$\ell_p$-minimization does not necessarily outperform $\ell_1$-minimization?

Le Zheng

Department of Electrical Engineering, Columbia University

Joint work with Arian Maleki and Xiaodong Wang

Columbia University
Model

\[ y = Ax_o + w \]

- \( x_o \): \( k \)-sparse vector in \( \mathbb{R}^N \)
- \( A \): \( n \times N \) design matrix
- \( y \): measurement vector in \( \mathbb{R}^n \)
- \( w \): measurement noise in \( \mathbb{R}^n \)
$\ell_p$-regularized least squares

Many useful heuristic approaches:

- **LPLS**
  - minimize $\min_x \frac{1}{2} \|y - Ax\|_2^2 + \lambda \|x\|_p^p$ \hspace{1cm} $0 \leq p \leq 1$

- **Facts:**
  - NP-hard except for $p = 1$
  - Its performance is of great interest

Chen, Donoho, Saunders (96), Tibshirani (96), Ge, Jiang, Ye (11)
Global minimum of LPLS for $p < 1$ outperforms LASSO

- Lots of empirical result
- Some theoretical results

Our goal: Evaluating the validity scope of this folklore
Related work

Nonasymptotic analysis:

- R. Gribonval, M. Nielsen: $\ell_p$ is better than $\ell_1$
- Chartrand et al, Gribnoval et al., Saab et al., Foucart et al., Davies et al: Sufficient conditions
- Peng, Yue, and Li: Equivalence of $\ell_0$ and $\ell_p$ (noiseless setting)

Asymptotic analysis:

- Stojnic, Wang et al. : Nice analysis; Sharp only for $\delta \to 1$.
- Rangan et al., Kabashima et al: Replica analysis.
Our analysis framework

Setup:

- \( \delta = \frac{n}{N} \)
- \( x_{o,i} \sim (1 - \epsilon)\delta(x_{o,i}) + \epsilon G(x_{o,i}). \)
  - \( k \approx \epsilon N = \rho n \)
- \( A_{i,j} \sim \mathcal{N}(0, \frac{1}{n}) \)

Asymptotic setting: \( N \to \infty \)
What do we know about LASSO?
Noiseless setting

Fix:
\[ \delta = \frac{n}{N} \quad \text{and} \quad \epsilon = \frac{\|x_o\|_0}{N} \]

Let:
\[ \lambda \to 0 \quad \text{and} \quad N \to \infty \]

Noiseless measurements:
\[ y = Ax_o \]

Question:
For what values of \((\delta, \epsilon)\), \(\ell_1\) recovers \(k\)-sparse solution exactly?

Donoho (05), Donoho-Tanner (08), Donoho, M., Montanari (09), Stojnic (09), Amelunxen et al. (13)
Noisy observations

Noise:

- effect of noise on phase transition curve

Noisy setup:

- $x_0$: $k$-sparse
- $A_{i,j} \sim \mathcal{N}(0, 1/n)$
- $y = Ax_0 + w$
- $w \sim \mathcal{N}(0, \sigma_w^2 I)$
- $\text{MSE} = \lim_{N \to \infty} \frac{\|\hat{x} - x_0\|^2}{N}$ almost surely

LPLS

- minimize \( \min_x \frac{1}{2} \| y - Ax \|_2^2 + \lambda \| x \|_p^p \) \quad 0 \leq p \leq 1

Disclaimer:

- Analysis is based on
  
  - Approximate message passing (Rigorous)
  - Replica (Nonrigorous)
Noiseless setting: global minimum

Fix:
\[ \delta = \frac{n}{N} \text{ and } \epsilon = \frac{\|x_0\|_0}{N} \]

Phase transition of LPLS:
- \( \epsilon = \delta \)

Main features:
- Much higher than LASSO
- Same for every \( 0 \leq p < 1 \)
- Same for every \( G \)

Zheng, Maleki, Wang (15)
How about noisy setting?

LPLS: \[ \hat{x}_p(\lambda) = \arg\min_x \frac{1}{2} \| y - Ax \|_2^2 + \lambda \| x \|_p^p \]

How to compare different \( p \)s?

- Given \( G \) and \( \sigma_w \)
  
  \[ \lambda_p^* \equiv \arg\min \lambda \lim_{N \to \infty} \frac{\| \hat{x}_p(\lambda) - x_o \|_2^2}{N} \]

  \[ \text{Compare MSE} \equiv \lim_{N \to \infty} \frac{\| \hat{x}_p(\lambda_p^*) - x_o \|_2^2}{N} \text{ for different } p \]
How about noisy setting?

LPLS: \[ \hat{x}_p(\lambda) = \arg \min_x \frac{1}{2} \| y - Ax \|^2_2 + \lambda \| x \|^p_p \]

Given \( G \) and \( \sigma_w \)

- \[ \lambda^*_p \triangleq \arg \min_{\lambda} \lim_{N \to \infty} \frac{\| \hat{x}_p(\lambda) - x_o \|^2}{N} \]

- Compare MSE \( \triangleq \lim_{N \to \infty} \frac{\| \hat{x}_p(\lambda^*_p) - x_o \|^2}{N} \) for different \( p \)
How about noisy setting?

Under the assumption of Replica:

**Theorem**

There exists $\sigma_h > 0$ s.t. $\forall \sigma_w < \sigma_h$ optimal-$\lambda$ LPLS with $p = 0$ outperforms the other values of $p$. Furthermore, there exists $\sigma_u > 0$ s.t. for $\forall \sigma_w > \sigma_u$, optimal-$\lambda$ LASSO outperforms every $0 \leq p < 1$. 

![Graph](image-url)
How do we get the results?
How do we get the results?

Under the assumption of Replica:

**Theorem**

As $N \to \infty$, $(\hat{x}_j(\lambda, p), x_j)$ converges in distribution to $(\eta_p(X + \sigma_\ell Z; \lambda), X)$ where $X \sim p_X$ and $Z \sim \mathcal{N}(0, 1)$, then the following holds:

$$\sigma^2_\ell = \sigma^2_w + \frac{1}{\delta} \mathbb{E}(\eta_p(X + \sigma_\ell Z; \lambda) - X)^2.$$
How do we get the results?

Replica: \( \sigma^2_{\ell} = \sigma^2_w + \frac{1}{\delta} \mathbb{E}(\eta_p(X + \sigma\ell Z; \lambda) - X)^2 \)

Define \( \Psi_{\lambda,p}(\sigma^2) = \sigma^2_w + \frac{1}{\delta} \mathbb{E}(\eta_p(X + \sigma Z; \lambda) - X)^2 \)

- \( \text{MSE} = \mathbb{E}(\eta_p(X + \sigma\ell Z; \lambda) - X)^2 \)
- Smaller \( \sigma^2_{\ell} \): smaller MSE
- Best performance: find \( \lambda \) that has the smallest stable fixed point

\[
\begin{align*}
\tilde{\sigma}^2 & = \sigma^2 \\
\tilde{\sigma}^2 & = \Psi_{\lambda,p}(\sigma^2)
\end{align*}
\]
How do we get the results

Define:

$$\lambda^*(\sigma^2) = \arg \min_\lambda \mathbb{E}(\eta_p(X + \sigma Z; \lambda) - X)^2$$

$$\Psi_{\lambda^*,p}(\sigma^2) = \sigma_w^2 + \frac{1}{\delta} \mathbb{E}(\eta_p(X + \sigma Z; \lambda^*_p(\sigma)) - X)^2$$

Lemma:

The stable fixed point of $$\Psi_{\lambda^*,p}(\sigma^2)$$ is $$\inf_\lambda \sigma^2_{\ell}(\lambda)$$

Challenges:

- Existence of multiple stable fixed points
- $$\eta_p$$ does not have explicit form for $$0 < p < 1$$
Conclusions

Noiseless setting:

- The global minimum of $\ell_p$-minimization ($0 \leq p < 1$) performs much better than that of $\ell_1$-minimization.
- The global minimum of $\ell_p$-minimization ($0 \leq p < 1$) is not affected by $p$ or $G$.

Noisy setting:

- For small $\sigma_w$, $\ell_0$-minimization outperforms the other $\ell_p$-minimization ($0 < p \leq 1$).
- For large $\sigma_w$, LASSO outperforms $\ell_p$-minimization ($0 \leq p < 1$).
- The global minimum of $\ell_p$-minimization ($0 \leq p \leq 1$) is affected by $p$ and $G$. 
Practical and analyzable algorithms? (To some extent addressed in our paper)

http://arxiv.org/abs/1501.03704
State evolution for $\ell_1$-minimization:
State evolution for $\ell_p$-minimization ($p < 1$):
State evolution for $\ell_p$-minimization ($p < 1$):
Comparison of state evolution for $\ell_p$-minimization:

![Graph showing the comparison of state evolution for $\ell_p$-minimization]
Comparison of $\ell_p$-minimization:
Comparison of $\ell_p$-minimization: