Subspace embeddings

Daniel Hsu

COMS 4772

Supremum of simple stochastic processes
**Recap: JL lemma**

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

\[
(1-\epsilon)\|x-y\|_2^2 \leq \|f(x)-f(y)\|_2^2 \leq (1+\epsilon)\|x-y\|_2^2 \quad \text{for all } x, y \in S.
\]

**Main probabilistic lemma**

\( \exists \) random linear map \( M : \mathbb{R}^d \rightarrow \mathbb{R}^k \) such that, for any \( u \in S^{d-1} \),

\[
P\left(\|Mu\|_2^2 - 1 > \epsilon\right) \leq 2 \exp\left(-\Omega(\epsilon^2 k^2)\right).
\]

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):

\[
P\left(\max_{u \in T} \|Mu\|_2^2 - 1 > \epsilon\right) \leq |T| \cdot 2 \exp\left(-\Omega(\epsilon^2 k^2)\right).
\]

**Related question**

For \( T \subset S^{d-1} \), expected maximum deviation

\[
\mathbb{E} \max_{u \in T} \|Mu\|_2^2 - 1 \leq ?
\]

**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

\[
\mathbb{E} \max_{t \in T} X_t \leq \sqrt{2v \ln |T|}.
\]

- 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 (\( \lambda > 0 \)):

\[
\mathbb{E} \max_{t \in T} X_t = \frac{1}{\lambda} \ln \exp \left( \mathbb{E} \max_{t \in T} \lambda X_t \right)
\]

- Apply Jensen’s inequality:

\[
\leq \frac{1}{\lambda} \ln \mathbb{E} \exp \left( \max_{t \in T} \lambda X_t \right) = \frac{1}{\lambda} \ln \mathbb{E} \left( \max_{t \in T} \exp(\lambda X_t) \right)
\]

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

\[
\leq \frac{1}{\lambda} \ln \sum_{t \in T} \mathbb{E} \exp(\lambda X_t)
\]

- Exploit \( v \)-subgaussian property:

\[
\leq \frac{1}{\lambda} \ln \sum_{t \in T} \exp \left( v \lambda^2 / 2 \right) = \frac{\ln |T|}{\lambda} + \frac{v \lambda}{2}
\]

- Choose appropriate \( \lambda \) to conclude.
Alternative proof

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

$$
E(Y) = \int_0^\infty P(Y \geq y) \, dy .
$$

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

Infinite collections

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

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

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

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

\[
\mathbb{E} \max_{\tilde{w} \in \text{conv}(T)} \langle \tilde{w}, X \rangle \lesssim \sqrt{2 \nu \ln |T|}.
\]

Proof:

▶ Write \( \tilde{w} \in \text{conv}(T) \) as \( \tilde{w} = \sum_{w \in T} \rho_w w \) for some \( \rho_w \geq 0 \) that sum to one.
▶ Observe that

\[
\langle \tilde{w}, x \rangle = \sum_{w \in T} \rho_w \langle w, x \rangle \leq \max_{w \in T} \langle w, x \rangle.
\]

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

Euclidean norm

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

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

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 \( u \in S^{d-1} \), there exists \( u_0 \in \mathcal{N} \) with

\[
\|u - u_0\|_2 \leq \varepsilon.
\]

▶ Such a set \( \mathcal{N} \) is called an \( \varepsilon \)-net for \( S^{d-1} \).
▶ We need a 1/2-net, of cardinality at most \( 5^d \).
Proof

- Write $u \in S^{d-1}$ as
  $$u = u_0 + \delta q,$$
  where $u_0 \in \mathcal{N}$, $q \in S^{d-1}$, $\delta \in [0, 1/2]$, so
  $$\langle u, X \rangle = \langle u_0, X \rangle + \delta \langle q, X \rangle.$$

- Observe that
  $$\max_{u \in S^{d-1}} \langle u, X \rangle \leq \max_{u_0 \in \mathcal{N}} \langle u_0, X \rangle + \max_{\delta \in [0,1/2]} \max_{q \in S^{d-1}} \delta \langle q, X \rangle \leq \max_{u_0 \in \mathcal{N}} \langle u_0, X \rangle + \frac{1}{2} \max_{q \in S^{d-1}} \langle q, X \rangle.$$

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

\(\varepsilon\)-nets for unit sphere

There is an $\varepsilon$-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 $\varepsilon$ 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 $u \in S^{d-1}$ is at distance at most $\varepsilon$ from selected points.
  - I.e., selected points form an $\varepsilon$-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
  $$N \ vol((\varepsilon/2)B^d) \leq \ vol((1 + \varepsilon/2)B^d).$$
  - 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 $M$ whose entries are iid $N(0, 1/k)$. For a $W \subseteq \mathbb{R}^d$ be a subspace of dimension $r$,

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

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

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

Proof of subspace JL lemma

Let columns of $Q$ be ONB for $W$. Then

$$\max_{u \in S^{d-1} \cap W} \left| \|Mu\|_2^2 - 1 \right| = \max_{u \in S^{r-1}} \left| u^T Q^T (M^T M - I) Qu \right| = \max_{u, v \in S^{r-1}} u^T Q^T (M^T M - I) Qv.$$

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

$$X_{u,v} := u^T Q^T (M^T M - I) Qv$$

is $(O(1/k), O(1/k))$-subexponential.
Proof of subspace JL lemma (continued)

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

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

- Write $u, v \in S^{r-1}$ as
  \[ u = u_0 + \varepsilon p, \quad v = v_0 + \delta q, \]
  where $u_0, v_0 \in \mathcal{N}$, $p, q \in S^{r-1}$ and $\varepsilon, \delta \in [0, 1/4]$, so
  \[ X_{u,v} = X_{u_0,v_0} + \varepsilon X_{p,v} + \delta X_{u_0,q}. \]

- Therefore
  \[ \max_{u,v \in S^{r-1}} X_{u,v} \leq \max_{u_0,v_0 \in \mathcal{N}} X_{u_0,v_0} + \frac{1}{2} \max_{p,q \in S^{r-1}} X_{p,q}, \]
  which implies
  \[ \max_{u,v \in S^{r-1}} X_{u,v} \leq 2 \max_{u_0,v_0 \in \mathcal{N}} X_{u_0,v_0}. \]

- Conclude by applying previous result for finite collections. □

Application to least squares
Big data least squares

- **Input**: matrix $A \in \mathbb{R}^{n \times d}$, vector $b \in \mathbb{R}^n$ ($n \gg d$).
- **Goal**: find $x \in \mathbb{R}^d$ so as to (approx.) minimize $\|Ax - b\|_2^2$.
- Computation time: $O(nd^2)$.
- Can we speed this up?

Simple approach

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

Simple (somewhat loose) analysis

- Let $W$ be subspace spanned by columns of $A$ and $b$.
  - Dimension is at most $d + 1$.
- If $m \geq O(d/\varepsilon^2)$, then $M$ is subspace embedding for $W$:
  $$
  (1 - \varepsilon)\|x\|_2^2 \leq \|Mx\|_2^2 \leq (1 + \varepsilon)\|x\|_2^2
  $$
  for all $x \in W$.
- Let $x^* := \arg\min_{x \in \mathbb{R}^d} \|Ax - b\|_2^2$.
- $$
  \|A\hat{x} - b\|_2^2 \leq \frac{1}{1 - \varepsilon} \|M(A\hat{x} - b)\|_2^2
  \leq \frac{1}{1 - \varepsilon} \|M(Ax^* - b)\|_2^2
  \leq \frac{1 + \varepsilon}{1 - \varepsilon} \|Ax^* - b\|_2^2.
  $$
- Running time (using FJLT): $O\left((m + n)d \log n + md^2\right)$.  

Another perspective: random sampling

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

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

\[
\|A\hat{x} - b\|_2^2 \leq \left(1 + O\left(\frac{\kappa}{m}\right)\right) \cdot \|Ax^* - b\|_2^2
\]

(up to lower-order terms), where

\[
\kappa := n \cdot \max_{i \in [n]} \|(A^T A)^{-1/2}A^T e_i\|_2^2
\]

and $e_i$ is $i$-th coordinate basis vector.

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

\[
(A^T A)^{-1/2}A^T = (VS^2V^T)^{-1/2}VSU^T = VU^T.
\]

- So $\kappa = n \cdot \max_{i \in [n]} \|U^T e_i\|_2^2$.

- $\|U^T e_i\|_2^2$ is statistical leverage score for $i$-th row of $A$: measures how much “influence” $i$-th row has on least squares solution.

Statistical leverage

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

\[
U = \begin{bmatrix} I_{d \times d} & \text{0}_{(n-d) \times d} \end{bmatrix} \quad \Rightarrow \quad n \cdot \max_{i \in [n]} \ell_i = n.
\]

\[
U = \frac{1}{\sqrt{n}} \begin{bmatrix} H_n e_1 & H_n e_2 & \cdots & H_n e_d \end{bmatrix} \quad \Rightarrow \quad n \cdot \max_{i \in [n]} \ell_i = d,
\]

where $H_n$ 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 $A$ and $b$.
  - Randomly mixes up rows of $(A, b)$ 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.,
  $$\|A\hat{x} - b\|_2^2 \leq (1 + \varepsilon) \cdot \|Ax^\star - b\|_2^2,$$

suffices to have

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

Application to compressed sensing
Under-determined least squares

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

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

Null space property

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

- **Proof.** $(\Rightarrow)$ Take any $k$-sparse vectors $x$ and $y$ with $Ax = Ay$. Want to show $x = y$.
  - Then $x - y$ is 2k-sparse, and $A(x - y) = 0$.
  - By assumption, null space of $A$ does not contain any non-zero 2k-sparse vectors.
  - So $x - y = 0$, i.e., $x = y$.

- $(\Leftarrow)$ Take any 2k-sparse vector $z$ in the null space of $A$. Want to show $z = 0$.
  - Write it as $z = x - y$ for some $k$-sparse vectors $x$ and $y$ with disjoint supports.
  - Then $A(x - y) = 0$, and hence $x = y$ by assumption.
  - But $x$ and $y$ have disjoint support, so it must be that $x = y = 0$, so $z = 0$. $\square$
Null space property from subspace embeddings

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

- Want: for any $2k$-sparse vector $z$, $Az \neq 0$, i.e., $\|Az\|_2^2 > 0$.
- Consider a particular choice $I \subseteq [d]$ of $|I| = 2k$ coordinates, and the corresponding subspace $W_I$ spanned by $\{e_i : i \in I\}$.
  - Every $2k$-sparse $z$ is in $W_I$ for some $I$.
- Sufficient for $A$ to be $1/2$-subspace embedding for $W_I$ for all $I$:
  \[
  \frac{1}{2} \|z\|_2^2 \leq \|Az\|_2^2 \leq \frac{3}{2} \|z\|_2^2
  \]
  for all $2k$-sparse $z$.

Null space property from subspace embeddings (continued)

- Say $A$ fails for $I$ if it is not a $1/2$-subspace embedding for $W_I$.
- Subspace JL lemma:
  \[
  P(A \text{ fails for } I) \leq 2^{O(k)} \exp(-\Omega(n)).
  \]
- Union bound over all choices of $I$ with $|I| = 2k$:
  \[
  P(A \text{ fails for some } I) \leq \binom{d}{2k} 2^{O(k)} \exp(-\Omega(n)).
  \]
- To ensure this is, say, at most $1/2$, just need
  \[
  n \geq O\left(k + \log \left(\frac{d}{2k}\right)\right) = O(k + k \log(d/k)).
  \]
Restricted isometry property

$(\ell, \delta)$-restricted isometry property (RIP):

$$(1 - \delta)\|z\|_2^2 \leq \|Az\|_2^2 \leq (1 + \delta)\|z\|_2^2$$

for all $\ell$-sparse $z$.

- 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.