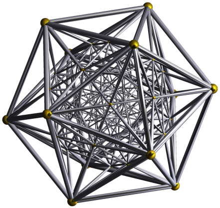
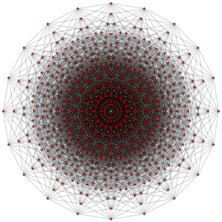
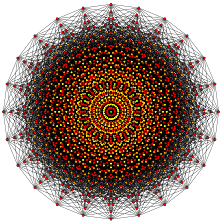
Computational Principles for High-dim Data Analysis

(Lecture Seven)

Yi Ma

EECS Department, UC Berkeley

September 21, 2021

Convex Methods for Sparse Signal Recovery
(Noisy Observations or Approximated Sparsity)

"Algebra is but written geometry; geometry is but drawn algebra."
– Sophie Germain

Problem Formulation

The observation $\boldsymbol{y}$ is perturbed by a small amount of noise $\boldsymbol{z}$:

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_o + \boldsymbol{z}, \quad \|\boldsymbol{z}\|_2 \le \epsilon. \tag{1}$$

Three typical scenarios (or combination of them):

- **Deterministic error**: $\boldsymbol{z}$ is bounded: $\|\boldsymbol{z}\|_2 \le \epsilon$, and ϵ is known.
- **Stochastic noise**: entries of $\boldsymbol{z} \sim_{iid} \mathcal{N}(0, \frac{\sigma^2}{m})$ hence $\|\boldsymbol{z}\|_2 \approx \sigma$.
- **Inexact sparsity**: $\boldsymbol{x}_o$ is not perfectly sparse with $\|\boldsymbol{x}_o - [\boldsymbol{x}_o]_k\|$ small.

Problem Formulation

The observation $\boldsymbol{y}$ is perturbed by a small amount of noise $\boldsymbol{z}$:

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_o + \boldsymbol{z}, \quad \|\boldsymbol{z}\|_2 \leq \epsilon. \tag{2}$$

Three typical tasks (or combination of them):

- **Estimation**: Is $\|\hat{\boldsymbol{x}} - \boldsymbol{x}_o\|_2$ small?
- **Prediction**: Is $\boldsymbol{A}\hat{\boldsymbol{x}} \approx \boldsymbol{A}\boldsymbol{x}_o$?
- **Identification**: Is $\mathrm{supp}(\hat{\boldsymbol{x}}) = \mathrm{supp}(\boldsymbol{x}_o)$?

Lasso versus Basis Pursuit Denoising

To find a sparse $\boldsymbol{x}_o$ from noisy measurements:

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_o + \boldsymbol{z}, \quad \|\boldsymbol{z}\|_2 \leq \epsilon. \tag{3}$$

I. BPDN (basis pursuit denoising):

$$\min \|\boldsymbol{x}\|_1 \quad \text{subject to} \quad \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2 \leq \epsilon. \tag{4}$$

II. LASSO (least absolute shrinkage and selection operator):

$$\min \lambda \|\boldsymbol{x}\|_1 + \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2. \tag{5}$$

$\exists \lambda \leftrightarrow \epsilon$ **such that BPDN and LASSO have the same optimal solution.**

Stable Recovery: Bounded Error (Best Scenario)

Knowing the support I of $\boldsymbol{x}_o$, solve the least squares problem:

$$\min \|\boldsymbol{y} - \boldsymbol{A}_{\mathsf{I}}\boldsymbol{x}'(\mathsf{I})\|_2^2 \tag{6}$$

to obtain the "oracle" (best possible) estimate:

$$\begin{cases} \hat{\boldsymbol{x}}'(\mathsf{I}) = (\boldsymbol{A}_{\mathsf{I}}^*\boldsymbol{A}_{\mathsf{I}})^{-1}\boldsymbol{A}_{\mathsf{I}}^*\boldsymbol{y}, \\ \hat{\boldsymbol{x}}'(\mathsf{I}^c) = \mathbf{0}. \end{cases} \tag{7}$$

From $\|\boldsymbol{A}_{\mathsf{I}}\hat{\boldsymbol{x}}' - \boldsymbol{A}_{\mathsf{I}}\boldsymbol{x}_o\|_2 \leq \epsilon$, we have the (tight) error bound:

$$\left\|\hat{\boldsymbol{x}}' - \boldsymbol{x}_o\right\|_2 \leq \frac{\epsilon}{\sigma_{\min}(\boldsymbol{A}_{\mathsf{I}})} \sim c\epsilon. \tag{8}$$

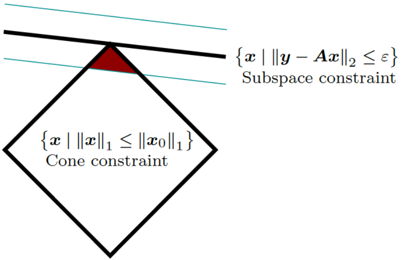
Stable Recovery: Bounded Error

Theorem (Stable Sparse Recovery via BPDN)

Suppose that $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_o + \boldsymbol{z}$, *with* $\|\boldsymbol{z}\|_2 \le \epsilon$, *and let* $k = \|\boldsymbol{x}_o\|_0$. *If* $\delta_{2k}(\boldsymbol{A}) < \sqrt{2} - 1$, *then any solution* $\hat{\boldsymbol{x}}$ *to the optimization problem:* $\min \|\boldsymbol{x}\|_1$ *s.t.* $\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2 \le \epsilon$ *satisfies*

$$\|\hat{\boldsymbol{x}} - \boldsymbol{x}_o\|_2 \le C\epsilon. \tag{9}$$

Here, C *is a constant which depends only on* $\delta_{2k}(\boldsymbol{A})$ *(and not on* ϵ*).*



Stable Recovery: Bounded Error

Proof.

From feasibility of the solutions,

$$\begin{aligned} \|\boldsymbol{A}(\hat{\boldsymbol{x}} - \boldsymbol{x}_o)\|_2 &= \|(\boldsymbol{y} - \boldsymbol{A}\hat{\boldsymbol{x}}) - (\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}_o)\|_2 \\ &\leq \|\boldsymbol{y} - \boldsymbol{A}\hat{\boldsymbol{x}}\|_2 + \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}_o\|_2 \\ &\leq 2\epsilon. \end{aligned}$$

Let $\boldsymbol{h} = \hat{\boldsymbol{x}} - \boldsymbol{x}_o$, from optimality of $\hat{\boldsymbol{x}}$: $\|\hat{\boldsymbol{x}}\|_1 \leq \|\boldsymbol{x}_o\|_1$, we have

$$\|\boldsymbol{h}_{\mathsf{I}^c}\|_1 \leq \|\boldsymbol{h}_{\mathsf{I}}\|_1 .$$

With $\delta_{2k} < \sqrt{2} - 1$, $\boldsymbol{A}$ satisfies the RSC property on $\boldsymbol{h}$ above. Therefore, we have

$$\|\boldsymbol{A}\boldsymbol{h}\|_2^2 \geq \mu \|\boldsymbol{h}\|_2^2. \tag{10}$$

□

Stable Recovery: Random Noise

Model: $\boldsymbol{x}_o$ is k-sparse, and the matrix $\boldsymbol{A} \sim \mathcal{N}(0, \frac{1}{m})$ and $\boldsymbol{z} \sim \mathcal{N}(0, \frac{\sigma^2}{m})$:

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_o + \boldsymbol{z} \quad \in \mathbb{R}^n. \tag{11}$$

Solve the Lasso program for an estimate $\hat{\boldsymbol{x}}$:

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \lambda_m \|\boldsymbol{x}\|_1 . \tag{12}$$

Let $\boldsymbol{h} = \hat{\boldsymbol{x}} - \boldsymbol{x}_o \in \mathbb{R}^n$ and $L(\boldsymbol{x}) \doteq \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2$. Notice that $\nabla L(\boldsymbol{x}) = -\boldsymbol{A}^*(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x})$ and in particular:

$$\nabla L(\boldsymbol{x}_o) = -\boldsymbol{A}^*(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}_o) = -\boldsymbol{A}^*\boldsymbol{z}.$$

$$L(\hat{\boldsymbol{x}}) = L(\boldsymbol{x}_o) + \langle \nabla L(\boldsymbol{x}_o), \hat{\boldsymbol{x}} - \boldsymbol{x}_o \rangle + \frac{1}{2}\|\boldsymbol{A}(\hat{\boldsymbol{x}} - \boldsymbol{x}_o)\|_2^2.$$

Stable Recovery: Random Noise

Since $\hat{\boldsymbol{x}}$ minimizes the objective function, we have:

$$\begin{aligned} 0 &\geq L(\hat{\boldsymbol{x}}) + \lambda_m \|\hat{\boldsymbol{x}}\|_1 - L(\boldsymbol{x}_o) - \lambda_m \|\boldsymbol{x}_o\|_1 \\ &\geq \langle \nabla L(\boldsymbol{x}_o), \hat{\boldsymbol{x}} - \boldsymbol{x}_o \rangle + \lambda_m (\|\hat{\boldsymbol{x}}\|_1 - \|\boldsymbol{x}_o\|_1) \\ &\geq -\left| \langle \boldsymbol{A}^* \boldsymbol{z}, \boldsymbol{h} \rangle \right| + \lambda_m (\|\hat{\boldsymbol{x}}\|_1 - \|\boldsymbol{x}_o\|_1) \\ &\geq -\|\boldsymbol{A}^* \boldsymbol{z}\|_\infty \|\boldsymbol{h}\|_1 + \lambda_m (\|\boldsymbol{h}_{\mathsf{I}^c}\|_1 - \|\boldsymbol{h}_{\mathsf{I}}\|_1). \end{aligned} \tag{13}$$

This is almost the cone condition: $\|\boldsymbol{h}_{\mathsf{I}^c}\|_1 \leq \|\boldsymbol{h}_{\mathsf{I}}\|_1$, given the first term is very small.

Need a slightly relaxed version of the cone condition.

Stable Recovery: Random Noise

Lemma

For the lasso problem (12), if we choose $\lambda_m \geq c \cdot 2\sigma\sqrt{\frac{\log n}{m}}$*, then with high probability,* $\boldsymbol{h} = \hat{\boldsymbol{x}} - \boldsymbol{x}_o$ *satisfies the cone condition:*

$$\|\boldsymbol{h}_{\mathsf{I}^c}\|_1 \leq \frac{c+1}{c-1} \cdot \|\boldsymbol{h}_{\mathsf{I}}\|_1. \tag{14}$$

Proof (Sketch):

As $\boldsymbol{a}_i^* \boldsymbol{z}$ is a Gaussian random variable of variance σ^2/m, we have

$$\mathbb{P}\left[|\boldsymbol{a}_i^* \boldsymbol{z}| \geq t\right] \leq 2\exp\left(-\frac{mt^2}{2\sigma^2}\right). \tag{15}$$

By union bound on the n columns, we have

$$\mathbb{P}\left[\|\boldsymbol{A}^* \boldsymbol{z}\|_\infty \geq t\right] \leq 2\exp\left(-\frac{mt^2}{2\sigma^2} + \log n\right). \tag{16}$$

Stable Recovery: Random Noise

Proof (continued): Choose $t^2 = 4\frac{\sigma^2 \log n}{m}$, then with high probability at least $1 - cn^{-1}$, we have

$$\|\boldsymbol{A}^* \boldsymbol{z}\|_\infty \le 2\sigma \sqrt{\frac{\log n}{m}}.$$

choose $\lambda_m \ge c \cdot 2\sigma \sqrt{\frac{\log n}{m}}$ for some $c > 0$. Then from the last inequality of (13), we have

$$\begin{aligned} 0 &\ge -\|\boldsymbol{A}^* \boldsymbol{z}\|_\infty \|\boldsymbol{h}\|_1 + \lambda_m (\|\hat{\boldsymbol{x}}\|_1 - \|\boldsymbol{x}_o\|_1) \\ &\ge -\frac{\lambda_m}{c}\|\boldsymbol{h}_{\mathsf{I}}\|_1 - \frac{\lambda_m}{c}\|\boldsymbol{h}_{\mathsf{I}^c}\|_1 + \lambda_m \|\boldsymbol{h}_{\mathsf{I}^c}\|_1 - \lambda_m \|\boldsymbol{h}_{\mathsf{I}}\|_1 \\ &= \lambda_m \left(\left(1 - \frac{1}{c}\right) \|\boldsymbol{h}_{\mathsf{I}^c}\|_1 - \left(1 + \frac{1}{c}\right) \|\boldsymbol{h}_{\mathsf{I}}\|_1 \right). \end{aligned} \tag{17}$$

Stable Recovery: Random Noise

Theorem (Stable Sparse Recovery via Lasso)

Suppose that $\boldsymbol{A} \sim_{iid} \mathcal{N}(0, \frac{1}{m})$, *and* $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_o + \boldsymbol{z}$, *with* $\boldsymbol{x}_o$ k*-sparse and* $\boldsymbol{z} \sim_{iid} \mathcal{N}(0, \frac{\sigma^2}{m})$. *Solve the Lasso*

$$\min \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \lambda_m \|\boldsymbol{x}\|_1, \tag{18}$$

with regularization parameter $\lambda_m = c \cdot 2\sigma\sqrt{\frac{\log n}{m}}$ *for a large enough* c. *Then with high probability,*

$$\|\hat{\boldsymbol{x}} - \boldsymbol{x}_o\|_2 \;\leq\; C'\sigma\sqrt{\frac{k \log n}{m}}. \tag{19}$$

Compared to (9), $C'\sqrt{\frac{k\log n}{m}}$ **can be very small as** $k/m \to 0$**!**

Stable Recovery: Random Noise

Proof.

From the optimality of $\hat{\boldsymbol{x}}$:

$$\begin{aligned} 0 &\geq L(\hat{\boldsymbol{x}}) + \lambda_m \|\hat{\boldsymbol{x}}\|_1 - L(\boldsymbol{x}_o) - \lambda_m \|\boldsymbol{x}_o\|_1 \\ &\geq \frac{1}{2}\|\boldsymbol{A}(\hat{\boldsymbol{x}} - \boldsymbol{x}_o)\|_2^2 + \langle \nabla L(\boldsymbol{x}_o), \hat{\boldsymbol{x}} - \boldsymbol{x}_o \rangle + \lambda_m (\|\hat{\boldsymbol{x}}\|_1 - \|\boldsymbol{x}_o\|_1) \\ &\geq \frac{1}{2}\|\boldsymbol{A}\boldsymbol{h}\|_2^2 + \lambda_m \left(\left(1 - \frac{1}{c}\right) \|\boldsymbol{h}_{\mathsf{I}^c}\|_1 - \left(1 + \frac{1}{c}\right) \|\boldsymbol{h}_{\mathsf{I}}\|_1 \right), \end{aligned} \tag{20}$$

Hence

$$\frac{1}{2}\|\boldsymbol{A}\boldsymbol{h}\|_2^2 \leq \lambda_m \left(1 + \frac{1}{c}\right) \|\boldsymbol{h}_{\mathsf{I}}\|_1 \leq \lambda_m \left(1 + \frac{1}{c}\right) \sqrt{k} \|\boldsymbol{h}\|_2.$$

W.H.P., random $\boldsymbol{A}$ satisfies the RSC property: $\|\boldsymbol{A}\boldsymbol{h}\|_2^2 \geq \mu \|\boldsymbol{h}\|_2^2$. □

Stable Recovery: Random Noise

The above bound is **nearly optimal** in the sense:[1]

Theorem

Suppose that we will observe $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{z}$*. Set*

$$M^\star(\boldsymbol{A}) = \inf_{\hat{\boldsymbol{x}}} \sup_{\|\boldsymbol{x}\|_0 \le k} \mathbb{E}\,\|\hat{\boldsymbol{x}}(\boldsymbol{y}) - \boldsymbol{x}\|_2^2 \,. \tag{21}$$

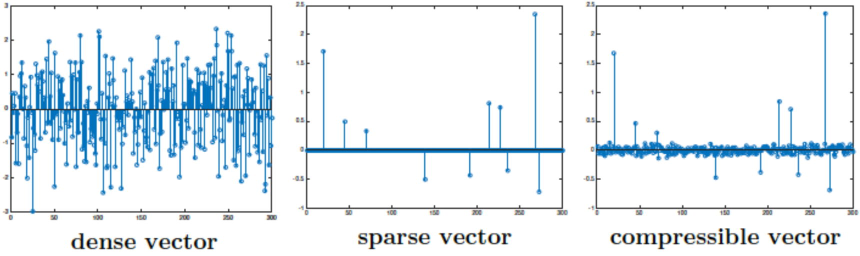
Then for any $\boldsymbol{A}$ *with* $\|\boldsymbol{e}_i^* \boldsymbol{A}\|_2 \le \sqrt{n}$ *for each* i*, we have*

$$M^\star(\boldsymbol{A}) \;\ge\; C\sigma^2 \frac{k \log(n/k)}{m}. \tag{22}$$

Difference in bound $\|\hat{\boldsymbol{x}}(\boldsymbol{y}) - \boldsymbol{x}\|_2^2$ **is only** $O(\sigma^2 \frac{k \log k}{m}) \searrow 0$ **as** $k/m \searrow 0$**.**

[1] *How well can we estimate a sparse vector?* E. Candes and M. Davenport, 2013.

Approximate Sparsity



$\boldsymbol{x}_o$ is not perfectly k-sparse. Let $[\boldsymbol{x}_o]_k$ be the best k-sparse signal that approximates $\boldsymbol{x}_o$. Then we can rewrite the observation model as:

$$\boldsymbol{y} = \boldsymbol{A}[\boldsymbol{x}_o]_k + \boldsymbol{A}(\boldsymbol{x}_o - [\boldsymbol{x}_o]_k) + \boldsymbol{z}.$$

How well does ℓ^1 minimization recover such signals?

Approximate Sparsity

Theorem

Let $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_o + \boldsymbol{z}$, with $\|\boldsymbol{z}\|_2 \le \epsilon$. Let $\hat{\boldsymbol{x}}$ solve the basis pursuit denoising problem

$$\min \|\boldsymbol{x}\|_1 \quad \text{subject to} \quad \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2 \le \epsilon. \tag{23}$$

Then for any k such that $\delta_{2k}(\boldsymbol{A}) < \sqrt{2} - 1$,

$$\|\hat{\boldsymbol{x}} - \boldsymbol{x}_o\|_2 \le C \frac{\|\boldsymbol{x}_o - [\boldsymbol{x}_o]_k\|_1}{\sqrt{k}} + C'\epsilon \tag{24}$$

for some constants C and C' which only depend on $\delta_{2k}(\boldsymbol{A})$.

Notice: When $\boldsymbol{x}_o - [\boldsymbol{x}_o]_k = \boldsymbol{0}$, this reduces to previous result on stable recovery.

Approximate Sparsity

Sketch of Proof.

From feasibility of the solution $\hat{\boldsymbol{x}}$:

$$\|\boldsymbol{A}\boldsymbol{h}\|_2 = \|\boldsymbol{A}(\hat{\boldsymbol{x}} - \boldsymbol{x}_o)\|_2 \leq 2\epsilon.$$

From optimality of the solution $\hat{\boldsymbol{x}}$:

$$0 \leq \|\boldsymbol{x}_o\|_1 - \|\hat{\boldsymbol{x}}\|_1 \quad \Longleftrightarrow \quad \|\boldsymbol{h}_{\mathsf{I}^c}\|_1 \leq \|\boldsymbol{h}_{\mathsf{I}}\|_1 + 2\|\boldsymbol{x}_{o\mathsf{I}^c}\|_1 . \tag{25}$$

Follow the same proof of RIP for the clean case. The only difference is to replace the condition $\|\boldsymbol{h}_{\mathsf{I}^c}\|_1 \leq \|\boldsymbol{h}_{\mathsf{I}}\|_1$ with the new one. We obtain:

$$\|\boldsymbol{A}\boldsymbol{h}\|_2 \geq \frac{1-(1+\sqrt{2})\delta_{2k}}{(1+\delta_{2k})^{1/2}}\|\boldsymbol{h}_{\mathsf{I}\cup\mathsf{J}_1}\|_2 - \frac{2\sqrt{2}\delta_{2k}}{(1+\delta_{2k})^{1/2}}\frac{\|\boldsymbol{x}_{o\mathsf{I}^c}\|_1}{\sqrt{k}}. \tag{26}$$

Combing with $\|\boldsymbol{h}\|_2 \leq 2\|\boldsymbol{h}_{\mathsf{I}\cup\mathsf{J}_1}\|_2 + 2\frac{\|\boldsymbol{x}_{o\mathsf{I}^c}\|_1}{\sqrt{k}}$ gives the result. □

Conclusions

ℓ^1 minimization

$$\min \|\boldsymbol{x}\|_1 \quad \text{subject to} \quad \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2 \le \epsilon.$$

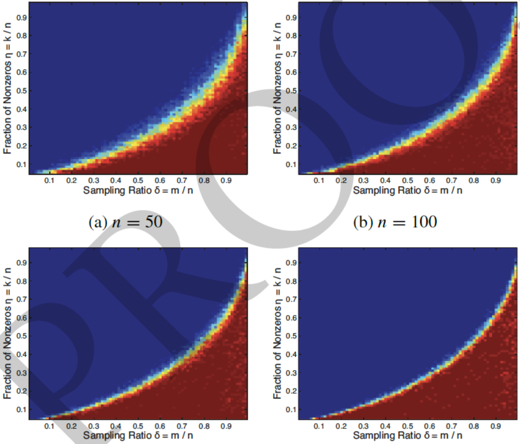
finds a stable estimate $\hat{\boldsymbol{x}}$ to the k-sparse $\boldsymbol{x}_o$:

$$\hat{\boldsymbol{x}} : \|\hat{\boldsymbol{x}} - \boldsymbol{x}_o\|_2 \le C\epsilon.$$

For a random matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, we need:

- **mutual coherence:**
$$m = O(k^2).$$
- **restricted isometry:**
$$m = O\big(k \log(n/k)\big).$$

Next: the Phase Transition Phenomenon



(a) $n = 50$ (b) $n = 100$

(c) $n = 200$ (d) $n = 400$

Can we characterize this phenomenon mathematically?

Assignments

- Reading: Section 3.5 of Chapter 3.
- Written Homework # 2.