Compressed Sensing and Sparse Recovery

Yuxin Chen

Wharton Statistics & Data Science, Spring 2022

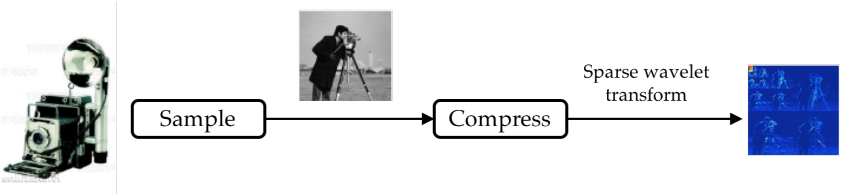
Outline

- Compressed sensing
- Restricted isometry property (RIP)
- A RIPless theory

Motivation: wastefulness of data acquisition

Conventional paradigms for data acquisition:

- Measure full data
- Compress (by discarding a large fraction of coefficients)



Problem: data are often highly compressible

- Most of acquired data can be thrown away without any perceptual loss

Blind sensing

Ideally, if we know *a priori* which coefficients are worth estimating, then we can simply measure these coefficients

- Unfortunately, we often have no idea which coefficients are relevant

Compressed sensing: compression on the fly

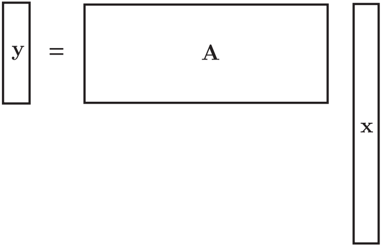
- mimic the behavior of the above ideal situation without pre-computing all coefficients
- often achieved by *random* sensing mechanism

Why go to so much effort to acquire all the data when most of what we get will be thrown away?

Can't we just directly measure the part that won't end up being thrown away?

— David Donoho

Setup: sparse recovery



Recover $\boldsymbol{x} \in \mathbb{R}^p$ given $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$

where $\boldsymbol{A} = [\boldsymbol{a}_1, \cdots, \boldsymbol{a}_n]^\top \in \mathbb{R}^{n \times p}$ ($n \ll p$): sampling matrix;
$\boldsymbol{a}_i$: sampling vector; $\boldsymbol{x}$: sparse signal

Restricted isometry properties

Optimality for ℓ_0 minimization

$$\text{minimize}_{\boldsymbol{x}\in\mathbb{R}^p} \ \|\boldsymbol{x}\|_0 \quad \text{s.t. } \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}$$

If instead $\exists$ a sparser feasible $\widetilde{\boldsymbol{x}} \neq \boldsymbol{x}$ s.t. $\|\widetilde{\boldsymbol{x}}\|_0 \leq \|\boldsymbol{x}\|_0 = k$, then

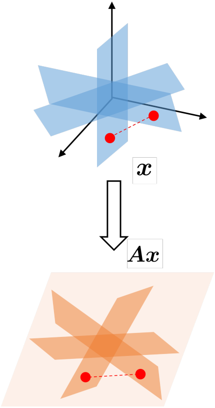
$$\boldsymbol{A}\left(\boldsymbol{x} - \widetilde{\boldsymbol{x}}\right) = \boldsymbol{0}. \qquad (9.1)$$

We don't want (9.1) to happen, so we hope

$$\boldsymbol{A}(\underbrace{\boldsymbol{x} - \widetilde{\boldsymbol{x}}}_{2k\text{-sparse}}) \neq \boldsymbol{0}, \qquad \forall \widetilde{\boldsymbol{x}} \ \text{ with } \|\widetilde{\boldsymbol{x}}\|_0 \leq k$$

To simultaneously account for all k-sparse $\boldsymbol{x}$, we hope $\boldsymbol{A}_T$ $(|T| \leq 2k)$ to have full rank, where $\boldsymbol{A}_T$ consists of all columns of $\boldsymbol{A}$ at indices from T

Restricted isometry property (RIP)



Definition 9.1 (Restricted isometry constant (Candès & Tao '06))

Restricted isometry constant δ_k of $\boldsymbol{A}$ is the smallest quantity s.t.

$$(1-\delta_k)\|\boldsymbol{x}\|_2^2 \leq \|\boldsymbol{A}\boldsymbol{x}\|_2^2 \leq (1+\delta_k)\|\boldsymbol{x}\|_2^2 \tag{9.2}$$

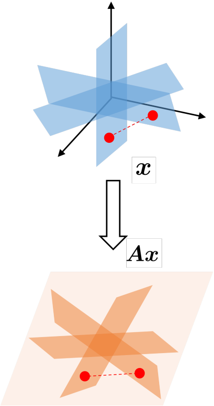
holds for all k-sparse vector $\boldsymbol{x} \in \mathbb{R}^p$

- (check) equivalently, (9.2) says

$$\max_{S:|S|=k} \underbrace{\|\boldsymbol{A}_S^\top \boldsymbol{A}_S - \boldsymbol{I}_k\|}_{\text{near orthonormality}} = \delta_k$$

where $\boldsymbol{A}_S$ consists of all columns of $\boldsymbol{A}$ at indices from S

Restricted isometry property (RIP)



Definition 9.1 (Restricted isometry constant (Candès & Tao '06))

Restricted isometry constant δ_k of $\boldsymbol{A}$ is the smallest quantity s.t.

$$(1-\delta_k)\|\boldsymbol{x}\|_2^2 \le \|\boldsymbol{A}\boldsymbol{x}\|_2^2 \le (1+\delta_k)\|\boldsymbol{x}\|_2^2 \tag{9.2}$$

holds for all k-sparse vector $\boldsymbol{x} \in \mathbb{R}^p$

- (Homework) For any $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ that are supported on disjoint subsets S_1, S_2 with $|S_1| \le s_1$ and $|S_2| \le s_2$:

$$\underbrace{|\langle \boldsymbol{A}\boldsymbol{x}_1, \boldsymbol{A}\boldsymbol{x}_2\rangle| \le \delta_{s_1+s_2}\|\boldsymbol{x}_1\|_2\|\boldsymbol{x}_2\|_2}_{\text{approximately preserves the inner product}} \tag{9.3}$$

RIP and ℓ_0 minimization

$$\text{minimize}_{\boldsymbol{x}\in\mathbb{R}^p} \ \|\boldsymbol{x}\|_0 \quad \text{s.t. } \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}$$

Fact 9.2

(Exercise) Suppose a feasible $\boldsymbol{x}$ is k-sparse. If $\delta_{2k} < 1$, then $\boldsymbol{x}$ is the unique solution to ℓ_0 minimization

RIP and ℓ_1 minimization

$$\text{minimize}_{\boldsymbol{x}\in\mathbb{R}^p} \ \|\boldsymbol{x}\|_1 \quad \text{s.t. } \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}$$

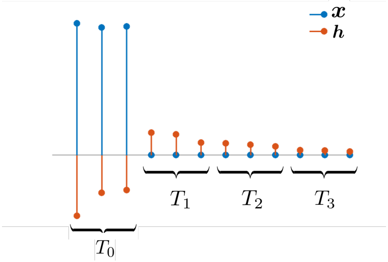
Theorem 9.3 (Candès & Tao '06, Candès '08)

Suppose a feasible $\boldsymbol{x}$ is k-sparse. If $\delta_{2k} < \sqrt{2} - 1$, then $\boldsymbol{x}$ is the unique solution to ℓ_1 minimization

- RIP implies the success of ℓ_1 minimization
- A universal result: works simultaneously for all k-sparse signals
- As we will see later, many random designs satisfy this condition with *near-optimal sample complexity*

Proof of Theorem 9.3

Suppose $\boldsymbol{x}+\boldsymbol{h}$ is feasible and obeys $\|\boldsymbol{x}+\boldsymbol{h}\|_1 \le \|\boldsymbol{x}\|_1$. The goal is to show that $\boldsymbol{h}=\boldsymbol{0}$ under RIP.



The key is to decompose $\boldsymbol{h}$ into $\boldsymbol{h}_{T_0}+\boldsymbol{h}_{T_1}+\ldots$

- T_0: locations of the k largest entries of $\boldsymbol{x}$
- T_1: locations of the k largest entries of $\boldsymbol{h}$ in $T_0{}^{\mathsf{c}}$
- T_2: locations of the k largest entries of $\boldsymbol{h}$ in $(T_0 \cup T_1)^{\mathsf{c}}$
- ...

Proof of Theorem 9.3

Informally, the proof proceeds by showing that

1. $\boldsymbol{h}_{T_0 \cup T_1}$ "dominates" $\boldsymbol{h}_{(T_0 \cup T_1)^{\mathsf{c}}}$ (by objective function) — *see Step 1*

2. (converse) $\boldsymbol{h}_{(T_0 \cup T_1)^{\mathsf{c}}}$ "dominates" $\boldsymbol{h}_{T_0 \cup T_1}$ (by RIP + feasibility) — *see Step 2*

These cannot happen simultaneously unless $\boldsymbol{h}$ vanishes

Proof of Theorem 9.3

Step 1 (depending only on the objective function). Show that

$$\sum_{j\geq 2}\|\boldsymbol{h}_{T_j}\|_2 \leq \frac{1}{\sqrt{k}}\|\boldsymbol{h}_{T_0}\|_1 \tag{9.4}$$

This follows immediately by combining the following 2 observations:

(i) Since $\boldsymbol{x}+\boldsymbol{h}$ is assumed to be a better estimate:

$$\|\boldsymbol{x}\|_1 \geq \|\boldsymbol{x}+\boldsymbol{h}\|_1 = \underbrace{\|\boldsymbol{x}+\boldsymbol{h}_{T_0}\|_1 + \|\boldsymbol{h}_{T_0^c}\|_1}_{\text{since } T_0 \text{ is support of } \boldsymbol{x}} \geq \underbrace{\|\boldsymbol{x}\|_1 - \|\boldsymbol{h}_{T_0}\|_1}_{\text{triangle inequality}} + \|\boldsymbol{h}_{T_0^c}\|_1$$

$$\Longrightarrow \quad \|\boldsymbol{h}_{T_0^c}\|_1 \leq \|\boldsymbol{h}_{T_0}\|_1 \tag{9.5}$$

(ii) Since entries of $\boldsymbol{h}_{T_{j-1}}$ uniformly dominate those of $\boldsymbol{h}_{T_j}$ $(j\geq 2)$:

$$\|\boldsymbol{h}_{T_j}\|_2 \leq \sqrt{k}\|\boldsymbol{h}_{T_j}\|_\infty \leq \sqrt{k}\frac{\|\boldsymbol{h}_{T_{j-1}}\|_1}{k} = \frac{1}{\sqrt{k}}\|\boldsymbol{h}_{T_{j-1}}\|_1$$

$$\Longrightarrow \quad \sum_{j\geq 2}\|\boldsymbol{h}_{T_j}\|_2 \leq \frac{1}{\sqrt{k}}\sum_{j\geq 2}\|\boldsymbol{h}_{T_{j-1}}\|_1 = \frac{1}{\sqrt{k}}\|\boldsymbol{h}_{T_0^c}\|_1 \tag{9.6}$$

Proof of Theorem 9.3

Step 2 (using feasibility + RIP). Show that $\exists \rho < 1$ s.t.

$$\|\boldsymbol{h}_{T_0 \cup T_1}\|_2 \leq \rho \sum\nolimits_{j \geq 2} \|\boldsymbol{h}_{T_j}\|_2 \tag{9.7}$$

If this claim holds, then

$$\begin{aligned} \|\boldsymbol{h}_{T_0 \cup T_1}\|_2 &\leq \rho \sum\nolimits_{j \geq 2} \|\boldsymbol{h}_{T_j}\|_2 \overset{(9.4)}{\leq} \rho \frac{1}{\sqrt{k}} \|\boldsymbol{h}_{T_0}\|_1 \\ &\leq \rho \frac{1}{\sqrt{k}} \left(\sqrt{k} \|\boldsymbol{h}_{T_0}\|_2 \right) = \rho \|\boldsymbol{h}_{T_0}\|_2 \ \leq \ \rho \|\boldsymbol{h}_{T_0 \cup T_1}\|_2 \end{aligned} \tag{9.8}$$

Since $\rho < 1$, we necessarily have $\boldsymbol{h}_{T_0 \cup T_1} = \boldsymbol{0}$, which together with (9.5) yields $\boldsymbol{h} = \boldsymbol{0}$

Proof of Theorem 9.3

We now prove (9.7). To connect $\boldsymbol{h}_{T_0 \cup T_1}$ with $\boldsymbol{h}_{(T_0 \cup T_1)^c}$, we use feasibility:

$$\boldsymbol{A}\boldsymbol{h} = \boldsymbol{0} \quad \Longleftrightarrow \quad \boldsymbol{A}\boldsymbol{h}_{T_0 \cup T_1} = -\sum\nolimits_{j \geq 2} \boldsymbol{A}\boldsymbol{h}_{T_j},$$

which taken collectively with RIP yields

$$(1 - \delta_{2k}) \|\boldsymbol{h}_{T_0 \cup T_1}\|_2^2 \leq \|\boldsymbol{A}\boldsymbol{h}_{T_0 \cup T_1}\|_2^2 = \left| \langle \boldsymbol{A}\boldsymbol{h}_{T_0 \cup T_1}, \sum\nolimits_{j \geq 2} \boldsymbol{A}\boldsymbol{h}_{T_j} \rangle \right|$$

It follows from (9.3) that for all $j \geq 2$,

$$\begin{aligned} |\langle \boldsymbol{A}\boldsymbol{h}_{T_0 \cup T_1}, \boldsymbol{A}\boldsymbol{h}_{T_j} \rangle| &\leq |\langle \boldsymbol{A}\boldsymbol{h}_{T_0}, \boldsymbol{A}\boldsymbol{h}_{T_j} \rangle| + |\langle \boldsymbol{A}\boldsymbol{h}_{T_1}, \boldsymbol{A}\boldsymbol{h}_{T_j} \rangle| \\ &\overset{(9.3)}{\leq} \delta_{2k} (\|\boldsymbol{h}_{T_0}\|_2 + \|\boldsymbol{h}_{T_1}\|_2) \|\boldsymbol{h}_{T_j}\|_2 \ \leq \ \delta_{2k} \sqrt{2} \|\boldsymbol{h}_{T_0 \cup T_1}\|_2 \cdot \|\boldsymbol{h}_{T_j}\|_2, \end{aligned}$$

which gives

$$\begin{aligned} (1 - \delta_{2k}) \|\boldsymbol{h}_{T_0 \cup T_1}\|_2^2 &\leq \sum\nolimits_{j \geq 2} |\langle \boldsymbol{A}\boldsymbol{h}_{T_0 \cup T_1}, \boldsymbol{A}\boldsymbol{h}_{T_j} \rangle| \\ &\leq \sqrt{2} \delta_{2k} \|\boldsymbol{h}_{T_0 \cup T_1}\|_2 \sum\nolimits_{j \geq 2} \|\boldsymbol{h}_{T_j}\|_2 \end{aligned}$$

This establishes (9.7) if $\rho := \frac{\sqrt{2}\delta_{2k}}{1 - \delta_{2k}} < 1$ (or equivalently, $\delta_{2k} < \sqrt{2} - 1$).

Robustness for compressible signals

Theorem 9.4 (Candès & Tao '06, Candès '08)

If $\delta_{2k} < \sqrt{2} - 1$*, then the solution* $\widehat{\boldsymbol{x}}$ *to* ℓ_1 *minimization obeys*

$$\|\widehat{\boldsymbol{x}} - \boldsymbol{x}\|_2 \lesssim \frac{\|\boldsymbol{x} - \boldsymbol{x}_k\|_1}{\sqrt{k}},$$

where $\boldsymbol{x}_k$ *is the best* k*-term approximation of* $\boldsymbol{x}$

- Suppose the l^{th} largest entry of $\boldsymbol{x}$ is $1/l^{\alpha}$ for some $\alpha > 1$, then

$$\frac{1}{\sqrt{k}}\|\boldsymbol{x} - \boldsymbol{x}_k\|_1 \approx \frac{1}{\sqrt{k}} \sum\nolimits_{l>k} l^{-\alpha} \approx k^{-\alpha+0.5} \ll 1$$

- ℓ_1-min works well in recovering compressible signals
- Follows similar arguments as in the proof of Theorem 9.3

Proof of Theorem 9.4

Step 1 (depending only on objective function). Show that

$$\sum\nolimits_{j\geq 2} \|\boldsymbol{h}_{T_j}\|_2 \leq \frac{1}{\sqrt{k}} \|\boldsymbol{h}_{T_0}\|_1 + \frac{2}{\sqrt{k}} \|\boldsymbol{x} - \boldsymbol{x}_{T_0}\|_1 \tag{9.9}$$

This follows immediately by combining the following 2 observations:

(i) Since $\boldsymbol{x} + \boldsymbol{h}$ is assumed to be a better estimate:

$$\begin{aligned}
\|\boldsymbol{x}_{T_0}\|_1 + \|\boldsymbol{x}_{T_0^c}\|_1 = \|\boldsymbol{x}\|_1 \geq \|\boldsymbol{x} + \boldsymbol{h}\|_1 &= \|\boldsymbol{x}_{T_0} + \boldsymbol{h}_{T_0}\|_1 + \|\boldsymbol{x}_{T_0^c} + \boldsymbol{h}_{T_0^c}\|_1 \\
&\geq \|\boldsymbol{x}_{T_0}\|_1 - \|\boldsymbol{h}_{T_0}\|_1 + \|\boldsymbol{h}_{T_0^c}\|_1 - \|\boldsymbol{x}_{T_0^c}\|_1
\end{aligned}$$

$$\Longrightarrow \quad \|\boldsymbol{h}_{T_0^c}\|_1 \leq \|\boldsymbol{h}_{T_0}\|_1 + 2\|\boldsymbol{x}_{T_0^c}\|_1 \tag{9.10}$$

(ii) Recall from (9.6) that $\sum_{j\geq 2} \|\boldsymbol{h}_{T_j}\|_2 \leq \frac{1}{\sqrt{k}} \|\boldsymbol{h}_{T_0^c}\|_1$

We highlight in red the part different from the proof of Theorem 9.3.

Proof of Theorem 9.4

Step 2 (using feasibility + RIP). Recall from (9.7) that $\exists \rho < 1$ s.t.

$$\|\boldsymbol{h}_{T_0 \cup T_1}\|_2 \leq \rho \sum\nolimits_{j \geq 2} \|\boldsymbol{h}_{T_j}\|_2 \tag{9.11}$$

If this claim holds, then

$$\begin{aligned} \|\boldsymbol{h}_{T_0 \cup T_1}\|_2 &\leq \rho \sum\nolimits_{j \geq 2} \|\boldsymbol{h}_{T_j}\|_2 \overset{(9.10) \text{ and } (9.6)}{\leq} \rho \frac{1}{\sqrt{k}} \left\{ \|\boldsymbol{h}_{T_0}\|_1 + 2\|\boldsymbol{x}_{T_0^c}\|_1 \right\} \\ &\leq \rho \frac{1}{\sqrt{k}} \left(\sqrt{k} \|\boldsymbol{h}_{T_0}\|_2 + 2\|\boldsymbol{x}_{T_0^c}\|_1 \right) = \rho \|\boldsymbol{h}_{T_0}\|_2 + \frac{2\rho}{\sqrt{k}} \|\boldsymbol{x}_{T_0^c}\|_1 \\ &\leq \rho \|\boldsymbol{h}_{T_0 \cup T_1}\|_2 + \frac{2\rho}{\sqrt{k}} \|\boldsymbol{x}_{T_0^c}\|_1 \end{aligned}$$

$$\Longrightarrow \quad \|\boldsymbol{h}_{T_0 \cup T_1}\|_2 \leq \frac{2\rho}{1 - \rho} \frac{\|\boldsymbol{x}_{T_0^c}\|_1}{\sqrt{k}} \tag{9.12}$$

We highlight in red the part different from the proof of Theorem 9.3.

Proof of Theorem 9.4

Finally, putting the above together yields

$$
\begin{aligned}
\|\boldsymbol{h}\|_2 &\leq \|\boldsymbol{h}_{T_0 \cup T_1}\|_2 + \|\boldsymbol{h}_{(T_0 \cup T_1)^{\mathsf{c}}}\|_2 \\
&\overset{(9.9)}{\leq} \|\boldsymbol{h}_{T_0 \cup T_1}\|_2 + \frac{1}{\sqrt{k}}\|\boldsymbol{h}_{T_0}\|_1 + \frac{2}{\sqrt{k}}\|\boldsymbol{x} - \boldsymbol{x}_{T_0}\|_1 \\
&\leq \|\boldsymbol{h}_{T_0 \cup T_1}\|_2 + \|\boldsymbol{h}_{T_0}\|_2 + \frac{2}{\sqrt{k}}\|\boldsymbol{x} - \boldsymbol{x}_{T_0}\|_1 \\
&\leq 2\|\boldsymbol{h}_{T_0 \cup T_1}\|_2 + \frac{2}{\sqrt{k}}\|\boldsymbol{x} - \boldsymbol{x}_{T_0}\|_1 \\
&\overset{(9.12)}{\leq} \frac{2(1+\rho)}{1-\rho}\frac{\|\boldsymbol{x} - \boldsymbol{x}_{T_0}\|_1}{\sqrt{k}}
\end{aligned}
$$

We highlight in red the part different from the proof of Theorem 9.3.

Which design matrix satisfies RIP?

First example: i.i.d. Gaussian design

Lemma 9.5

A random matrix $\boldsymbol{A} \in \mathbb{R}^{n \times p}$ *with i.i.d.* $\mathcal{N}\left(0, \frac{1}{n}\right)$ *entries satisfies* $\delta_k < \delta$ *with high prob., as long as*

$$n \gtrsim \frac{1}{\delta^2} k \log \frac{p}{k}$$

- This is where non-asymptotic random matrix theory comes into play

Gaussian random matrices

Lemma 9.6 (See Vershynin '10)

Suppose $\boldsymbol{B} \in \mathbb{R}^{n \times k}$ *is composed of i.i.d.* $\mathcal{N}(0,1)$ *entries. Then*

$$\begin{cases} \mathbb{P}\left(\frac{1}{\sqrt{n}}\sigma_{\max}(\boldsymbol{B}) > 1 + \sqrt{\frac{k}{n}} + t\right) & \leq e^{-nt^2/2} \\ \mathbb{P}\left(\frac{1}{\sqrt{n}}\sigma_{\min}(\boldsymbol{B}) < 1 - \sqrt{\frac{k}{n}} - t\right) & \leq e^{-nt^2/2} \end{cases}$$

- When $n \gg k$, one has $\frac{1}{n}\boldsymbol{B}^\top \boldsymbol{B} \approx \boldsymbol{I}_k$
- Similar results (up to different constants) hold for i.i.d. sub-Gaussian matrices

Proof of Lemma 9.5

1. Fix any index subset $S \subseteq \{1, \cdots, \}$, $|S| = k$, then $\boldsymbol{A}_S$ (submatrix of $\boldsymbol{A}$ consisting of columns at indices from S) obeys

$$\|\boldsymbol{A}_S^\top \boldsymbol{A}_S - \boldsymbol{I}_k\| \leq O\Big(\sqrt{k/n}\Big) + t$$

with prob. exceeding $1 - 2e^{-c_1 n t^2}$, where $c_1 > 0$ is constant.

2. Taking a union bound over all $S \subseteq \{1, \cdots, p\}$, $|S| = k$ yields

$$\delta_k = \max_{S:|S|=k} \|\boldsymbol{A}_S^\top \boldsymbol{A}_S - \boldsymbol{I}_k\| \leq O\Big(\sqrt{k/n}\Big) + t$$

with prob. exceeding $1 - 2\binom{p}{k} e^{-c_1 n t^2} \geq 1 - 2e^{k \log(ep/k) - c_1 n t^2}$. Thus, $\delta_k < \delta$ with high prob. as long as $n \gtrsim \delta^{-2} k \log(p/k)$.

Other design matrices that satisfy RIP

- Random matrices with i.i.d. sub-Gaussian entries, as long as

$$n \gtrsim k \log(p/k)$$

- Random partial DFT matrices with

$$n \gtrsim k \log^4 p,$$

  where the rows of $\boldsymbol{A}$ are independently sampled from the rows of the DFT matrix $\boldsymbol{F}$ (Rudelson & Vershynin '08)

  - If you have learned entropy methods or generic chaining, check out Rudelson & Vershynin '08 and Candès & Plan '11

Other design matrices that satisfy RIP

- Random convolution matrices with

$$n \gtrsim k \log^4 p,$$

where the rows of $\boldsymbol{A}$ are independently sampled from the rows of

$$\boldsymbol{G} = \begin{bmatrix} g_0 & g_1 & g_2 & \cdots & g_{p-1} \\ g_{p-1} & g_0 & g_1 & \cdots & g_{p-2} \\ g_{p-2} & g_{p-1} & g_0 & \cdots & g_{p-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ g_1 & g_2 & g_3 & \cdots & g_0 \end{bmatrix}$$

with $\mathbb{P}(g_i = \pm 1) = 0.5$ (Krahmer, Mendelson, & Rauhut '14)

RIP guarantees success of many other methods

Example: projected gradient descent (iterative hard thresholding)

alternates between

- **gradient descent:**

$$\boldsymbol{z}^t \leftarrow \boldsymbol{x}^t - \mu_t \underbrace{\boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{x}^t - \boldsymbol{y})}_{\text{gradient of } \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2^2}$$

- **projection:** keep only the k largest (in magnitude) entries

Iterative hard thresholding (IHT)

Algorithm 9.1 Projected gradient descent / iterative hard thresholding

for $t = 0, 1, \cdots$**:**

$$\boldsymbol{x}^{t+1} = \mathcal{P}_k\left(\boldsymbol{x}^t - \mu_t \boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{x}^t - \boldsymbol{y})\right)$$

where $\mathcal{P}_k(\boldsymbol{x}) := \arg\min_{\|\boldsymbol{z}\|_0 = k} \|\boldsymbol{z} - \boldsymbol{x}\|_2$ is the best k-term approximation of $\boldsymbol{x}$

Geometric convergence of IHT under RIP

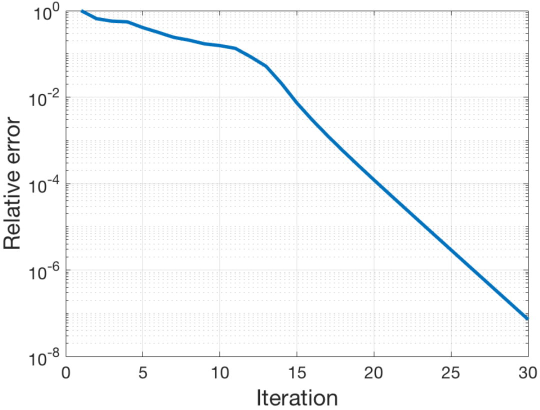
Theorem 9.7 (Blumensath & Davies '09)

Suppose $\boldsymbol{x}$ *is* k*-sparse, and the RIP constant* $\delta_{3k} < 1/2$*. Then taking* $\mu_t \equiv 1$ *gives*

$$\|\boldsymbol{x}^t - \boldsymbol{x}\|_2 \leq (2\delta_{3k})^t \, \|\boldsymbol{x}^0 - \boldsymbol{x}\|_2$$

- Under RIP, IHT attains ε-accuracy within $O\left(\log \frac{1}{\varepsilon}\right)$ iterations
- Each iteration takes time proportional to a matrix-vector product
- Drawback: need prior knowledge on k

Numerical performance of IHT



Relative error $\frac{\|\boldsymbol{x}^t - \boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2}$ vs. iteration count t

$(n = 100,\ k = 5,\ p = 1000,\ A_{i,j} \sim \mathcal{N}(0, 1/n))$

Proof of Theorem 9.7

Let $\boldsymbol{z} := \boldsymbol{x}^t - \boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{x}^t - \boldsymbol{y}) = \boldsymbol{x}^t - \boldsymbol{A}^\top\boldsymbol{A}(\boldsymbol{x}^t - \boldsymbol{x})$. By definition of $\mathcal{P}_k$,

$$\|\underbrace{\boldsymbol{x}}_{k\text{-sparse}} - \boldsymbol{z}\|_2^2 \geq \|\underbrace{\boldsymbol{x}^{t+1}}_{\text{best } k\text{-sparse}} - \boldsymbol{z}\|_2^2 = \|\boldsymbol{x}^{t+1} - \boldsymbol{x} - (\boldsymbol{z} - \boldsymbol{x})\|_2^2$$

$$= \|\boldsymbol{x}^{t+1} - \boldsymbol{x}\|_2^2 - 2\langle \boldsymbol{x}^{t+1} - \boldsymbol{x}, \boldsymbol{z} - \boldsymbol{x}\rangle + \|\boldsymbol{z} - \boldsymbol{x}\|_2^2$$

$$\begin{aligned} \Longrightarrow \quad \|\boldsymbol{x}^{t+1} - \boldsymbol{x}\|_2^2 &\leq 2\langle \boldsymbol{x}^{t+1} - \boldsymbol{x},\ \boldsymbol{z} - \boldsymbol{x}\rangle \\ &= 2\left\langle \boldsymbol{x}^{t+1} - \boldsymbol{x},\ (\boldsymbol{I} - \boldsymbol{A}^\top\boldsymbol{A})(\boldsymbol{x}^t - \boldsymbol{x})\right\rangle \\ &\leq 2\delta_{3k}\|\boldsymbol{x}^{t+1} - \boldsymbol{x}\|_2 \cdot \|\boldsymbol{x}^t - \boldsymbol{x}\|_2 \end{aligned} \tag{9.13}$$

which gives

$$\|\boldsymbol{x}^{t+1} - \boldsymbol{x}\|_2 \leq 2\delta_{3k}\|\boldsymbol{x}^t - \boldsymbol{x}\|_2$$

as claimed. Here, (9.13) follows from the following fact (homework)

$$|\langle \boldsymbol{u},\ (\boldsymbol{I} - \boldsymbol{A}^\top\boldsymbol{A})\boldsymbol{v}\rangle| \leq \delta_s\|\boldsymbol{u}\|_2 \cdot \|\boldsymbol{v}\|_2 \quad \text{with } s = |\mathrm{supp}\,(\boldsymbol{u}) \cup \mathrm{supp}\,(\boldsymbol{v})|$$

A RIPless theory

Is RIP necessary?

- RIP leads to a universal result holding simultaneously for all k-sparse $\boldsymbol{x}$
  - Universality is often not needed as we might only care about a particular $\boldsymbol{x}$

- There may be a gap between the regime where RIP holds and the regime in which one has minimal measurements

- Certifying RIP is hard

Can we develop a non-universal RIPless theory?

A standard recipe

1. Write out Karush-Kuhn-Tucker (KKT) optimality conditions
   - typically involves certain dual variables

2. Construct dual variables satisfying KKT conditions

Karush-Kuhn-Tucker (KKT) conditions

Consider a convex problem

$$\begin{aligned} \text{minimize}_{\boldsymbol{x}} \quad & f(\boldsymbol{x}) \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} - \boldsymbol{y} = \boldsymbol{0} \end{aligned}$$

Lagrangian:

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\nu}) := f(\boldsymbol{x}) + \boldsymbol{\nu}^\top (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}) \qquad (\boldsymbol{\nu} : \text{Lagrangian multiplier})$$

If $\boldsymbol{x}$ is the optimizer, then the KKT optimality conditions read

$$\begin{cases} \boldsymbol{0} = \nabla_{\boldsymbol{v}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{v}) \\ \boldsymbol{0} \in \underbrace{\partial_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{v})}_{\text{subdifferential}} \end{cases}$$

Karush-Kuhn-Tucker (KKT) conditions

Consider a convex problem

$$\begin{aligned} \text{minimize}_{\boldsymbol{x}} \quad & f(\boldsymbol{x}) \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} - \boldsymbol{y} = \boldsymbol{0} \end{aligned}$$

Lagrangian:

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\nu}) := f(\boldsymbol{x}) + \boldsymbol{\nu}^\top (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}) \qquad (\boldsymbol{\nu} : \text{Lagrangian multiplier})$$

If $\boldsymbol{x}$ is the optimizer, then the KKT optimality conditions read

$$\begin{cases} \boldsymbol{A}\boldsymbol{x} - \boldsymbol{y} = \boldsymbol{0} \\ \boldsymbol{0} \in \partial f(\boldsymbol{x}) + \boldsymbol{A}^\top \boldsymbol{\nu} \quad (\text{no constraint on } \boldsymbol{\nu}) \end{cases}$$

KKT condition for ℓ_1 minimization

$$\begin{array}{rl} \text{minimize}_{\boldsymbol{x}} & \|\boldsymbol{x}\|_1 \\ \text{s.t.} & \boldsymbol{A}\boldsymbol{x} - \boldsymbol{y} = \boldsymbol{0} \end{array}$$

If $\boldsymbol{x}$ is the optimizer, then KKT optimality condition reads

$$\begin{cases} \boldsymbol{A}\boldsymbol{x} - \boldsymbol{y} = \boldsymbol{0}, & \text{(naturally satisfied as } \boldsymbol{x} \text{ is the truth)} \\ \boldsymbol{0} \in \partial \|\boldsymbol{x}\|_1 + \boldsymbol{A}^\top \boldsymbol{\nu} & \text{(no constraint on } \boldsymbol{\nu}\text{)} \end{cases}$$

$$\Longleftrightarrow \quad \exists \boldsymbol{u} \in \text{range}(\boldsymbol{A}^\top) \quad \text{s.t.} \quad \underbrace{\begin{cases} u_i = \text{sign}(x_i), & \text{if } x_i \neq 0 \\ u_i \in [-1, 1], & \text{else} \end{cases}}_{\text{subgradient of } \|\boldsymbol{x}\|_1}$$

Depends only on the signs of x_i's, irrespective of their magnitudes

Uniqueness

Theorem 9.8 (A sufficient — and almost necessary — condition)

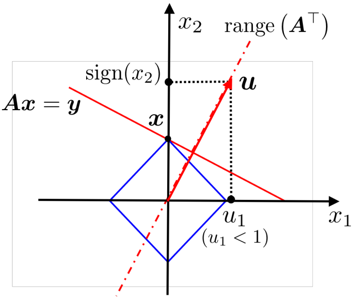
Let $T := \mathsf{supp}(\boldsymbol{x})$. Suppose $\boldsymbol{A}_T$ has full rank. If

$$\exists \boldsymbol{u} = \boldsymbol{A}^\top \boldsymbol{\nu} \text{ for some } \boldsymbol{\nu} \in \mathbb{R}^n \quad \text{s.t.} \quad \begin{cases} u_i & = \mathsf{sign}(x_i), \quad \text{if } x_i \neq 0 \\ u_i & \in (-1, 1), \quad \text{else} \end{cases},$$

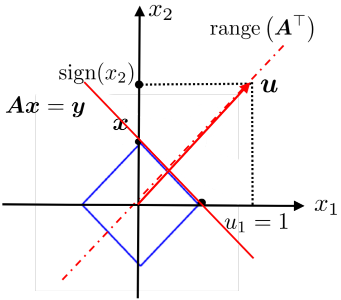
then $\boldsymbol{x}$ is the unique solution to ℓ_1 minimization

- Only slightly stronger than KKT!
- $\boldsymbol{\nu}$ is said to be a dual certificate
  - recall that $\boldsymbol{\nu}$ is the Lagrangian multiplier
- *Finding $\boldsymbol{\nu}$ comes down to solving another convex problem*

Geometric interpretation of the dual certificate



When $|u_1| < 1$, solution is unique



When $|u_1| = 1$, solution is non-unique

When we are able to find $\boldsymbol{u} \in \text{range}(\boldsymbol{A}^\top)$ s.t. $u_2 = \text{sign}(x_2)$ and $|u_1| < 1$, then $\boldsymbol{x}$ (with $x_1 = 0$) is the unique solution to ℓ_1-min

Proof of Theorem 9.8

Suppose that $\boldsymbol{x}+\boldsymbol{h}$ is the optimizer. Let $\boldsymbol{w} \in \partial\|\boldsymbol{x}\|_1$ be

$$\begin{cases} w_i = \mathsf{sign}(x_i), & \text{if } i \in T \text{ (support of } \boldsymbol{x}\text{)}; \\ w_i = \mathsf{sign}(h_i), & \text{else.} \end{cases}$$

If $\boldsymbol{x}+\boldsymbol{h}$ obeys $\boldsymbol{h}_{T^c} \neq \boldsymbol{0}$, then

$$\begin{aligned} \|\boldsymbol{x}\|_1 \geq \|\boldsymbol{x}+\boldsymbol{h}\|_1 &\overset{\text{by convexity}}{\geq} \|\boldsymbol{x}\|_1 + \langle \boldsymbol{w}, \boldsymbol{h} \rangle = \|\boldsymbol{x}\|_1 + \langle \boldsymbol{u}, \boldsymbol{h} \rangle + \langle \boldsymbol{w}-\boldsymbol{u}, \boldsymbol{h} \rangle \\ &= \|\boldsymbol{x}\|_1 + \langle \underbrace{\boldsymbol{A}^\top \boldsymbol{\nu}}_{\text{assumption on } \boldsymbol{u}}, \boldsymbol{h} \rangle + \sum_{i \notin T} (\text{sign}(h_i) h_i - u_i h_i) \\ &= \|\boldsymbol{x}\|_1 + \langle \boldsymbol{\nu}, \underbrace{\boldsymbol{A}\boldsymbol{h}}_{=\boldsymbol{0} \text{ (feasibility)}} \rangle + \sum_{i \notin T} (|h_i| - u_i h_i) \\ &\geq \|\boldsymbol{x}\|_1 + \sum\nolimits_{i \notin T} (1 - |u_i|)\, |h_i| \; > \; \|\boldsymbol{x}\|_1, \end{aligned}$$

resulting in contradiction. Therefore, $\boldsymbol{h}_{T^c} = \boldsymbol{0}$.

Proof of Theorem 9.8 (cont.)

Further, $\boldsymbol{h}_{T^c} = \boldsymbol{0}$ and $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}_T \boldsymbol{x}_T = \boldsymbol{y}$ imply that $\boldsymbol{A}_T(\boldsymbol{x}_T + \boldsymbol{h}_T) = \boldsymbol{y}$, and hence

$$\boldsymbol{A}_T \boldsymbol{h}_T = \boldsymbol{0}$$

From left-invertibility of $\boldsymbol{A}_T$, one must have $\boldsymbol{h}_T = \boldsymbol{0}$.

As a result, $\boldsymbol{h} = \boldsymbol{h}_T + \boldsymbol{h}_{T^c} = \boldsymbol{0}$. This concludes the proof.

Constructing dual certificates under Gaussian design

We illustrate how to construct dual certificates for the following setup

- $\boldsymbol{x} \in \mathbb{R}^p$ is k-sparse
- Entries of $\boldsymbol{A} \in \mathbb{R}^{n \times p}$ are i.i.d. standard Gaussian
- The sample size n obeys

$$n \gtrsim k \log p$$

Constructing dual certificates under Gaussian design

$$
\begin{aligned}
\text{Find} \quad & \boldsymbol{\nu} \in \mathbb{R}^n \\
\text{s.t.} \quad & (\boldsymbol{A}^\top \boldsymbol{\nu})_T = \mathsf{sign}(\boldsymbol{x}_T) && (9.14) \\
& |(\boldsymbol{A}^\top \boldsymbol{\nu})_i| < 1, \qquad i \notin T && (9.15)
\end{aligned}
$$

Constructing dual certificates under Gaussian design

Step 1: propose a $\boldsymbol{\nu}$ compatible with linear constraints (9.14). One candidate is the least squares (LS) solution:

$$\boldsymbol{\nu} = \boldsymbol{A}_T(\boldsymbol{A}_T^\top \boldsymbol{A}_T)^{-1}\mathsf{sign}(\boldsymbol{x}_T) \qquad \text{(explicit expression)}$$

- The LS solution minimizes $\|\boldsymbol{\nu}\|_2$, which will also be helpful when bounding $|(\boldsymbol{A}^\top \boldsymbol{\nu})_i|$
- From Lemma 9.6, $\boldsymbol{A}_T^\top \boldsymbol{A}_T$ is invertible with high prob. when $n \gtrsim k \log p$

Constructing dual certificates under Gaussian design

Step 2: verify (9.15), which amounts to controlling

$$\max_{i \notin T} \left| \left\langle \underbrace{\boldsymbol{A}_{:,i}}_{i\text{th column of } \boldsymbol{A}}, \underbrace{\boldsymbol{A}_T (\boldsymbol{A}_T^\top \boldsymbol{A}_T)^{-1} \mathsf{sign}(\boldsymbol{x}_T)}_{\boldsymbol{\nu}} \right\rangle \right|$$

- Since $\boldsymbol{A}_{:,i} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$ and $\boldsymbol{\nu}$ are independent for any $i \notin T$,

$$\max_{i \notin T} |\langle \boldsymbol{A}_{:,i}, \ \boldsymbol{\nu} \rangle| \lesssim \|\boldsymbol{\nu}\|_2 \sqrt{\log p} \qquad \text{with high prob.}$$

- $\|\boldsymbol{\nu}\|_2$ can be bounded by

$$\begin{aligned} \|\boldsymbol{\nu}\|_2 &\leq \|\boldsymbol{A}_T (\boldsymbol{A}_T^\top \boldsymbol{A}_T)^{-1}\| \cdot \|\mathsf{sgn}(\boldsymbol{x}_T)\|_2 \\ &= \|(\underbrace{\boldsymbol{A}_T^\top \boldsymbol{A}_T}_{\text{eigenvalues } \asymp n \text{ (Lemma 9.6)}})^{-1/2}\| \cdot \sqrt{k} \ \lesssim \ \sqrt{k/n} \end{aligned}$$

- When $n/(k \log p)$ is sufficiently large, $\max_{i \notin T} |\langle \boldsymbol{A}_{:,i}, \ \boldsymbol{\nu} \rangle| < 1$
- Exerciese: fill in missing details

More general random sampling

Consider a random design: each sampling vector $\boldsymbol{a}_i$ is *independently* drawn from a distribution F

$$\boldsymbol{a}_i \sim F$$

Incoherence sampling:

- *Isotropy:*

$$\mathbb{E}[\boldsymbol{a}\boldsymbol{a}^\top] = \boldsymbol{I}, \qquad \boldsymbol{a} \sim F$$

  - components of $\boldsymbol{a}$: (i) unit variance; (ii) uncorrelated

- *Incoherence:* let $\mu(F)$ be the smallest quantity s.t. for $\boldsymbol{a} \sim F$,

$$\|\boldsymbol{a}\|_\infty^2 \le \mu(F) \quad \text{with high prob.}$$

Incoherence

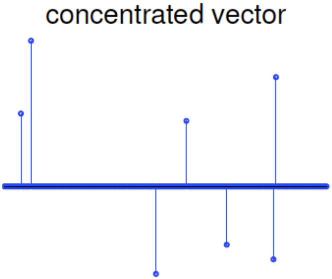
We want $\mu(F)$ (resp. $\boldsymbol{A}$) to be small (resp. dense)!

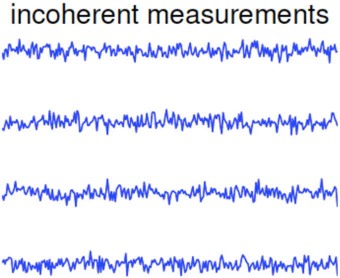
What happen if sampling vectors $\boldsymbol{a}_i$ are sparse?

- Example: $\boldsymbol{a}_i \sim$ Uniform$(\{\sqrt{p}\,\boldsymbol{e}_1, \cdots, \sqrt{p}\,\boldsymbol{e}_p\})$

$$\underbrace{\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}}_{\substack{\boldsymbol{y} \\ \text{no information}}} = \underbrace{\sqrt{p} \begin{bmatrix} 1 & & & & & & & & & \\ & & & & 1 & & & & & \\ & & & & & 1 & & & & \\ & & & & & & & & & 1 \end{bmatrix}}_{\boldsymbol{A}} \underbrace{\begin{bmatrix} 0 \\ 0 \\ 0 \\ 3 \\ 0 \\ 0 \\ 0 \\ 5 \\ 0 \\ 0 \end{bmatrix}}_{\boldsymbol{x}}$$

Incoherent random sampling





A general RIPless theory

Theorem 9.9 (Candès & Plan '11)

Suppose $\boldsymbol{x} \in \mathbb{R}^p$ *is* k*-sparse, and* $\boldsymbol{a}_i \overset{\text{ind.}}{\sim} F$ *is isotropic. Then* ℓ_1 *minimization is exact and unique with high prob., provided that*

$$n \gtrsim \mu(F) k \log p$$

- Near-optimal even for highly structured sampling matrices
- Proof idea: produce an (approximate) dual certificate by a clever *golfing scheme* pioneered by David Gross

Examples of incoherent sampling

- Binary sensing: $\mathbb{P}(a[i] = \pm 1) = 0.5$:

$$\mathbb{E}[\boldsymbol{a}\boldsymbol{a}^\top] = \boldsymbol{I}, \qquad \|\boldsymbol{a}\|_\infty^2 = 1, \qquad \mu = 1$$

$$\Longrightarrow \quad \ell_1\text{-min succeeds if } n \gtrsim k \log p$$

- Partial Fourier transform: pick a random frequency $f \sim \mathsf{Unif}\{0, \frac{1}{p}, \cdots, \frac{p-1}{p}\}$ or $f \sim \mathsf{Unif}[0,1]$ and set $a[i] = e^{j2\pi f i}$:

$$\mathbb{E}[\boldsymbol{a}\boldsymbol{a}^\top] = \boldsymbol{I}, \qquad \|\boldsymbol{a}\|_\infty^2 = 1, \qquad \mu = 1$$

$$\Longrightarrow \quad \ell_1\text{-min succeeds if } n \gtrsim k \log p$$

  - Improves upon the RIP-based result ($n \gtrsim k \log^4 p$)

Examples of incoherent sampling

- Random convolution matrices: rows of $\boldsymbol{A}$ are independently sampled from rows of

$$\boldsymbol{G} = \begin{bmatrix} g_0 & g_1 & g_2 & \cdots & g_{p-1} \\ g_{p-1} & g_0 & g_1 & \cdots & g_{p-2} \\ g_{p-2} & g_{p-1} & g_0 & \cdots & g_{p-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ g_1 & g_2 & g_3 & \cdots & g_0 \end{bmatrix}$$

  with $\mathbb{P}(g_i = \pm 1) = 0.5$. One has

$$\mathbb{E}[\boldsymbol{a}\boldsymbol{a}^\top] = \boldsymbol{I}, \qquad \|\boldsymbol{a}\|_\infty^2 = 1, \qquad \mu = 1$$

$$\Longrightarrow \quad \ell_1\text{-min succeeds if } n \gtrsim k \log p$$

  - Improves upon RIP-based result ($n \gtrsim k \log^4 p$)

A general scheme for dual construction (optional)

$$
\begin{aligned}
\text{Find} \quad & \boldsymbol{\nu} \in \mathbb{R}^n \\
\text{s.t.} \quad & \boldsymbol{A}_T^\top \boldsymbol{\nu} = \mathsf{sign}(\boldsymbol{x}_T) && (9.16) \\
& \|\boldsymbol{A}_{T^c}^\top \boldsymbol{\nu}\|_\infty < 1 && (9.17)
\end{aligned}
$$

A candidate: the least squares solution w.r.t. (9.16)

$$\boldsymbol{\nu} = \boldsymbol{A}_T(\boldsymbol{A}_T^\top \boldsymbol{A}_T)^{-1}\mathsf{sign}(\boldsymbol{x}_T) \qquad \text{(explicit expression)}$$

To verify (9.17), we need to control $\boldsymbol{A}_{T^c}^\top \boldsymbol{A}_T(\boldsymbol{A}_T^\top \boldsymbol{A}_T)^{-1}\mathsf{sign}(\boldsymbol{x}_T)$

- Issue 1: in general, $\boldsymbol{A}_{T^c}$ and $\boldsymbol{A}_T$ are dependent
- Issue 2: $(\boldsymbol{A}_T^\top \boldsymbol{A}_T)^{-1}$ is hard to deal with

A general scheme for dual construction (optional)

$$\begin{aligned} \text{Find} \quad & \boldsymbol{\nu} \in \mathbb{R}^n \\ \text{s.t.} \quad & \boldsymbol{A}_T^\top \boldsymbol{\nu} = \mathsf{sign}(\boldsymbol{x}_T) \qquad (9.16) \\ & \|\boldsymbol{A}_{T^c}^\top \boldsymbol{\nu}\|_\infty < 1 \qquad (9.17) \end{aligned}$$

Key idea 1: use iterative scheme (e.g. gradient descent) to solve minimize$_{\boldsymbol{\nu}}$ $\frac{1}{2}\|\boldsymbol{A}_T^\top \boldsymbol{\nu} - \mathsf{sign}(\boldsymbol{x}_T)\|_2^2$

for $t = 1, 2, \cdots$

$$\boldsymbol{\nu}^{(t)} = \boldsymbol{\nu}^{(t-1)} - \underbrace{\boldsymbol{A}_T \left(\boldsymbol{A}_T^\top \boldsymbol{\nu}^{(t-1)} - \mathrm{sign}(\boldsymbol{x}_T)\right)}_{\text{grad of } \frac{1}{2}\|\boldsymbol{A}_T^\top \boldsymbol{v} - \mathrm{sign}(\boldsymbol{x}_T)\|_2^2}$$

- Converges to a solution obeying (9.16); no inversion involved
- Issue: complicated dependency across iterations

Golfing scheme (Gross '11) (optional)

Key idea 2: sample splitting — use independent samples for each iteration to decouple statistical dependency

- Partition $\boldsymbol{A}$ into L row blocks $\underbrace{\boldsymbol{A}^{(1)} \in \mathbb{R}^{n_1 \times p}, \cdots, \boldsymbol{A}^{(L)} \in \mathbb{R}^{n_L \times p}}_{\text{independent}}$

- **for** $t = 1, 2, \cdots$ (stochastic gradient)

$$\boldsymbol{\nu}^{(t)} = \boldsymbol{\nu}^{(t-1)} - \underbrace{\mu_t \boldsymbol{A}_T^{(t)} \left(\boldsymbol{A}_T^{(t)\top} \boldsymbol{\nu}^{(t-1)} - \operatorname{sign}(\boldsymbol{x}_T) \right)}_{\in \mathbb{R}^{n_t} \text{ (but we need it in } \in \mathbb{R}^n)}$$

Golfing scheme (Gross '11) (optional)

Key idea 2: sample splitting — use independent samples for each iteration to decouple statistical dependency

- Partition $\boldsymbol{A}$ into L row blocks $\underbrace{\boldsymbol{A}^{(1)} \in \mathbb{R}^{n_1 \times p}, \cdots, \boldsymbol{A}^{(L)} \in \mathbb{R}^{n_L \times p}}_{\text{independent}}$

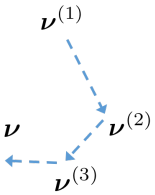
- **for** $t = 1, 2, \cdots$ (stochastic gradient)

$$\boldsymbol{\nu}^{(t)} = \boldsymbol{\nu}^{(t-1)} - \mu_t \widetilde{\boldsymbol{A}}_T^{(t)} \left(\boldsymbol{A}_T^{(t)\top} \boldsymbol{\nu}^{(t-1)} - \operatorname{sign}(\boldsymbol{x}_T) \right)$$

where $\widetilde{\boldsymbol{A}}^{(t)} = \begin{bmatrix} \mathbf{0} \\ \boldsymbol{A}^{(t)} \\ \mathbf{0} \end{bmatrix} \in \mathbb{R}^{n \times p}$ is obtained by zero-padding

Golfing scheme (Gross '11) (optional)

$$\boldsymbol{\nu}^{(t)} = \boldsymbol{\nu}^{(t-1)} - \mu_t \widetilde{\boldsymbol{A}}_T^{(t)} \Big(\boldsymbol{A}_T^{(t)\top} \underbrace{\boldsymbol{\nu}^{(t-1)}}_{\text{depends only on } \boldsymbol{A}^{(1)}, \cdots, \boldsymbol{A}^{(t-1)}} - \mathsf{sign}(\boldsymbol{x}_T) \Big)$$



- Statistical independence (fresh samples) across iterations, which significantly simplifies analysis
- Each iteration brings us closer to the target (like each golf shot brings us closer to the hole)

Reference

- "*Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information*," E. Candès, J. Romberg, and T. Tao, *IEEE Transactions on Information Theory*, 2006.
- "*Compressed sensing*," D. Donoho, *IEEE Transactions on Information Theory*, 2006
- "*Near-optimal signal recovery from random projections: Universal encoding strategies?*," E. Candès, and T. Tao, *IEEE Transactions on Information Theory*, 2006.
- "*Lecture notes, Advanced topics in signal processing (ECE 8201)*," Y. Chi, 2015.
- "*A mathematical introduction to compressive sensing*," S. Foucart and H. Rauhut, *Springer*, 2013.

Reference

- "*On sparse reconstruction from Fourier and Gaussian measurements*," M. Rudelson and R. Vershynin, *Communications on Pure and Applied Mathematics*, 2008.
- "*Decoding by linear programming*," E. Candès, and T. Tao, *IEEE Transactions on Information Theory*, 2006.
- "*The restricted isometry property and its implications for compressed sensing*," E. Candès, *Compte Rendus de l'Academie des Sciences*, 2008.
- "*Introduction to the non-asymptotic analysis of random matrices*," R. Vershynin, *Compressed Sensing: Theory and Applications*, 2010.
- "*Iterative hard thresholding for compressed sensing*," T. Blumensath, M. Davies, *Applied and computational harmonic analysis*, 2009.
- "*Recovering low-rank matrices from few coefficients in any basis*," D. Gross, *IEEE Transactions on Information Theory*, 2011.

Reference

- "*A probabilistic and RIPless theory of compressed sensing*," E. Candès and Y. Plan, *IEEE Transactions on Information Theory*, 2011.
- "*Suprema of chaos processes and the restricted isometry property*," F. Krahmer, S. Mendelson, and H. Rauhut, *Communications on Pure and Applied Mathematics*, 2014.
- "*High-dimensional data analysis with sparse models: Theory, algorithms, and applications*," J. Wright, Y. Ma, and A. Yang, 2018.
- "*Statistical learning with sparsity: the Lasso and generalizations*," T. Hastie, R. Tibshirani, and M. Wainwright, 2015.
- "*Statistical machine learning for high-dimensional data*,", J. Fan, R. Li, C. Zhang, H. Zou, 2018.