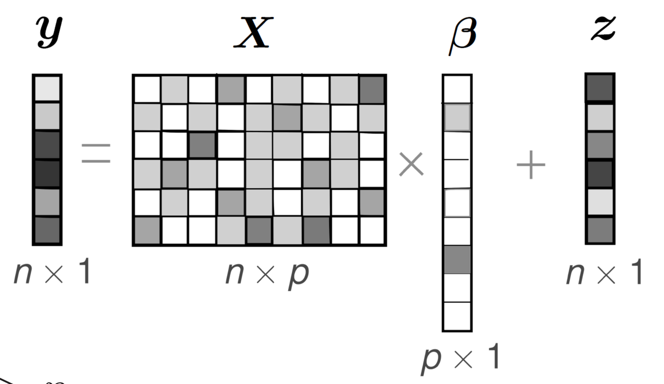
SLOPE MEETS AMP: DOES SLOPE OUTPERFORM LASSO?

Zhiqi Bu*, Cynthia Rush§, Jason M. Klusowski† and Weijie Su*

*University of Pennsylvania, §Columbia University, †Rutgers University

"Algorithmic Analysis and Statistical Estimation of SLOPE via Approximate Message Passing" (arxiv.org/abs/1907.07502)

SPARSE HIGH-DIM LINEAR REGRESSION



- Often $p > n$
- $\beta_j \neq 0$ means the jth variable is relevant
- Most of entries of β are zeros

EXISTING PROCEDURE: LASSO

In 1996, Tibshirani proposed LASSO problem:

$$\min_{\boldsymbol{b}} \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b}\|_2^2 + \underbrace{\lambda\|\boldsymbol{b}\|_1}_{\ell_1 \text{ norm}}$$

$$= \min_{\boldsymbol{b}} \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b}\|_2^2 + \lambda|b_1| + \cdots + \lambda|b_p|$$

- Useful in identifying which $\beta_j \neq 0$ (support recovery, feature selection, etc.)
- LASSO selects at most n variates before it saturates

CALIBRATION

For every iteration t, assign $\boldsymbol{\theta}_t = \boldsymbol{\alpha}\tau_t$, where $\boldsymbol{\alpha}$ is a vector in the same direction as $\boldsymbol{\lambda}$.

Theorem 1 *Under conditions on Λ, the state evolution recursion with calibration defined above, has a **unique fixed point** to which the convergence monotonic in t, for **any** initial condition.*

Calibration between $\boldsymbol{\lambda}$ and α:

$$\boldsymbol{\lambda} = \alpha\tau_*\left(1 - \lim_p \frac{1}{\delta p}\mathbb{E}\|\operatorname{prox}_{J_{\alpha\tau_*}}(\mathbf{B}+\tau_*\boldsymbol{Z})\|_0^*\right).$$

CHALLENGES

Framework: $\hat{\beta} \approx \beta^t \approx \operatorname{prox}_{J_{\alpha\tau_*}}(\beta + \tau_* Z)$

Although AMP for LASSO is well-studied [Bayati-Montanari '15], applying AMP to SLOPE is challenging because the proximal operator of sorted ℓ_1 norm is **non-separable**.

[Berthier-Montanari-Nguyen '17] showed for general non-separable functions

$$\beta^t \approx \operatorname{prox}(\beta + \tau Z)$$

The main challenge is to show AMP iterate

$$\hat{\beta} \approx \beta^t$$

SLOPE PROBLEM AND PROPERTIES

Bogdan et al. (2015) proprosed SLOPE problem as recovering:

$$\widehat{\boldsymbol{\beta}} := \operatorname*{argmin}_{\boldsymbol{b}} \tfrac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b}\|_2^2 + J_{\boldsymbol{\lambda}}(\boldsymbol{b})$$

where

$$J_{\boldsymbol{\lambda}}(\boldsymbol{b}) = \underbrace{\lambda_1|b|_{(1)} + \cdots + \lambda_p|b|_{(p)}}_{\text{sorted } \ell_1 \text{ norm}}$$

having $\lambda_1 \geq \ldots \geq \lambda_p \geq 0$, and $|b|_{(1)} \geq \ldots \geq |b|_{(p)}$ are the order statistics.

Clearly, when $\lambda_1 = \cdots = \lambda_p$, SLOPE reduces to LASSO.

- **Estimation:** SLOPE achieves minimax estimation properties under certain random designs without requiring knowledge of the sparsity degree of β [Su-Candès '16; Bellec-Lecué-Tsybakov '18]
- **Testing:** SLOPE controls the false discovery rate in the case of independent predictors [Bogdan-Berg-Sabatti-Su-Candès '15]
- **Optimization:** Since sorted ℓ_1 norm is a norm, cost remains convex and it can be efficiently solved by using standard methods like proximal gradient descent (see R package "SLOPE") [Bogdan-Berg-Sabatti-Su-Candès '15]

MAIN RESULTS OF SLOPE AMP

Theorem 2 *Under some assumptions,*

$$\lim_{p\to\infty} \frac{1}{p}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^t\|^2 = c_t, \quad \textit{where} \quad \lim_{t\to\infty} c_t = 0.$$

Theorem 3 *Under some assumptions, for any uniformly pseudo-Lipschitz sequence of functions ψ_p and for $\mathbf{Z} \sim \mathcal{N}(0, \mathbb{I}_p)$,*

$$\lim_p \psi_p(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = \lim_t \lim_p \mathbb{E}_{\mathbf{Z}}[\psi_p(\operatorname{prox}_{J_{\alpha\tau_t}}(\boldsymbol{\beta} + \tau_t\mathbf{Z}), \boldsymbol{\beta})].$$

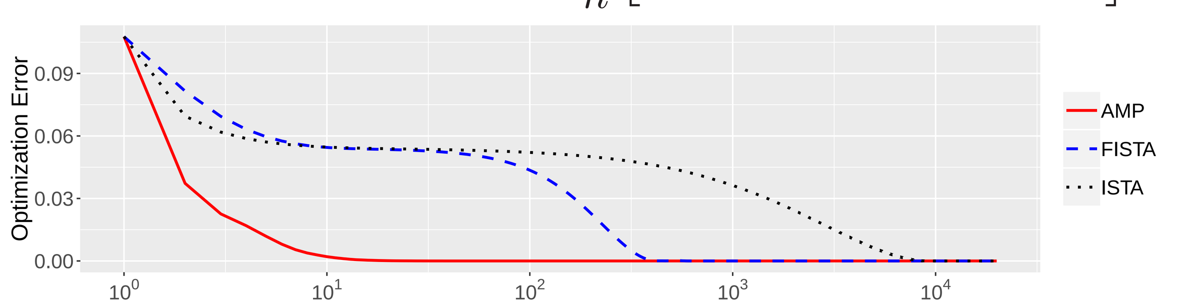
Corollary 3.1 *Under some assumptions,*

$$\lim_p \frac{1}{p}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2 = \delta(\tau_*^2 - \sigma_z^2).$$

COMPUTING SLOPE SOLUTION

Denote $\operatorname{prox}(\boldsymbol{y}; \boldsymbol{\lambda}) := \operatorname{argmin}_{\boldsymbol{b}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{b}\|_2^2 + J_{\boldsymbol{\lambda}}(\boldsymbol{b})$.
We may solve the SLOPE problem by

- **Subgradient method:**
$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - s_t \cdot g^{(t)}$$
where $g^{(t)}$ is a subgradient of objective function at $\boldsymbol{\beta}^{(t)}$.
- **ISTA:** Iterative Shrinkage Thresholding Algorithm,
$$\boldsymbol{\beta}^{(t+1)} = \operatorname{prox}\left(\boldsymbol{\beta}^{(t)} + s_t X^\top(y - X\boldsymbol{\beta}^{(t)}); \lambda s_t\right)$$
- **FISTA:** Fast ISTA, with $s_{t+1} = (1 + \sqrt{1 + 4s_t^2})/2$
$$\boldsymbol{\beta}^{(t+1)} = \operatorname{prox}\left(M^{(t)} + s_t X^\top(y - XM^{(t)}); \lambda s_t\right)$$
$$M^{(t+1)} = \boldsymbol{\beta}^{(t)} + \frac{s_t - 1}{s_{t+1}} \cdot (\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^{(t-1)})$$
- **AMP:** Approximate Message Passing [Donoho-Maleki-Montanari '10],
$$\boldsymbol{\beta}^{t+1} = \operatorname{prox}(\mathbf{X}^T\mathbf{r}^t + \boldsymbol{\beta}^t; \alpha\tau_t),$$
$$\mathbf{r}^{t+1} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{t+1} + \frac{\mathbf{r}^t}{n}\left[\nabla \operatorname{prox}(\mathbf{X}^\top\mathbf{r}^t + \boldsymbol{\beta}^t; \alpha\tau_t)\right].$$



PROXIMAL OPERATORS

- For any function h, prox_h is defined as
$$\operatorname{prox}_h(\boldsymbol{x}) := \operatorname*{argmin}_{\boldsymbol{b}} \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{b}\|^2 + h(\boldsymbol{b}),$$
and $\nabla \operatorname{prox}_h$ is divergence of the proximal operator
- There exists an algorithm to compute the proximal operator when $h = J_{\boldsymbol{\theta}}$.[Bogdan-Berg-Sabatti-Su-Candès '15]
- For SLOPE, $\nabla \operatorname{prox}_{J_{\boldsymbol{\theta}_t}}(\boldsymbol{b}) = \|\operatorname{prox}_{J_{\boldsymbol{\theta}_t}}(\boldsymbol{b})\|_0^*$, where $\|\boldsymbol{b}\|_0^*$ counts the **unique** non-zero magnitudes in $\boldsymbol{b}$.

STATE EVOLUTION

SLOPE AMP (Approximate Message Passing):

$$\boldsymbol{\beta}^{t+1} = \operatorname{prox}(\mathbf{X}^T\mathbf{r}^t + \boldsymbol{\beta}^t; \boldsymbol{\theta}); \mathbf{r}^{t+1} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{t+1} + \frac{\mathbf{r}^t}{n}\|\boldsymbol{\beta}^{(t+1)}\|_0^*$$

The dynamics of the AMP iterations are tracked by a recursive sequence referred to as the **state evolution**.

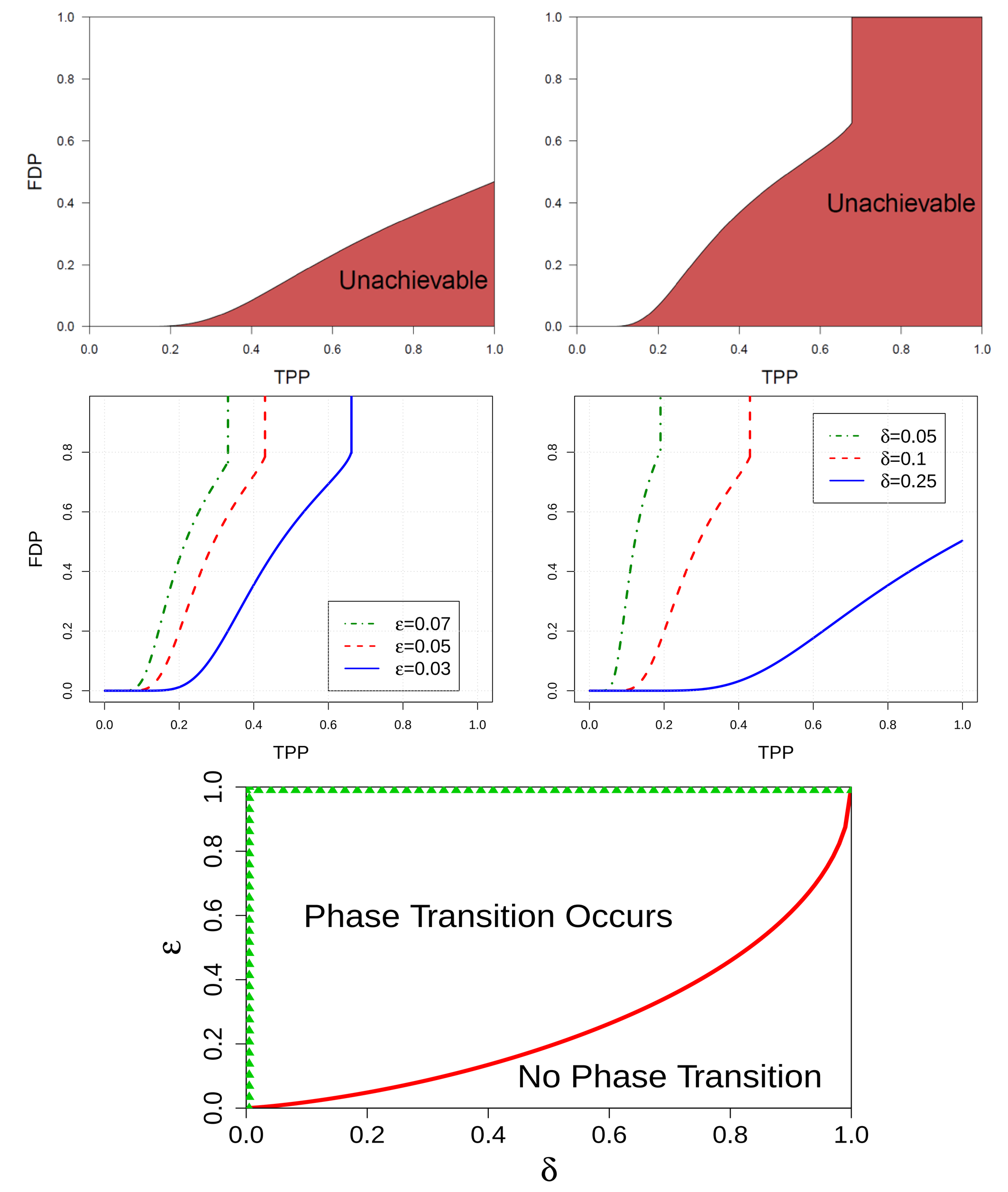
State Evolution ($n/p \to \delta$):

$$\tau^2 = F(\tau^2, \boldsymbol{\alpha}\tau) := \sigma_z^2 + \lim_p \frac{1}{\delta p}\mathbb{E}\|\operatorname{prox}(\mathbf{B} + \tau\mathbf{Z}; \boldsymbol{\alpha}\tau) - \mathbf{B}\|^2.$$

which is solved iteratively via $\tau_{t+1}^2 = F(\tau_t^2, \boldsymbol{\alpha}\tau_t)$.

INFERENCE: TPP,FDP & MSE

For large enough $\epsilon := |\operatorname{supp}(\boldsymbol{\beta})|/p$ or small enough $\delta := n/p$, LASSO suffers from Donoho-Tanner phase transition: TPP is bounded away from 1 [Donoho-Tanner '09; Su-Bogdan-Candès '17 (image source)].



However, SLOPE overcomes the phase transition. Specifically we can charaterize one of the SLOPE path as a Mobius transformation: for TPP$= u$, the minimum FDP is at most $\frac{au+b}{cu+d}$ for some constants a, b, c, d.

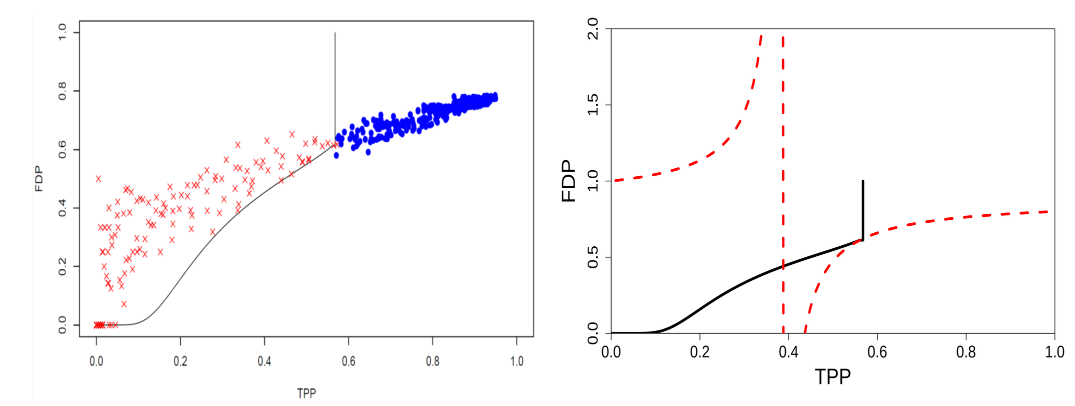


Figure 1: Red dot: LASSO; Blue dot: SLOPE; Black solid line: LASSO trade-off; Red dashed line: SLOPE trade-off.

In addition, fixing the signal prior and under some assumptions, we show that switching from LASSO to SLOPE gives better paths in the sense of achieving smaller FDP, larger TPP and smaller mean squared error at the same time.