Subspace embeddings

Daniel Hsu

COMS 4772

Supremum of simple stochastic processes

Recap: JL lemma

JL lemma. For any $\varepsilon \in (0, 1/2)$, point set $S \subset \mathbb{R}^d$ of cardinality $|S| = n$, and $k \in \mathbb{N}$ such that $k \geq \frac{16 \ln n}{\varepsilon^2}$, there exists a linear map $f \colon \mathbb{R}^d \to \mathbb{R}^k$ such that

$$(1-\varepsilon)\|\boldsymbol{x}-\boldsymbol{y}\|_2^2 \;\leq\; \|f(\boldsymbol{x})-f(\boldsymbol{y})\|_2^2 \;\leq\; (1+\varepsilon)\|\boldsymbol{x}-\boldsymbol{y}\|_2^2 \quad \text{for all } \boldsymbol{x}, \boldsymbol{y} \in S\,.$$

Main probabilistic lemma

$\exists$ random linear map $\boldsymbol{M} \colon \mathbb{R}^d \to \mathbb{R}^k$ such that, for any $\boldsymbol{u} \in S^{d-1}$,

$$\mathbb{P}\Big(\Big|\|\boldsymbol{M}\boldsymbol{u}\|_2^2 - 1\Big| > \varepsilon\Big) \;\leq\; 2\exp\Big(-\Omega(k\varepsilon^2)\Big)\,.$$

JL lemma is consequence of main probabilistic lemma as applied to collection $T \subset S^{d-1}$ of $|T| = \binom{n}{2}$ unit vectors (+ union bound):

$$\mathbb{P}\Big(\max_{\boldsymbol{u} \in T}\Big|\|\boldsymbol{M}\boldsymbol{u}\|_2^2 - 1\Big| > \varepsilon\Big) \;\leq\; |T| \cdot 2\exp\Big(-\Omega(k\varepsilon^2)\Big)\,.$$

Related question

For $T \subseteq S^{d-1}$, *expected* maximum deviation

$$\mathbb{E}\max_{\boldsymbol{u} \in T}\Big|\|\boldsymbol{M}\boldsymbol{u}\|_2^2 - 1\Big| \;\leq\; ?$$

General questions

For arbitrary collection of zero-mean random variables $\{X_t : t \in T\}$:

$$\mathbb{E}\max_{t \in T} X_t \;\leq\; ?$$

$$\mathbb{E}\max_{t \in T} |X_t| \;\leq\; ?$$

Finite collections

Let $\{X_t : t \in T\}$ be a *finite* collection of v-subgaussian and mean-zero random variables. Then

$$\mathbb{E} \max_{t \in T} X_t \;\le\; \sqrt{2v \ln |T|}\,.$$

- Doesn't assume independence of $\{X_t : t \in T\}$.
  - (Independent case is the worst.)
- Get bound on $\mathbb{E} \max_{t \in T} |X_t|$ as corollary.
  - Apply result to collection

$$\{X_t : t \in T\} \cup \{-X_t : t \in T\}\,.$$

Proof

Starting point is identity from two invertible operations ($\lambda > 0$):

$$\mathbb{E} \max_{t \in T} X_t \;=\; \frac{1}{\lambda} \ln \exp\Big(\mathbb{E} \max_{t \in T} \lambda X_t\Big)$$

- Apply Jensen's inequality:

$$\le\; \frac{1}{\lambda} \ln \mathbb{E} \exp\Big(\max_{t \in T} \lambda X_t\Big) \;=\; \frac{1}{\lambda} \ln \mathbb{E}\Big(\max_{t \in T} \exp(\lambda X_t)\Big)$$

- Bound max with sum, and use linearity of expectation:

$$\le\; \frac{1}{\lambda} \ln \sum_{t \in T} \mathbb{E} \exp(\lambda X_t)$$

- Exploit v-subgaussian property:

$$\le\; \frac{1}{\lambda} \ln \sum_{t \in T} \exp\Big(v\lambda^2/2\Big) \;=\; \frac{\ln |T|}{\lambda} + \frac{v\lambda}{2}$$

- Choose appropriate λ to conclude. $\square$

Alternative proof

Integrate tail bound: for any non-negative random variable Y,

$$\mathbb{E}(Y) \;=\; \int_0^\infty \mathbb{P}(Y \geq y)\,\mathrm{d}y\,.$$

For $Y := \max_{t \in T} |X_t|$, gives same result up to constants.

Infinite collections

For *infinite* collection of zero-mean random variables $\{X_t : t \in T\}$:

$$\mathbb{E} \sup_{t \in T} X_t \;\leq\; ?$$

- In general, can go $\to \infty$.
- To bound, must exploit *correlations* among the X_t.
  - E.g., in $\left\{ \left| \|\boldsymbol{M}\boldsymbol{u}\|_2^2 - 1 \right| : \boldsymbol{u} \in T \right\}$ for $T \subseteq S^{d-1}$, the random variables for $\boldsymbol{u}$ and $\boldsymbol{u} + \boldsymbol{\delta}$, for small $\boldsymbol{\delta}$, are highly correlated.

Convex hulls of linear functionals

Let $T \subset \mathbb{R}^d$ be a finite set of vectors, and let $\boldsymbol{X}$ be a random vector in $\mathbb{R}^d$ such that $\langle \boldsymbol{w}, \boldsymbol{X} \rangle$ is v-subgaussian for every $\boldsymbol{w} \in T$. Then

$$\mathbb{E} \max_{\tilde{\boldsymbol{w}} \in \mathrm{conv}(T)} \langle \tilde{\boldsymbol{w}}, \boldsymbol{X} \rangle \ \leq \ \sqrt{2v \ln |T|} \,.$$

Proof:

- Write $\tilde{\boldsymbol{w}} \in \mathrm{conv}(T)$ as $\tilde{\boldsymbol{w}} = \sum_{\boldsymbol{w} \in T} p_{\boldsymbol{w}} \boldsymbol{w}$ for some $p_{\boldsymbol{w}} \geq 0$ that sum to one.
- Observe that

$$\langle \tilde{\boldsymbol{w}}, \boldsymbol{x} \rangle \ = \ \sum_{\boldsymbol{w} \in T} p_{\boldsymbol{w}} \langle \boldsymbol{w}, \boldsymbol{x} \rangle \ \leq \ \max_{\boldsymbol{w} \in T} \langle \boldsymbol{w}, \boldsymbol{x} \rangle \,.$$

- So max over $\tilde{\boldsymbol{w}} \in \mathrm{conv}(T)$ is at most max over $\boldsymbol{w} \in T$.
- Conclude by applying previous result for finite collections. $\square$

Euclidean norm

Let $\boldsymbol{X}$ be a random vector such that $\langle \boldsymbol{u}, \boldsymbol{X} \rangle$ is v-subgaussian for every $\boldsymbol{u} \in S^{d-1}$. Then

$$\mathbb{E} \|\boldsymbol{X}\|_2 \ = \ \mathbb{E} \max_{\boldsymbol{u} \in S^{d-1}} \langle \boldsymbol{u}, \boldsymbol{X} \rangle \ \leq \ 2\sqrt{2v \ln 5^d} \ = \ O\left(\sqrt{vd}\right) .$$

Key step of proof:

- For any $\varepsilon > 0$, there is a finite subset $\mathcal{N} \subset S^{d-1}$ of cardinality $|\mathcal{N}| \leq (1 + 2/\varepsilon)^d$ such that, for every $\boldsymbol{u} \in S^{d-1}$, there exists $\boldsymbol{u}_0 \in \mathcal{N}$ with
$$\|\boldsymbol{u} - \boldsymbol{u}_0\|_2 \ \leq \ \varepsilon \,.$$
  - Such a set $\mathcal{N}$ is called an ε-*net* for S^{d-1}.
  - We need a $1/2$-*net*, of cardinality at most 5^d.

Proof

- Write $\boldsymbol{u} \in S^{d-1}$ as

$$\boldsymbol{u} \;=\; \boldsymbol{u}_0 + \delta \boldsymbol{q}\,,$$

where $\boldsymbol{u}_0 \in \mathcal{N}$, $\boldsymbol{q} \in S^{d-1}$, $\delta \in [0, 1/2]$, so

$$\langle \boldsymbol{u}, \boldsymbol{X} \rangle \;=\; \langle \boldsymbol{u}_0, \boldsymbol{X} \rangle + \delta \langle \boldsymbol{q}, \boldsymbol{X} \rangle\,.$$

- Observe that

$$\begin{aligned} \max_{\boldsymbol{u} \in S^{d-1}} \langle \boldsymbol{u}, \boldsymbol{X} \rangle \;&\leq\; \max_{\boldsymbol{u}_0 \in \mathcal{N}} \langle \boldsymbol{u}_0, \boldsymbol{X} \rangle + \max_{\delta \in [0,1/2]} \max_{\boldsymbol{q} \in S^{d-1}} \delta \langle \boldsymbol{q}, \boldsymbol{X} \rangle \\ &\leq\; \max_{\boldsymbol{u}_0 \in \mathcal{N}} \langle \boldsymbol{u}_0, \boldsymbol{X} \rangle + \frac{1}{2} \max_{\boldsymbol{q} \in S^{d-1}} \langle \boldsymbol{q}, \boldsymbol{X} \rangle\,. \end{aligned}$$

- So max over S^{d-1} is at most twice max over $\mathcal{N}$.
- Conclude by applying previous result for finite collections. $\square$

ε-nets for unit sphere

There is an ε-net for S^{d-1} of cardinality at most $(1 + 2/\varepsilon)^d$.

Proof:

- Repeatedly select points from S^{d-1} so that each selected point has distance more than ε from all previously selected points.
- Equivalent: repeatedly select points from S^{d-1} as long as balls of radius $\varepsilon/2$, centered at selected points, are disjoint.
  - (Process must eventually stop.)
- When process stops, every $\boldsymbol{u} \in S^{d-1}$ is at distance at most ε from selected points.
  - I.e., selected points form an ε-net for S^{d-1}.
- If select N points, then the N balls of radius $\varepsilon/2$ are disjoint, and they are contained in a ball of radius $1 + \varepsilon/2$. So

$$N \,\mathrm{vol}((\varepsilon/2)B^d) \;\leq\; \mathrm{vol}((1 + \varepsilon/2)B^d)\,.$$

  - This implies $N \leq (1 + 2/\varepsilon)^d$. $\square$

Remarks

- All previous results also hold with random variables are (v, c)-*subexponential* (possibly with $c > 0$), with a slightly different bound: e.g.,

$$\mathbb{E} \max_{t \in T} X_t \;\leq\; \max\left\{ \sqrt{2v \ln |T|},\, 2c \ln |T| \right\}.$$

- Also easy to get probability tail bounds (rather than expectation bounds).

Subspace embeddings

Subspace JL lemma

Consider $k \times d$ random matrix $\boldsymbol{M}$ whose entries are iid $\mathrm{N}(0, 1/k)$. For a $W \subseteq \mathbb{R}^d$ be a subspace of dimension r,

$$\mathbb{E} \max_{\boldsymbol{u} \in S^{d-1} \cap W} \left| \|\boldsymbol{M}\boldsymbol{u}\|_2^2 - 1 \right| \;\leq\; O\left(\sqrt{\frac{r}{k}} + \frac{r}{k} \right) .$$

Bound is at most ε when $k \geq O\left(\frac{r}{\varepsilon^2}\right)$.

Implies existence of mapping $\boldsymbol{M} \colon \mathbb{R}^d \to \mathbb{R}^k$ that approximately preserves all distances between points in W.

Proof of subspace JL lemma

Let columns of $\boldsymbol{Q}$ be ONB for W. Then

$$\begin{aligned} \max_{\boldsymbol{u} \in S^{d-1} \cap W} \left| \|\boldsymbol{M}\boldsymbol{u}\|_2^2 - 1 \right| \;&=\; \max_{\boldsymbol{u} \in S^{r-1}} \left| \boldsymbol{u}^\top \boldsymbol{Q}^\top (\boldsymbol{M}^\top \boldsymbol{M} - \boldsymbol{I}) \boldsymbol{Q} \boldsymbol{u} \right| \\ &=\; \max_{\boldsymbol{u}, \boldsymbol{v} \in S^{r-1}} \boldsymbol{u}^\top \boldsymbol{Q}^\top (\boldsymbol{M}^\top \boldsymbol{M} - \boldsymbol{I}) \boldsymbol{Q} \boldsymbol{v} . \end{aligned}$$

Lemma. For any $\boldsymbol{u}, \boldsymbol{v} \in S^{r-1}$,

$$X_{\boldsymbol{u},\boldsymbol{v}} \;:=\; \boldsymbol{u}^\top \boldsymbol{Q}^\top (\boldsymbol{M}^\top \boldsymbol{M} - \boldsymbol{I}) \boldsymbol{Q} \boldsymbol{v}$$

is $(O(1/k), O(1/k))$-subexponential.

Proof of subspace JL lemma (continued)

For $\boldsymbol{u}, \boldsymbol{v} \in S^{r-1}$, $X_{\boldsymbol{u},\boldsymbol{v}} := \boldsymbol{u}^\top \boldsymbol{Q}^\top (\boldsymbol{M}^\top \boldsymbol{M} - \boldsymbol{I}) \boldsymbol{Q} \boldsymbol{v}$.

Let $\mathcal{N}$ be $1/4$-net for S^{r-1}.

- Write $\boldsymbol{u}, \boldsymbol{v} \in S^{r-1}$ as

$$\boldsymbol{u} \;=\; \boldsymbol{u}_0 + \varepsilon \boldsymbol{p} \,, \quad \boldsymbol{v} \;=\; \boldsymbol{v}_0 + \delta \boldsymbol{q} \,,$$

  where $\boldsymbol{u}_0, \boldsymbol{v}_0 \in \mathcal{N}$, $\boldsymbol{p}, \boldsymbol{q} \in S^{r-1}$ and $\varepsilon, \delta \in [0, 1/4]$, so

$$X_{\boldsymbol{u},\boldsymbol{v}} \;=\; X_{\boldsymbol{u}_0,\boldsymbol{v}_0} + \varepsilon X_{\boldsymbol{p},\boldsymbol{v}} + \delta X_{\boldsymbol{u}_0,\boldsymbol{q}} \,.$$

- Therefore

$$\max_{\boldsymbol{u},\boldsymbol{v} \in S^{r-1}} X_{\boldsymbol{u},\boldsymbol{v}} \;\leq\; \max_{\boldsymbol{u}_0,\boldsymbol{v}_0 \in \mathcal{N}} X_{\boldsymbol{u}_0,\boldsymbol{v}_0} + \frac{1}{2} \max_{\boldsymbol{p},\boldsymbol{q} \in S^{r-1}} X_{\boldsymbol{p},\boldsymbol{q}} \,,$$

  which implies

$$\max_{\boldsymbol{u},\boldsymbol{v} \in S^{r-1}} X_{\boldsymbol{u},\boldsymbol{v}} \;\leq\; 2 \max_{\boldsymbol{u}_0,\boldsymbol{v}_0 \in \mathcal{N}} X_{\boldsymbol{u}_0,\boldsymbol{v}_0} \,.$$

- Conclude by applying previous result for finite collections. □

Application to least squares

Big data least squares

- **Input**: matrix $\boldsymbol{A} \in \mathbb{R}^{n \times d}$, vector $\boldsymbol{b} \in \mathbb{R}^n$ ($n \gg d$).
- **Goal**: find $\boldsymbol{x} \in \mathbb{R}^d$ so as to (approx.) minimize $\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2$.

- Computation time: $O(nd^2)$.
- Can we speed this up?

Simple approach

- Pick $m \ll n$.
- Let $\boldsymbol{M}$ be random $m \times n$ matrix (e.g., entries iid $\mathrm{N}(0, 1/m)$, Fast JL Transform).
- Let $\widetilde{\boldsymbol{A}} := \boldsymbol{M}\boldsymbol{A}$ and $\tilde{\boldsymbol{b}} := \boldsymbol{M}\boldsymbol{b}$.
- Obtain solution $\hat{\boldsymbol{x}}$ to least squares problem on $(\widetilde{\boldsymbol{A}}, \tilde{\boldsymbol{b}})$.

Simple (somewhat loose) analysis

- Let W be subspace spanned by columns of $\boldsymbol{A}$ and $\boldsymbol{b}$.
  - Dimension is at most $d+1$.
- If $m \geq O(d/\varepsilon^2)$, then $\boldsymbol{M}$ is subspace embedding for W:

$$(1-\varepsilon)\|\boldsymbol{x}\|_2^2 \;\leq\; \|\boldsymbol{M}\boldsymbol{x}\|_2^2 \;\leq\; (1+\varepsilon)\|\boldsymbol{x}\|_2^2 \quad \text{for all } \boldsymbol{x} \in W\,.$$

- Let $\boldsymbol{x}_\star := \arg\min_{\boldsymbol{x}\in\mathbb{R}^d} \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2$.
-

$$\begin{aligned}
\|\boldsymbol{A}\hat{\boldsymbol{x}} - \boldsymbol{b}\|_2^2 &\leq \frac{1}{1-\varepsilon}\|\boldsymbol{M}(\boldsymbol{A}\hat{\boldsymbol{x}} - \boldsymbol{b})\|_2^2 \\
&\leq \frac{1}{1-\varepsilon}\|\boldsymbol{M}(\boldsymbol{A}\boldsymbol{x}_\star - \boldsymbol{b})\|_2^2 \\
&\leq \frac{1+\varepsilon}{1-\varepsilon}\|\boldsymbol{A}\boldsymbol{x}_\star - \boldsymbol{b}\|_2^2\,.
\end{aligned}$$

- Running time (using FJLT): $O\Big((m+n)d\log n + md^2\Big)$. $\square$

Another perspective: random sampling

- Pick random sample of $m \ll n$ of rows of $(\boldsymbol{A}, \boldsymbol{b})$; obtain solution $\hat{\boldsymbol{x}}$ for least squares problem on the sample.
- Hope $\hat{\boldsymbol{x}}$ is also good for the original problem.
- In statistics, this is the *random design* setting for regression.
  - Random sample of covariates $\widetilde{\boldsymbol{A}} \in \mathbb{R}^{m\times d}$ and responses $\tilde{\boldsymbol{b}} \in \mathbb{R}^m$ from full population $(\boldsymbol{A}, \boldsymbol{b})$.
  - Least squares solution $\hat{\boldsymbol{x}}$ on $(\widetilde{\boldsymbol{A}}, \tilde{\boldsymbol{b}})$ is *MLE* for linear regression coefficients under linear model with Gaussian noise.
  - Can also regard $\hat{\boldsymbol{x}}$ as *empirical risk minimizer* among all linear predictors under squared loss.

Simple random design analysis

- Let $\boldsymbol{x}_\star := \arg\min_{\boldsymbol{x} \in \mathbb{R}^d} \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2$.
- With high probability over choice of random sample,

$$\|\boldsymbol{A}\hat{\boldsymbol{x}} - \boldsymbol{b}\|_2^2 \;\leq\; \left(1 + O\left(\frac{\kappa}{m}\right)\right) \cdot \|\boldsymbol{A}\boldsymbol{x}_\star - \boldsymbol{b}\|_2^2$$

  (up to lower-order terms), where

$$\kappa \;:=\; n \cdot \max_{i \in [n]} \|(\boldsymbol{A}^\top \boldsymbol{A})^{-1/2} \boldsymbol{A}^\top \boldsymbol{e}_i\|_2^2$$

  and $\boldsymbol{e}_i$ is i-th coordinate basis vector.

- Write thin SVD of $\boldsymbol{A}$ as $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top$, where $\boldsymbol{U} \in \mathbb{R}^{n \times d}$. Then

$$(\boldsymbol{A}^\top \boldsymbol{A})^{-1/2} \boldsymbol{A}^\top \;=\; (\boldsymbol{V}\boldsymbol{S}^2\boldsymbol{V}^\top)^{-1/2} \boldsymbol{V}\boldsymbol{S}\boldsymbol{U}^\top \;=\; \boldsymbol{V}\boldsymbol{U}^\top .$$

- So $\kappa = n \cdot \max_{i \in [n]} \|\boldsymbol{U}^\top \boldsymbol{e}_i\|_2^2$.
  - $\|\boldsymbol{U}^\top \boldsymbol{e}_i\|_2^2$ is *statistical leverage score* for i-th row of $\boldsymbol{A}$: measures how much "influence" i-th row has on least squares solution.

Statistical leverage

- i-th *statistical leverage score*: $\ell_i := \|\boldsymbol{U}^\top \boldsymbol{e}_i\|_2^2$, where $\boldsymbol{U} \in \mathbb{R}^{n \times d}$ is matrix of left singular vectors of $\boldsymbol{A}$.
- Two extreme cases:

$$\boldsymbol{U} \;=\; \begin{bmatrix} \boldsymbol{I}_{d \times d} \\ \boldsymbol{0}_{(n-d) \times d} \end{bmatrix} \qquad \Rightarrow \qquad n \cdot \max_{i \in [n]} \ell_i \;=\; n \,.$$

$$\boldsymbol{U} \;=\; \frac{1}{\sqrt{n}} \begin{bmatrix} \boldsymbol{H}_n \boldsymbol{e}_1 & \boldsymbol{H}_n \boldsymbol{e}_2 & \cdots & \boldsymbol{H}_n \boldsymbol{e}_d \end{bmatrix} \;\Rightarrow\; n \cdot \max_{i \in [n]} \ell_i \;=\; d \,,$$

  where $\boldsymbol{H}_n$ is $n \times n$ Hadamard matrix.
- First case: first d rows are the only rows that matter.
- Second case: all n rows equally important.

Ensuring small statistical leverage

- To ensure situation is more like second case, apply random rotation (e.g., randomized Hadamard transform) to $\boldsymbol{A}$ and $\boldsymbol{b}$.
  - Randomly mixes up rows of $(\boldsymbol{A}, \boldsymbol{b})$ so no single row is (much) more important than another.
  - Get $n \cdot \max_{i \in [n]} \ell_i = O(d + \log n)$ with high probability.
- To get $1 + \varepsilon$ approximation ratio, i.e.,

$$\|\boldsymbol{A}\hat{\boldsymbol{x}} - \boldsymbol{b}\|_2^2 \;\leq\; (1 + \varepsilon) \cdot \|\boldsymbol{A}\boldsymbol{x}_\star - \boldsymbol{b}\|_2^2 \,,$$

  suffices to have

$$m \;\geq\; O\left(\frac{d + \log n}{\varepsilon}\right) .$$

Application to compressed sensing

Under-determined least squares

- **Input**: matrix $\boldsymbol{A} \in \mathbb{R}^{n \times d}$, vector $\boldsymbol{b} \in \mathbb{R}^n$ ($n \ll d$).
- **Goal**: find *sparsest* $\boldsymbol{x} \in \mathbb{R}^d$ so as to minimize $\|\boldsymbol{Ax} - \boldsymbol{b}\|_2^2$.

- NP-hard in general.
- Suppose $\boldsymbol{b} = \boldsymbol{A}\bar{\boldsymbol{x}}$ for some $\bar{\boldsymbol{x}} \in \mathbb{R}^d$ with $\operatorname{nnz}(\bar{\boldsymbol{x}}) \le k$.
  - I.e., $\bar{\boldsymbol{x}}$ is *k-sparse*.
  - Is $\bar{\boldsymbol{x}}$ the (unique) sparsest solution?
  - If so, how to find it?

Null space property

Lemma. Null space of $\boldsymbol{A}$ does not contain any non-zero $2k$-sparse vectors $\iff$ every k-sparse vector $\bar{\boldsymbol{x}} \in \mathbb{R}^d$ is the unique solution to $\boldsymbol{Ax} = \boldsymbol{A}\bar{\boldsymbol{x}}$.

- **Proof**. ($\Rightarrow$) Take any k-sparse vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ with $\boldsymbol{Ax} = \boldsymbol{Ay}$. Want to show $\boldsymbol{x} = \boldsymbol{y}$.
  - Then $\boldsymbol{x} - \boldsymbol{y}$ is $2k$-sparse, and $\boldsymbol{A}(\boldsymbol{x} - \boldsymbol{y}) = \boldsymbol{0}$.
  - By assumption, null space of $\boldsymbol{A}$ does not contain any non-zero $2k$-sparse vectors.
  - So $\boldsymbol{x} - \boldsymbol{y} = \boldsymbol{0}$, i.e., $\boldsymbol{x} = \boldsymbol{y}$.
- ($\Leftarrow$) Take any $2k$-sparse vector $\boldsymbol{z}$ in the null space of $\boldsymbol{A}$. Want to show $\boldsymbol{z} = \boldsymbol{0}$.
  - Write it as $\boldsymbol{z} = \boldsymbol{x} - \boldsymbol{y}$ for some k-sparse vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ with disjoint supports.
  - Then $\boldsymbol{A}(\boldsymbol{x} - \boldsymbol{y}) = \boldsymbol{0}$, and hence $\boldsymbol{x} = \boldsymbol{y}$ by assumption.
  - But $\boldsymbol{x}$ and $\boldsymbol{y}$ have disjoint support, so it must be that $\boldsymbol{x} = \boldsymbol{y} = \boldsymbol{0}$, so $\boldsymbol{z} = \boldsymbol{0}$. $\square$

Null space property from subspace embeddings

If $\boldsymbol{A}$ is $n \times d$ random matrix with iid $\mathrm{N}(0,1)$ entries, then under what conditions is there no non-zero $2k$-sparse vector in its null space?

- Want: for any $2k$-sparse vector $\boldsymbol{z}$, $\boldsymbol{Az} \neq \boldsymbol{0}$, i.e., $\|\boldsymbol{Az}\|_2^2 > 0$.
- Consider a particular choice $\mathcal{I} \subseteq [d]$ of $|\mathcal{I}| = 2k$ coordinates, and the corresponding subspace $W_{\mathcal{I}}$ spanned by $\{\boldsymbol{e}_i : i \in \mathcal{I}\}$.
  - Every $2k$-sparse $\boldsymbol{z}$ is in $W_{\mathcal{I}}$ for some $\mathcal{I}$.
- Sufficient for $\boldsymbol{A}$ to be $1/2$-subspace embedding for $W_{\mathcal{I}}$ for all $\mathcal{I}$:

$$\frac{1}{2}\|\boldsymbol{z}\|_2^2 \;\leq\; \|\boldsymbol{Az}\|_2^2 \;\leq\; \frac{3}{2}\|\boldsymbol{z}\|_2^2 \quad \text{for all } 2k\text{-sparse } \boldsymbol{z}\,.$$

Null space property from subspace embeddings (continued)

- Say $\boldsymbol{A}$ *fails* for $\mathcal{I}$ if it is not a $1/2$-subspace embedding for $W_{\mathcal{I}}$.
- Subspace JL lemma:

$$\mathbb{P}(\boldsymbol{A} \text{ fails for } \mathcal{I}) \;\leq\; 2^{O(k)} \exp(-\Omega(n))\,.$$

- Union bound over all choices of $\mathcal{I}$ with $|\mathcal{I}| = 2k$:

$$\mathbb{P}(\boldsymbol{A} \text{ fails for some } \mathcal{I}) \;\leq\; \binom{d}{2k} 2^{O(k)} \exp(-\Omega(n))\,.$$

- To ensure this is, say, at most $1/2$, just need

$$n \;\geq\; O\left(k + \log\binom{d}{2k}\right) \;=\; O(k + k\log(d/k))\,.$$

Restricted isometry property

(ℓ, δ)-*restricted isometry property (RIP)*:

$$(1-\delta)\|\boldsymbol{z}\|_2^2 \;\leq\; \|\boldsymbol{A}\boldsymbol{z}\|_2^2 \;\leq\; (1+\delta)\|\boldsymbol{z}\|_2^2 \quad \text{for all } \ell\text{-sparse } \boldsymbol{z}\,.$$

- Many algorithms can recover unique sparsest solution under RIP (with $\ell = O(k)$ and $\delta = \Omega(1)$).
  - E.g., Basis pursuit, Lasso, orthogonal matching pursuit.