Recovery Theory

Fallstudien der mathematischen Modellbildung, Teil 2

20.10.2023 - 21.11.2023,

paul.catala@tum.de

- $y = Ax_0 + e$, $\|e\| \leqslant \delta$, $A \in \mathbb{R}^{m \times n}$ $m < n$. Assume x_0 is k-sparse, and consider the Lasso

$$x = \mathbf{argmin}_x \frac{1}{2}\|y - Ax\|^2 + \lambda\|x\|_1$$

- Questions: is $x = x_0$ when $\delta = 0$, $\lambda = 0$? how close is x from x_0?
  $\rightarrow$ depends on properties of A, and on k.

- $y = Ax_0 + e$, $\|e\| \leqslant \delta$, $A \in \mathbb{R}^{m \times n}$ $m < n$. Assume x_0 is k-sparse, and consider the Lasso

$$x = \textbf{argmin}_x \frac{1}{2}\|y - Ax\|^2 + \lambda\|x\|_1$$

- Questions: is $x = x_0$ when $\delta = 0$, $\lambda = 0$? how close is x from x_0?
  $\rightarrow$ depends on properties of A, and on k.
- Noiseless, constrained problem basis pursuit

$$\textbf{min}\ \|x\|_1 \quad \text{s.t.} \quad Ax = y$$

We will always assume that the problem is feasible, *i.e.* there exists $x_0 \in \mathbb{R}^n$ such that $Ax_0 = y$

Dual Certificate

- For f_0 is differentiable, consider

$$\min f_0(x) \quad \text{s.t.} \quad Ax = y$$

Is x_0 a minimizer? The (sufficient) KKT conditions for optimality give

$$\begin{cases} \nabla f_0(x) - A^\top \nu = 0 \\ Ax - y = 0 \end{cases}$$

Therefore, if there exists ν such that $\nabla f_0(x_0) = A^\top \nu$, then x_0 is a solution. $A^\top \nu$ is called a *dual certificate*. In that case, ν is also a solution of the dual problem. The dual certificate is not necessary unique.

Dual Certificate

- For f_0 is differentiable, consider

$$\min f_0(x) \quad \text{s.t.} \quad Ax = y$$

Is x_0 a minimizer? The (sufficient) KKT conditions for optimality give

$$\begin{cases} \nabla f_0(x) - A^\top \nu = 0 \\ Ax - y = 0 \end{cases}$$

Therefore, if there exists ν such that $\nabla f_0(x_0) = A^\top \nu$, then x_0 is a solution. $A^\top \nu$ is called a *dual certificate*. In that case, ν is also a solution of the dual problem. The dual certificate is not necessary unique.

- For basis pursuit, KKT at x_0 gives

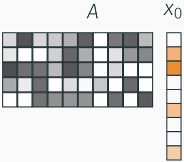
$$\begin{cases} A^\top \nu \in \partial \|x\|_1 \\ Ax - y = 0 \end{cases}$$

They are also sufficient: if $A^\top \nu \in \partial \|x_0\|_1$, then for all x

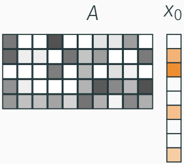
$$\|x\|_1 \geqslant \|x_0\|_1 + \langle A^\top \nu, x - x_0 \rangle \iff \|x\|_1 \geqslant \|x_0\|_1 + \langle \nu, Ax - Ax_0 \rangle$$

which directly shows that $\|x\|_1 \geqslant \|x_0\|_1$ if x is feasible.

- Let $I = \mathrm{Supp}\, x_0$, I^c its complement, x_I the restriction of $x \in \mathbb{R}^n$ to I and A_I the column restriction of A to I



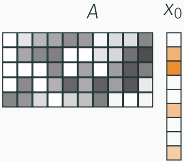
- Let $I = \operatorname{Supp} x_0$, I^c its complement, x_I the restriction of $x \in \mathbb{R}^n$ to I and A_I the column restriction of A to I



- $\eta = A^\top \nu \in \partial \|x_0\|_1$ when
  (i) $\eta_I = \mathbf{sign}(x_0)_I$
  (ii) $\|\eta_{I^c}\|_\infty \leqslant 1$

Sparse Recovery

- Let $I = \mathrm{Supp}\, x_0$, I^c its complement, x_I the restriction of $x \in \mathbb{R}^n$ to I and A_I the column restriction of A to I



- $\eta = A^\top \nu \in \partial \|x_0\|_1$ when
  (i) $\eta_I = \mathbf{sign}(x_0)_I$
  (ii) $\|\eta_{I^c}\|_\infty \leqslant 1$

- ***Theorem (Fuchs, 2004).*** x_0 is the unique minimizer of

$$\min \|x\|_1 \quad \text{s.t.} \quad Ax = Ax_0$$

if there exists $\eta \in \mathrm{Ran}\, A^\top$ such that

(i) $\eta_I = \mathbf{sign}(x_0)_I$

(ii) $\|\eta_{I^c}\|_\infty < 1$ ("non-degeneracy")

and furthermore A is injective on I.

Proof

Let $x \in \mathbb{R}^n$ such that $Ax = Ax_0$. We have

$$\begin{aligned} \|x_0\|_1 &= \langle \eta, x_0 \rangle \\ &= \langle \eta, x \rangle \\ &= \langle \eta_I, x_I \rangle + \langle \eta_{I^c}, x_{I^c} \rangle \\ &\leqslant \|x_I\|_1 + \|\eta_{I^c}\|_\infty \|x_{I^c}\|_1 \end{aligned}$$

If $x_{I^c} \neq 0$, then $\|x_0\|_1 < \|x\|_1$

If $x_{I^c} = 0$ (*i.e.* x has the same support as x_0), then $A(x - x_0) = A_I(x_I - (x_0)_I) = 0$, so by injectivity $x_I = (x_0)_I$, and therefore $x = x_0$.

Hence, $\|x_0\|_1 < \|x\|_1$ unless $x = x_0$.

$\square$

How to Build η

- Classical way: look for minimal norm certificate

$$\min \|\nu\|_2^2 \quad \text{s.t.} \quad (A^\top \nu)_I = \mathbf{sign}(x_0)_I$$

- Note that $(A^\top \nu)_I = \mathbf{sign}(x_0)_I \iff (A_I)^\top \nu = \mathbf{sign}(x_0)_I$: underdetermined system for sufficiently sparse vector
  The minimal norm solution is given by the pseudo-inverse, *i.e.*

$$\eta_0 = A^\top (A_I^\top)^\dagger \, \mathbf{sign}(x_0)_I$$

  Since A_I is injective, $\eta_I = A_I^\top (A_I)^\dagger \, \mathbf{sign}(x_0)_I = \mathbf{sign}(x_0)_I$
- It remains to show that $\|\eta_{I^c}\|_\infty < 1$

Example: Gaussian Random Matrix

- Assume $A_{ij} \sim \mathcal{N}(0,1)$ i.i.d.. This means in particular

$$\mathbb{P}(|A_{ij}| \geqslant u) \propto 2\int_u^\infty \exp\left(-\frac{t^2}{2}\right)\mathrm{d}t \leqslant \exp\left(-\frac{u^2}{2}\right)$$

$$\int_u^\infty e^{-t^2/2}\mathrm{d}t = \int_0^\infty e^{-(t+u)^2/2}\mathrm{d}t = e^{-u^2/2}\int_0^\infty e^{-tu}e^{-t^2/2}\mathrm{d}t$$

$$\begin{cases} \leqslant e^{-u^2/2}\int_0^\infty e^{-t^2/2}\mathrm{d}t = \sqrt{\frac{\pi}{2}}e^{-u^2/2} \\ \leqslant e^{-u^2/2}\int_0^\infty e^{-tu}\mathrm{d}t \leqslant \frac{1}{u}e^{-u^2/2} \end{cases}$$

Example: Gaussian Random Matrix

- Assume $A_{ij} \sim \mathcal{N}(0,1)$ i.i.d.. This means in particular

$$\mathbb{P}(|A_{ij}| \geqslant u) \propto 2 \int_u^\infty \exp\left(-\frac{t^2}{2}\right) \mathrm{d}t \leqslant \exp\left(-\frac{u^2}{2}\right)$$

- Suppose $A \in \mathbb{R}^{m\times n}$ (m is the number of measurements), $A_I \in \mathbb{R}^{m\times k}$ and $x_0 \in \mathbb{R}^n$ with $\|x_0\|_0 = k$, $k < m$. The entries of the dual certificate are

$$\eta_i = \sum_{j=1}^{m} A_{ji} \left[(A_I^\top)^\dagger \operatorname{sign}(x_0)_I\right]_j =: \sum A_{ji} cj$$

The goal is to find conditions on m, n, k such that $|\eta_i| < 1$, $i \notin I$ with high probability.

Example: Gaussian Random Matrix

- Assume $A_{ij} \sim \mathcal{N}(0,1)$ i.i.d.. This means in particular

$$\mathbb{P}(|A_{ij}| \geqslant u) \propto 2 \int_u^\infty \exp\left(-\frac{t^2}{2}\right) \mathrm{d}t \leqslant \exp\left(-\frac{u^2}{2}\right)$$

- Suppose $A \in \mathbb{R}^{m \times n}$ (m is the number of measurements), $A_I \in \mathbb{R}^{m \times k}$ and $x_0 \in \mathbb{R}^n$ with $\|x_0\|_0 = k$, $k < m$. The entries of the dual certificate are

$$\eta_i = \sum_{j=1}^{m} A_{ji} \left[(A_I^\top)^\dagger \operatorname{sign}(x_0)_I\right]_j =: \sum A_{ji} cj$$

The goal is to find conditions on m, n, k such that $|\eta_i| < 1$, $i \notin I$ with high probability.

- ***Lemma (1).*** If $X_i \sim \mathcal{N}(0,1)$ i.i.d., then $\mathbb{P}(|\sum c_i X_i| \geqslant u) \leqslant \exp\left(-\frac{u^2}{2\|c\|_2^2}\right)$

Lemma (2, admitted). If $A_{ij} \sim \mathcal{N}(0,1)$ i.i.d., $i = 1, \ldots, m, j = 1, \ldots, k$ $k < m$, then

$$\mathbb{P}\left(\sigma_{\max}\left(\frac{A}{\sqrt{m}}\right) \geqslant 1 + \sqrt{\frac{k}{m}} + t\right) \leqslant \exp\left(-\frac{mt^2}{2}\right)$$

$$\mathbb{P}\left(\sigma_{\min}\left(\frac{A}{\sqrt{m}}\right) \leqslant 1 - \sqrt{\frac{k}{m}} + t\right) \leqslant \exp\left(-\frac{mt^2}{2}\right)$$

Failure Probability (1/2)

Estimate $P = \mathbb{P}(\exists i \notin I : |\eta_i| \geqslant 1)$. Consider the event E: $\|c\|_2 \leqslant \alpha$ for some $\alpha > 0$.

- Then

$$P = \mathbb{P}(\exists i \notin I : |\eta_i| \geqslant 1 | E)\mathbb{P}(E) + \mathbb{P}(\exists i \notin I : |\eta_i| \geqslant 1 | E^c)\mathbb{P}(E^c)$$
$$\leqslant \mathbb{P}(\exists i \notin I : |\eta_i| \geqslant 1 | E) + \mathbb{P}(E^c)$$

Failure Probability (1/2)

Estimate $P = \mathbb{P}(\exists i \notin I : |\eta_i| \geqslant 1)$. Consider the event E: $\|c\|_2 \leqslant \alpha$ for some $\alpha > 0$.

- Then
$$P = \mathbb{P}(\exists i \notin I : |\eta_i| \geqslant 1|E)\mathbb{P}(E) + \mathbb{P}(\exists i \notin I : |\eta_i| \geqslant 1|E^c)\mathbb{P}(E^c)$$
$$\leqslant \mathbb{P}(\exists i \notin I : |\eta_i| \geqslant 1|E) + \mathbb{P}(E^c)$$
- For a fixed $i \notin I$, using Lemma 1,
$$\mathbb{P}(|\sum A_{ij}c_j| \geqslant 1|E) \leqslant \exp\left(-\frac{1}{2\|c\|_2^2}\right) \leqslant \exp\left(-\frac{1}{2\alpha^2}\right),$$
and since there are $(n-k)$ possible values of $i \notin I$, we obtain
$$\mathbb{P}(\exists i \notin I : |\eta_i| \geqslant 1|E) \leqslant (n-k)\exp\left(-\frac{1}{2\alpha^2}\right)$$

Failure Probability (1/2)

Estimate $P = \mathbb{P}(\exists i \notin I : |\eta_i| \geqslant 1)$. Consider the event E: $\|c\|_2 \leqslant \alpha$ for some $\alpha > 0$.

- Then
$$P = \mathbb{P}(\exists i \notin I : |\eta_i| \geqslant 1 | E)\mathbb{P}(E) + \mathbb{P}(\exists i \notin I : |\eta_i| \geqslant 1 | E^c)\mathbb{P}(E^c)$$
$$\leqslant \mathbb{P}(\exists i \notin I : |\eta_i| \geqslant 1 | E) + \mathbb{P}(E^c)$$
- For a fixed $i \notin I$, using Lemma 1,
$$\mathbb{P}(|\sum A_{ij} c_j| \geqslant 1 | E) \leqslant \exp\left(-\frac{1}{2\|c\|_2^2}\right) \leqslant \exp\left(-\frac{1}{2\alpha^2}\right),$$
and since there are $(n-k)$ possible values of $i \notin I$, we obtain
$$\mathbb{P}(\exists i \notin I : |\eta_i| \geqslant 1 | E) \leqslant (n-k) \exp\left(-\frac{1}{2\alpha^2}\right)$$
- Assume $\|c\|_2 \leqslant B$. Then $\mathbb{P}(E^c) = \mathbb{P}(\|c\|_2 \geqslant \alpha) \leqslant \mathbb{P}(\alpha \leqslant B)$. But
$$\|c\|_2 = \|(A_I)^\dagger \,\mathbf{sign}(x_0)_I\|_2 \leqslant (\sigma_{\min}(A_I))^{-1}\sqrt{k} = \sigma_{\min}^{-1}(A_I/\sqrt{k})$$

Failure Probability (1/2)

Estimate $P = \mathbb{P}(\exists i \notin I : |\eta_i| \geqslant 1)$. Consider the event E: $\|c\|_2 \leqslant \alpha$ for some $\alpha > 0$.

- Then
$$P = \mathbb{P}(\exists i \notin I : |\eta_i| \geqslant 1 | E)\mathbb{P}(E) + \mathbb{P}(\exists i \notin I : |\eta_i| \geqslant 1 | E^c)\mathbb{P}(E^c)$$
$$\leqslant \mathbb{P}(\exists i \notin I : |\eta_i| \geqslant 1 | E) + \mathbb{P}(E^c)$$
- For a fixed $i \notin I$, using Lemma 1,
$$\mathbb{P}(|\sum A_{ij} c_j| \geqslant 1 | E) \leqslant \exp\left(-\frac{1}{2\|c\|_2^2}\right) \leqslant \exp\left(-\frac{1}{2\alpha^2}\right),$$
and since there are $(n-k)$ possible values of $i \notin I$, we obtain
$$\mathbb{P}(\exists i \notin I : |\eta_i| \geqslant 1 | E) \leqslant (n-k)\exp\left(-\frac{1}{2\alpha^2}\right)$$
- Assume $\|c\|_2 \leqslant B$. Then $\mathbb{P}(E^c) = \mathbb{P}(\|c\|_2 \geqslant \alpha) \leqslant \mathbb{P}(\alpha \leqslant B)$. But
$$\|c\|_2 = \|(A_I)^\dagger \operatorname{sign}(x_0)_I\|_2 \leqslant (\sigma_{\min}(A_I))^{-1}\sqrt{k} = \sigma_{\min}^{-1}(A_I/\sqrt{k})$$
and hence
$$\mathbb{P}(E^c) \leqslant \mathbb{P}(\sigma_{\min}^{-1}(A_I/\sqrt{k}) \geqslant \alpha) = \mathbb{P}\left(\sigma_{\min}(A_I/\sqrt{m}) \leqslant \frac{1}{\alpha}\sqrt{\frac{k}{m}}\right)$$
$$\leqslant \exp\left(-\frac{m}{2}\left(1 - (1 + \frac{1}{\alpha})\sqrt{\frac{k}{m}}\right)^2\right)$$

Let $\varepsilon > 0$, we want (a) + (b) $\leqslant \varepsilon$

- Set

$$\exp\left(-\frac{1}{2\alpha^2}\right) = \frac{\varepsilon}{n}$$

i.e. $\alpha^{-2} = 2\ln(n/\varepsilon)$, then $(a) = (n-k)\varepsilon/n \leqslant \frac{n-1}{n}\varepsilon$

Let $\varepsilon > 0$, we want (a) + (b) $\leqslant \varepsilon$

- Set
$$\exp\left(-\frac{1}{2\alpha^2}\right) = \frac{\varepsilon}{n}$$
i.e. $\alpha^{-2} = 2\ln(n/\varepsilon)$, then $(a) = (n-k)\varepsilon/n \leqslant \frac{n-1}{n}\varepsilon$
- We would like $(b) \leqslant \varepsilon/n$

$$\iff m\left(1\left(1+\frac{1}{\alpha}\right)\sqrt{\frac{k}{m}}\right)^2 \geqslant 2\ln\left(\frac{n}{\varepsilon}\right) = \alpha^{-2}$$
$$\iff \sqrt{m} - \left(1+\frac{1}{\alpha}\right)\sqrt{k} \geqslant \alpha^{-1}$$
$$\iff \sqrt{m} \geqslant \alpha^{-1} + (1+\alpha^{-1})\sqrt{k}$$
$$\iff \sqrt{m} \geqslant 3\alpha^{-1}\sqrt{k}$$
$$\iff m \geqslant 18k\ln\left(\frac{n}{\varepsilon}\right)$$
$$\iff \#\text{ measurements} \gtrsim \text{sparsity} \times \ln(n)$$

Theorem (Donoho; Candès-Romberg-Tao). Let $x_0 \in \mathbb{R}^n$ be k-sparse, $y = Ax_0$ for $A \in \mathbb{R}^{m \times n}$, $A_{ij} \sim \mathcal{N}(0,1)$ i.i.d. Then

$$\min \|x\|_1 \quad \text{s.t.} \quad Ax = y$$

recovers x_0 with probability $\geqslant 1 - \varepsilon$, provided

$$m \geqslant Ck \ln\left(\frac{n}{\varepsilon}\right)$$

Remark (Tighter analysis). A sharper condition can be derived

$$m \geqslant 2k \ln\left(\frac{en}{k}\right) - \ln \varepsilon$$

in general for sub-gaussian matrices.

- In practice: noise in measurements, and inexact knowledge of the sparsity

[1] *A Mathematical Introduction to Compressive Sensing, 1993*

- In practice: noise in measurements, and inexact knowledge of the sparsity
- ***Definition (Best k-term approximation).*** For $p > 0$, $x \in \mathbb{R}^n$,

$$\sigma_k(x)_p := \inf \{\|x - y\|_p \; ; \; y \text{ is } k\text{-sparse}\}$$

inf is realised by taking $y = x_I$ where I is the support of the k largest entries of x

- In practice: noise in measurements, and inexact knowledge of the sparsity
- *Definition (Best k-term approximation).* For $p > 0$, $x \in \mathbb{R}^n$,

$$\sigma_k(x)_p := \inf \{\|x - y\|_p \ ; \ y \text{ is } k\text{-sparse}\}$$

- Setting $y = Ax_0 + w$ with
  - $\|w\|_2 \leqslant \varepsilon$
  - injectivity of A controlled on I with $|I| = k$

  then any recovered x obeys

$$\|x - x_0\|_2 \lesssim \varepsilon + \sigma_k(x_0)_2$$

[1] *A Mathematical Introduction to Compressive Sensing, 1993*

- In practice: noise in measurements, and inexact knowledge of the sparsity
- ***Definition (Best k-term approximation).*** For $p > 0$, $x \in \mathbb{R}^n$,

$$\sigma_k(x)_p := \mathbf{inf}\,\{\|x - y\|_p \;;\; y \text{ is } k\text{-sparse}\}$$

- Setting $y = Ax_0 + w$ with
  - $\|w\|_2 \leqslant \varepsilon$
  - injectivity of A controlled on I with $|I| = k$

  then any recovered x obeys

$$\|x - x_0\|_2 \lesssim \varepsilon + \sigma_k(x_0)_2$$

- Let $A = \left[a_1, \ldots, a_n\right]$ and

$$x = \mathbf{argmin}\,\|x\|_1 \quad \text{s.t.} \quad \|Ax - y\| \leqslant \varepsilon$$

Proposition (Dual Certification for Inexact Data[1]). Let $x_0 \in \mathbb{R}^n$ with k largest components on I. Let $y = Ax_0 + w$ with $\|w\| \leqslant \varepsilon$. Assume $\exists \delta, \beta, \gamma, \theta, \tau$ ($\delta < 1$, $\theta < 1$) and $\eta = A^\top \nu$ such that

- $\|A_I^\top A_I - I_n\|_{2,2} \leqslant \delta$
- $\|\eta_I - \mathbf{sign}(x_0)_I\|_2 \leqslant \gamma$
- $\|\nu\|_2 \leqslant \tau\sqrt{k}$.
- $\mathbf{max}_{i \notin I} \|A_I^\top a_i\|_2 \leqslant \beta$
- $\|\eta_{I^c}\|_\infty \leqslant \theta$

If $\theta + \beta\gamma/(1-\delta) < 1$, then

$$\|x - x_0\|_2 \leqslant C_1 \sigma_k(x)_1 + C_2\sqrt{k}\varepsilon$$

[1] *A Mathematical Introduction to Compressive Sensing*, 1993

- General conditions on A to ensure that every k-sparse x_0 is the unique minimizer of

$$\min \|x\|_1 \quad \text{s.t.} \quad Ax = Ax_0?$$

.

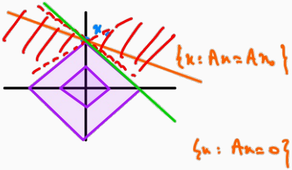
Uniform Recovery: Nullspace Property

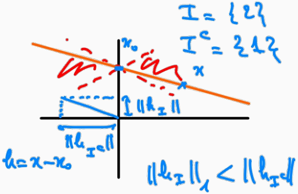
- General conditions on A to ensure that every k-sparse x_0 is the unique minimizer of

$$\min \|x\|_1 \quad \text{s.t.} \quad Ax = Ax_0?$$

.

- Recoverability of $x_0 \iff$ favorable orientation of $\operatorname{Ker} A$





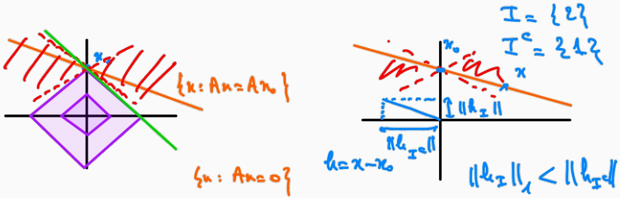
Uniform Recovery: Nullspace Property

- General conditions on A to ensure that every k-sparse x_0 is the unique minimizer of

$$\min \|x\|_1 \quad \text{s.t.} \quad Ax = Ax_0?$$

.

- Recoverability of x_0 $\iff$ favorable orientation of $\operatorname{Ker} A$



- ***Definition (NSP).*** A obeys the nullspace property relative to I if for all $h \in \operatorname{Ker} A \setminus 0$,

$$\|h_I\|_1 < \|h_{I^c}\|_1$$

it obeys the NSP of order k if it satisfies the NSP for every I such that $|I| = k$.

- ***Theorem (Uniform Recovery).*** Every k-sparse x_0 such that $Ax_0 = y$ is the unique solution of (Basis Pursuit) if and only if A obeys NSP(k).

- ***Definition (Robust NSP).*** A obeys the robust NSP of order k if $\exists 0 < \rho < 1, \tau > 0$, such that, for all $I : |I| = k$, for all $h \in \operatorname{Ker} A \setminus 0$

$$\|h_I\|_1 \leqslant \rho \|h_{I^c}\|_1 + \tau \|Ah\|_2$$

- ***Theorem (Uniform Robust Recovery).*** If A obeys the robust NSP(k), then for every k-sparse x_0 such that $Ax_0 = y$, any solution of (Basis Pursuit-ε) satisfies

$$\|x - x_0\|_1 \leqslant 2\frac{1+\rho}{1-\rho}\sigma_k(x_0)_1 + 4\frac{\tau}{1-\rho}\varepsilon$$

- Limitation of NSP: number of measurements (m) required has bad dependency on k (*e.g.* $m \sim k^2$ for $A = \begin{bmatrix} I & F \end{bmatrix}$)

- Limitation of NSP: number of measurements (m) required has bad dependency on k (*e.g.* $m \sim k^2$ for $A = \begin{bmatrix} I & F \end{bmatrix}$)
- Compressed Sensing: stronger conditions on A to ensure uniform recovery with fewer measurements ($m \sim k \log n$)

Definition (Restricted Isometry Property (RIP)). A is said to satisfy the RIP if there exists δ_k, such that for all k-sparse x,

$$(1 - \delta_k)\|x\|_2^2 \leqslant \|Ax\|^2 \leqslant (1 + \delta_k)\|x\|^2$$

$$\delta_k = \max_{|S|=k} \|A_S^\top A_S - I_k\|_{2,2}$$