Dimensionality Reduction

Jelani Nelson
UC Berkeley

August 1, 2022

General setup

We have high-dimensional data, e.g.

- Machine learning. Database of e-mails featurized as high-dimensional vectors; we want to learn a spam classifier.
- Bioinformatics. Motif discovery in DNA sequences.
- Computational geometry. Fingerprint matching in a large database.
- Data mining. Clustering similar featurized objects.
- Compression and fast image acquisition. Compressed sensing.
- Large-scale linear algebra. Low-rank approximation or regression on a huge matrix.

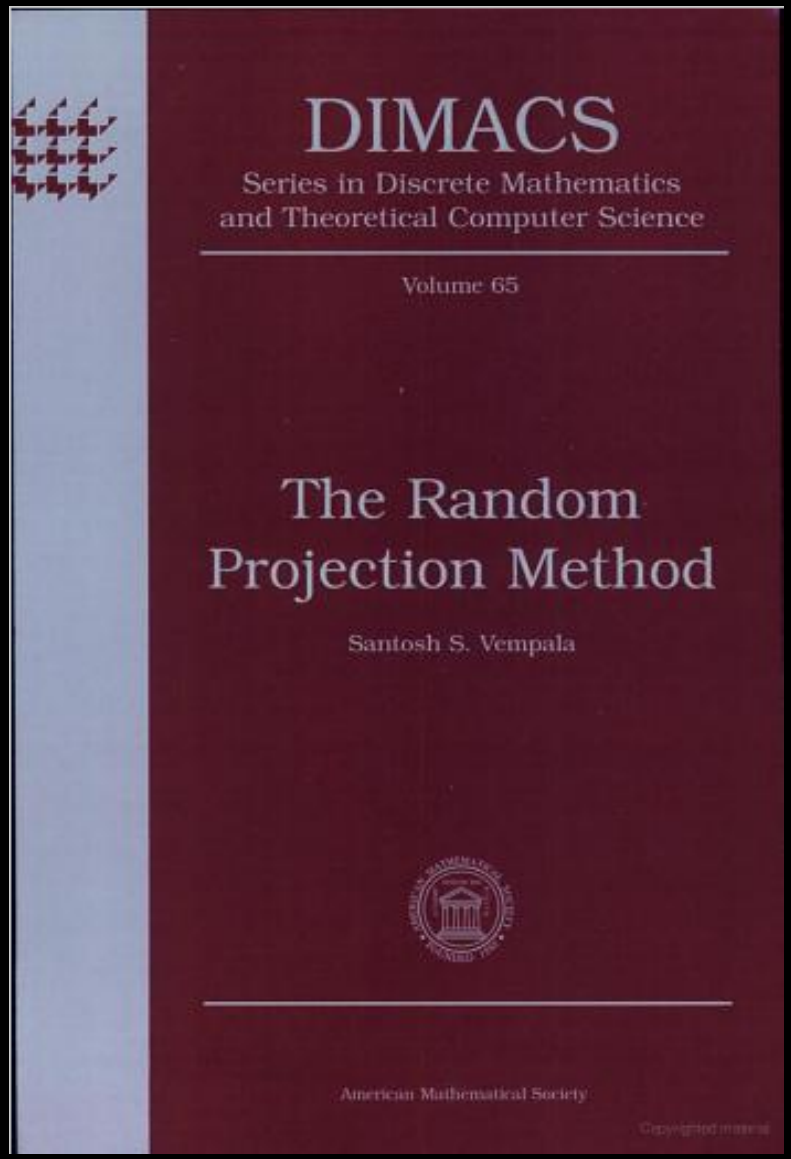
General setup

We have high-dimensional data, e.g.

- Machine learning. Database of e-mails featurized as high-dimensional vectors; we want to learn a spam classifier.
- Bioinformatics. Motif discovery in DNA sequences.
- Computational geometry. Fingerprint matching in a large database.
- Data mining. Clustering similar featurized objects.
- Compression and fast image acquisition. Compressed sensing.
- Large-scale linear algebra. Low-rank approximation or regression on a huge matrix.

Can we reduce dimensionality of the data in a pre-processing step, in a way that doesn't disrupt downstream applications?

- Faster running times (lower dimension $=$ faster algorithms)
- Save space
- Minimize communication for distributed applications



Random projection method: pick some random linear map, $x \mapsto \Pi x$, and apply Π to input as a pre-processing step

Other dimensionality reduction methods?

- Principal component analysis (PCA)
- Kernel PCA
- Multidimensional scaling
- ISOMAP
- Hessian Eigenmaps
- ...

Other dimensionality reduction methods?

- Principal component analysis (PCA)
- Kernel PCA
- Multidimensional scaling
- ISOMAP
- Hessian Eigenmaps
- ...

Random projection is orthogonal to, and complements, other dimensionality reduction methods. Its purpose is to make other algorithms more efficient, not be the data analysis algorithm. (will say more soon)

Cornerstone dim. reduction/random projections result

JL lemma [Johnson, Lindenstrauss '84]

For every $X \subset \ell_2$ of size n, there is an embedding $f : X \to \ell_2^m$ for $m = O(\varepsilon^{-2} \log n)$ with distortion $1 + \varepsilon$. That is, for each $x, y \in X$,

$$(1 - \varepsilon)\|x - y\|_2^2 \leq \|f(x) - f(y)\|_2^2 \leq (1 + \varepsilon)\|x - y\|_2^2$$

Cornerstone dim. reduction/random projections result

JL lemma [Johnson, Lindenstrauss '84]

For every $X \subset \ell_2$ of size n, there is an embedding $f : X \to \ell_2^m$ for $m = O(\varepsilon^{-2} \log n)$ with distortion $1 + \varepsilon$. That is, for each $x, y \in X$,

$$(1 - \varepsilon)\|x - y\|_2^2 \leq \|f(x) - f(y)\|_2^2 \leq (1 + \varepsilon)\|x - y\|_2^2$$

Summary: For any n vectors in arbitrary dimension, can map to $O(\log n)$ dim. while approximately preserving Euclidean geometry.

How to prove the JL lemma

Distributional JL (DJL) lemma

Lemma (DJL lemma [Johnson, Lindenstrauss '84])

For any $0 < \varepsilon, \delta < 1/2$ *and* $d \geq 1$ *there exists a distribution* $\mathcal{D}_{\varepsilon,\delta}$ *on* $\mathbb{R}^{m \times d}$ *for* $m = O(\varepsilon^{-2} \log(1/\delta))$ *such that for any* $z \in \mathbb{R}^d$

$$\mathop{\mathbb{P}}_{\Pi \sim \mathcal{D}_{\varepsilon,\delta}} \left(\|\Pi z\|_2^2 \notin [(1-\varepsilon)\|z\|_2^2, (1+\varepsilon)\|z\|_2^2] \right) < \delta.$$

How to prove the JL lemma

Distributional JL (DJL) lemma

Lemma (DJL lemma [Johnson, Lindenstrauss '84])

For any $0 < \varepsilon, \delta < 1/2$ *and* $d \geq 1$ *there exists a distribution* $\mathcal{D}_{\varepsilon,\delta}$ *on* $\mathbb{R}^{m \times d}$ *for* $m = O(\varepsilon^{-2} \log(1/\delta))$ *such that for any* $z \in \mathbb{R}^d$

$$\mathbb{P}_{\Pi \sim \mathcal{D}_{\varepsilon,\delta}} \left(\|\Pi z\|_2^2 \notin [(1-\varepsilon)\|z\|_2^2, (1+\varepsilon)\|z\|_2^2] \right) < \delta.$$

Proof of JL: Set $\delta = 1/n^2$ in DJL and z as the difference vector of some pair of points. Union bound over the $\binom{n}{2}$ pairs. Thus the map $f : X \to \ell_2^m$ can be linear: $f(x) = \Pi x$.

How to prove the JL lemma

Distributional JL (DJL) lemma

Lemma (DJL lemma [Johnson, Lindenstrauss '84])

For any $0 < \varepsilon, \delta < 1/2$ *and* $d \geq 1$ *there exists a distribution* $\mathcal{D}_{\varepsilon,\delta}$ *on* $\mathbb{R}^{m \times d}$ *for* $m = O(\varepsilon^{-2} \log(1/\delta))$ *such that for any* $z \in \mathbb{R}^d$

$$\mathbb{P}_{\Pi \sim \mathcal{D}_{\varepsilon,\delta}} \left(\|\Pi z\|_2^2 \notin [(1-\varepsilon)\|z\|_2^2, (1+\varepsilon)\|z\|_2^2] \right) < \delta.$$

Proof of JL: Set $\delta = 1/n^2$ in DJL and z as the difference vector of some pair of points. Union bound over the $\binom{n}{2}$ pairs. Thus the map $f : X \to \ell_2^m$ can be linear: $f(x) = \Pi x$.

First proof of DJL in [JL'84] took $\mathcal{D}_{\varepsilon,\delta}$ as (scaled) orthogonal projection onto a random m-dimensional subspace.

This talk

This talk

- What about other, non-Euclidean norms?
- Proof of DJL.
- What is the optimal target dimension m for DJL?
- What is the optimal target dimension m for JL?
- How fast can we apply the map $x \mapsto \Pi x$?
- Isn't storing the random $m \times d$ matrix Π expensive?
- What can we get for specific applications and instances?
- Generalizing JL: terminal embeddings.

This talk

- What about other, non-Euclidean norms?
- Proof of DJL.
- What is the optimal target dimension m for DJL?
- What is the optimal target dimension m for JL?
- How fast can we apply the map $x \mapsto \Pi x$?
- Isn't storing the random $m \times d$ matrix Π expensive?
- What can we get for specific applications and instances?
- Generalizing JL: terminal embeddings.

Non-Euclidean norms

Non-Euclidean norms

(Informal) Theorem [Johnson-Naor'10]. Any norm enjoying JL-type dimensionality reduction must be "almost" Euclidean.

Non-Euclidean norms

(Informal) Theorem [Johnson-Naor'10]. Any norm enjoying JL-type dimensionality reduction must be "almost" Euclidean.

(Formal) Theorem [Johnson-Naor'10]. Suppose Z is a normed space satisfying the following property: for every n points $x_1, \ldots, x_n \in Z$ there is a linear subspace $F \subset Z$ of dimension $O(\log n)$ and a linear map $L : Z \to F$ such that $\|x_i - x_j\| \leq \|L(x_i) - L(x_j)\| \leq O(1) \cdot \|x_i - x_j\|$ for all $1 \leq i, j \leq n$. Then every k-dimensional subspace of Z embeds into Euclidean space with distortion $2^{2^{O(\log^* k)}}$.

Non-Euclidean norms

(Informal) Theorem [Johnson-Naor'10]. Any norm enjoying JL-type dimensionality reduction must be "almost" Euclidean.

(Formal) Theorem [Johnson-Naor'10]. Suppose Z is a normed space satisfying the following property: for every n points $x_1, \ldots, x_n \in Z$ there is a linear subspace $F \subset Z$ of dimension $O(\log n)$ and a linear map $L : Z \to F$ such that $\|x_i - x_j\| \leq \|L(x_i) - L(x_j)\| \leq O(1) \cdot \|x_i - x_j\|$ for all $1 \leq i, j \leq n$. Then every k-dimensional subspace of Z embeds into Euclidean space with distortion $2^{2^{O(\log^* k)}}$.

A lower bound is also shown, that the $2^{2^{O(\log^* k)}}$ term must be $\omega(1)$ (specifically $2^{\Omega(\alpha(k))}$).

This talk

- What about other, non-Euclidean norms?
- Proof of DJL.
- What is the optimal target dimension m for DJL?
- What is the optimal target dimension m for JL?
- How fast can we apply the map $x \mapsto \Pi x$?
- Isn't storing the random $m \times d$ matrix Π expensive?
- What can we get for specific applications and instances?
- Generalizing JL: terminal embeddings.

Lemma (DJL lemma [Johnson, Lindenstrauss '84])

For any $0 < \varepsilon, \delta < 1/2$ *and* $d \geq 1$ *there exists a distribution* $\mathcal{D}_{\varepsilon,\delta}$ *on* $\mathbb{R}^{m \times d}$ *for* $m = O(\varepsilon^{-2} \log(1/\delta))$ *such that for any* $z \in \mathbb{R}^d$

$$\mathbb{P}_{\Pi \sim \mathcal{D}_{\varepsilon,\delta}} \left(\|\Pi z\|_2^2 \notin [(1-\varepsilon)\|z\|_2^2, (1+\varepsilon)\|z\|_2^2] \right) < \delta.$$

Sketch of short proof of DJL

For random variable X, define $\psi_X(\lambda) = \ln \mathbb{E}\, e^{\lambda(X - \mathbb{E}\, X)}$

Sketch of short proof of DJL

For random variable X, define $\psi_X(\lambda) = \ln \mathbb{E}\, e^{\lambda(X - \mathbb{E}\, X)}$

Say X is (σ, B)-subgamma if $\forall |\lambda| < 1/B$, $\psi_X(\lambda) \leq \frac{\lambda^2 \sigma^2}{2(1 - B\lambda)}$

Sketch of short proof of DJL

For random variable X, define $\psi_X(\lambda) = \ln \mathbb{E}\, e^{\lambda(X - \mathbb{E}\, X)}$

Say X is (σ, B)-subgamma if $\forall |\lambda| < 1/B$, $\psi_X(\lambda) \leq \frac{\lambda^2 \sigma^2}{2(1 - B\lambda)}$

- Can show if X_i are indep. (σ_i, B_i)-subgamma, then $\sum_i X_i$ is $(\sqrt{\sum_i \sigma_i^2}, \min_i B_i)$-subgamma.
- Can show Bernstein-type inequality X subgamma $\implies$ $\mathbb{P}(|X| > t) \lesssim e^{-\frac{t^2}{2\sigma^2}} + e^{-\frac{t}{2B}}$

Sketch of short proof of DJL

For random variable X, define $\psi_X(\lambda) = \ln \mathbb{E}\, e^{\lambda(X - \mathbb{E}\, X)}$

Say X is (σ, B)-subgamma if $\forall |\lambda| < 1/B$, $\psi_X(\lambda) \leq \frac{\lambda^2 \sigma^2}{2(1 - B\lambda)}$

- Can show if X_i are indep. (σ_i, B_i)-subgamma, then $\sum_i X_i$ is $(\sqrt{\sum_i \sigma_i^2}, \min_i B_i)$-subgamma.
- Can show Bernstein-type inequality X subgamma $\implies$ $\mathbb{P}(|X| > t) \lesssim e^{-\frac{t^2}{2\sigma^2}} + e^{-\frac{t}{2B}}$

Suppose $\mathcal{D}_{\varepsilon,\delta}$ is such that $\Pi_{r,i}$ is $\frac{g_{r,i}}{\sqrt{m}}$, where $g_{r,i}$ is subgaussian.

Sketch of short proof of DJL

For random variable X, define $\psi_X(\lambda) = \ln \mathbb{E}\, e^{\lambda(X - \mathbb{E}\, X)}$

Say X is (σ, B)-subgamma if $\forall |\lambda| < 1/B$, $\psi_X(\lambda) \leq \frac{\lambda^2 \sigma^2}{2(1 - B\lambda)}$

- Can show if X_i are indep. (σ_i, B_i)-subgamma, then $\sum_i X_i$ is $(\sqrt{\sum_i \sigma_i^2}, \min_i B_i)$-subgamma.
- Can show Bernstein-type inequality X subgamma $\implies$ $\mathbb{P}(|X| > t) \lesssim e^{-\frac{t^2}{2\sigma^2}} + e^{-\frac{t}{2B}}$

Suppose $\mathcal{D}_{\varepsilon,\delta}$ is such that $\Pi_{r,i}$ is $\frac{g_{r,i}}{\sqrt{m}}$, where $g_{r,i}$ is subgaussian.

Then $y_r := \sum_i \Pi_{r,i} x_i$ is subgaussian with appropriate parameters, which implies y_r^2 subgamma with appropriate parameters.

Sketch of short proof of DJL

For random variable X, define $\psi_X(\lambda) = \ln \mathbb{E}\, e^{\lambda(X - \mathbb{E}\, X)}$

Say X is (σ, B)-subgamma if $\forall |\lambda| < 1/B$, $\psi_X(\lambda) \leq \frac{\lambda^2 \sigma^2}{2(1 - B\lambda)}$

- Can show if X_i are indep. (σ_i, B_i)-subgamma, then $\sum_i X_i$ is $(\sqrt{\sum_i \sigma_i^2}, \min_i B_i)$-subgamma.
- Can show Bernstein-type inequality X subgamma $\implies$ $\mathbb{P}(|X| > t) \lesssim e^{-\frac{t^2}{2\sigma^2}} + e^{-\frac{t}{2B}}$

Suppose $\mathcal{D}_{\varepsilon,\delta}$ is such that $\Pi_{r,i}$ is $\frac{g_{r,i}}{\sqrt{m}}$, where $g_{r,i}$ is subgaussian.

Then $y_r := \sum_i \Pi_{r,i} x_i$ is subgaussian with appropriate parameters, which implies y_r^2 subgamma with appropriate parameters.

Then $\sum_r y_r^2$ is subgamma by first bullet above, so satisfies Bernstein-type inequality from second bullet.

Longer, less simple proof (but extendable to other settings)

Strategy outline: will show . . .

- Khintchine's inequality (tail bound for linear forms)
- Decoupling (convert tail bounds for quadratic forms to ones for linear forms)
- Hanson-Wright inequality (concentration for quadratic forms)

Longer, less simple proof (but extendable to other settings)

Strategy outline: will show ...

- Khintchine's inequality (tail bound for linear forms)
- Decoupling (convert tail bounds for quadratic forms to ones for linear forms)
- Hanson-Wright inequality (concentration for quadratic forms)

Note if $\Pi_{r,i} = \frac{1}{\sqrt{m}} g_{r,i}$, then

$$\Pi z = B_z g$$

where B_z is the $md \times md$ matrix

$$B_z = \frac{1}{\sqrt{m}} \begin{bmatrix} z^\top & 0 & 0 & \dots & 0 \\ 0 & z^\top & 0 & \dots & 0 \\ 0 & 0 & z^\top & \dots & 0 \\ & & & \ddots & \\ 0 & 0 & 0 & & 0 \\ 0 & 0 & 0 & \dots & z^\top \end{bmatrix}$$

Longer, less simple proof (but extendable to other settings)

Strategy outline: will show ...

- Khintchine's inequality (tail bound for linear forms)
- Decoupling (convert tail bounds for quadratic forms to ones for linear forms)
- Hanson-Wright inequality (concentration for quadratic forms)

Note if $\Pi_{r,i} = \frac{1}{\sqrt{m}} g_{r,i}$, then

$$\Pi z = B_z g$$

where B_z is the $md \times md$ matrix

$$B_z = \frac{1}{\sqrt{m}} \begin{bmatrix} z^\top & 0 & 0 & \dots & 0 \\ 0 & z^\top & 0 & \dots & 0 \\ 0 & 0 & z^\top & \dots & 0 \\ & & & \ddots & \\ 0 & 0 & 0 & & 0 \\ 0 & 0 & 0 & \dots & z^\top \end{bmatrix}$$

so that $\|\Pi z\|_2^2 = \underbrace{g^\top B_z^\top B_z g}_{\text{quadratic form}}$

Khintchine inequality

Lemma (Khintchine). Recall $\|X\|_p := (\mathbb{E}\,|X|^p)^{1/p}$. Then if $\sigma \in \{-1,1\}^N$ is uniformly at random, then for any $x \in \mathbb{R}^N$,

$$\|\langle \sigma, x\rangle\|_p \lesssim \|x\|_2 \sqrt{p}$$

Khintchine inequality

Lemma (Khintchine). Recall $\|X\|_p := (\mathbb{E}\,|X|^p)^{1/p}$. Then if $\sigma \in \{-1,1\}^N$ is uniformly at random, then for any $x \in \mathbb{R}^N$,

$$\|\langle \sigma, x\rangle\|_p \lesssim \|x\|_2 \sqrt{p}$$

Proof. Let $g, g_1, \ldots, g_N$ be i.i.d. $\mathcal{N}(0,1)$. Then

Khintchine inequality

Lemma (Khintchine). Recall $\|X\|_p := (\mathbb{E}\,|X|^p)^{1/p}$. Then if $\sigma \in \{-1,1\}^N$ is uniformly at random, then for any $x \in \mathbb{R}^N$,

$$\|\langle \sigma, x\rangle\|_p \lesssim \|x\|_2 \sqrt{p}$$

Proof. Let $g, g_1, \ldots, g_N$ be i.i.d. $\mathcal{N}(0,1)$. Then

$$\begin{aligned}
\|\sum_i \sigma_i x_i\|_p &= \sqrt{\frac{\pi}{2}}\,\|\mathop{\mathbb{E}}_g \sum_i \sigma_i |g_i| x_i\|_p \\
&\lesssim \|\sum_i \sigma_i |g_i| x_i\|_p && \text{(Jensen)} \\
&= \|\sum_i g_i x_i\|_p \\
&= \|x\|_2 \cdot \|g\|_p && \text{(gaussian 2-stability)} \\
&\simeq \|x\|_2 \sqrt{p}
\end{aligned}$$

Decoupling

Lemma. If $\sigma, \sigma' \in \{-1, 1\}^N$ are independent and each uniform, and $(a_{i,j})$ are reals, then

$$\| \sum_{i \neq j} a_{i,j} \sigma_i \sigma_j \|_p \lesssim \| \sum_{i,j} a_{i,j} \sigma_i \sigma'_j \|_p$$

Decoupling

Lemma. If $\sigma, \sigma' \in \{-1, 1\}^N$ are independent and each uniform, and $(a_{i,j})$ are reals, then

$$\| \sum_{i \neq j} a_{i,j} \sigma_i \sigma_j \|_p \lesssim \| \sum_{i,j} a_{i,j} \sigma_i \sigma'_j \|_p$$

Proof. Let (η_i) be indep. Bernoulli. Then there exists some η^* s.t.

Decoupling

Lemma. If $\sigma, \sigma' \in \{-1,1\}^N$ are independent and each uniform, and $(a_{i,j})$ are reals, then

$$\|\sum_{i\neq j} a_{i,j}\sigma_i\sigma_j\|_p \lesssim \|\sum_{i,j} a_{i,j}\sigma_i\sigma'_j\|_p$$

Proof. Let (η_i) be indep. Bernoulli. Then there exists some η^* s.t.

$$\begin{aligned}
\|\sum_{i\neq j} a_{i,j}\sigma_i\sigma_j\|_p &= 4\|\mathop{\mathbb{E}}_{\eta}\sum_{i\neq j}\eta_i(1-\eta_j)a_{i,j}\sigma_i\sigma_j\|_p \\
&\leq 4\|\sum_{i\neq j}\eta_i(1-\eta_j)a_{i,j}\sigma_i\sigma_j\|_p \qquad \text{(Jensen)} \\
&= 4(\mathop{\mathbb{E}}_{\eta}\mathop{\mathbb{E}}_{\sigma}|\sum_{i\neq j}\eta_i(1-\eta_j)a_{i,j}\sigma_i\sigma_j|^p)^{1/p} \\
&\leq 4(\mathop{\mathbb{E}}_{\sigma}|\sum_{i\neq j}\eta_i^*(1-\eta_j^*)a_{i,j}\sigma_i\sigma_j|^p)^{1/p}
\end{aligned}$$

Define $S = \{i : \eta_i^* = 1\}$. Then

$$
\begin{aligned}
(\mathop{\mathbb{E}}_{\sigma} |\sum_{i \neq j} \eta_i^*(1-\eta_j^*) a_{i,j}\sigma_i\sigma_j|^p)^{1/p} &= \|\sum_{i\in S}\sum_{j\in \bar{S}} a_{i,j}\sigma_i\sigma_j\|_p \\
&= \|\sum_{i\in S}\sum_{j\in \bar{S}} a_{i,j}\sigma_i\sigma'_j\|_p \\
&= \|\sum_{i\in S}\sum_{j\in \bar{S}} a_{i,j}\sigma_i\sigma'_j\|_{L^p(\sigma_S,\sigma'_{\bar{S}})} \\
&= \|\mathop{\mathbb{E}}_{\sigma_{\bar{S}}}\mathop{\mathbb{E}}_{\sigma'_S}\sum_{i,j} a_{i,j}\sigma_i\sigma'_j\|_{L^p(\sigma_S,\sigma'_{\bar{S}})} \\
&\leq \|\sum_{i,j} a_{i,j}\sigma_i\sigma'_j\|_p \qquad \text{(Jensen)}
\end{aligned}
$$

■

Hanson-Wright

Lemma (Hanson-Wright). $\sigma \in \{-1, 1\}^N$ uniformly random and $A \in \mathbb{R}^{N \times N}$. Then

$$\|\sigma^\top A\sigma - \mathbb{E}\,\sigma^\top A\sigma\|_p \lesssim \sqrt{p}\|A\|_F + p\|A\|$$

Hanson-Wright

Lemma (Hanson-Wright). $\sigma \in \{-1,1\}^N$ uniformly random and $A \in \mathbb{R}^{N\times N}$. Then

$$\|\sigma^\top A\sigma - \mathbb{E}\,\sigma^\top A\sigma\|_p \lesssim \sqrt{p}\|A\|_F + p\|A\|$$

Proof.

$$\begin{aligned}
\|\sigma^\top A\sigma - \mathbb{E}\,\sigma^\top A\sigma\|_p &\le 4\|\sigma^\top A\sigma'\|_p \text{ (decoupling)}\\
&\lesssim \sqrt{p}\|\|A\sigma'\|_2\|_p \text{ (Khintchine)}\\
&= \sqrt{p}\|\|A\sigma'\|_2^2\|_{p/2}^{1/2}\\
&\le \sqrt{p}\left(\mathbb{E}\,\|A\sigma'\|_2^2 + \|\|A\sigma'\|_2^2 - \mathbb{E}\,\|A\sigma'\|_2^2\|_{p/2}\right)^{1/2} \text{ (triangle inequality)}\\
&\le \sqrt{p}\|A\|_F + \sqrt{p}\|\|A\sigma'\|_2^2 - \mathbb{E}\,\|A\sigma'\|_2^2\|_{p/2}^{1/2}
\end{aligned}$$

Hanson-Wright

Lemma (Hanson-Wright). $\sigma \in \{-1,1\}^N$ uniformly random and $A \in \mathbb{R}^{N\times N}$. Then

$$\|\sigma^\top A\sigma - \mathbb{E}\,\sigma^\top A\sigma\|_p \lesssim \sqrt{p}\|A\|_F + p\|A\|$$

Proof.

$$\begin{aligned}
\|\sigma^\top A\sigma - \mathbb{E}\,\sigma^\top A\sigma\|_p &\le 4\|\sigma^\top A\sigma'\|_p \text{ (decoupling)}\\
&\lesssim \sqrt{p}\|\|A\sigma'\|_2\|_p \text{ (Khintchine)}\\
&= \sqrt{p}\|\|A\sigma'\|_2^2\|_{p/2}^{1/2}\\
&\le \sqrt{p}\left(\mathbb{E}\,\|A\sigma'\|_2^2 + \|\|A\sigma'\|_2^2 - \mathbb{E}\,\|A\sigma'\|_2^2\|_{p/2}\right)^{1/2} \text{ (triangle inequality)}\\
&\le \sqrt{p}\|A\|_F + \sqrt{p}\|\|A\sigma'\|_2^2 - \mathbb{E}\,\|A\sigma'\|_2^2\|_{p/2}^{1/2}
\end{aligned}$$

For $p = 2^k$, can do induction on k. Then for q not a power of 2, bound $\|X\|_q \le \|X\|_p$ where we round up to nearest power of 2.

Proof of DJL (finally)

Hanson-Wright. $\|\sigma^\top A\sigma - \mathbb{E}\,\sigma^\top A\sigma\|_p \lesssim \sqrt{p}\|A\|_F + p\|A\|$

Proof of DJL (finally)

Hanson-Wright. $\|\sigma^\top A\sigma - \mathbb{E}\,\sigma^\top A\sigma\|_p \lesssim \sqrt{p}\|A\|_F + p\|A\|$

Equivalent to

$\mathbb{P}(|\sigma^\top A\sigma - \mathbb{E}\,\sigma^\top A\sigma| > t) \lesssim \exp(-Ct^2/\|A\|_F^2) + \exp(-Ct/\|A\|)$

Proof of DJL (finally)

Hanson-Wright. $\|\sigma^\top A\sigma - \mathbb{E}\,\sigma^\top A\sigma\|_p \lesssim \sqrt{p}\|A\|_F + p\|A\|$
Equivalent to
$\mathbb{P}(|\sigma^\top A\sigma - \mathbb{E}\,\sigma^\top A\sigma| > t) \lesssim \exp(-Ct^2/\|A\|_F^2) + \exp(-Ct/\|A\|)$

Recall

$$\Pi z = B_z \sigma$$

where B_z is the $md \times md$ matrix

$$B_z = \frac{1}{\sqrt{m}} \begin{bmatrix} z^\top & 0 & 0 & \dots & 0 \\ 0 & z^\top & 0 & \dots & 0 \\ 0 & 0 & z^\top & \dots & 0 \\ 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \dots & z^\top \end{bmatrix}$$

Proof of DJL (finally)

Hanson-Wright. $\|\sigma^\top A\sigma - \mathbb{E}\,\sigma^\top A\sigma\|_p \lesssim \sqrt{p}\|A\|_F + p\|A\|$
Equivalent to
$\mathbb{P}(|\sigma^\top A\sigma - \mathbb{E}\,\sigma^\top A\sigma| > t) \lesssim \exp(-Ct^2/\|A\|_F^2) + \exp(-Ct/\|A\|)$

Recall

$$\Pi z = B_z \sigma$$

where B_z is the $md \times md$ matrix

$$B_z = \frac{1}{\sqrt{m}} \begin{bmatrix} z^\top & 0 & 0 & \dots & 0 \\ 0 & z^\top & 0 & \dots & 0 \\ 0 & 0 & z^\top & \dots & 0 \\ 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \dots & z^\top \end{bmatrix}$$

so that $\|\Pi z\|_2^2 = \underbrace{\sigma^\top B_z^\top B_z \sigma}_{\text{quadratic form}}$

Proof of DJL (finally)

Hanson-Wright.

$$\mathbb{P}(|\sigma^\top A\sigma - \mathbb{E}\,\sigma^\top A\sigma| > t) \lesssim \exp(-Ct^2/\|A\|_F^2) + \exp(-Ct/\|A\|)$$

$$\|\Pi z\|_2^2 = \sigma^\top B_z^\top B_z \sigma = \sigma^\top A_z \sigma$$

Proof of DJL (finally)

Hanson-Wright.

$$\mathbb{P}(|\sigma^\top A\sigma - \mathbb{E}\,\sigma^\top A\sigma| > t) \lesssim \exp(-Ct^2/\|A\|_F^2) + \exp(-Ct/\|A\|)$$

$$\|\Pi z\|_2^2 = \sigma^\top B_z^\top B_z \sigma = \sigma^\top A_z \sigma$$

A_z is block-diagonal with m blocks, each equal to $\frac{1}{m} zz^\top$

Proof of DJL (finally)

Hanson-Wright.
$\mathbb{P}(|\sigma^\top A\sigma - \mathbb{E}\,\sigma^\top A\sigma| > t) \lesssim \exp(-Ct^2/\|A\|_F^2) + \exp(-Ct/\|A\|)$

$\|\Pi z\|_2^2 = \sigma^\top B_z^\top B_z \sigma = \sigma^\top A_z \sigma$

A_z is block-diagonal with m blocks, each equal to $\frac{1}{m} zz^\top$

- $\|A_z\|_F^2 = \frac{1}{m}\|z\|_2^4$
- $\|A_z\| = \frac{1}{m}\|z\|_2^2$
- Apply Hanson-Wright with $t = \varepsilon\|z\|_2^2, m = C\varepsilon^{-2}\log(1/\delta)$

This talk

- What about other, non-Euclidean norms?
- Proof of DJL.
- What is the optimal target dimension m for DJL?
- What is the optimal target dimension m for JL?
- How fast can we apply the map $x \mapsto \Pi x$?
- Isn't storing the random $m \times d$ matrix Π expensive?
- What can we get for specific applications and instances?
- Generalizing JL: terminal embeddings.

Distributional JL (DJL) lemma

Lemma (DJL lemma [Johnson, Lindenstrauss '84])

For any $0 < \varepsilon, \delta < 1/2$ *and* $d \geq 1$ *there exists a distribution* $\mathcal{D}_{\varepsilon,\delta}$ *on* $\mathbb{R}^{m \times d}$ *for* $m = O(\varepsilon^{-2} \log(1/\delta))$ *such that for any* $z \in S^{d-1}$

$$\mathbb{P}_{\Pi \sim \mathcal{D}_{\varepsilon,\delta}} \left(\|\Pi z\|_2^2 \notin [1-\varepsilon, 1+\varepsilon] \right) < \delta.$$

Distributional JL (DJL) lemma

Lemma (DJL lemma [Johnson, Lindenstrauss '84])

For any $0 < \varepsilon, \delta < 1/2$ *and* $d \geq 1$ *there exists a distribution* $\mathcal{D}_{\varepsilon,\delta}$ *on* $\mathbb{R}^{m \times d}$ *for* $m = O(\varepsilon^{-2} \log(1/\delta))$ *such that for any* $z \in S^{d-1}$

$$\mathop{\mathbb{P}}_{\Pi \sim \mathcal{D}_{\varepsilon,\delta}} \left(\|\Pi z\|_2^2 \notin [1-\varepsilon, 1+\varepsilon] \right) < \delta.$$

Theorem (Jayram-Woodruff, 2011; Kane-Meka-N., 2011)

For DJL, $m = \min\{d, \Theta(\varepsilon^{-2} \log(1/\delta))\}$ *is optimal.*

DJL lower bound proof idea [Kane-Meka-N. 2011]

Suppose $\mathcal{D}_{\varepsilon,\delta}$ is a good DJL distribution on $\mathbb{R}^{m \times d}$.

DJL lower bound proof idea [Kane-Meka-N. 2011]

Suppose $\mathcal{D}_{\varepsilon,\delta}$ is a good DJL distribution on $\mathbb{R}^{m\times d}$.

$$\forall z \in S^{d-1} \quad \mathop{\mathbb{P}}_{\Pi\sim\mathcal{D}_{\varepsilon,\delta}}(|\|\Pi z\|_2^2 - 1| > \varepsilon) < \delta$$

$$\implies \mathop{\mathbb{P}}_{z\sim\mathcal{F}} \mathop{\mathbb{P}}_{\Pi\sim\mathcal{D}_{\varepsilon,\delta}}(|\|\Pi z\|_2^2 - 1| > \varepsilon) < \delta$$

$$\implies \mathop{\mathbb{P}}_{\Pi\sim\mathcal{D}_{\varepsilon,\delta}} \mathop{\mathbb{P}}_{z\sim\mathcal{F}}(|\|\Pi z\|_2^2 - 1| > \varepsilon) < \delta$$

$$\implies \exists \Pi \in \mathbb{R}^{m\times d} \mathop{\mathbb{P}}_{z\sim\mathcal{F}}(|\|\Pi z\|_2^2 - 1| > \varepsilon) < \delta$$

DJL lower bound proof idea [Kane-Meka-N. 2011]

Suppose $\mathcal{D}_{\varepsilon,\delta}$ is a good DJL distribution on $\mathbb{R}^{m\times d}$.

$$\forall z \in S^{d-1} \quad \mathop{\mathbb{P}}_{\Pi\sim\mathcal{D}_{\varepsilon,\delta}}(|\|\Pi z\|_2^2 - 1| > \varepsilon) < \delta$$

$$\implies \mathop{\mathbb{P}}_{z\sim\mathcal{F}} \mathop{\mathbb{P}}_{\Pi\sim\mathcal{D}_{\varepsilon,\delta}}(|\|\Pi z\|_2^2 - 1| > \varepsilon) < \delta$$

$$\implies \mathop{\mathbb{P}}_{\Pi\sim\mathcal{D}_{\varepsilon,\delta}} \mathop{\mathbb{P}}_{z\sim\mathcal{F}}(|\|\Pi z\|_2^2 - 1| > \varepsilon) < \delta$$

$$\implies \exists \Pi \in \mathbb{R}^{m\times d} \mathop{\mathbb{P}}_{z\sim\mathcal{F}}(|\|\Pi z\|_2^2 - 1| > \varepsilon) < \delta$$

(easy direction of "Yao's minimax principle")

DJL lower bound proof idea [Kane-Meka-N. 2011]

Suppose $\mathcal{D}_{\varepsilon,\delta}$ is a good DJL distribution on $\mathbb{R}^{m\times d}$.

$$\forall z \in S^{d-1} \quad \mathop{\mathbb{P}}_{\Pi\sim\mathcal{D}_{\varepsilon,\delta}}(|\|\Pi z\|_2^2 - 1| > \varepsilon) < \delta$$

$$\implies \mathop{\mathbb{P}}_{z\sim\mathcal{F}} \mathop{\mathbb{P}}_{\Pi\sim\mathcal{D}_{\varepsilon,\delta}}(|\|\Pi z\|_2^2 - 1| > \varepsilon) < \delta$$

$$\implies \mathop{\mathbb{P}}_{\Pi\sim\mathcal{D}_{\varepsilon,\delta}} \mathop{\mathbb{P}}_{z\sim\mathcal{F}}(|\|\Pi z\|_2^2 - 1| > \varepsilon) < \delta$$

$$\implies \exists \Pi \in \mathbb{R}^{m\times d} \mathop{\mathbb{P}}_{z\sim\mathcal{F}}(|\|\Pi z\|_2^2 - 1| > \varepsilon) < \delta$$

(easy direction of "Yao's minimax principle")

Then show that if $\mathcal{F}$ is the uniform distribution on the sphere and $m < d/2$, then the probability any fixed $\Pi \in \mathbb{R}^{m\times d}$ fails to preserve z is $\exp(-O(\varepsilon^2 m + 1)) \implies m = \Omega(\varepsilon^{-2}\log(1/\delta))$ to succeed.

This talk

- What about other, non-Euclidean norms?
- Proof of DJL.
- What is the optimal target dimension m for DJL?
- What is the optimal target dimension m for JL?
- How fast can we apply the map $x \mapsto \Pi x$?
- Isn't storing the random $m \times d$ matrix Π expensive?
- What can we get for specific applications and instances?
- Generalizing JL: terminal embeddings.

Lower bound techniques over time

Lower bounds over time

- **Volume argument.** $m = \Omega(\log n)$ [Johnson, Lindenstrauss '84]

Lower bounds over time

- **Volume argument.** $m = \Omega(\log n)$ [Johnson, Lindenstrauss '84]
- **Incoherence + tensor trick.** $m = \Omega(\frac{1}{\varepsilon^2} \frac{\log n}{\log(1/\varepsilon)})$ [Alon '03]

Lower bounds over time

- **Volume argument.** $m = \Omega(\log n)$ [Johnson, Lindenstrauss '84]
- **Incoherence + tensor trick.** $m = \Omega(\frac{1}{\varepsilon^2} \frac{\log n}{\log(1/\varepsilon)})$ [Alon '03]
- **Net argument + probabilistic method.** $m = \Omega(\frac{1}{\varepsilon^2} \log n)$ (only against linear maps $f(x) = \Pi x$) [Larsen, Nelson '16]

Lower bounds over time

- **Volume argument.** $m = \Omega(\log n)$ [Johnson, Lindenstrauss '84]
- **Incoherence + tensor trick.** $m = \Omega(\frac{1}{\varepsilon^2} \frac{\log n}{\log(1/\varepsilon)})$ [Alon '03]
- **Net argument + probabilistic method.** $m = \Omega(\frac{1}{\varepsilon^2} \log n)$ (only against linear maps $f(x) = \Pi x$) [Larsen, Nelson '16]
- **Encoding argument.** $m = \Omega(\frac{1}{\varepsilon^2} \log n)$ [Larsen, Nelson '17]

Optimal lower bound

Theorem ([Larsen, Nelson '17])

For any integers $d, n \geq 2$ and any $\frac{1}{(\min\{n,d\})^{0.4999}} < \varepsilon < 1$, there exists a set $X \subset \ell_2^d$ such that any embedding $f : X \to \ell_2^m$ with distortion at most $1 + \varepsilon$ must have

$$m = \Omega(\varepsilon^{-2} \log n)$$

Optimal lower bound

Theorem ([Larsen, Nelson '17])

For any integers $d, n \geq 2$ and any $\frac{1}{(\min\{n,d\})^{0.4999}} < \varepsilon < 1$, there exists a set $X \subset \ell_2^d$ such that any embedding $f : X \to \ell_2^m$ with distortion at most $1 + \varepsilon$ must have

$$m = \Omega(\varepsilon^{-2} \log n)$$

- **Can always achieve $m = d$:** f is the identity map.
- **Can always achieve $m = n - 1$:** translate so one vector is 0. Then all vectors live in $(n-1)$-dimensional subspace. Project.

Optimal lower bound

Theorem ([Larsen, Nelson '17])

For any integers $d, n \geq 2$ and any $\frac{1}{(\min\{n,d\})^{0.4999}} < \varepsilon < 1$, there exists a set $X \subset \ell_2^d$ such that any embedding $f : X \to \ell_2^m$ with distortion at most $1 + \varepsilon$ must have

$$m = \Omega(\varepsilon^{-2} \log n)$$

- **Can always achieve $m = d$:** f is the identity map.
- **Can always achieve $m = n - 1$:** translate so one vector is 0. Then all vectors live in $(n-1)$-dimensional subspace. Project.
- So can only hope JL optimal for $\varepsilon^{-2} \log n \leq \min\{n, d\}$, can view theorem assumption as $\varepsilon^{-2} \log n \ll \min\{n, d\}^{0.999}$.

Optimal lower bound:

An Encoding Argument.

JL is optimal even against non-linear maps

- We don't give explicit hard X; we just show one exists (compression argument / pigeonhole principle)

JL is optimal even against non-linear maps

- We don't give explicit hard X; we just show one exists (compression argument / pigeonhole principle)
- We define large collection $\mathcal{X}$ of n-sized sets $X \subset \mathbb{R}^d$ s.t. if all $X \in \mathcal{X}$ can be embedded into dimension $\leq 10^{-10} \cdot \varepsilon^{-2} \log_2 n$, then there is an encoding of elements of $\mathcal{X}$ into $< \log_2 |\mathcal{X}|$ bits (i.e. an injection from $\mathcal{X}$ to $\{0,1\}^t$ for $t < \log_2 |\mathcal{X}|$). **Contradiction.**

Encoding argument.

[Larsen, Nelson '17]

Encoding argument.

[Larsen, Nelson '17]

For now: assume $d = n/\lg(1/\varepsilon)$

Observation

Preserving distances implies preserving dot products.

Observation

Preserving distances implies preserving dot products.

- Say $\|x\|_2 = \|y\|_2 = 1$.

Observation

Preserving distances implies preserving dot products.

- Say $\|x\|_2 = \|y\|_2 = 1$.
- Can show that if ...
  - $\|f(x)\|_2 = (1 \pm \varepsilon)\|x\|_2$
  - $\|f(y)\|_2 = (1 \pm \varepsilon)\|y\|_2$
  - $\|f(x) - f(y))\|_2 = (1 \pm \varepsilon)\|x - y\|_2$

  Then $\langle f(x), f(y)\rangle = \langle x, y\rangle \pm O(\varepsilon)$

Observation

Preserving distances implies preserving dot products.

- Say $\|x\|_2 = \|y\|_2 = 1$.
- Can show that if ...
  - $\|f(x)\|_2 = (1 \pm \varepsilon)\|x\|_2$
  - $\|f(y)\|_2 = (1 \pm \varepsilon)\|y\|_2$ (true if dist to 0 must be preserved)
  - $\|f(x) - f(y))\|_2 = (1 \pm \varepsilon)\|x - y\|_2$

  Then $\langle f(x), f(y)\rangle = \langle x, y\rangle \pm O(\varepsilon)$

JL lower bound outline

- Pick $k = \frac{1}{100\varepsilon^2}$.
- For $S \subset [d]$ of size k, define indicator vector $y_S = \frac{1}{\sqrt{k}} \sum_{j \in S} e_j$.

$$\langle y_S, e_i \rangle = \begin{cases} 10\varepsilon, & i \in S \\ 0, & \text{otherwise} \end{cases}$$

- **Idea:** low-distortion embedding preserves dot products up to $\pm\varepsilon$, which is enough to distinguish $i \in S$ vs. $i \notin S$

JL lower bound outline

- Pick $k = \frac{1}{100\varepsilon^2}$.
- For $S \subset [d]$ of size k, define indicator vector $y_S = \frac{1}{\sqrt{k}} \sum_{j \in S} e_j$.

$$\langle y_S, e_i \rangle = \begin{cases} 10\varepsilon, & i \in S \\ 0, & \text{otherwise} \end{cases}$$

- **Idea:** low-distortion embedding preserves dot products up to $\pm\varepsilon$, which is enough to distinguish $i \in S$ vs. $i \notin S$
- $\mathcal{X}$ is set of all ordered tuples of points, possibly with repetition $X = (0, e_1, \ldots, e_d, y_{S_1}, \ldots, y_{S_{n-d-1}})$ with the $S_i \in \binom{[d]}{k}$.

JL lower bound outline

- Pick $k = \frac{1}{100\varepsilon^2}$.
- For $S \subset [d]$ of size k, define indicator vector $y_S = \frac{1}{\sqrt{k}} \sum_{j \in S} e_j$.

$$\langle y_S, e_i \rangle = \begin{cases} 10\varepsilon, & i \in S \\ 0, & \text{otherwise} \end{cases}$$

- **Idea:** low-distortion embedding preserves dot products up to $\pm\varepsilon$, which is enough to distinguish $i \in S$ vs. $i \notin S$
- $\mathcal{X}$ is set of all ordered tuples of points, possibly with repetition $X = (0, e_1, \ldots, e_d, y_{S_1}, \ldots, y_{S_{n-d-1}})$ with the $S_i \in \binom{[d]}{k}$.
- $|\mathcal{X}| = \binom{d}{k}^{n-d-1}$, thus any encoding of $X \in \mathcal{X}$ requires $\geq (n-d-1) \lg \binom{d}{k} = (1 - o_\varepsilon(1)) nk \lg(d/k)$ bits.

JL lower bound outline

- Pick $k = \frac{1}{100\varepsilon^2}$.
- For $S \subset [d]$ of size k, define indicator vector $y_S = \frac{1}{\sqrt{k}} \sum_{j \in S} e_j$.

$$\langle y_S, e_i \rangle = \begin{cases} 10\varepsilon, & i \in S \\ 0, & \text{otherwise} \end{cases}$$

- **Idea:** low-distortion embedding preserves dot products up to $\pm\varepsilon$, which is enough to distinguish $i \in S$ vs. $i \notin S$
- $\mathcal{X}$ is set of all ordered tuples of points, possibly with repetition $X = (0, e_1, \ldots, e_d, y_{S_1}, \ldots, y_{S_{n-d-1}})$ with the $S_i \in \binom{[d]}{k}$.
- $|\mathcal{X}| = \binom{d}{k}^{n-d-1}$, thus any encoding of $X \in \mathcal{X}$ requires $\geq (n-d-1) \lg \binom{d}{k} = (1 - o_\varepsilon(1)) nk \lg(d/k)$ bits.
- Will show any $(1+\varepsilon)$-distortion embedding into ℓ_2^m implies encoding into $O(nm)$ bits, hence $nm = \Omega(nk \log(d/k))$ $\Rightarrow m = \Omega(k \log(d/k)) = \Omega(\varepsilon^{-2} \log n)$ for ε not too small.

Problem: Encoding of $X \in \mathcal{X}$ can't just be a description of $f(0), f(e_1), \ldots, f(e_d), f(y_{S_1}), \ldots, f(y_{S_{n-d-1}})$.

Why not?

Problem: Encoding of $X \in \mathcal{X}$ can't just be a description of $f(0), f(e_1), \ldots, f(e_d), f(y_{S_1}), \ldots, f(y_{S_{n-d-1}})$.

Why not? Want to violate pigeonhole principle, so range of the encoding must be of size $< \lg |\mathcal{X}|$. But $f(x)$ has real entries, so the range is infinite!

Problem: Encoding of $X \in \mathcal{X}$ can't just be a description of $f(0), f(e_1), \ldots, f(e_d), f(y_{S_1}), \ldots, f(y_{S_{n-d-1}})$.

Why not? Want to violate pigeonhole principle, so range of the encoding must be of size $< \lg |\mathcal{X}|$. But $f(x)$ has real entries, so the range is infinite!

Fix attempt 1: Round entries of $f(x)$ to integer multiples of γ. Can show $\gamma = O(\frac{\varepsilon}{\sqrt{m}})$ suffices $\Longrightarrow$ $O(nm \log(m/\varepsilon))$ bit encoding

Problem: Encoding of $X \in \mathcal{X}$ can't just be a description of $f(0), f(e_1), \ldots, f(e_d), f(y_{S_1}), \ldots, f(y_{S_{n-d-1}})$.

Why not? Want to violate pigeonhole principle, so range of the encoding must be of size $< \lg |\mathcal{X}|$. But $f(x)$ has real entries, so the range is infinite!

Fix attempt 1: Round entries of $f(x)$ to integer multiples of γ. Can show $\gamma = O(\frac{\varepsilon}{\sqrt{m}})$ suffices $\implies O(nm \log(m/\varepsilon))$ bit encoding $\implies m = \Omega(\varepsilon^{-2} \frac{\log n}{\log(m/\varepsilon)})$ final lower bound

Problem: Encoding of $X \in \mathcal{X}$ can't just be a description of $f(0), f(e_1), \ldots, f(e_d), f(y_{S_1}), \ldots, f(y_{S_{n-d-1}})$.

Why not? Want to violate pigeonhole principle, so range of the encoding must be of size $< \lg |\mathcal{X}|$. But $f(x)$ has real entries, so the range is infinite!

Fix attempt 1: Round entries of $f(x)$ to integer multiples of γ. Can show $\gamma = O(\frac{\varepsilon}{\sqrt{m}})$ suffices $\Longrightarrow O(nm \log(m/\varepsilon))$ bit encoding $\Longrightarrow m = \Omega(\varepsilon^{-2} \frac{\log n}{\log(m/\varepsilon)})$ final lower bound

Slightly better fix: Round each $f(x)$ to a point $\widetilde{f(x)}$ in an ε-net under ℓ_2.

Problem: Encoding of $X \in \mathcal{X}$ can't just be a description of $f(0), f(e_1), \ldots, f(e_d), f(y_{S_1}), \ldots, f(y_{S_{n-d-1}})$.

Why not? Want to violate pigeonhole principle, so range of the encoding must be of size $< \lg |\mathcal{X}|$. But $f(x)$ has real entries, so the range is infinite!

Fix attempt 1: Round entries of $f(x)$ to integer multiples of γ. Can show $\gamma = O(\frac{\varepsilon}{\sqrt{m}})$ suffices $\Longrightarrow O(nm \log(m/\varepsilon))$ bit encoding $\Longrightarrow m = \Omega(\varepsilon^{-2} \frac{\log n}{\log(m/\varepsilon)})$ final lower bound

Slightly better fix: Round each $f(x)$ to a point $\widetilde{f(x)}$ in an ε-net under ℓ_2. ε-net has size $O(1/\varepsilon)^m$, so $O(nm \log(1/\varepsilon))$ bit encoding $\Longrightarrow m = \Omega(\varepsilon^{-2} \frac{\log n}{\log(1/\varepsilon)})$

Problem: Encoding of $X \in \mathcal{X}$ can't just be a description of $f(0), f(e_1), \ldots, f(e_d), f(y_{S_1}), \ldots, f(y_{S_{n-d-1}})$.

Why not? Want to violate pigeonhole principle, so range of the encoding must be of size $< \lg |\mathcal{X}|$. But $f(x)$ has real entries, so the range is infinite!

Fix attempt 1: Round entries of $f(x)$ to integer multiples of γ. Can show $\gamma = O(\frac{\varepsilon}{\sqrt{m}})$ suffices $\Longrightarrow O(nm \log(m/\varepsilon))$ bit encoding $\Longrightarrow m = \Omega(\varepsilon^{-2} \frac{\log n}{\log(m/\varepsilon)})$ final lower bound

Slightly better fix: Round each $f(x)$ to a point $\widetilde{f(x)}$ in an ε-net under ℓ_2. ε-net has size $O(1/\varepsilon)^m$, so $O(nm \log(1/\varepsilon))$ bit encoding $\Longrightarrow m = \Omega(\varepsilon^{-2} \frac{\log n}{\log(1/\varepsilon)})$
(same lower bound as [Alon'03] but totally different argument!)

Problem: Encoding of $X \in \mathcal{X}$ can't just be a description of $f(0), f(e_1), \ldots, f(e_d), f(y_{S_1}), \ldots, f(y_{S_{n-d-1}})$.

Why not? Want to violate pigeonhole principle, so range of the encoding must be of size $< \lg |\mathcal{X}|$. But $f(x)$ has real entries, so the range is infinite!

Fix attempt 1: Round entries of $f(x)$ to integer multiples of γ. Can show $\gamma = O(\frac{\varepsilon}{\sqrt{m}})$ suffices $\Longrightarrow O(nm \log(m/\varepsilon))$ bit encoding $\Longrightarrow m = \Omega(\varepsilon^{-2} \frac{\log n}{\log(m/\varepsilon)})$ final lower bound

Slightly better fix: Round each $f(x)$ to a point $\widetilde{f(x)}$ in an ε-net under ℓ_2. ε-net has size $O(1/\varepsilon)^m$, so $O(nm \log(1/\varepsilon))$ bit encoding $\Longrightarrow m = \Omega(\varepsilon^{-2} \frac{\log n}{\log(1/\varepsilon)})$
(same lower bound as [Alon'03] but totally different argument!)

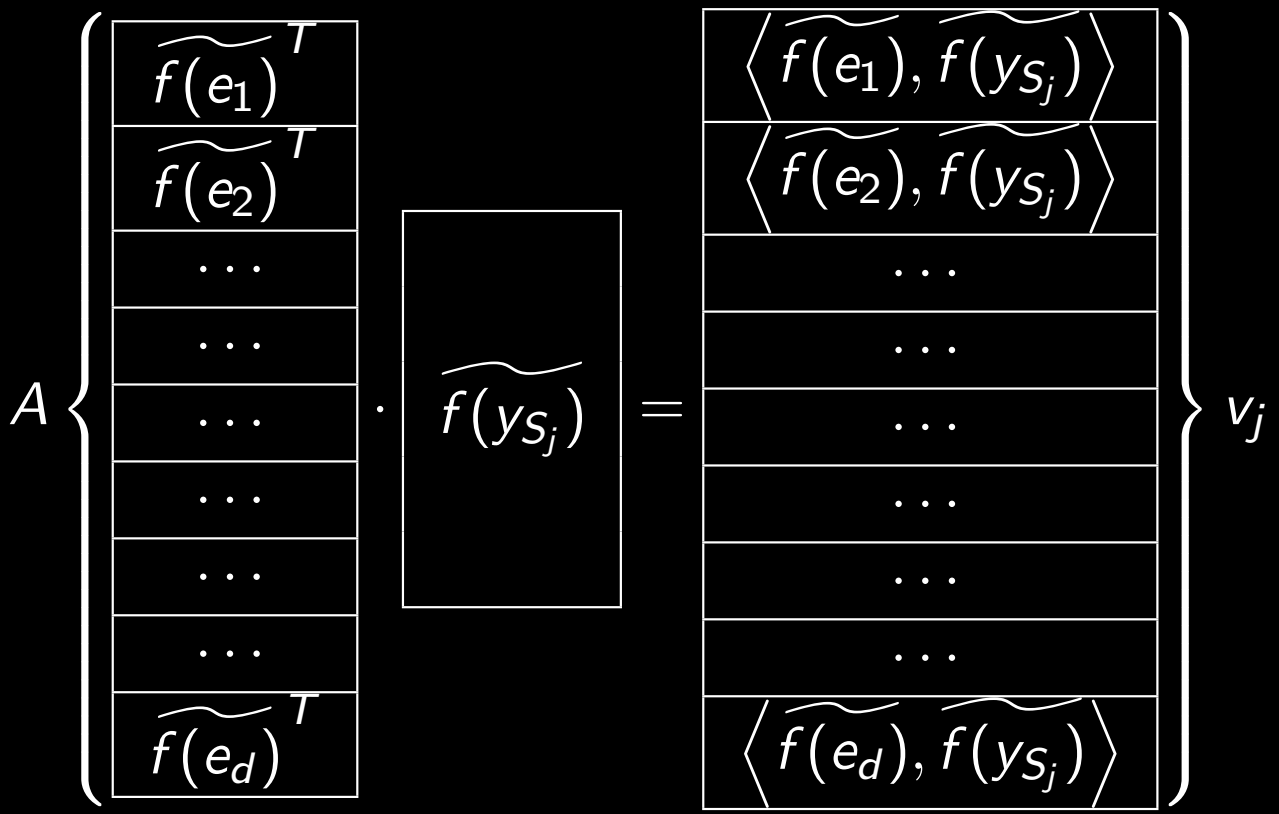
Next: a better encoding (when $d = n/\log(1/\varepsilon)$)

An encoding of X into $O(nm)$ bits

Sufficed for decoding X: knowing $\left\langle \widetilde{f(e_i)}, \widetilde{f(y_{S_j})} \right\rangle$ for each i, j

An encoding of X into $O(nm)$ bits

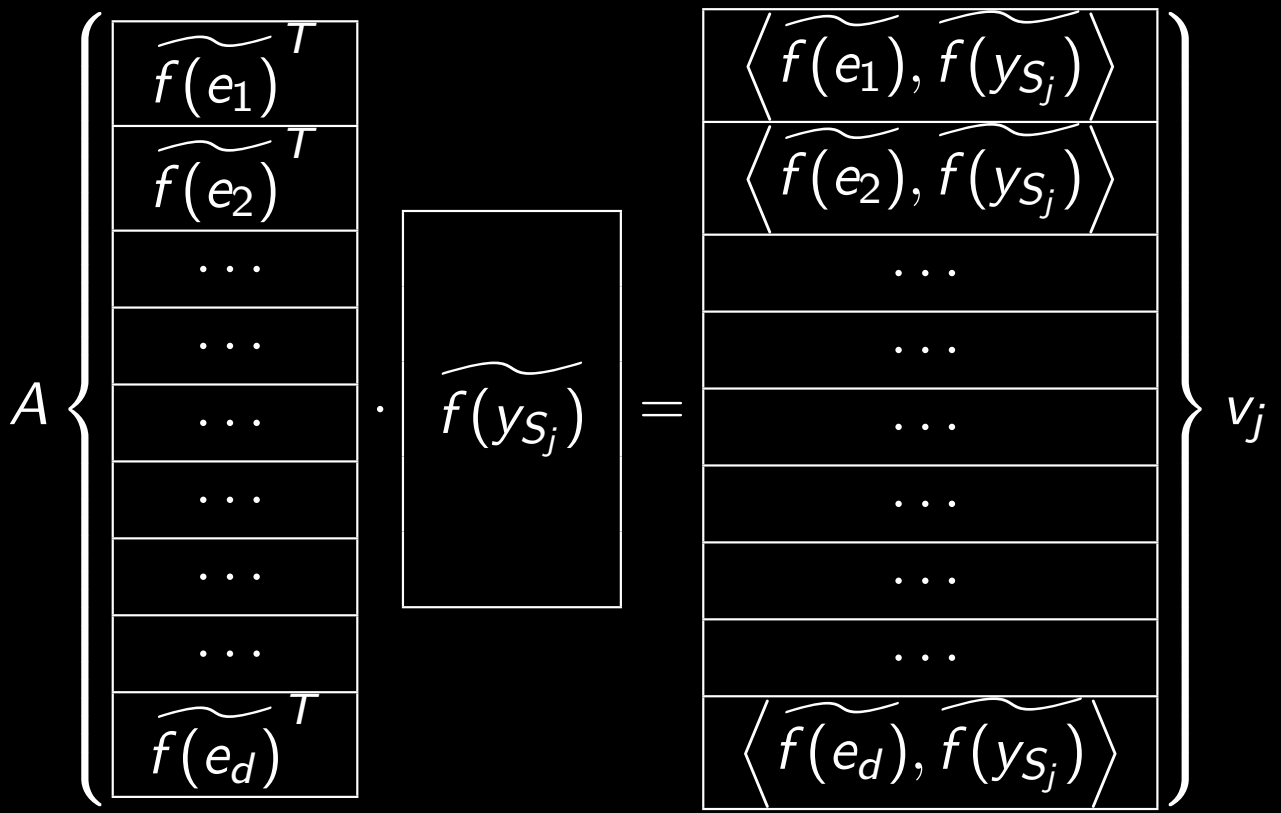
Sufficed for decoding X: knowing $\left\langle \widetilde{f(e_i)}, \widetilde{f(y_{S_j})} \right\rangle$ for each i, j



- Knowing $v_1, \ldots, v_{n-d-1}$ would allow us to decode.

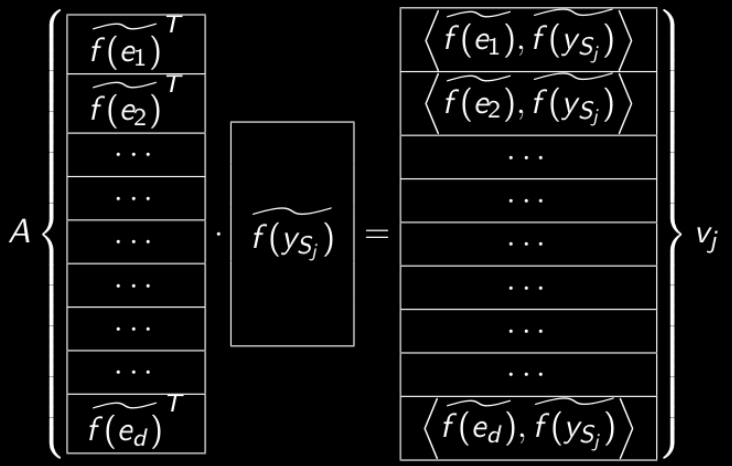
An encoding of X into $O(nm)$ bits

Sufficed for decoding X: knowing $\left\langle \widetilde{f(e_i)}, \widetilde{f(y_{S_j})} \right\rangle$ for each i, j



- Knowing $v_1, \ldots, v_{n-d-1}$ would allow us to decode.
- In fact, suffices to know $\tilde{v}_j$ such that $\|v_j - \tilde{v}_j\|_\infty < \varepsilon$.
  (then each entry of $\tilde{v}_j$ is $< 3\varepsilon$ or $> 7\varepsilon$ in magnitude)

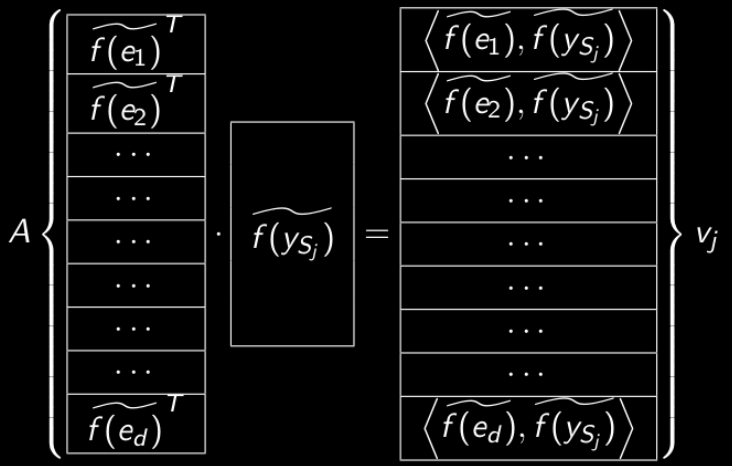
An encoding of X into $O(nm)$ bits



- Let E denote the column space of A $dim(E) \leq m$.

- $A \in \mathbb{R}^{d \times m}$
- $\widetilde{f(ys_j)} \in \mathbb{R}^m$
- $v_j \in \mathbb{R}^d$

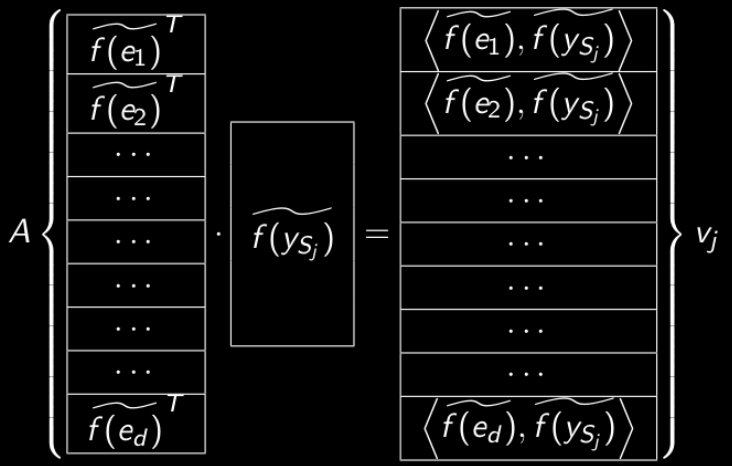
An encoding of X into $O(nm)$ bits



- Let E denote the column space of A $dim(E) \leq m$.
- Define $K = E \cap (13\varepsilon B_{\ell_\infty^d})$, $\forall j$ $v_j \in K$

- $A \in \mathbb{R}^{d \times m}$
- $\widetilde{f(ys_j)} \in \mathbb{R}^m$
- $v_j \in \mathbb{R}^d$

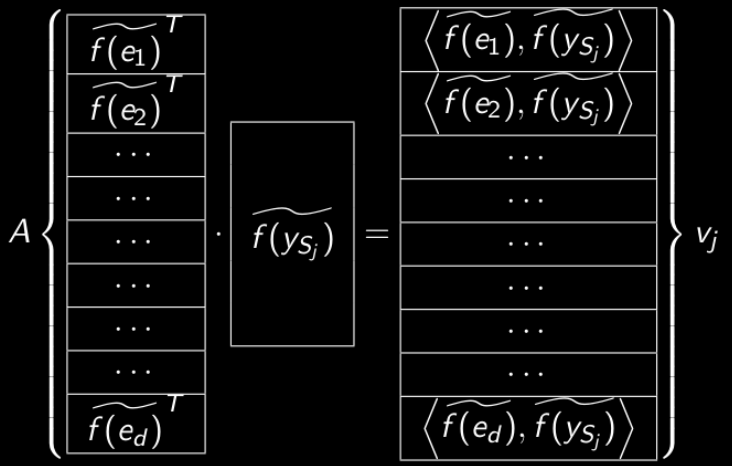
An encoding of X into $O(nm)$ bits



- Let E denote the column space of A $dim(E) \leq m$.
- Define $K = E \cap (13\varepsilon B_{\ell_\infty^d})$, $\forall j$ $v_j \in K$
- K has $\frac{1}{13}$-net in K-norm of size $\leq 2^{O(m)}$

- $A \in \mathbb{R}^{d \times m}$
- $\widetilde{f(y_{S_j})} \in \mathbb{R}^m$
- $v_j \in \mathbb{R}^d$

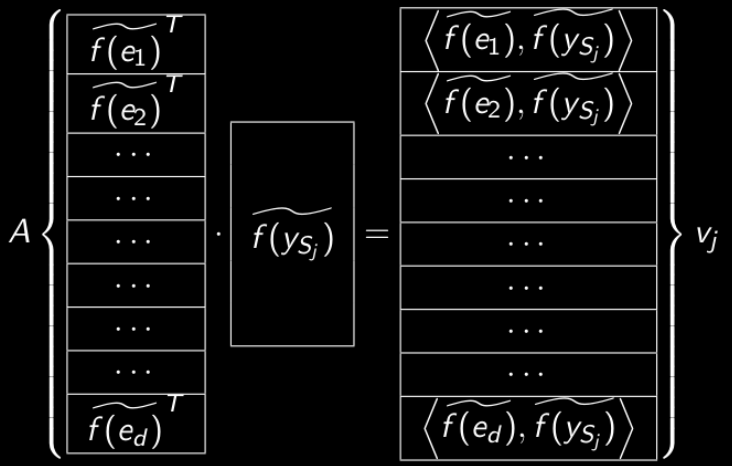
An encoding of X into $O(nm)$ bits



- $A \in \mathbb{R}^{d \times m}$
- $\widetilde{f(y_{S_j})} \in \mathbb{R}^m$
- $v_j \in \mathbb{R}^d$

- Let E denote the column space of A
  $dim(E) \leq m$.
- Define $K = E \cap (13\varepsilon B_{\ell_\infty^d})$, $\forall j$ $v_j \in K$
- K has $\frac{1}{13}$-net in K-norm of size $\leq 2^{O(m)}$
- Define $\tilde{v}_j$ as closest point in this net to v_j
  $\implies \|v_j - \tilde{v}_j\|_\infty < \varepsilon$.
  $O(m)$ bits to specify $\tilde{v}_j$.

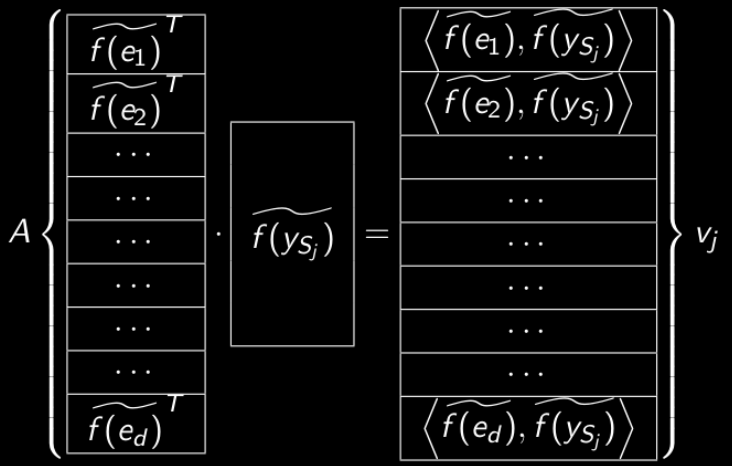
An encoding of X into $O(nm)$ bits



- $A \in \mathbb{R}^{d \times m}$
- $\widetilde{f(y_{S_j})} \in \mathbb{R}^m$
- $v_j \in \mathbb{R}^d$

- Let E denote the column space of A
  $dim(E) \leq m$.
- Define $K = E \cap (13\varepsilon B_{\ell_\infty^d})$, $\forall j$ $v_j \in K$
- K has $\frac{1}{13}$-net in K-norm of size $\leq 2^{O(m)}$
- Define $\tilde{v}_j$ as closest point in this net to v_j
  $\Longrightarrow \|v_j - \tilde{v}_j\|_\infty < \varepsilon$.
  $O(m)$ bits to specify $\tilde{v}_j$.
- Encoding needs to specify E (i.e. A).
  Encode $\widetilde{f(e_i)}$ using $O(m \log(1/\varepsilon))$ bits
  For the e_i: $O(dm \lg(1/\varepsilon)) = O(nm)$ bits

An encoding of X into $O(nm)$ bits



- $A \in \mathbb{R}^{d \times m}$
- $\widetilde{f(y_{S_j})} \in \mathbb{R}^m$
- $v_j \in \mathbb{R}^d$

- Let E denote the column space of A
  $dim(E) \leq m$.
- Define $K = E \cap (13\varepsilon B_{\ell_\infty^d})$, $\forall j$ $v_j \in K$
- K has $\frac{1}{13}$-net in K-norm of size $\leq 2^{O(m)}$
- Define $\tilde{v}_j$ as closest point in this net to v_j
  $\implies \|v_j - \tilde{v}_j\|_\infty < \varepsilon$.
  $O(m)$ bits to specify $\tilde{v}_j$.
- Encoding needs to specify E (i.e. A).
  Encode $\widetilde{f(e_i)}$ using $O(m \log(1/\varepsilon))$ bits
  For the e_i: $O(dm \lg(1/\varepsilon)) = O(nm)$ bits
- **Total:** $O(nm)$ bit encoding

QED

QED

What about when $d \neq n/\lg(1/\varepsilon)$?

Extending to arbitrary d, n

- Suppose $X \subset \ell_2^{d'}$, $|X| = n$, is a hard set for some ε where $d' = \Theta(n/\log(1/\varepsilon))$ (X has $\Omega(\varepsilon^{-2} \log n)$ lower bound).

Extending to arbitrary d, n

- Suppose $X \subset \ell_2^{d'}$, $|X| = n$, is a hard set for some ε where $d' = \Theta(n/\log(1/\varepsilon))$ (X has $\Omega(\varepsilon^{-2} \log n)$ lower bound).
- For $d > d'$: zero-pad vectors in X; still hard

Extending to arbitrary d, n

- Suppose $X \subset \ell_2^{d'}$, $|X| = n$, is a hard set for some ε where $d' = \Theta(n/\log(1/\varepsilon))$ (X has $\Omega(\varepsilon^{-2}\log n)$ lower bound).
- For $d > d'$: zero-pad vectors in X; still hard
- For $C\varepsilon^{-2}\log n \leq d < d'$: let $f : X \to \ell_2^d$ be low-distortion map for X (JL lemma). Then $f(X)$ must be hard to embed into $o(\varepsilon^{-2}\log n)$ dimension, else X wouldn't be hard.

OPEN: later [Alon-Klartag'17] showed lower bound of $\Omega(\min\{n, d, \varepsilon^{-2}\log(\varepsilon^2 n)\})$ for full range of ε. Is there a matching upper bound for $\varepsilon \to 1/\sqrt{n}$?

This talk

- What about other, non-Euclidean norms?
- Proof of DJL.
- What is the optimal target dimension m for DJL?
- What is the optimal target dimension m for JL?
- How fast can we apply the map $x \mapsto \Pi x$?
- Isn't storing the random $m \times d$ matrix Π expensive?
- What can we get for specific applications and instances?
- Generalizing JL: terminal embeddings.

The DJL distribution over Π

Older proofs

- [Johnson-Lindenstrauss, 1984], [Frankl-Maehara, 1988], [Dasgupta-Gupta, 2003]: Random rotation, then projection onto first m coordinates.
- [Indyk-Motwani, 1998]:
  Random matrix with independent Gaussian entries.
- [Achlioptas, 2001]: Independent ± 1 entries.
- [Clarkson-Woodruff, 2009]:
  $O(\log(1/\delta))$-wise independent ± 1 entries.
- [Matousek, 2008]:
  Independent entries having mean 0, variance $1/m$, and subGaussian tails

The DJL distribution over Π

Older proofs

- [Johnson-Lindenstrauss, 1984], [Frankl-Maehara, 1988], [Dasgupta-Gupta, 2003]: Random rotation, then projection onto first m coordinates.
- [Indyk-Motwani, 1998]:
  Random matrix with independent Gaussian entries.
- [Achlioptas, 2001]: Independent ± 1 entries.
- [Clarkson-Woodruff, 2009]:
  $O(\log(1/\delta))$-wise independent ± 1 entries.
- [Matousek, 2008]:
  Independent entries having mean 0, variance $1/m$, and subGaussian tails

Downside: Performing embedding is dense matrix-vector multiplication, $O(m \cdot \|x\|_0)$ time

Fast JL Transforms

[Ailon-Chazelle'06]: Starting point: if you create a vector y by sampling m coordinates of x, then $\mathbb{E}\,\frac{1}{m}\|y\|_2^2 = \|x\|_2^2$.

$$\underbrace{\begin{bmatrix} 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}}_{S} \begin{bmatrix} \\ \\ x \\ \\ \\ \end{bmatrix} = \begin{bmatrix} \\ \\ y \\ \\ \end{bmatrix}$$

Fast JL Transforms

[Ailon-Chazelle'06]: Starting point: if you create a vector y by sampling m coordinates of x, then $\mathbb{E}\,\frac{1}{m}\|y\|_2^2 = \|x\|_2^2$.

$$\underbrace{\begin{bmatrix} 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}}_{S} \begin{bmatrix} \\ x \\ \\ \end{bmatrix} = \begin{bmatrix} \\ y \\ \\ \end{bmatrix}$$

Problem: High variance. What if x has only one non-zero entry?

Uncertainty Principle

$$\sigma_x \sigma_p \geq \frac{\hbar}{2}$$

Both x and $\hat{x}$ cannot be concentrated in few coordinates

Example: If $x = (1, 0, \ldots, 0)$ then $|\hat{x}_i| = \frac{1}{\sqrt{d}}$ for all i

Both x and $\hat{x}$ cannot be concentrated in few coordinates

Example: If $x = (1, 0, \ldots, 0)$ then $|\hat{x}_i| = \frac{1}{\sqrt{d}}$ for all i

$$y = SFx$$

(F is the DFT, or any "bounded orthonormal system"))

??

Both x and $\hat{x}$ cannot be concentrated in few coordinates

Example: If $x = (1, 0, \ldots, 0)$ then $|\hat{x}_i| = \frac{1}{\sqrt{d}}$ for all i

$$y = SFx$$

(F is the DFT, or any "bounded orthonormal system"))

??

New problem: What if x already well-spread?

Both x and $\hat{x}$ cannot be concentrated in few coordinates

Example: If $x = (1, 0, \ldots, 0)$ then $|\hat{x}_i| = \frac{1}{\sqrt{d}}$ for all i

$$y = SFx$$

(F is the DFT, or any "bounded orthonormal system"))

??

New problem: What if x already well-spread?

Ailon-Chazelle'06 fix:

$$y = SFDx$$

$$D = \begin{bmatrix} \pm 1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & \pm 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & \pm 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \pm 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \pm 1 \end{bmatrix}$$

Can show: with high probability, FDx is well-spread

Both x and $\hat{x}$ cannot be concentrated in few coordinates

Example: If $x = (1, 0, \ldots, 0)$ then $|\hat{x}_i| = \frac{1}{\sqrt{d}}$ for all i

$$y = SFx$$

(F is the DFT, or any "bounded orthonormal system"))

??

New problem: What if x already well-spread?

Ailon-Chazelle'06 fix:

$$y = SFDx$$

$$D = \begin{bmatrix} \pm 1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & \pm 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & \pm 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \pm 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \pm 1 \end{bmatrix}$$

Can show: with high probability, FDx is well-spread

End result: $x \mapsto f(x)$ in $O(d \log d)$ time (FFT)

- [Ailon-Chazelle, 2006]: $x \mapsto \frac{1}{\sqrt{m}} SFDx$, $O(d \log d + m^3)$ time

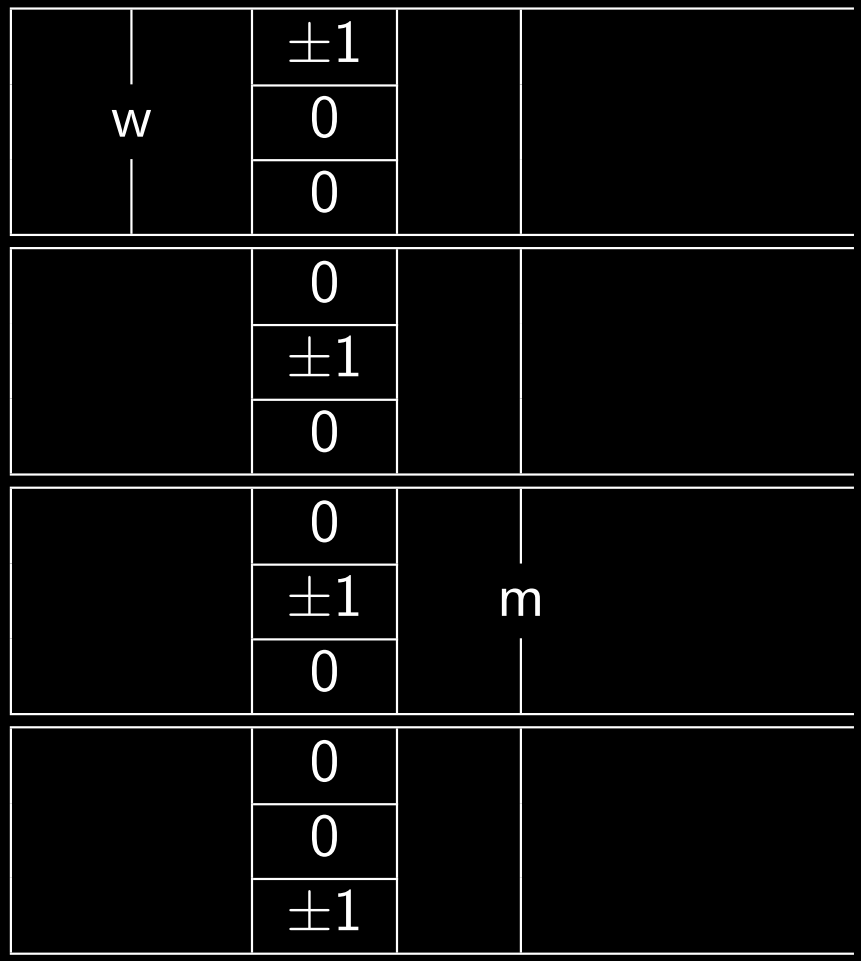
- [Ailon-Chazelle, 2006]: $x \mapsto \frac{1}{\sqrt{m}} SFDx$, $O(d \log d + m^3)$ time
- Several follow-up works improving the m^3: [Ailon-Liberty'08], [Ailon-Liberty'11], [Krahmer-Ward'11], [Rudelson-Vershynin'08], [Cheraghchi-Guruswami-Velingker'13], [N.-Price-Wootters'14], [Bourgain'14], [Haviv-Regev'16]

- [Ailon-Chazelle, 2006]: $x \mapsto \frac{1}{\sqrt{m}} SFDx$, $O(d \log d + m^3)$ time
- Several follow-up works improving the m^3: [Ailon-Liberty'08], [Ailon-Liberty'11], [Krahmer-Ward'11], [Rudelson-Vershynin'08], [Cheraghchi-Guruswami-Velingker'13], [N.-Price-Wootters'14], [Bourgain'14], [Haviv-Regev'16]

Downside: Slow to embed sparse vectors: running time is $\Omega(\min\{m \cdot \|x\|_0, d \log d\})$.

CountSketch [Charikar-Chen-FarachColton'02]

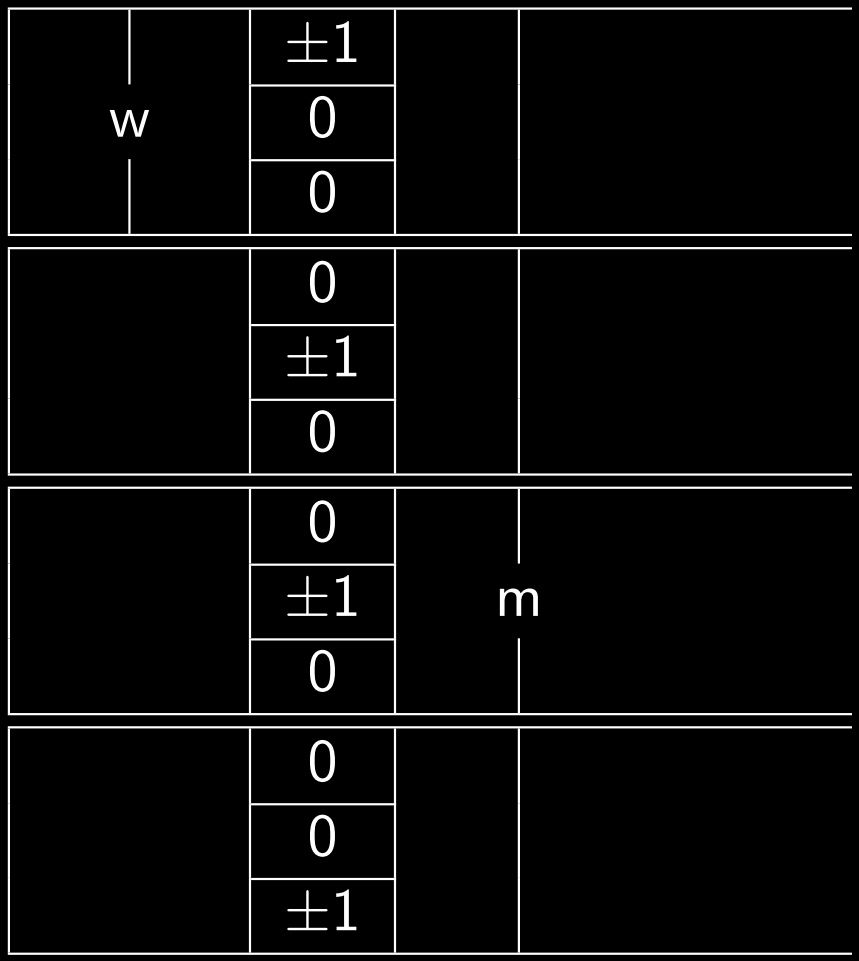
$$\Pi = \frac{1}{\sqrt{s}} \times$$



- partition m rows into s blocks of size $w = \frac{m}{s}$ each
- each column has exactly one $\frac{\pm 1}{\sqrt{s}}$ per block in random location

CountSketch [Charikar-Chen-FarachColton'02]

$$\Pi = \frac{1}{\sqrt{s}} \times$$



- **Note:** can map $x \mapsto \Pi x$ in time $O(s \cdot \|x\|_0)$.
- [Kane-N.'14] shows $m = O(\varepsilon^{-2} \log n)$, $s = O(\varepsilon m)$ suffices. [N.-Nguyễn'13] shows for this m, such s is *almost necessary*.
- See also [Bourgain-Dirksen-N.'15].

Sparse JL transforms

$s = \#$non-zero entries per column in embedding matrix
(so embedding time is $s \cdot \|x\|_0$)

reference	value of s	type
[JL84], [FM88], [IM98], ...	$m \approx 4\varepsilon^{-2} \log n$	dense
[Achlioptas'01]	$m/3$	sparse Bernoulli
[WeinbergerDALS'09]	no proof	hashing
[Dasgupta-Kumar-Sarlós'10]	$\tilde{O}(\varepsilon^{-1} \log^3 n)$	hashing
[Kane-Nelson10]*, [BravermanOR'10]*	$\tilde{O}(\varepsilon^{-1} \log^2 n)$	"
[Kane-Nelson'14]	$O(\varepsilon^{-1} \log n)$	CountSketch

* see also recent improvements by [Dahlgaard-Knudsen-Thorup'17], [Freksen-Kamma-Larsen'18], [Jagadeesan'19].

Analyzing Sparse JL

Analyzing Sparse JL

Recall for non-sparse JL:

$$\Pi z = B_z \sigma$$

where B_z is the $md \times md$ matrix

$$B_z = \frac{1}{\sqrt{m}} \begin{bmatrix} z^\top & 0 & 0 & \dots & 0 \\ 0 & z^\top & 0 & \dots & 0 \\ 0 & 0 & z^\top & \dots & 0 \\ 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \dots & z^\top \end{bmatrix}$$

Analyzing Sparse JL

Recall for non-sparse JL:

$$\Pi z = B_z \sigma$$

where B_z is the $md \times md$ matrix

$$B_z = \frac{1}{\sqrt{m}} \begin{bmatrix} z^\top & 0 & 0 & \dots & 0 \\ 0 & z^\top & 0 & \dots & 0 \\ 0 & 0 & z^\top & \dots & 0 \\ 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \dots & z^\top \end{bmatrix}$$

so that $\|\Pi z\|_2^2 = \underbrace{\sigma^\top B_z^\top B_z \sigma}_{\text{quadratic form}}$

Analyzing Sparse JL

Recall for non-sparse JL:

$$\Pi z = B_z \sigma$$

where B_z is the $md \times md$ matrix

$$B_z = \frac{1}{\sqrt{m}} \begin{bmatrix} z^\top & 0 & 0 & \dots & 0 \\ 0 & z^\top & 0 & \dots & 0 \\ 0 & 0 & z^\top & \dots & 0 \\ 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \dots & z^\top \end{bmatrix}$$

so that $\|\Pi z\|_2^2 = \underbrace{\sigma^\top B_z^\top B_z \sigma}_{\text{quadratic form}}$

For SparseJL, let $\eta_{r,i}$ be indicator random variable for event $\Pi_{r,i} \neq 0$. Then in ith row of B_z, replace $z^\top$ by $(\eta_r \odot z)^\top$ ($\odot$ is entry-wise product).

Analyzing Sparse JL

Recall for non-sparse JL:

$$\Pi z = B_z \sigma$$

where B_z is the $md \times md$ matrix

$$B_z = \frac{1}{\sqrt{m}} \begin{bmatrix} z^\top & 0 & 0 & \dots & 0 \\ 0 & z^\top & 0 & \dots & 0 \\ 0 & 0 & z^\top & \dots & 0 \\ 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \dots & z^\top \end{bmatrix}$$

so that $\|\Pi z\|_2^2 = \underbrace{\sigma^\top B_z^\top B_z \sigma}_{\text{quadratic form}}$

For SparseJL, let $\eta_{r,i}$ be indicator random variable for event $\Pi_{r,i} \neq 0$. Then in ith row of B_z, replace $z^\top$ by $(\eta_r \odot z)^\top$ ($\odot$ is entry-wise product).

Do Hanson-Wright as before and can still bound $\|A_z\| \leq \frac{1}{m}\|z\|_2^2$. $\|A_z\|_F^2$ is now a random variable, but can bound its moments.

This talk

- What about other, non-Euclidean norms?
- Proof of DJL.
- What is the optimal target dimension m for DJL?
- What is the optimal target dimension m for JL?
- How fast can we apply the map $x \mapsto \Pi x$?
- Isn't storing the random $m \times d$ matrix Π expensive?
- What can we get for specific applications and instances?
- Generalizing JL: terminal embeddings.

Representing $\Pi \in \mathbb{R}^{m \times d}$ space-efficiently

- [Achlioptas'01]: $\Pi_{i,j}$ can be i.i.d. $\sigma_{i,j}/\sqrt{m}$ for $\sigma_{i,j} = \pm 1$.

Representing $\Pi \in \mathbb{R}^{m \times d}$ space-efficiently

- [Achlioptas'01]: $\Pi_{i,j}$ can be i.i.d. $\sigma_{i,j}/\sqrt{m}$ for $\sigma_{i,j} = \pm 1$.
- [Clarkson-Woodruff'09]: in fact the $\sigma_{i,j}$ only need be $O(\log(1/\delta))$-wise independent. Combined with [Carter-Wegman'79], this implies Π can be stored using $O(\log(1/\delta)\log(md)) = O(\log(1/\delta)\log d)$ bits.

Representing $\Pi \in \mathbb{R}^{m \times d}$ space-efficiently

- [Achlioptas'01]: $\Pi_{i,j}$ can be i.i.d. $\sigma_{i,j}/\sqrt{m}$ for $\sigma_{i,j} = \pm 1$.
- [Clarkson-Woodruff'09]: in fact the $\sigma_{i,j}$ only need be $O(\log(1/\delta))$-wise independent. Combined with [Carter-Wegman'79], this implies Π can be stored using $O(\log(1/\delta)\log(md)) = O(\log(1/\delta)\log d)$ bits.
- [Kane-Meka-N.'11]: can construct Π as a product of $O(\log\log(1/\delta) + \log(1/\varepsilon))$ matrices using growing amounts of independence and gradually decreasing number of rows

Representing $\Pi \in \mathbb{R}^{m \times d}$ space-efficiently

- [Achlioptas'01]: $\Pi_{i,j}$ can be i.i.d. $\sigma_{i,j}/\sqrt{m}$ for $\sigma_{i,j} = \pm 1$.
- [Clarkson-Woodruff'09]: in fact the $\sigma_{i,j}$ only need be $O(\log(1/\delta))$-wise independent. Combined with [Carter-Wegman'79], this implies Π can be stored using $O(\log(1/\delta)\log(md)) = O(\log(1/\delta)\log d)$ bits.
- [Kane-Meka-N.'11]: can construct Π as a product of $O(\log\log(1/\delta) + \log(1/\varepsilon))$ matrices using growing amounts of independence and gradually decreasing number of rows

  Total number of bits:
  $O(\log d + \log(1/\delta)(\log\log(1/\delta) + \log(1/\varepsilon)))$.

  OPEN: $O(\log d + \log(1/\delta))$**?**

This talk

- What about other, non-Euclidean norms?
- Proof of DJL.
- What is the optimal target dimension m for DJL?
- What is the optimal target dimension m for JL?
- How fast can we apply the map $x \mapsto \Pi x$?
- Isn't storing the random $m \times d$ matrix Π expensive?
- What can we get for specific applications and instances?
- Generalizing JL: terminal embeddings.

Sketch-and-solve

[Sarlós'06]

Idea: input is a huge matrix

- **Least squares:** $\beta^{LS} = \mathit{argmin} \|X\beta - y\|_2$ (X is a big matrix)
- **PCA:** Find rank-k projection P^* minimizing $\|(I - P)A\|_F^2$
- **Matrix-matrix multiplication:** Compute $C = A^\top B$

Idea: input is a huge matrix

- **Least squares:** $\beta^{LS} = \mathit{argmin} \|X\beta - y\|_2$ (X is a big matrix)
- **PCA:** Find rank-k projection P^* minimizing $\|(I - P)A\|_F^2$
- **Matrix-matrix multiplication:** Compute $C = A^\top B$

Sketch-and-solve: Π is sketching matrix with m rows, m small

- **Least squares:** $\tilde{\beta}^{LS} = \mathit{argmin} \|\Pi X\beta - \Pi y\|_2$
- **PCA:** Find rank-k projection $\tilde{P}^*$ minimizing $\|(I - P)A\Pi^\top\|_F^2$
- **Matrix-matrix multiplication:** Compute $\tilde{C} = (\Pi A)^\top(\Pi B)$

Idea: input is a huge matrix

- **Least squares:** $\beta^{LS} = \mathit{argmin} \|X\beta - y\|_2$ (X is a big matrix)
- **PCA:** Find rank-k projection P^* minimizing $\|(I - P)A\|_F^2$
- **Matrix-matrix multiplication:** Compute $C = A^\top B$

Sketch-and-solve: Π is sketching matrix with m rows, m small

- **Least squares:** $\tilde{\beta}^{LS} = \mathit{argmin} \|\Pi X\beta - \Pi y\|_2$
- **PCA:** Find rank-k projection $\tilde{P}^*$ minimizing $\|(I - P)A\Pi^\top\|_F^2$
- **Matrix-matrix multiplication:** Compute $\tilde{C} = (\Pi A)^\top(\Pi B)$

Show that if Π chosen appropriately, then whp

- $\|X\tilde{\beta}^{LS} - y\|_2^2 \leq (1+\varepsilon)\|X\beta^{LS} - y\|_2^2$ [Sarlós'06]
- $\|(I - \tilde{P}^*)A\|_F^2 \leq (1+\varepsilon)\|(I - P^*)A\|_F^2$ [Cohen-Elder-Musco-Musco-Persu'15]
- $\|C - \tilde{C}\|_F \leq \varepsilon\|A\|_F\|B\|_F$ [Sarlós'06]

Idea: input is a huge matrix

- **Least squares:** $\beta^{LS} = \mathit{argmin} \|X\beta - y\|_2$ (X is a big matrix)
- **PCA:** Find rank-k projection P^* minimizing $\|(I-P)A\|_F^2$
- **Matrix-matrix multiplication:** Compute $C = A^\top B$

Sketch-and-solve: Π is sketching matrix with m rows, m small

- **Least squares:** $\tilde{\beta}^{LS} = \mathit{argmin} \|\Pi X\beta - \Pi y\|_2$
- **PCA:** Find rank-k projection $\tilde{P}^*$ minimizing $\|(I-P)A\Pi^\top\|_F^2$
- **Matrix-matrix multiplication:** Compute $\tilde{C} = (\Pi A)^\top(\Pi B)$

Show that if Π chosen appropriately, then whp

- $\|X\tilde{\beta}^{LS} - y\|_2^2 \le (1+\varepsilon)\|X\beta^{LS} - y\|_2^2$ [Sarlós'06]
- $\|(I-\tilde{P}^*)A\|_F^2 \le (1+\varepsilon)\|(I-P^*)A\|_F^2$ [Cohen-Elder-Musco-Musco-Persu'15]
- $\|C - \tilde{C}\|_F \le \varepsilon\|A\|_F\|B\|_F$ [Sarlós'06]

Example appropriate sketching matrices: i.i.d. subgaussian entries [Sarlós'06], *SFD* [Sarlós'06], [Tropp'11], [Cohen-Nelson-Woodruff'16], CountSketch [Clarkson-Woodruff'13], [Meng-Mahoney'13], [Nelson-Nguyen'13], [Bourgain-Dirksen-Nelson'15], [Cohen'16]

Analysis example (regression)

Analysis example (regression)

Definition. For $E \subset \mathbb{R}^n$ a d-dim. linear subspace, we say Π is an *ε-subspace embedding for E* if $\forall x \in E,\ \|\Pi x\|_2^2 = (1 \pm \varepsilon)\|x\|_2^2$

Analysis example (regression)

Definition. For $E \subset \mathbb{R}^n$ a d-dim. linear subspace, we say Π is an *ε-subspace embedding for E* if $\forall x \in E,\ \|\Pi x\|_2^2 = (1 \pm \varepsilon)\|x\|_2^2$

Claim. If Π is an ε-subspace embedding for $span\{cols(X), y\}$, then

$$\|X\tilde{\beta}^{LS} - y\|_2^2 \leq \left(\frac{1+\varepsilon}{1-\varepsilon}\right) \|X\beta^{LS} - y\|_2^2$$

Analysis example (regression)

Definition. For $E \subset \mathbb{R}^n$ a d-dim. linear subspace, we say Π is an *ε-subspace embedding for E* if $\forall x \in E,\ \|\Pi x\|_2^2 = (1 \pm \varepsilon)\|x\|_2^2$

Claim. If Π is an ε-subspace embedding for $span\{cols(X), y\}$, then

$$\|X\tilde{\beta}^{LS} - y\|_2^2 \leq \left(\frac{1+\varepsilon}{1-\varepsilon}\right) \|X\beta^{LS} - y\|_2^2$$

Proof.

$$\| \underbrace{\Pi X\tilde{\beta}^{LS} - \Pi y}_{\Pi(X\tilde{\beta}^{LS} - y)} \|_2^2 \leq \|\Pi X\beta^{LS} - \Pi y\|_2^2$$

Analysis example (regression)

Definition. For $E \subset \mathbb{R}^n$ a d-dim. linear subspace, we say Π is an *ε-subspace embedding for E* if $\forall x \in E,\ \|\Pi x\|_2^2 = (1 \pm \varepsilon)\|x\|_2^2$

Claim. If Π is an ε-subspace embedding for $span\{cols(X), y\}$, then

$$\|X\tilde{\beta}^{LS} - y\|_2^2 \leq \left(\frac{1+\varepsilon}{1-\varepsilon}\right) \|X\beta^{LS} - y\|_2^2$$

Proof.

$$(1-\varepsilon)\|X\tilde{\beta}^{LS} - y\|_2^2 \leq \|\underbrace{\Pi X\tilde{\beta}^{LS} - \Pi y}_{\Pi(X\tilde{\beta}^{LS} - y)}\|_2^2 \leq \|\Pi X\beta^{LS} - \Pi y\|_2^2$$

Analysis example (regression)

Definition. For $E \subset \mathbb{R}^n$ a d-dim. linear subspace, we say Π is an *ε-subspace embedding for E* if $\forall x \in E,\ \|\Pi x\|_2^2 = (1 \pm \varepsilon)\|x\|_2^2$

Claim. If Π is an ε-subspace embedding for $span\{cols(X), y\}$, then

$$\|X\tilde{\beta}^{LS} - y\|_2^2 \leq \left(\frac{1+\varepsilon}{1-\varepsilon}\right) \|X\beta^{LS} - y\|_2^2$$

Proof.

$$(1-\varepsilon)\|X\tilde{\beta}^{LS} - y\|_2^2 \leq \|\underbrace{\Pi X\tilde{\beta}^{LS} - \Pi y}_{\Pi(X\tilde{\beta}^{LS} - y)}\|_2^2 \leq \|\Pi X\beta^{LS} - \Pi y\|_2^2 \leq (1+\varepsilon)\|X\beta^{LS} - y\|_2^2$$

Obtaining (oblivious) subspace embeddings

If we organize orthonormal basis for E as columns of $U \in \mathbb{R}^{n \times d}$,

Π is ε-subspace embedding for $E \iff \|(\Pi U)^\top (\Pi U) - I\| \leq \varepsilon$

Obtaining (oblivious) subspace embeddings

If we organize orthonormal basis for E as columns of $U \in \mathbb{R}^{n \times d}$,

Π is ε-subspace embedding for $E \iff \|(\Pi U)^\top (\Pi U) - I\| \leq \varepsilon$

- **net argument:** apply JL to a net of the unit ball of E
- **moment method:**

$$\mathbb{P}(\|(\Pi U)^\top (\Pi U) - I\| > \varepsilon) < \frac{\mathbb{E}\,\|(\Pi U)^\top (\Pi U) - I\|^\ell}{\varepsilon^\ell} \leq \frac{\mathbb{E}\,tr(((\Pi U)^\top (\Pi U) - I)^\ell)}{\varepsilon^\ell}$$

(can also use moment-generating function — see [Tropp'12])

Another sketch-and-solve application: k-means

k-means: given k and $x_1, \ldots, x_n \in \mathbb{R}^d$, find $y_1, \ldots, y_k$ minimizing

$$\sum_{i=1}^{n} \min_{1 \le j \le k} \|x_i - y_j\|_2^2$$

k-means: given k and $x_1, \ldots, x_n \in \mathbb{R}^d$, find $y_1, \ldots, y_k$ minimizing

$$\sum_{i=1}^{n} \min_{1 \le j \le k} \|x_i - y_j\|_2^2$$

Clustering induces a k-partition $\mathcal{P}$ on $[n]$, so want to find best $\mathcal{P} = (P_1, \ldots, P_k)$. For fixed $\mathcal{P}$, best choice of y_j is centroid of P_j.

$$\begin{aligned} cost(\mathcal{P}) &= \sum_{j=1}^{k} \sum_{i \in P_j} \|x_i - \frac{\sum_{t \in P_j} x_t}{|P_j|}\|_2^2 \\ &= \sum_{j=1}^{k} \frac{1}{|P_j|} \sum_{i < i' \in P_j} \|x_i - x_{i'}\|_2^2. \end{aligned}$$

k-means: given k and $x_1, \ldots, x_n \in \mathbb{R}^d$, find $y_1, \ldots, y_k$ minimizing

$$\sum_{i=1}^{n} \min_{1 \le j \le k} \|x_i - y_j\|_2^2$$

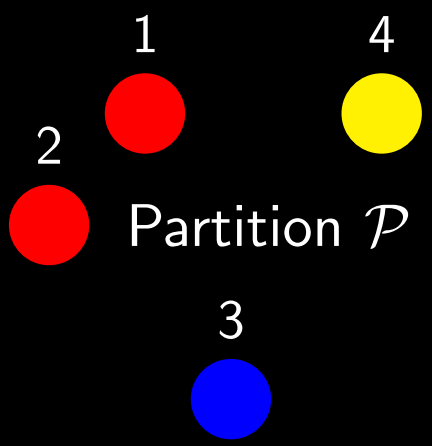
Clustering induces a k-partition $\mathcal{P}$ on $[n]$, so want to find best $\mathcal{P} = (P_1, \ldots, P_k)$. For fixed $\mathcal{P}$, best choice of y_j is centroid of P_j.

$$\begin{aligned} cost(\mathcal{P}) &= \sum_{j=1}^{k} \sum_{i \in P_j} \|x_i - \frac{\sum_{t \in P_j} x_t}{|P_j|}\|_2^2 \\ &= \sum_{j=1}^{k} \frac{1}{|P_j|} \sum_{i < i' \in P_j} \|x_i - x_{i'}\|_2^2. \end{aligned}$$

Thus JL embedding f preserves $cost(\mathcal{P})$ for all $\mathcal{P}$, so can optimize over $f(X)$ ($X = \{x_i\}_{i=1}^n$). Can reduce to dimension $O(\varepsilon^{-2} \log n)$.

Better dimension reduction for k-means

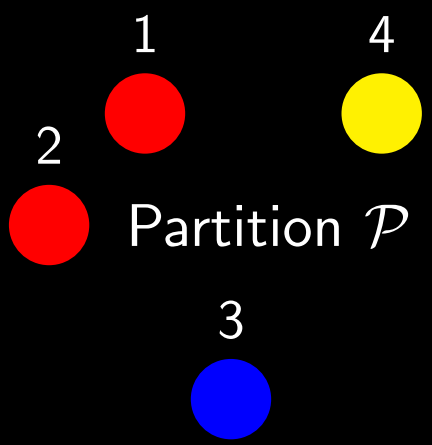
[Boutsidis-Zouzias-Mahoney-Drineas'11]: can reformulate k-means as a constrained low-rank approximation problem



$$X_{\mathcal{P}} = \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & 0 \\ \frac{1}{\sqrt{2}} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, X_{\mathcal{P}} X_{\mathcal{P}}^{\top} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Better dimension reduction for k-means

[Boutsidis-Zouzias-Mahoney-Drineas'11]: can reformulate k-means as a constrained low-rank approximation problem

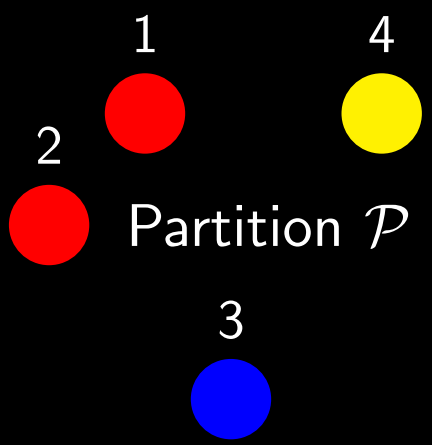


$$X_{\mathcal{P}} = \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & 0 \\ \frac{1}{\sqrt{2}} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, X_{\mathcal{P}} X_{\mathcal{P}}^{\top} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$X_{\mathcal{P}} X_{\mathcal{P}}^{\top}$ is a rank-k orthogonal projection, and if we put points as rows of a matrix A, then $X_{\mathcal{P}} X_{\mathcal{P}}^{\top} A$ maps each point (i.e. each row of A) to the centroid of its partition.

Better dimension reduction for k-means

[Boutsidis-Zouzias-Mahoney-Drineas'11]: can reformulate k-means as a constrained low-rank approximation problem



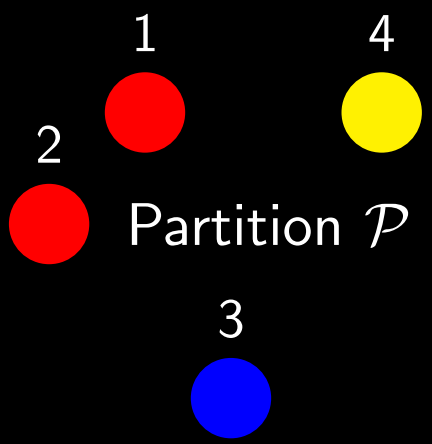
$$X_{\mathcal{P}} = \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & 0 \\ \frac{1}{\sqrt{2}} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, X_{\mathcal{P}} X_{\mathcal{P}}^\top = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$X_{\mathcal{P}} X_{\mathcal{P}}^\top$ is a rank-k orthogonal projection, and if we put points as rows of a matrix A, then $X_{\mathcal{P}} X_{\mathcal{P}}^\top A$ maps each point (i.e. each row of A) to the centroid of its partition.

$$cost(\mathcal{P}) = \|A - X_{\mathcal{P}} X_{\mathcal{P}}^\top A\|_F^2$$

Better dimension reduction for k-means

[Boutsidis-Zouzias-Mahoney-Drineas'11]: can reformulate k-means as a constrained low-rank approximation problem



$$X_{\mathcal{P}} = \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & 0 \\ \frac{1}{\sqrt{2}} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, X_{\mathcal{P}} X_{\mathcal{P}}^\top = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$X_{\mathcal{P}} X_{\mathcal{P}}^\top$ is a rank-k orthogonal projection, and if we put points as rows of a matrix A, then $X_{\mathcal{P}} X_{\mathcal{P}}^\top A$ maps each point (i.e. each row of A) to the centroid of its partition.

$$cost(\mathcal{P}) = \|A - X_{\mathcal{P}} X_{\mathcal{P}}^\top A\|_F^2$$

$\mathcal{Q} = \{X_{\mathcal{P}} X_{\mathcal{P}}^\top : \mathcal{P} \text{ a } k\text{-partition}\}$, constrained low-rank approx!: want $Q_{opt} = argmin_{Q \in \mathcal{Q}} \|A - QA\|_F^2$,

Better dimension reduction for k-means

$\mathcal{Q} = \{X_{\mathcal{P}} X_{\mathcal{P}}^\top : \mathcal{P}$ a k-partition$\}$, want

$$Q_{opt} = \underset{Q \in \mathcal{Q}}{\operatorname{argmin}} \|(I - Q)A\|_F^2.$$

Better dimension reduction for k-means

$\mathcal{Q} = \{X_{\mathcal{P}} X_{\mathcal{P}}^\top : \mathcal{P}$ a k-partition$\}$, want

$$Q_{opt} = \underset{Q \in \mathcal{Q}}{argmin} \, \|(I - Q)A\|_F^2.$$

Dimensionality reduction for optimization:

$$\tilde{Q}_{opt} = \underset{Q \in \mathcal{Q}}{argmin} \, \|(I - Q)A\Pi^\top\|_F^2 \text{ (sketch-and-solve)}$$

Better dimension reduction for k-means

$\mathcal{Q} = \{X_{\mathcal{P}} X_{\mathcal{P}}^\top : \mathcal{P} \text{ a } k\text{-partition}\}$, want

$$Q_{opt} = \underset{Q \in \mathcal{Q}}{\operatorname{argmin}} \|(I - Q)A\|_F^2.$$

Dimensionality reduction for optimization:

$$\tilde{Q}_{opt} = \underset{Q \in \mathcal{Q}}{\operatorname{argmin}} \|(I - Q)A\Pi^\top\|_F^2 \text{ (sketch-and-solve)}$$

To optimize up to $1 + \varepsilon$, suffices for sketching matrix Π to only have $O(k/\varepsilon^2)$ rows [Cohen-Elder-Musco-Musco-Persu'15] (see also [Cohen-N.-Woodruff'16]).

Better dimension reduction for k-means

$\mathcal{Q} = \{X_{\mathcal{P}} X_{\mathcal{P}}^\top : \mathcal{P} \text{ a } k\text{-partition}\}$, want

$$Q_{opt} = \underset{Q \in \mathcal{Q}}{argmin} \|(I - Q)A\|_F^2.$$

Dimensionality reduction for optimization:

$$\tilde{Q}_{opt} = \underset{Q \in \mathcal{Q}}{argmin} \|(I - Q)A\Pi^\top\|_F^2 \text{ (sketch-and-solve)}$$

To optimize up to $1 + \varepsilon$, suffices for sketching matrix Π to only have $O(k/\varepsilon^2)$ rows [Cohen-Elder-Musco-Musco-Persu'15] (see also [Cohen-N.-Woodruff'16]).

Recently bound was improved to $O(\log(k/\varepsilon)/\varepsilon^2)$ rows [Makarychev-Makarychev-Razenshteyn'19].

Better dimension reduction for k-means

$\mathcal{Q} = \{X_{\mathcal{P}} X_{\mathcal{P}}^\top : \mathcal{P} \text{ a } k\text{-partition}\}$, want

$$Q_{opt} = \underset{Q \in \mathcal{Q}}{argmin} \|(I - Q)A\|_F^2.$$

Dimensionality reduction for optimization:

$$\tilde{Q}_{opt} = \underset{Q \in \mathcal{Q}}{argmin} \|(I - Q)A\Pi^\top\|_F^2 \text{ (sketch-and-solve)}$$

To optimize up to $1 + \varepsilon$, suffices for sketching matrix Π to only have $O(k/\varepsilon^2)$ rows [Cohen-Elder-Musco-Musco-Persu'15] (see also [Cohen-N.-Woodruff'16]).

Recently bound was improved to $O(\log(k/\varepsilon)/\varepsilon^2)$ rows [Makarychev-Makarychev-Razenshteyn'19].

Better dimension reduction for k-means

$\mathcal{Q} = \{X_{\mathcal{P}} X_{\mathcal{P}}^\top : \mathcal{P} \text{ a } k\text{-partition}\}$, want

$$Q_{opt} = \operatorname*{argmin}_{Q \in \mathcal{Q}} \|(I - Q)A\|_F^2.$$

Dimensionality reduction for optimization:

$$\tilde{Q}_{opt} = \operatorname*{argmin}_{Q \in \mathcal{Q}} \|(I - Q)A\Pi^\top\|_F^2 \text{ (sketch-and-solve)}$$

To optimize up to $1 + \varepsilon$, suffices for sketching matrix Π to only have $O(k/\varepsilon^2)$ rows [Cohen-Elder-Musco-Musco-Persu'15] (see also [Cohen-N.-Woodruff'16]).

Recently bound was improved to $O(\log(k/\varepsilon)/\varepsilon^2)$ rows [Makarychev-Makarychev-Razenshteyn'19].

Projection-cost preserving sketch: Π satisfies that for *all* rank-k orthogonal projections P, $\|(I - P)A\Pi^\top\|_F^2 = (1 \pm \varepsilon)\|(I - P)A\|_F^2$.

Non-oblivious subspace embeddings

Non-oblivious subspace embeddings: sampling

$$\text{Want } \|Ax\|_2^2 = x^\top A^\top A x \approx x^\top \tilde{A}^\top \tilde{A} x = \|\tilde{A}x\|_2^2$$
$$(\tilde{A} = \Pi A)$$

Non-oblivious subspace embeddings: sampling

$$\text{Want } \|Ax\|_2^2 = x^\top A^\top A x \approx x^\top \tilde{A}^\top \tilde{A} x = \|\tilde{A}x\|_2^2$$
$$(\tilde{A} = \Pi A)$$

What about row-sampling? (each row of Π has exactly one non-zero entry)

$$A^\top A = \sum_{i=1}^{n} a_i a_i^\top, \qquad \tilde{A}^\top \tilde{A} = \sum_{i=1}^{n} \frac{\eta_i}{p_i} a_i a_i^\top$$

keep row i with probability p_i, and η_i is Bernoulli(p_i)

Non-oblivious subspace embeddings: sampling

$$\text{Want } \|Ax\|_2^2 = x^\top A^\top A x \approx x^\top \tilde{A}^\top \tilde{A} x = \|\tilde{A}x\|_2^2$$
$$(\tilde{A} = \Pi A)$$

What about row-sampling? (each row of Π has exactly one non-zero entry)

$$A^\top A = \sum_{i=1}^{n} a_i a_i^\top, \qquad \tilde{A}^\top \tilde{A} = \sum_{i=1}^{n} \frac{\eta_i}{p_i} a_i a_i^\top$$

keep row i with probability p_i, and η_i is Bernoulli(p_i)

Define sensitivity (coreset terminology) s_i as $s_i := \sup_x \frac{\langle a_i, x\rangle^2}{\|Ax\|_2^2}$.
s_i also known as leverage score ℓ_i of row i

Non-oblivious subspace embeddings: sampling

$$\text{Want } \|Ax\|_2^2 = x^\top A^\top A x \approx x^\top \tilde{A}^\top \tilde{A} x = \|\tilde{A}x\|_2^2$$
$$(\tilde{A} = \Pi A)$$

What about row-sampling? (each row of Π has exactly one non-zero entry)

$$A^\top A = \sum_{i=1}^{n} a_i a_i^\top, \qquad \tilde{A}^\top \tilde{A} = \sum_{i=1}^{n} \frac{\eta_i}{p_i} a_i a_i^\top$$

keep row i with probability p_i, and η_i is Bernoulli(p_i)

Define sensitivity (coreset terminology) s_i as $s_i := \sup_x \frac{\langle a_i, x\rangle^2}{\|Ax\|_2^2}$.
s_i also known as leverage score ℓ_i of row i

$\sum_i \ell_i = d$. Also $p_i < \frac{s_i}{2}$ doesn't make sense, so $\sum_i p_i \geq \frac{d}{2}$.

Non-oblivious subspace embeddings: sampling

$$\text{Want } \|Ax\|_2^2 = x^\top A^\top A x \approx x^\top \tilde{A}^\top \tilde{A} x = \|\tilde{A}x\|_2^2$$
$$(\tilde{A} = \Pi A)$$

What about row-sampling? (each row of Π has exactly one non-zero entry)

$$A^\top A = \sum_{i=1}^{n} a_i a_i^\top, \qquad \tilde{A}^\top \tilde{A} = \sum_{i=1}^{n} \frac{\eta_i}{p_i} a_i a_i^\top$$

keep row i with probability p_i, and η_i is Bernoulli(p_i)

Define sensitivity (coreset terminology) s_i as $s_i := \sup_x \frac{\langle a_i, x\rangle^2}{\|Ax\|_2^2}$.
s_i also known as leverage score ℓ_i of row i

$\sum_i \ell_i = d$. Also $p_i < \frac{s_i}{2}$ doesn't make sense, so $\sum_i p_i \geq \frac{d}{2}$.

Theorem [Spielman-Srivastava'08]. $p_i = \min\{1, C\frac{\log(d/\delta)}{\varepsilon^2}\}$ works.

Sketching for iterative methods

[Avron-Maymounkov-Toledo'10], [Clarkson-Woodruff'13], [Pilanci-Wainwright'16]

Main idea

Let Π be an α-subspace embedding and write $\Pi A = U\Sigma V^\top$ (SVD)

Main idea

Let Π be an α-subspace embedding and write $\Pi A = U\Sigma V^\top$ (SVD)

$$\forall x,\ \|x\|_2^2 = \|\Pi A \underbrace{V\Sigma^{-1}}_{R} x\|_2^2 = (1 \pm \alpha)\|A \underbrace{V\Sigma^{-1}}_{R} x\|_2^2.$$

Main idea

Let Π be an α-subspace embedding and write $\Pi A = U\Sigma V^\top$ (SVD)

$$\forall x,\ \|x\|_2^2 = \|\Pi A \underbrace{V\Sigma^{-1}}_{R} x\|_2^2 = (1 \pm \alpha)\|A \underbrace{V\Sigma^{-1}}_{R} x\|_2^2.$$

Thus AR is well-conditioned: condition number $\kappa \le \frac{1+\alpha}{1-\alpha} = O(1)$

Main idea

Let Π be an α-subspace embedding and write $\Pi A = U\Sigma V^\top$ (SVD)

$$\forall x,\ \|x\|_2^2 = \|\Pi A \underbrace{V\Sigma^{-1}}_{R} x\|_2^2 = (1 \pm \alpha)\|A \underbrace{V\Sigma^{-1}}_{R} x\|_2^2.$$

Thus AR is well-conditioned: condition number $\kappa \leq \frac{1+\alpha}{1-\alpha} = O(1)$

Gradient descent will get ε-error in $O(\kappa \log(1/\varepsilon))$ iterations
(and conjugate gradient in $O(\sqrt{\kappa}\log(1/\varepsilon))$)

now κ is constant!

Instance-wise bounds

Dimensionality reduction beyond worst-case analysis

Suppose we have $T \subset S^{d-1}$ and want matrix $\Pi \in \mathbb{R}^{m \times d}$ such that

$$\forall x \in T,\ (1-\varepsilon)\|x\|_2^2 \leq \|\Pi x\|_2^2 \leq (1+\varepsilon)\|x\|_2^2.$$

Dimensionality reduction beyond worst-case analysis

Suppose we have $T \subset S^{d-1}$ and want matrix $\Pi \in \mathbb{R}^{m \times d}$ such that

$$\forall x \in T,\ (1-\varepsilon)\|x\|_2^2 \leq \|\Pi x\|_2^2 \leq (1+\varepsilon)\|x\|_2^2.$$

Equivalent to

$$\sup_{x \in T} \left| \|\Pi x\|_2^2 - 1 \right| < \varepsilon$$

Dimensionality reduction beyond worst-case analysis

Suppose we have $T \subset S^{d-1}$ and want matrix $\Pi \in \mathbb{R}^{m \times d}$ such that

$$\forall x \in T,\ (1-\varepsilon)\|x\|_2^2 \le \|\Pi x\|_2^2 \le (1+\varepsilon)\|x\|_2^2.$$

Equivalent to

$$\sup_{x \in T} \left| \|\Pi x\|_2^2 - 1 \right| < \varepsilon$$

- JL lemma'84: $m \gtrsim \varepsilon^{-2} \log |T|$ suffices.

Dimensionality reduction beyond worst-case analysis

Suppose we have $T \subset S^{d-1}$ and want matrix $\Pi \in \mathbb{R}^{m \times d}$ such that

$$\forall x \in T,\ (1-\varepsilon)\|x\|_2^2 \le \|\Pi x\|_2^2 \le (1+\varepsilon)\|x\|_2^2.$$

Equivalent to

$$\sup_{x \in T} \left| \|\Pi x\|_2^2 - 1 \right| < \varepsilon$$

- JL lemma'84: $m \gtrsim \varepsilon^{-2} \log |T|$ suffices.
- Gordon'88: $m \gtrsim \varepsilon^{-2}(w^2(T)+1)$ suffices, where $w(T)$ is the Gaussian mean width of T. $w(T) := \mathbb{E} \sup_{x \in T} \langle g, x \rangle$.

Dimensionality reduction beyond worst-case analysis

Suppose we have $T \subset S^{d-1}$ and want matrix $\Pi \in \mathbb{R}^{m \times d}$ such that

$$\forall x \in T,\ (1-\varepsilon)\|x\|_2^2 \le \|\Pi x\|_2^2 \le (1+\varepsilon)\|x\|_2^2.$$

Equivalent to

$$\sup_{x \in T} |\|\Pi x\|_2^2 - 1| < \varepsilon$$

- JL lemma'84: $m \gtrsim \varepsilon^{-2} \log |T|$ suffices.
- Gordon'88: $m \gtrsim \varepsilon^{-2}(w^2(T)+1)$ suffices, where $w(T)$ is the Gaussian mean width of T. $w(T) := \mathbb{E} \sup_{x \in T} \langle g, x \rangle$.

  Note: $w(T) \lesssim \sqrt{\log |T|}$ always by union bound (tight if T vectors are orthogonal). Can be much smaller if gain vectors are close, since $|\langle g, x \rangle - \langle g, y \rangle| = |\langle g, x-y \rangle|$.

Dimensionality reduction beyond worst-case analysis

Suppose we have $T \subset S^{d-1}$ and want matrix $\Pi \in \mathbb{R}^{m \times d}$ such that

$$\forall x \in T,\ (1-\varepsilon)\|x\|_2^2 \le \|\Pi x\|_2^2 \le (1+\varepsilon)\|x\|_2^2.$$

Equivalent to

$$\sup_{x \in T} |\|\Pi x\|_2^2 - 1| < \varepsilon$$

- JL lemma'84: $m \gtrsim \varepsilon^{-2} \log |T|$ suffices.
- Gordon'88: $m \gtrsim \varepsilon^{-2}(w^2(T)+1)$ suffices, where $w(T)$ is the Gaussian mean width of T. $w(T) := \mathbb{E} \sup_{x \in T} \langle g, x \rangle$.

  Note: $w(T) \lesssim \sqrt{\log |T|}$ always by union bound (tight if T vectors are orthogonal). Can be much smaller if gain vectors are close, since $|\langle g, x \rangle - \langle g, y \rangle| = |\langle g, x-y \rangle|$.
- Gordon showed result for Π having i.i.d. gaussian entries. But what about other Π?

Dimensionality reduction beyond worst-case analysis

Using Π other than i.i.d. gaussian entries:

- [Klartag-Mendelson'05], [Mendelson-Pajor-TomczakJaegermann'07], [Dirksen'16] i.i.d. subgaussian entries suffice (e.g. $\pm 1/\sqrt{m}$).

Dimensionality reduction beyond worst-case analysis

Using Π other than i.i.d. gaussian entries:

- [Klartag-Mendelson'05], [Mendelson-Pajor-TomczakJaegermann'07], [Dirksen'16] i.i.d. subgaussian entries suffice (e.g. $\pm 1/\sqrt{m}$).
- [Bourgain-Dirksen-N.'15] Sparse JL Transform works with a similar number of rows, with low sparsity, under technical conditions concerning the point set to be reduced. Qualitatively recovers all known results for applications of sparse JL to specific domains, like subspace embeddings (next slide) up to $\log d$ factors.

Dimensionality reduction beyond worst-case analysis

Using Π other than i.i.d. gaussian entries:

- [Klartag-Mendelson'05], [Mendelson-Pajor-TomczakJaegermann'07], [Dirksen'16] i.i.d. subgaussian entries suffice (e.g. $\pm 1/\sqrt{m}$).
- [Bourgain-Dirksen-N.'15] Sparse JL Transform works with a similar number of rows, with low sparsity, under technical conditions concerning the point set to be reduced. Qualitatively recovers all known results for applications of sparse JL to specific domains, like subspace embeddings (next slide) up to $\log d$ factors.
- [Oymak-Recht-Soltanokotabi'17] Fast JL Transform of Ailon-Chazelle works with similar number of rows, up to $\log d$ factors.

This talk

- What about other, non-Euclidean norms?
- Proof of DJL.
- What is the optimal target dimension m for DJL?
- What is the optimal target dimension m for JL?
- How fast can we apply the map $x \mapsto \Pi x$?
- Isn't storing the random $m \times d$ matrix Π expensive?
- What can we get for specific applications and instances?
- Generalizing JL: terminal embeddings.

Johnson-Lindenstrauss lemma

JL Lemma, 1984

For every set of n points X in $\mathbb{R}^d$ and $\varepsilon \in (0, 1/2)$, there exists $f : X \to \ell_2^m$ for $m = O(\varepsilon^{-2} \log n)$ s.t $\forall x \in X, \forall y \in X$,

$$(1 - \varepsilon)\|x - y\|_2 \leq \|f(x) - f(y)\|_2 \leq (1 + \varepsilon)\|x - y\|_2$$

Furthermore, f can be taken as a linear map: $x \mapsto \Pi x$ for some linear sketching matrix $\Pi \in \mathbb{R}^{m \times d}$.

New theorem

[Narayanan-Nelson'19 ("Terminal" dimensionality reduction)]
For every set of n points X in $\mathbb{R}^d$ and $\varepsilon \in (0, 1/2)$, there exists $f : X \to \ell_2^m$ for $m = O(\varepsilon^{-2} \log n)$ s.t $\forall x \in X, \forall y \in \mathbb{R}^d$,

$$(1-\varepsilon)\|x-y\|_2 \leq \|f(x)-f(y)\|_2 \leq (1+\varepsilon)\|x-y\|_2$$

~~Furthermore, f can be taken as a linear map: $x \mapsto \Pi x$ for some linear sketching matrix $\Pi \in \mathbb{R}^{m \times d}$.~~

New theorem

[Narayanan-Nelson'19 ("Terminal" dimensionality reduction)]
For every set of n points X in $\mathbb{R}^d$ and $\varepsilon \in (0, 1/2)$, there exists $f : X \to \ell_2^m$ for $m = O(\varepsilon^{-2} \log n)$ s.t $\forall x \in X, \forall y \in \mathbb{R}^d$,

$$(1-\varepsilon)\|x-y\|_2 \leq \|f(x)-f(y)\|_2 \leq (1+\varepsilon)\|x-y\|_2$$

~~Furthermore, f can be taken as a linear map: $x \mapsto \Pi x$ for some linear sketching matrix $\Pi \in \mathbb{R}^{m \times d}$.~~

Improves bound of [Mahabadi-Makarychev-Makarychev-Razenshteyn'18], which achieved $m = O(\varepsilon^{-4} \log n)$.

New theorem

[Narayanan-Nelson'19 ("Terminal" dimensionality reduction)]
For every set of n points X in $\mathbb{R}^d$ and $\varepsilon \in (0, 1/2)$, there exists $f : X \to \ell_2^m$ for $m = O(\varepsilon^{-2} \log n)$ s.t $\forall x \in X, \forall y \in \mathbb{R}^d$,

$$(1-\varepsilon)\|x-y\|_2 \leq \|f(x)-f(y)\|_2 \leq (1+\varepsilon)\|x-y\|_2$$

~~Furthermore, f can be taken as a linear map: $x \mapsto \Pi x$ for some linear sketching matrix $\Pi \in \mathbb{R}^{m \times d}$.~~

Improves bound of [Mahabadi-Makarychev-Makarychev-Razenshteyn'18], which achieved $m = O(\varepsilon^{-4} \log n)$.

Optimal since $\Omega(\varepsilon^{-2} \log n)$ is a lower bound even for non-terminal embeddings [Larsen-Nelson'17].

Computer Science Motivation

Computer Science Motivation

Static data structures

- Lay out a static database D, i.e. D cannot be updated, in memory in such a way as to
  (1) use little memory
  (2) answer queries from a known family of queries quickly
  e.g. $\texttt{nns}(q)$: return $\mathit{argmin}_{x \in D}\, d(x, q)$ (nearest neighbor search)

Computer Science Motivation

Static data structures

- Lay out a static database D, i.e. D cannot be updated, in memory in such a way as to
  (1) use little memory
  (2) answer queries from a known family of queries quickly
  e.g. $\texttt{nns}(q)$: return $\mathit{argmin}_{x \in D}\, d(x, q)$ (nearest neighbor search)
- Problem: Don't know query q when picking embedding f

Computer Science Motivation

Static data structures

- Lay out a static database D, i.e. D cannot be updated, in memory in such a way as to
  (1) use little memory
  (2) answer queries from a known family of queries quickly
  e.g. $\texttt{nns}(q)$: return $\mathit{argmin}_{x \in D}\, d(x, q)$ (nearest neighbor search)
- Problem: Don't know query q when picking embedding f
- Standard solution: Oblivious dim. reduction (randomized)

Computer Science Motivation

Static data structures

- Lay out a static database D, i.e. D cannot be updated, in memory in such a way as to
  (1) use little memory
  (2) answer queries from a known family of queries quickly
  e.g. $\texttt{nns}(q)$: return $\mathit{argmin}_{x \in D}\, d(x, q)$ (nearest neighbor search)
- Problem: Don't know query q when picking embedding f
- Standard solution: Oblivious dim. reduction (randomized)
- Terminal Dim. Red.: Don't need to know q

  f satisfies terminal guarantee $\implies$ correctness is deterministic

  in usual compressed sensing language: "for each" vs. "for all" guarantee

  (better for handling *adaptive* queries)

History

(terminal embeddings from ℓ_2^d)

Distortion ρ, target dimension m:

History

(terminal embeddings from ℓ_2^d)

Distortion ρ, target dimension m:

- [Elkin-Filtser-Neiman '15]: $\rho = \sqrt{10} + \varepsilon$, $m = O(\varepsilon^{-2} \log n)$

History

(terminal embeddings from ℓ_2^d)

Distortion ρ, target dimension m:

- [Elkin-Filtser-Neiman '15]: $\rho = \sqrt{10} + \varepsilon$, $m = O(\varepsilon^{-2} \log n)$

  $f(q) = (g(x_q), \|x_q - q\|_2)$, where x_q is closest point to q in X

  g is low distortion embedding from X to ℓ_2^m (use JL)

History

(terminal embeddings from ℓ_2^d)

Distortion ρ, target dimension m:

- [Elkin-Filtser-Neiman '15]: $\rho = \sqrt{10} + \varepsilon$, $m = O(\varepsilon^{-2} \log n)$

  $f(q) = (g(x_q), \|x_q - q\|_2)$, where x_q is closest point to q in X

  g is low distortion embedding from X to ℓ_2^m (use JL)
- [Mahabadi et al.'18]: $\rho = 1 + \varepsilon$, $m = O(\varepsilon^{-4} \log n)$

History

(terminal embeddings from ℓ_2^d)

Distortion ρ, target dimension m:

- [Elkin-Filtser-Neiman '15]: $\rho = \sqrt{10} + \varepsilon$, $m = O(\varepsilon^{-2} \log n)$

  $f(q) = (g(x_q), \|x_q - q\|_2)$, where x_q is closest point to q in X

  g is low distortion embedding from X to ℓ_2^m (use JL)

- [Mahabadi et al.'18]: $\rho = 1 + \varepsilon$, $m = O(\varepsilon^{-4} \log n)$

  $f(q) = (g(x_q) + z, \sqrt{\|q - x_q\|_2^2 - \|z\|_2^2})$

  g is low-distortion embedding (now with distortion $1 + \varepsilon^2$)

  $z \in \mathbb{R}^m$ is found via convex programming

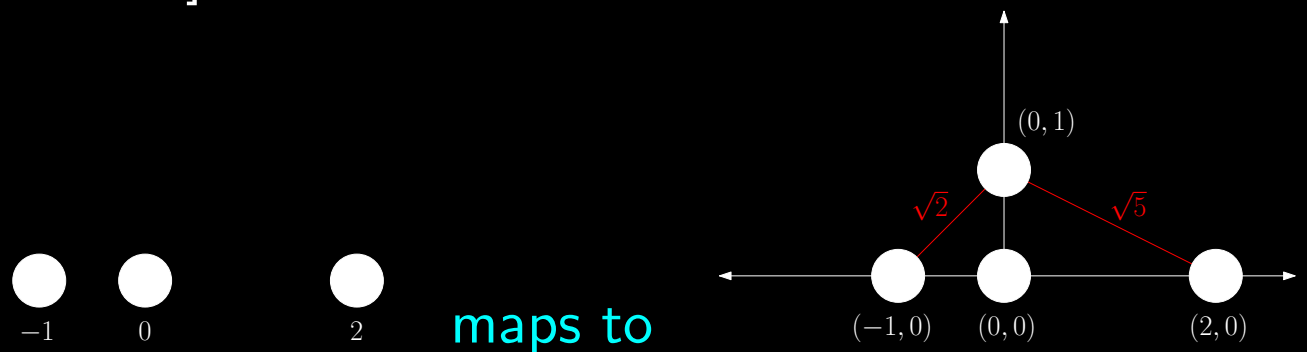
Bad News

- [EFN'15]:

$X = \{-1, 0, 2\}$, new point is $q = 1$. EFN maps 1 to $(0, 1)$.

Bad News

- [EFN'15]:



  $X = \{-1, 0, 2\}$, new point is $q = 1$. EFN maps 1 to $(0, 1)$.

- [MMMR'18]: "extending a map with distortion $1 + \varepsilon$ by one point might require blowing up the distortion to $1 + \Omega(\sqrt{\varepsilon})$"

[Narayanan-Nelson'19]: Not extending just *any* $(1+\varepsilon)$-distortion embedding, but rather one taken in a specific way (JL, e.g. random projection). Can use more properties of this map to get analysis of MMMR'18 embedding needing only $m = O(\varepsilon^{-2} \log n)$.

Analysis

Convex hull distortion

Definition. Π provides ε-convex hull distortion for T if

$$\forall z \in \mathrm{conv}(T),\ |\|\Pi z\|_2 - \|z\|_2| < \varepsilon.$$

Convex hull distortion

Definition. Π provides ε-convex hull distortion for T if

$$\forall z \in \mathrm{conv}(T),\ |\|\Pi z\|_2 - \|z\|_2| < \varepsilon.$$

How do we get good convex hull distortion?

Convex hull distortion

Definition. Π provides ε-convex hull distortion for T if

$$\forall z \in \mathrm{conv}(T),\ |\|\Pi z\|_2 - \|z\|_2| < \varepsilon.$$

How do we get good convex hull distortion?

Theorem [Gordon'88; Klartag-Mendelson'05, Mendelson-Pajor-TomczakJaegermann'07, Dirksen'15]
Let $T = \{\frac{x-y}{\|x-y\|_2} : x, y \in X\}$. Suppose $\Pi \in \mathbb{R}^{m\times d}$ has i.i.d. subgaussian entries with variance $1/m$. Then $m = \Omega(\varepsilon^{-2} g^2(T))$ implies $\mathbb{E}_\Pi \sup_{t\in T} |\|\Pi t\|_2^2 - 1| < \varepsilon$. Here $g(T)$ is the gaussian mean width $\mathbb{E}_g \sup_{t\in T} \langle g, t\rangle$.

Convex hull distortion

Definition. Π provides ε-convex hull distortion for T if

$$\forall z \in \mathrm{conv}(T),\ |\|\Pi z\|_2 - \|z\|_2| < \varepsilon.$$

How do we get good convex hull distortion?

Theorem [Gordon'88; Klartag-Mendelson'05, Mendelson-Pajor-TomczakJaegermann'07, Dirksen'15]
Let $T = \{\frac{x-y}{\|x-y\|_2} : x, y \in X\}$. Suppose $\Pi \in \mathbb{R}^{m\times d}$ has i.i.d. subgaussian entries with variance $1/m$. Then $m = \Omega(\varepsilon^{-2} g^2(T))$ implies $\mathbb{E}_\Pi \sup_{t\in T} |\|\Pi t\|_2^2 - 1| < \varepsilon$. Here $g(T)$ is the gaussian mean width $\mathbb{E}_g \sup_{t\in T} \langle g, t\rangle$.
Note: $g(\mathrm{conv}(T)) = g(T) = O(\sqrt{\log |T|})$

Convex hull distortion

Definition. Π provides ε-convex hull distortion for T if

$$\forall z \in \mathrm{conv}(T),\ |\|\Pi z\|_2 - \|z\|_2| < \varepsilon.$$

How do we get good convex hull distortion?

Theorem [Gordon'88; Klartag-Mendelson'05, Mendelson-Pajor-TomczakJaegermann'07, Dirksen'15]
Let $T = \{\frac{x-y}{\|x-y\|_2} : x, y \in X\}$. Suppose $\Pi \in \mathbb{R}^{m\times d}$ has i.i.d. subgaussian entries with variance $1/m$. Then $m = \Omega(\varepsilon^{-2} g^2(T))$ implies $\mathbb{E}_\Pi \sup_{t\in T} |\|\Pi t\|_2^2 - 1| < \varepsilon$. Here $g(T)$ is the gaussian mean width $\mathbb{E}_g \sup_{t\in T} \langle g, t\rangle$.
Note: $g(\mathrm{conv}(T)) = g(T) = O(\sqrt{\log |T|})$
$\|\Pi z\|_2^2 = \|z\|_2^2 \pm \varepsilon$ for all $z \in \mathrm{conv}(T)$, but $\|\Pi z\|_2 \stackrel{?}{=} \|z\|_2 \pm \varepsilon$?

Convex hull distortion

Definition. Π provides ε-convex hull distortion for T if

$$\forall z \in \mathrm{conv}(T),\ |\|\Pi z\|_2 - \|z\|_2| < \varepsilon.$$

How do we get good convex hull distortion?

Theorem [Gordon'88; Klartag-Mendelson'05, Mendelson-Pajor-TomczakJaegermann'07, Dirksen'15]
Let $T = \{\frac{x-y}{\|x-y\|_2} : x, y \in X\}$. Suppose $\Pi \in \mathbb{R}^{m\times d}$ has i.i.d. subgaussian entries with variance $1/m$. Then $m = \Omega(\varepsilon^{-2} g^2(T))$ implies $\mathbb{E}_\Pi \sup_{t\in T} |\|\Pi t\|_2^2 - 1| < \varepsilon$. Here $g(T)$ is the gaussian mean width $\mathbb{E}_g \sup_{t\in T} \langle g, t\rangle$.
Note: $g(\mathrm{conv}(T)) = g(T) = O(\sqrt{\log |T|})$

$\|\Pi z\|_2^2 = \|z\|_2^2 \pm \varepsilon$ for all $z \in \mathrm{conv}(T)$, but $\|\Pi z\|_2 \stackrel{?}{=} \|z\|_2 \pm \varepsilon$?

$$\underbrace{|\|\Pi z\|_2^2 - \|z\|_2^2|}_{\text{small}} = |\|\Pi z\|_2 + \|z\|_2| \cdot |\|\Pi z\|_2 - \|z\|_2| \geq \|z\|_2 \cdot |\|\Pi z\|_2 - \|z\|_2|$$

Convex hull distortion

Definition. Π provides ε-convex hull distortion for T if

$$\forall z \in \mathrm{conv}(T),\ |\|\Pi z\|_2 - \|z\|_2| < \varepsilon.$$

How do we get good convex hull distortion?

Theorem [Gordon'88; Klartag-Mendelson'05, Mendelson-Pajor-TomczakJaegermann'07, Dirksen'15]
Let $T = \{\frac{x-y}{\|x-y\|_2} : x, y \in X\}$. Suppose $\Pi \in \mathbb{R}^{m\times d}$ has i.i.d. subgaussian entries with variance $1/m$. Then $m = \Omega(\varepsilon^{-2} g^2(T))$ implies $\mathbb{E}_\Pi \sup_{t\in T} |\|\Pi t\|_2^2 - 1| < \varepsilon$. Here $g(T)$ is the gaussian mean width $\mathbb{E}_g \sup_{t\in T} \langle g, t\rangle$.
Note: $g(\mathrm{conv}(T)) = g(T) = O(\sqrt{\log |T|})$

$\|\Pi z\|_2^2 = \|z\|_2^2 \pm \varepsilon$ for all $z \in \mathrm{conv}(T)$, but $\|\Pi z\|_2 \stackrel{?}{=} \|z\|_2 \pm \varepsilon$?

$$\underbrace{|\|\Pi z\|_2^2 - \|z\|_2^2|}_{\text{small}} = |\|\Pi z\|_2 + \|z\|_2| \cdot |\|\Pi z\|_2 - \|z\|_2| \geq \|z\|_2 \cdot |\|\Pi z\|_2 - \|z\|_2|$$

Dyadic partition on $\|z\|_2$ then union bound over $\log(1/\varepsilon)$ partitions

Convex hull distortion

Definition. Π provides ε-convex hull distortion for T if

$$\forall z \in \mathrm{conv}(T),\ |\|\Pi z\|_2 - \|z\|_2| < \varepsilon.$$

How do we get good convex hull distortion?

Theorem [Gordon'88; Klartag-Mendelson'05, Mendelson-Pajor-TomczakJaegermann'07, Dirksen'15]
Let $T = \{\frac{x-y}{\|x-y\|_2} : x, y \in X\}$. Suppose $\Pi \in \mathbb{R}^{m\times d}$ has i.i.d. subgaussian entries with variance $1/m$. Then $m = \Omega(\varepsilon^{-2} g^2(T))$ implies $\mathbb{E}_\Pi \sup_{t\in T} |\|\Pi t\|_2^2 - 1| < \varepsilon$. Here $g(T)$ is the gaussian mean width $\mathbb{E}_g \sup_{t\in T} \langle g, t\rangle$.
Note: $g(\mathrm{conv}(T)) = g(T) = O(\sqrt{\log |T|})$
$\|\Pi z\|_2^2 = \|z\|_2^2 \pm \varepsilon$ for all $z \in \mathrm{conv}(T)$, but $\|\Pi z\|_2 \stackrel{?}{=} \|z\|_2 \pm \varepsilon$?

$$\underbrace{|\|\Pi z\|_2^2 - \|z\|_2^2|}_{\text{small}} = |\|\Pi z\|_2 + \|z\|_2| \cdot |\|\Pi z\|_2 - \|z\|_2| \geq \|z\|_2 \cdot |\|\Pi z\|_2 - \|z\|_2|$$

Dyadic partition on $\|z\|_2$ then union bound over $\log(1/\varepsilon)$ partitions
Lemma. If Π is as above with $m = \Omega(\varepsilon^{-2} \log(\frac{n \log(1/\varepsilon)}{\delta}))$, it provides convex hull distortion w.p. $\geq 1 - \delta$.

Using convex hull distortion

We prove the following modification of a lemma in [MMMR'18], which only assumed g has distortion $1 + \varepsilon$ and concluded a terminal distortion bound of $1 + O(\sqrt{\varepsilon})$.

Using convex hull distortion

We prove the following modification of a lemma in [MMMR'18], which only assumed g has distortion $1+\varepsilon$ and concluded a terminal distortion bound of $1+O(\sqrt{\varepsilon})$.

Lemma. Define $T = \{\frac{x-y}{\|x-y\|_2} : x, y \in X\}$. Then if Π provides ε-convex hull distortion for T, for every $q \in \mathbb{R}^d \backslash X$ there exists $z \in \mathbb{R}^m$ such that $(g(x_q) + z, \sqrt{\|q - x_q\|_2^2 - \|z\|_2^2})$ preserves distance up to $1+\varepsilon$ between q and X.

Using convex hull distortion

We prove the following modification of a lemma in [MMMR'18], which only assumed g has distortion $1+\varepsilon$ and concluded a terminal distortion bound of $1+O(\sqrt{\varepsilon})$.

Lemma. Define $T = \{\frac{x-y}{\|x-y\|_2} : x, y \in X\}$. Then if Π provides ε-convex hull distortion for T, for every $q \in \mathbb{R}^d \backslash X$ there exists $z \in \mathbb{R}^m$ such that $(g(x_q)+z, \sqrt{\|q-x_q\|_2^2 - \|z\|_2^2})$ preserves distance up to $1+\varepsilon$ between q and X.

Corollary. Can achieve terminal dist. $1+\varepsilon$ with $m = O(\varepsilon^{-2}\log n)$.
Proof. Map $q \notin X$ as above, and $x \in X$ to $(g(x), 0)$.

Proof overview (inspired by [MMMR'18])

Lemma. Let $t_1, \ldots, t_N$ be arbitrary, $v_i := \frac{t_i}{\|t_i\|_2}$. Def. $V = \{\pm v_i\}_i$.
If Π provides ε-convex hull dist. for V, $\forall u \in \mathbb{R}^d$ $\exists z \in \mathbb{R}^m$ s.t.

(1) $\|z\|_2 \leq \|u\|_2$

(2) $\forall i,\ |\langle z, \Pi t_i \rangle - \langle u, t_i \rangle| \leq \varepsilon \|u\|_2 \cdot \|t_i\|_2$

Proof overview (inspired by [MMMR'18])

Lemma. Let $t_1, \ldots, t_N$ be arbitrary, $v_i := \frac{t_i}{\|t_i\|_2}$. Def. $V = \{\pm v_i\}_i$.
If Π provides ε-convex hull dist. for V, $\forall u \in \mathbb{R}^d$ $\exists z \in \mathbb{R}^m$ s.t.

(1) $\|z\|_2 \leq \|u\|_2$

(2) $\forall i,\ |\langle z, \Pi t_i \rangle - \langle u, t_i \rangle| \leq \varepsilon \|u\|_2 \cdot \|t_i\|_2$

Proof. Want $\forall i\ |\langle z, \Pi v_i \rangle - \langle u, v_i \rangle| \leq \varepsilon \|u\|_2$.

Proof overview (inspired by [MMMR'18])

Lemma. Let $t_1, \ldots, t_N$ be arbitrary, $v_i := \frac{t_i}{\|t_i\|_2}$. Def. $V = \{\pm v_i\}_i$.
If Π provides ε-convex hull dist. for V, $\forall u \in \mathbb{R}^d$ $\exists z \in \mathbb{R}^m$ s.t.

(1) $\|z\|_2 \leq \|u\|_2$

(2) $\forall i,\ |\langle z, \Pi t_i \rangle - \langle u, t_i \rangle| \leq \varepsilon \|u\|_2 \cdot \|t_i\|_2$

Proof. Want $\forall i\ |\langle z, \Pi v_i \rangle - \langle u, v_i \rangle| \leq \varepsilon \|u\|_2$.

- Want $\forall i\ |\langle z, \Pi v_i \rangle - \langle u, v_i \rangle| - \varepsilon \|u\|_2 \leq 0$

Proof overview (inspired by [MMMR'18])

Lemma. Let $t_1, \ldots, t_N$ be arbitrary, $v_i := \frac{t_i}{\|t_i\|_2}$. Def. $V = \{\pm v_i\}_i$.
If Π provides ε-convex hull dist. for V, $\forall u \in \mathbb{R}^d$ $\exists z \in \mathbb{R}^m$ s.t.

(1) $\|z\|_2 \le \|u\|_2$

(2) $\forall i,\ |\langle z, \Pi t_i\rangle - \langle u, t_i\rangle| \le \varepsilon \|u\|_2 \cdot \|t_i\|_2$

Proof. Want $\forall i\ |\langle z, \Pi v_i\rangle - \langle u, v_i\rangle| \le \varepsilon\|u\|_2$.

- Want $\forall i\ |\langle z, \Pi v_i\rangle - \langle u, v_i\rangle| - \varepsilon\|u\|_2 \le 0$
- **Convex relaxation**. Define $\Lambda := \ell_1^{|V|}$, $B := B^m_{\ell_2}(0, \|u\|_2)$.
  Define $\Phi(u', \lambda) := \sum_{v \in V}(\lambda_v(\langle u, v\rangle - \langle u', \Pi v\rangle) - \varepsilon|\lambda_v| \cdot \|u\|_2)$.
  want $\exists z\ \forall \lambda \in \Lambda,\ \Phi(z, \lambda) \le 0$

Proof overview (inspired by [MMMR'18])

Lemma. Let $t_1, \ldots, t_N$ be arbitrary, $v_i := \frac{t_i}{\|t_i\|_2}$. Def. $V = \{\pm v_i\}_i$.
If Π provides ε-convex hull dist. for V, $\forall u \in \mathbb{R}^d$ $\exists z \in \mathbb{R}^m$ s.t.

(1) $\|z\|_2 \le \|u\|_2$

(2) $\forall i,\ |\langle z, \Pi t_i\rangle - \langle u, t_i\rangle| \le \varepsilon \|u\|_2 \cdot \|t_i\|_2$

Proof. Want $\forall i\ |\langle z, \Pi v_i\rangle - \langle u, v_i\rangle| \le \varepsilon\|u\|_2$.

- Want $\forall i\ |\langle z, \Pi v_i\rangle - \langle u, v_i\rangle| - \varepsilon\|u\|_2 \le 0$
- **Convex relaxation**. Define $\Lambda := \ell_1^{|V|}$, $B := B^m_{\ell_2}(0, \|u\|_2)$.
  Define $\Phi(u', \lambda) := \sum_{v \in V}(\lambda_v(\langle u, v\rangle - \langle u', \Pi v\rangle) - \varepsilon|\lambda_v| \cdot \|u\|_2)$.
  want $\exists z\ \forall \lambda \in \Lambda,\ \Phi(z, \lambda) \le 0$

Minimax: $\min_{u' \in B} \max_{\lambda \in \Lambda} \Phi(u', \lambda) = \max_{\lambda \in \Lambda} \min_{u' \in B} \Phi(u', \lambda)$

Proof overview (inspired by [MMMR'18])

Lemma. Let $t_1, \ldots, t_N$ be arbitrary, $v_i := \frac{t_i}{\|t_i\|_2}$. Def. $V = \{\pm v_i\}_i$.
If Π provides ε-convex hull dist. for V, $\forall u \in \mathbb{R}^d$ $\exists z \in \mathbb{R}^m$ s.t.

(1) $\|z\|_2 \le \|u\|_2$

(2) $\forall i,\ |\langle z, \Pi t_i\rangle - \langle u, t_i\rangle| \le \varepsilon \|u\|_2 \cdot \|t_i\|_2$

Proof. Want $\forall i\ |\langle z, \Pi v_i\rangle - \langle u, v_i\rangle| \le \varepsilon\|u\|_2$.

- Want $\forall i\ |\langle z, \Pi v_i\rangle - \langle u, v_i\rangle| - \varepsilon\|u\|_2 \le 0$
- **Convex relaxation**. Define $\Lambda := \ell_1^{|V|}$, $B := B^m_{\ell_2}(0, \|u\|_2)$.
  Define $\Phi(u', \lambda) := \sum_{v \in V}(\lambda_v(\langle u, v\rangle - \langle u', \Pi v\rangle) - \varepsilon|\lambda_v| \cdot \|u\|_2)$.
  want $\exists z\ \forall \lambda \in \Lambda,\ \Phi(z, \lambda) \le 0$

Minimax: $\min_{u' \in B} \max_{\lambda \in \Lambda} \Phi(u', \lambda) = \max_{\lambda \in \Lambda} \min_{u' \in B} \Phi(u', \lambda)$
RHS nonpositive suffices. For fixed λ, define $P = \sum_v \lambda_v v$. Can show $u' = \|u\|_2 \cdot \frac{\Pi P}{\|\Pi P\|_2}$ works, using convex hull distortion.

Proof overview (inspired by [MMMR'18])

Lemma. Let $t_1, \ldots, t_N$ be arbitrary, $v_i := \frac{t_i}{\|t_i\|_2}$. Def. $V = \{\pm v_i\}_i$.
If Π provides ε-convex hull dist. for V, $\forall u \in \mathbb{R}^d$ $\exists z \in \mathbb{R}^m$ s.t.

(1) $\|z\|_2 \le \|u\|_2$

(2) $\forall i,\ |\langle z, \Pi t_i\rangle - \langle u, t_i\rangle| \le \varepsilon \|u\|_2 \cdot \|t_i\|_2$

Proof. Want $\forall i\ |\langle z, \Pi v_i\rangle - \langle u, v_i\rangle| \le \varepsilon\|u\|_2$.

- Want $\forall i\ |\langle z, \Pi v_i\rangle - \langle u, v_i\rangle| - \varepsilon\|u\|_2 \le 0$
- **Convex relaxation**. Define $\Lambda := \ell_1^{|V|}$, $B := B^m_{\ell_2}(0, \|u\|_2)$.
  Define $\Phi(u', \lambda) := \sum_{v \in V}(\lambda_v(\langle u, v\rangle - \langle u', \Pi v\rangle) - \varepsilon|\lambda_v| \cdot \|u\|_2)$.
  want $\exists z\ \forall \lambda \in \Lambda,\ \Phi(z, \lambda) \le 0$

Minimax: $\min_{u' \in B} \max_{\lambda \in \Lambda} \Phi(u', \lambda) = \max_{\lambda \in \Lambda} \min_{u' \in B} \Phi(u', \lambda)$
RHS nonpositive suffices. For fixed λ, define $P = \sum_v \lambda_v v$. **Can show $u' = \|u\|_2 \cdot \frac{\Pi P}{\|\Pi P\|_2}$ works, using convex hull distortion.**

Proof overview (inspired by [MMMR'18])

Recall, bounding $\max_\lambda \min_{u'} \Phi(u', \lambda) := \max_\lambda \min_{u'} \sum_{v \in V} (\lambda_v (\langle u, v \rangle - \langle u', \Pi v \rangle) - \varepsilon |\lambda_v| \cdot \|u\|_2)$

Proof overview (inspired by [MMMR'18])

Recall, bounding $\max_\lambda \min_{u'} \Phi(u', \lambda) := \max_\lambda \min_{u'} \sum_{v \in V} (\lambda_v (\langle u, v \rangle - \langle u', \Pi v \rangle) - \varepsilon |\lambda_v| \cdot \|u\|_2)$

- For fixed λ, pick $P = \sum_v \lambda_v v$ and $u' = \|u\|_2 \cdot \frac{\Pi P}{\|\Pi P\|_2}$.

Proof overview (inspired by [MMMR'18])

Recall, bounding $\max_\lambda \min_{u'} \Phi(u', \lambda) := \max_\lambda \min_{u'} \sum_{v \in V} (\lambda_v (\langle u, v \rangle - \langle u', \Pi v \rangle) - \varepsilon |\lambda_v| \cdot \|u\|_2)$

- For fixed λ, pick $P = \sum_v \lambda_v v$ and $u' = \|u\|_2 \cdot \frac{\Pi P}{\|\Pi P\|_2}$.
- $\Phi(u', \lambda) = \langle u, P \rangle - \langle u', \Pi P \rangle - \varepsilon \|\lambda\|_1 \|u\|_2$

Proof overview (inspired by [MMMR'18])

Recall, bounding $\max_\lambda \min_{u'} \Phi(u', \lambda) := \max_\lambda \min_{u'} \sum_{v \in V} (\lambda_v (\langle u, v \rangle - \langle u', \Pi v \rangle) - \varepsilon |\lambda_v| \cdot \|u\|_2)$

- For fixed λ, pick $P = \sum_v \lambda_v v$ and $u' = \|u\|_2 \cdot \frac{\Pi P}{\|\Pi P\|_2}$.
- $\Phi(u', \lambda) = \langle u, P \rangle - \langle u', \Pi P \rangle - \varepsilon \|\lambda\|_1 \|u\|_2$

  $= \langle u, P \rangle - \|u\|_2 \|\Pi P\|_2 - \varepsilon \|u\|_2$

Proof overview (inspired by [MMMR'18])

Recall, bounding $\max_\lambda \min_{u'} \Phi(u', \lambda) := \max_\lambda \min_{u'} \sum_{v \in V} (\lambda_v (\langle u, v \rangle - \langle u', \Pi v \rangle) - \varepsilon |\lambda_v| \cdot \|u\|_2)$

- For fixed λ, pick $P = \sum_v \lambda_v v$ and $u' = \|u\|_2 \cdot \frac{\Pi P}{\|\Pi P\|_2}$.
- $$\begin{aligned} \Phi(u', \lambda) &= \langle u, P \rangle - \langle u', \Pi P \rangle - \varepsilon \|\lambda\|_1 \|u\|_2 \\ &= \langle u, P \rangle - \|u\|_2 \|\Pi P\|_2 - \varepsilon \|u\|_2 \\ &\leq \|u_2\|_2 \cdot (\|P\|_2 - \|\Pi P\|_2 - \varepsilon) \end{aligned}$$

Proof overview (inspired by [MMMR'18])

Recall, bounding $\max_\lambda \min_{u'} \Phi(u', \lambda) := \max_\lambda \min_{u'} \sum_{v \in V} (\lambda_v (\langle u, v \rangle - \langle u', \Pi v \rangle) - \varepsilon |\lambda_v| \cdot \|u\|_2)$

- For fixed λ, pick $P = \sum_v \lambda_v v$ and $u' = \|u\|_2 \cdot \frac{\Pi P}{\|\Pi P\|_2}$.
- $\Phi(u', \lambda) = \langle u, P \rangle - \langle u', \Pi P \rangle - \varepsilon \|\lambda\|_1 \|u\|_2$

  $= \langle u, P \rangle - \|u\|_2 \|\Pi P\|_2 - \varepsilon \|u\|_2$

  $\leq \|u_2\|_2 \cdot (\|P\|_2 - \|\Pi P\|_2 - \varepsilon)$

  ≤ 0 (convex hull distortion)

Proof of main theorem

Theorem. MMMR'18 embedding

$$f(x) = \begin{cases} (\Pi x, 0), & x \in X \\ (\Pi x_q + z, \sqrt{\|q - x_q\|_2^2 - \|z\|_2^2}), & x \in \mathbb{R}^d \backslash X \end{cases}$$

has terminal distortion $1 + O(\varepsilon)$.

Proof of main theorem

$$f(x) = \begin{cases} (\Pi x, 0), & x \in X \\ (\Pi x_q + z, \sqrt{\|q - x_q\|_2^2 - \|z\|_2^2}), & x \in \mathbb{R}^d \setminus X \end{cases}$$

Lemma:

(1) $\|z\|_2 \leq \|u\|_2$

(2) $\forall i, \ |\langle z, \Pi t_i \rangle - \langle u, t_i \rangle| \leq \varepsilon \|u\|_2 \cdot \|t_i\|_2$

Proof of main theorem

$$f(x) = \begin{cases} (\Pi x, 0), & x \in X \\ (\Pi x_q + z, \sqrt{\|q - x_q\|_2^2 - \|z\|_2^2}), & x \in \mathbb{R}^d \setminus X \end{cases}$$

Lemma:

(1) $\|z\|_2 \leq \|u\|_2$

(2) $\forall i,\ |\langle z, \Pi t_i \rangle - \langle u, t_i \rangle| \leq \varepsilon \|u\|_2 \cdot \|t_i\|_2$

Proof. Lemma w/ $t_i := x_i - x_q$, $u = q - x_q$. Def. $w_i := x_i - x_q$.

Proof of main theorem

$$f(x) = \begin{cases} (\Pi x, 0), & x \in X \\ (\Pi x_q + z, \sqrt{\|q - x_q\|_2^2 - \|z\|_2^2}), & x \in \mathbb{R}^d \backslash X \end{cases}$$

Lemma:

(1) $\|z\|_2 \leq \|u\|_2$

(2) $\forall i,\ |\langle z, \Pi t_i \rangle - \langle u, t_i \rangle| \leq \varepsilon \|u\|_2 \cdot \|t_i\|_2$

Proof. Lemma w/ $t_i := x_i - x_q$, $u = q - x_q$. Def. $w_i := x_i - x_q$.

$$\|f(q) - f(x_i)\|_2^2 = \|q - x_k\|_2^2 + \|\Pi w_i\|_2^2 - 2\langle z, \Pi w_i \rangle$$

$$\|q - x_i\|_2^2 = \|q - x_k\|_2^2 + \|w_i\|_2^2 - 2\langle q - x_k, w_i \rangle$$

Proof of main theorem

$$f(x) = \begin{cases} (\Pi x, 0), & x \in X \\ (\Pi x_q + z, \sqrt{\|q - x_q\|_2^2 - \|z\|_2^2}), & x \in \mathbb{R}^d \setminus X \end{cases}$$

Lemma:

(1) $\|z\|_2 \leq \|u\|_2$

(2) $\forall i,\ |\langle z, \Pi t_i \rangle - \langle u, t_i \rangle| \leq \varepsilon \|u\|_2 \cdot \|t_i\|_2$

Proof. Lemma w/ $t_i := x_i - x_q$, $u = q - x_q$. Def. $w_i := x_i - x_q$.

$$\|f(q) - f(x_i)\|_2^2 = \|q - x_k\|_2^2 + \|\Pi w_i\|_2^2 - 2\langle z, \Pi w_i \rangle$$
$$\|q - x_i\|_2^2 = \|q - x_k\|_2^2 + \|w_i\|_2^2 - 2\langle q - x_k, w_i \rangle$$

Absolute value of difference is

$$\leq O(\varepsilon)\|w_i\|_2^2 + 2|\langle z, \Pi w_i \rangle - \langle q - x_q, w_i \rangle|$$
$$\leq O(\varepsilon)\|w_i\|_2^2 + 2\varepsilon\|q - x_q\|_2\|w_i\|_2 \text{ (lemma)}$$
$$\leq O(\varepsilon)(\|w_i\|_2^2 + \|q - x_q\|_2^2) \text{ (AM-GM)}$$
$$\leq O(\varepsilon)\|q - x_i\|_2^2 \text{ (triangle inequality)}$$

Algorithm to compute $f(q)$

Algorithm to compute $f(q)$

Analysis says we can set $f(q) := z$ as long as

(1) $\|z\|_2 \leq \|q - x_q\|_2$

(2) $\forall i \; |\langle z, \underbrace{\Pi(x_i - x_q)}_{\alpha}\rangle - \underbrace{\langle q - x_q, x_i - x_q\rangle}_{\beta}| \leq \underbrace{\varepsilon \|q - x_q\|_2 \|x_i - x_q\|_2}_{\gamma}$

Algorithm to compute $f(q)$

Analysis says we can set $f(q) := z$ as long as

(1) $\|z\|_2 \leq \|q - x_q\|_2$

(2) $\forall i \; |\langle z, \underbrace{\Pi(x_i - x_q)}_{\alpha}\rangle - \underbrace{\langle q - x_q, x_i - x_q\rangle}_{\beta}| \leq \underbrace{\varepsilon \|q - x_q\|_2 \|x_i - x_q\|_2}_{\gamma}$

For fixed i, can express (2) by two linear constraints:
$\langle z, \alpha\rangle \leq \beta + \gamma$ and $\langle z, \alpha\rangle \geq \beta - \gamma$.

Algorithm to compute $f(q)$

Analysis says we can set $f(q) := z$ as long as

(1) $\|z\|_2 \leq \|q - x_q\|_2$

(2) $\forall i \; |\langle z, \underbrace{\Pi(x_i - x_q)}_{\alpha}\rangle - \underbrace{\langle q - x_q, x_i - x_q\rangle}_{\beta}| \leq \underbrace{\varepsilon \|q - x_q\|_2 \|x_i - x_q\|_2}_{\gamma}$

For fixed i, can express (2) by two linear constraints:
$\langle z, \alpha\rangle \leq \beta + \gamma$ and $\langle z, \alpha\rangle \geq \beta - \gamma$.

Thus we are trying to find z in the intersection of a Euclidean ball and a polytope; separation oracle implementation is thus simple.
$\Longrightarrow$ Can use ellipsoid algorithm to get $poly(n, d)$ time.
(can also be expressed as a Semidefinite Program)

Algorithm to compute $f(q)$

Analysis says we can set $f(q) := z$ as long as

(1) $\|z\|_2 \leq \|q - x_q\|_2$

(2) $\forall i \; |\langle z, \underbrace{\Pi(x_i - x_q)}_{\alpha}\rangle - \underbrace{\langle q - x_q, x_i - x_q\rangle}_{\beta}| \leq \underbrace{\varepsilon \|q - x_q\|_2 \|x_i - x_q\|_2}_{\gamma}$

For fixed i, can express (2) by two linear constraints:
$\langle z, \alpha\rangle \leq \beta + \gamma$ and $\langle z, \alpha\rangle \geq \beta - \gamma$.

Thus we are trying to find z in the intersection of a Euclidean ball and a polytope; separation oracle implementation is thus simple.
$\implies$ Can use ellipsoid algorithm to get $poly(n, d)$ time.
(can also be expressed as a Semidefinite Program)

Different construction with faster embedding algorithm possible [Cherapanamjeri, Nelson '21].