- D-stat. path of (2) is piecewise affine in γ.
- To trace the path is to compute these pieces one by one.
- Each piece is associated with a "basis".
  - **Tuple of index sets to restrict the values of solution for (2).**

Pivoting method: high level ideas

$$\underset{\ell \le x \le u}{\textbf{minimize}} \quad \frac{1}{2} x^\top Q x + q^\top x + \gamma \sum_{i=1}^{n} p_i \min \left\{ \frac{|x_i|}{\delta}, 1 \right\} \tag{2}$$

- D-stat. path of (2) is piecewise affine in γ.
- To trace the path is to compute these pieces one by one.
- Each piece is associated with a “basis”.
- Fixed basis ⇒ reduced problem ⇒ reduced solution.
  - **Not necessarily d-stat. for** (2).

Pivoting method: high level ideas

$$\underset{\ell \leq x \leq u}{\textbf{minimize}} \quad \frac{1}{2} x^\top Q x + q^\top x + \gamma \sum_{i=1}^{n} p_i \min\left\{ \frac{|x_i|}{\delta}, 1 \right\} \tag{2}$$

- D-stat. path of (2) is piecewise affine in γ.
- To trace the path is to compute these pieces one by one.
- Each piece is associated with a “basis”.
- Fixed basis ⇒ reduced problem ⇒ reduced solution.
- Conditions for the reduced solution to be d-stat. of (2).
    - **Linear inequalities in γ.**
    - **Ratio test: the smallest γ for the current basis to be d-stat.**

Pivoting method: high level ideas

$$\underset{\ell \leq x \leq u}{\textbf{minimize}} \quad \frac{1}{2}x^\top Qx + q^\top x + \gamma \sum_{i=1}^{n} p_i \min\left\{\frac{|x_i|}{\delta}, 1\right\} \tag{2}$$

Pivoting method (informal)

- Compute the pieces/bases from right to the left.
- At the current piece/basis, do ratio test to get γ^*.
- When we go beyond γ^*, change the basis accordingly.

Pivoting method: high level ideas

$$\underset{\ell \le x \le u}{\textbf{minimize}} \quad \frac{1}{2}x^\top Q x + q^\top x + \gamma \sum_{i=1}^{n} p_i \min\left\{\frac{|x_i|}{\delta}, 1\right\} \tag{2}$$

Pivoting method (informal)

- Compute the pieces/bases from right to the left.
- At the current piece/basis, do ratio test to get γ^*.
- When we go beyond γ^*, change the basis accordingly.
- **Discontinuous at γ^* if taking** $\pm\delta$ (not allowed for (2)'s d-stat.).

Pivoting method: high level ideas

$$\underset{\ell \leq x \leq u}{\textbf{minimize}} \quad \frac{1}{2} x^\top Q x + q^\top x + \gamma \sum_{i=1}^{n} p_i \min \left\{ \frac{|x_i|}{\delta}, 1 \right\} \tag{2}$$

Pivoting method (informal)

- Compute the pieces/bases from right to the left.
- At the current piece/basis, do ratio test to get γ^*.
- When we go beyond γ^*, change the basis accordingly.
- Discontinuous at γ^* if taking $\pm\delta$ (not allowed for (2)'s d-stat.).
- We need an algorithm to restore a d-stat. solution at γ^* that:
    - **Doesn't need to compute from scratch.**
    - **Leverages the Stieltjes Q for faster computation.**

The GHP method (informal)

$$\underset{\ell \leq x \leq u}{\textbf{minimize}} \quad \frac{1}{2}x^\top Qx + q^\top x + \gamma \sum_{i=1}^{n} p_i \min\left\{\frac{|x_i|}{\delta}, 1\right\} \tag{2}$$

High level ideas

- Basis enumeration method, just for an alternative basis.
- Fixed basis ⇒ reduced problem ⇒ reduced solution.

The GHP method (informal)

$$\underset{\ell \leq x \leq u}{\textbf{minimize}} \quad \frac{1}{2}x^\top Q x + q^\top x + \gamma \sum_{i=1}^{n} p_i \min\left\{\frac{|x_i|}{\delta}, 1\right\} \tag{2}$$

High level ideas

- Basis enumeration method, just for an alternative basis.
- Fixed basis ⇒ reduced problem ⇒ reduced solution.
- Conditions for the reduced solution to be d-stat. for (2).
  - **Certain index sets should be empty.**
  - **If not, form a new basis by moving them accordingly.**
  - **Proceed to the next steps until such index sets are empty.**

The GHP method

Back to the pivoting method

- Suppose we are at a discontinuous γ^* where we need to restore.
- Initialize the GHP method with the current "non-d-stationary" solution, we can prove the following theorem.

The GHP method

Back to the pivoting method

- Suppose we are at a discontinuous γ^* which we need to restore.
- Initialize the GHP method with the current “non-d-stationary” solution, we can prove the following theorem.

Main theorem of GHP (informal)

GHP with this specialized initialization will terminate in $3n$ steps with a d-stationary solution of (2) at γ^*.

The GHP method

Back to the pivoting method

- Suppose we are at a discontinuous γ^* which we need to restore.
- Initialize the GHP method with the current "non-d-stationary" solution, we can prove the following theorem.

Main theorem of GHP (informal)

GHP with this specialized initialization will terminate in $3n$ steps with a d-stationary solution of (2) at γ^*.

Remark

- [At most $3n$ GHP subproblems] + [each $\mathcal{O}(n^3)$ by Q Stieltjes] $\Longrightarrow$ [total complexity $\mathcal{O}(n^4)$].
- It is inductively proved by leveraging a key property of Stieltjes matrices named the "least element property". [Pang, 1979].

Numerical experiments: GMRF

Gaussian Markov Random Field

The maximum a posteriori estimation of Gaussian Markov random field (GMRF) naturally gives rise to the Stieltjes structure.

Given graph (V, E):

$$\operatorname*{\mathbf{minimize}}_{x} \quad \sum_{i \in V} \frac{1}{\sigma_i^2}(y_i - x_i)^2 + \sum_{(i,j) \in E} \frac{1}{d_{ij}}(x_i - x_j)^2$$

Numerical experiments: GMRF

Summary: capped ℓ_1 vs. ℓ_1 vs. ℓ_0

- **Settings:**
    - $n \in \{100, 10000\}$, noise level $\in (0, 1]$
    - ℓ_0 (only for $n = 100$), ℓ_1, capped ℓ_1
    - $\delta \in \{10, 1, 10^{-1}, 10^{-4}\}$ for capped ℓ_1 **(controls its approx. to ℓ_0)**

Numerical experiments: GMRF

Summary: capped ℓ_1 vs. ℓ_1 vs. ℓ_0

- **Settings:**
    - $n \in \{100, 10000\}$, noise level $\in (0, 1]$
    - ℓ_0 (only for $n = 100$), ℓ_1, capped ℓ_1
    - $\delta \in \{10, 1, 10^{-1}, 10^{-4}\}$ for capped ℓ_1
- **Computation time:**
    - Capped ℓ_1 d-stat. path can be **60 - 3,000 times faster** than ℓ_0.
    - When δ is large, capped ℓ_1 is basically the same as ℓ_1.
    - When δ is small, capped ℓ_1 needs more effort **(can be 10 times slower, no free lunch).**

Theorem (no free lunch)

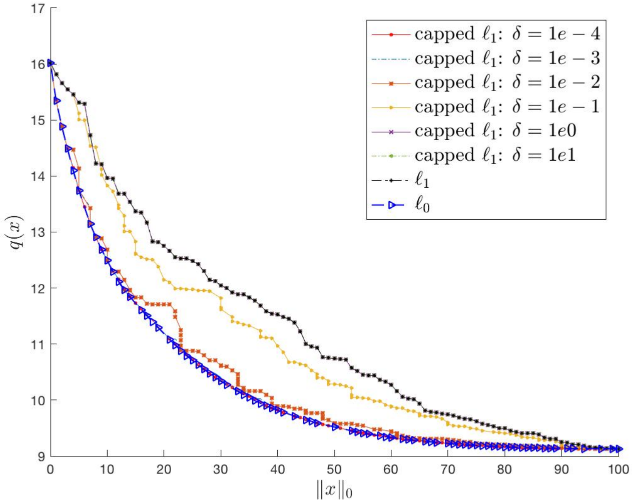
When $\delta < |\bar{x}_i^0|, \forall i \in [n]$ where $\bar{x}^0$ is the unique solution at $\gamma = 0$, then the unique continuous d-stat. path is $\bar{x}(\gamma) = \bar{x}^0, \forall \gamma \geq 0$.

Numerical experiments: GMRF

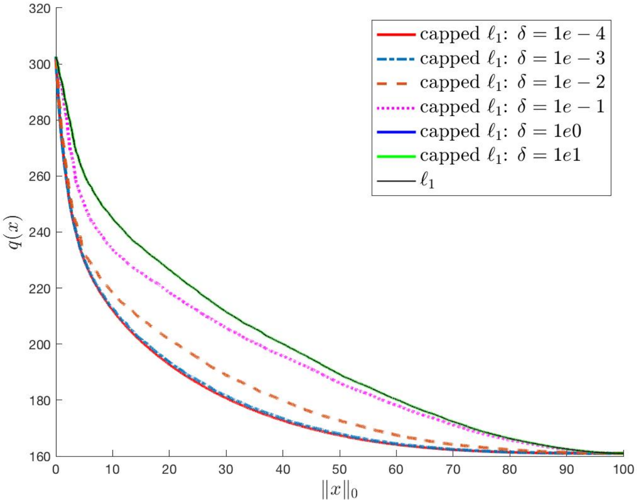
Summary: capped ℓ_1 vs. ℓ_1 vs. ℓ_0

- **Settings:**
  - $n \in \{100, 10000\}$, noise level $\in (0, 1]$
  - ℓ_0 (only for $n = 100$), ℓ_1, capped ℓ_1
  - $\delta \in \{10, 1, 10^{-1}, 10^{-4}\}$ for capped ℓ_1
- **Computation time:**
  - Capped ℓ_1 d-stat. path can be 60 - 3,000 times faster than ℓ_0.
  - When δ is large, capped ℓ_1 is basically the same as ℓ_1.
  - When δ is small, capped ℓ_1 needs more effort (can be 10 times slower, no free lunch).
- **Loss value:**
  - Capped ℓ_1 achieves better loss value ($\frac{1}{2}x^\top Qx + q^\top x$) than ℓ_1 when the solution sparsity is the same!

GMRF: loss vs. sparsity



(a) $n = 100$



(b) $n = 10000$

- When δ is small, capped ℓ_1 behaves like ℓ_0 (blue curve in the bottom).
- When δ is large, capped ℓ_1 behaves like ℓ_1 (black curve on the top).
- Capped ℓ_1 trade-off (acceptable) computation time to gain superior optimization and statistical (to be shown) performance.

Hyperparameter selection

- We care about the following quantities (X is some ground truth)

Signal recovery: $\sum_{i=1}^{p}\sum_{j=1}^{p} (x^*_{ij}(\gamma) - X_{ij})^2$

Support recovery: $\sum_{i=1}^{p}\sum_{j=1}^{p} \Big| \, | x^*_{ij}(\gamma) |_0 - | X_{ij} |_0 \Big|$

Hyperparameter selection

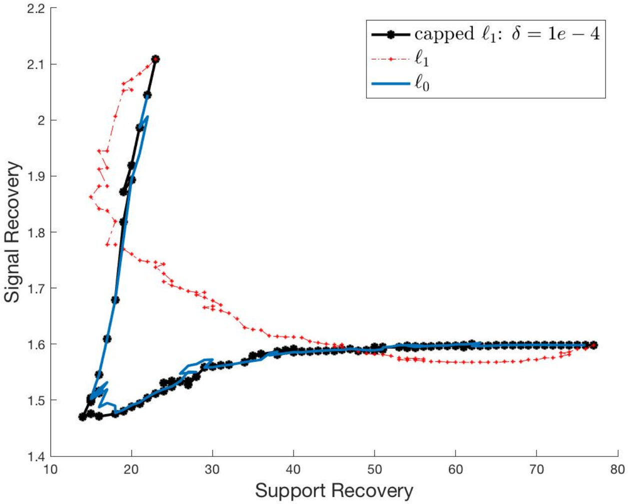
- We care about the following quantities (X is some ground truth)

$$\text{Signal recovery: } \sum_{i=1}^{p}\sum_{j=1}^{p}\left(x_{ij}^*(\gamma) - X_{ij} \right)^2$$

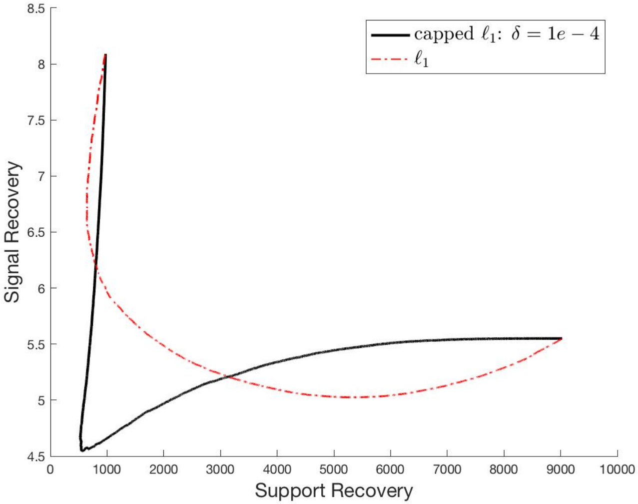
$$\text{Support recovery: } \sum_{i=1}^{p}\sum_{j=1}^{p}\left| \, | x_{ij}^*(\gamma) |_0 - | X_{ij} |_0 \right|$$

- Capped ℓ_1 d-stat. path is superior in hyperparameter selection
  - The best capped ℓ_1 solution dominates ℓ_1's in both quantities.
  - Capped ℓ_1 achieves the minimum of both quantities at the same γ.
  - ℓ_1 cannot achieve this: we always have to sacrifice one of them.

Signal vs. support recovery



(a) $n = 100$

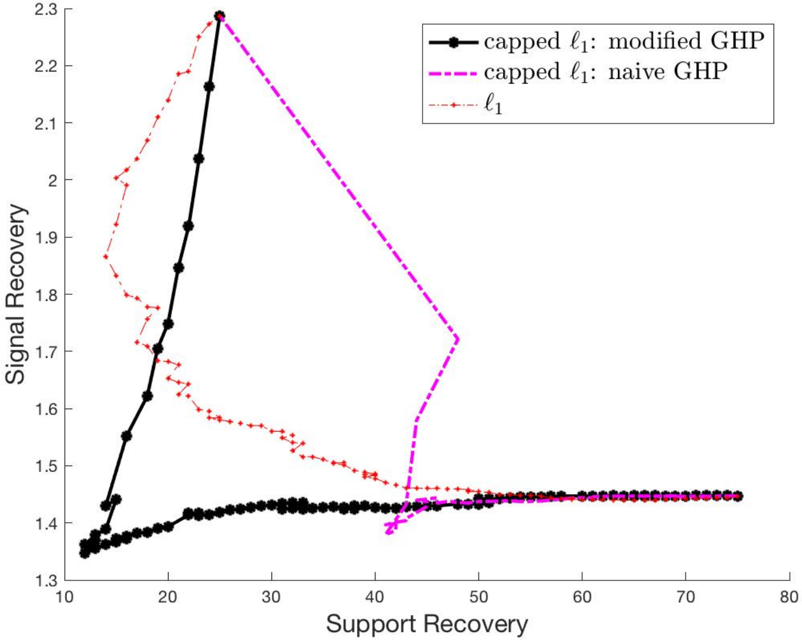


(b) $n = 10000$

- For ℓ_0, capped ℓ_1, good signal and support recovery are highly correlated.
- But for ℓ_1, we always have to sacrifice one quantity.

GHP restoration

Our specialized initialization for GHP restoration leverages the most recent basis, which turns out to be the key for all the nice properties.



Hyperparameter selection from the naïve initialization is worse than ℓ_1.

Numerical experiments: summary

- **Computation time:**
  - Capped ℓ_1 path is much more scalable than ℓ_0 path.
  - Capped ℓ_1 path (small δ) needs more (acceptable) effort than ℓ_1.

Numerical experiments: summary

- **Computation time:**
  - Capped ℓ_1 path is much more scalable than ℓ_0 path.
  - Capped ℓ_1 path (small δ) needs more (acceptable) effort than ℓ_1.
- **Optimization performance:**
  - Capped ℓ_1 path (small δ) is a better approximation to the ℓ_0 path.
  - For the same sparsity, capped ℓ_1 achieves better loss than ℓ_1.

Numerical experiments: summary

- **Computation time:**
  - Capped ℓ_1 path is much more scalable than ℓ_0 path.
  - Capped ℓ_1 path (small δ) needs more (acceptable) effort than ℓ_1.
- **Optimization performance:**
  - Capped ℓ_1 path (small δ) is a better approximation to the ℓ_0 path.
  - For the same sparsity, capped ℓ_1 achieves better loss than ℓ_1.
- **Hyperparameter selection:**
  - Capped ℓ_1: minimal signal and support recovery at the same γ.
  - ℓ_1 path does not have such γ.

Numerical experiments: summary

- **Computation time:**
  - Capped ℓ_1 path is much more scalable than ℓ_0 path.
  - Capped ℓ_1 path (small δ) needs more (acceptable) effort than ℓ_1.
- **Optimization performance:**
  - Capped ℓ_1 path (small δ) is a better approximation to the ℓ_0 path.
  - For the same sparsity, capped ℓ_1 always achieves better loss than ℓ_1.
- **Hyperparameter selection:**
  - Capped ℓ_1: minimal signal and support recovery at the same γ.
  - ℓ_1 path does not have such γ.
- The **GHP** restoration is critical for all the nice practical properties.

Reference

- Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *The Annals of statistics, 32*(2), 407-499.
- Mairal, J., & Yu, B. (2012). Complexity analysis of the lasso regularization path. *arXiv preprint arXiv:1205.0079*.
- Soussen, C., Idier, J., Duan, J., & Brie, D. (2015). Homotopy Based Algorithms for ℓ_0-Regularized Least-Squares. *IEEE Transactions on Signal Processing, 63*(13), 3301-3316.
- Yukawa, M., & Amari, S. I. (2015). ℓ_p-Regularized Least Squares $(0 < p < 1)$ and Critical Path. *IEEE Transactions on Information Theory, 62*(1), 488-502.
- Zhang, C. H. (2010). Nearly Unbiased Variable Selection Under Minimax Concave Penalty. *The Annals of statistics, 38*(2), 894-942.
- Pang, J. S. (1979). On a class of least-element complementarity problems. *Mathematical Programming, 16*(1), 111-126.
- Pang, J. S., & Chandrasekaran, R. (1985). Linear complementarity problems solvable by a polynomially bounded pivoting algorithm. In *Mathematical Programming Essays in Honor of George B. Dantzig Part II* (pp. 13-27). Springer, Berlin, Heidelberg.