Optimality of the Johnson-Lindenstrauss lemma

Jelani Nelson
Harvard

May 19, 2017

joint work with Kasper Green Larsen (Aarhus)
Johnson-Lindenstrauss (JL) lemma

JL lemma [Johnson, Lindenstrauss ’84]

For every set $X$ of $n$ points in Euclidean space, 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 $x, y \in X$,

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

**JL lemma [Johnson, Lindenstrauss ’84]**

For every set $X$ of $n$ points in Euclidean space, there is an embedding $f : X \rightarrow \ell^m_2$ for $m = O(\varepsilon^{-2} \log n)$ with distortion $1 + \varepsilon$. That is, for each $x, y \in X$,

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

**Uses in computer science:**

- Speed up geometric algorithms by first reducing dimension of input [Indyk, Motwani ’98], [Indyk ’01]
- Faster/streaming numerical linear algebra algorithms [Sarlós ’06], [LWMRT ’07], [Clarkson, Woodruff ’09]
- Essentially equivalent to RIP matrices from compressed sensing [Baraniuk et al. ’08], [Krahmer, Ward ’11] (used for recovery of sparse signals)
- Volume-preserving embeddings (applications to projective clustering) [Magen ’02]
How to prove the JL lemma

Distributional JL (DJL) lemma

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

For any $0 < \varepsilon, \delta < 1/2$ there exists a distribution $D_{\varepsilon,\delta}$ on $\mathbb{R}^{m \times n}$ for $m = O(\varepsilon^{-2} \log(1/\delta))$ so that for any $u$ of unit $\ell_2$ norm

$$\mathbb{P}_{\Pi \sim D_{\varepsilon,\delta}} \left( |\|\Pi u\|_2^2 - 1| > \varepsilon \right) < \delta.$$
How to prove the JL lemma

Distributional JL (DJL) lemma

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

For any $0 < \varepsilon, \delta < 1/2$ there exists a distribution $D_{\varepsilon,\delta}$ on $\mathbb{R}^{m \times n}$ for $m = O(\varepsilon^{-2} \log(1/\delta))$ so that for any $u$ of unit $\ell_2$ norm

$$\mathbb{P}_{\Pi \sim D_{\varepsilon,\delta}} (|\|\Pi u\|_2^2 - 1| > \varepsilon) < \delta.$$ 

Proof of JL: Set $\delta = 1/n^2$ in DJL and $u$ as the normalized difference vector of some pair of points. Union bound over the $\binom{n}{2}$ pairs. Thus, in fact, the map $f : X \rightarrow \ell_2^m$ can be linear.
How to prove the JL lemma

Distributional JL (DJL) lemma

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

For any $0 < \varepsilon, \delta < 1/2$ there exists a distribution $D_{\varepsilon, \delta}$ on $\mathbb{R}^{m \times n}$ for $m = O(\varepsilon^{-2} \log(1/\delta))$ so that for any $u$ of unit $\ell_2$ norm

$$\mathbb{P}_{\Pi \sim D_{\varepsilon, \delta}} (|\|\Pi u\|_2^2 - 1| > \varepsilon) < \delta.$$ 

Proof of JL: Set $\delta = 1/n^2$ in DJL and $u$ as the normalized difference vector of some pair of points. Union bound over the $\binom{n}{2}$ pairs. Thus, in fact, the map $f : X \to \ell_m^2$ can be linear.

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

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

Theorem (Alon, 2003)

For JL, $m = \Omega((\varepsilon^{-2} / \log(1/\varepsilon)) \log n)$ is required.

Theorem (Larsen, N. 2014)

For JL, $m = \Omega(\varepsilon^{-2} \log n)$ is required if $f$ must be a linear map.
How to prove the JL lemma

Distributional JL (DJL) lemma

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

For any $0 < \varepsilon, \delta < 1/2$ there exists a distribution $\mathcal{D}_{\varepsilon, \delta}$ on $\mathbb{R}^{m \times n}$ for $m = O(\varepsilon^{-2} \log(1/\delta))$ so that for any $u$ of unit $\ell_2$ norm

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

Proof of JL: Set $\delta = 1/n^2$ in DJL and $u$ as the normalized difference vector of some pair of points. Union bound over the $\binom{n}{2}$ pairs. Thus, in fact, the map $f : X \rightarrow \ell_2^m$ can be linear.

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

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

Theorem (Alon, 2003)

For JL, $m = \Omega((\varepsilon^{-2} / \log(1/\varepsilon)) \log n)$ is required.

Theorem (Larsen, N. 2014)

For JL, $m = \Omega(\varepsilon^{-2} \log n)$ is required if $f$ must be a linear map.
How to prove the JL lemma

Distributional JL (DJL) lemma

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

For any $0 < \varepsilon, \delta < 1/2$ there exists a distribution $D_{\varepsilon, \delta}$ on $\mathbb{R}^{m \times n}$ for $m = O(\varepsilon^{-2} \log(1/\delta))$ so that for any $u$ of unit $\ell_2$ norm

$$\mathbb{P}_{\Pi \sim D_{\varepsilon, \delta}} (|\|\Pi u\|_2^2 - 1| > \varepsilon) < \delta.$$

Proof of JL: Set $\delta = 1/n^2$ in DJL and $u$ as the normalized difference vector of some pair of points. Union bound over the $(n^2)$ pairs. Thus, in fact, the map $f : X \rightarrow \ell^m_2$ can be linear.

Theorem (Jayram-Woodruff, 2011; Kane-Meka-N., 2011) For DJL, $m = \Theta(\varepsilon^{-2} \log(1/\delta))$ is optimal.

Theorem (Alon, 2003) For JL, $m = \Omega((\varepsilon^{-2} / \log(1/\varepsilon)) \log n)$ is required.

Theorem (Larsen, N. 2016) For JL, $m = \Omega(\varepsilon^{-2} \log n)$ is required if $f$ must be a linear map.
Theorem ([Larsen, N. ’16])

For any integers \(d, n \geq 2\) and any \(\epsilon > \frac{1}{(\min\{n, d\})^{0.4999}}\) such that \(\epsilon < 1\), there exists a set \(X \subset \ell^d_2\) such that any embedding \(f : X \rightarrow \ell^m_2\) with distortion at most \(1 + \epsilon\) must have

\[
m = \Omega(\epsilon^{-2} \log n)
\]
JL lower bound

Theorem ([Larsen, N. ’16])
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 \(f : X \rightarrow \ell_2^m\) 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, so might as well be \(\text{span}(e_1, \ldots, e_{n-1})\).
JL lower bound

Theorem ([Larsen, N. '16])

For any integers \(d, n \geq 2\) and any \(\frac{1}{(\min\{n,d\})^{0.4999}} < \varepsilon < 1\), there exists a set \(X \subset \ell^d_2\) such that any embedding \(f : X \to \ell^m_2\) 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, so might as well be \(\text{span}(e_1, \ldots, e_{n-1})\).
- So can only hope JL optimal for \(\varepsilon^{-2} \log n \leq \min\{n, d\}\), can view theorem assumption as \(\varepsilon^{-2} \log n \ll \min\{n, d\}^{0.999}\).
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.** \( m = \Omega\left(\frac{1}{\varepsilon^2} \frac{\log n}{\log(1/\varepsilon)}\right) \) [Alon '03]
Lower bounds over time

- **Volume argument.** $m = \Omega(\log n)$ [Johnson, Lindenstrauss ’84]
- **Incoherence + tensor trick.** $m = \Omega\left(\frac{1}{\varepsilon^2} \frac{\log n}{\log(1/\varepsilon)}\right)$ [Alon ’03]
- **Net argument + probabilistic method.** $m = \Omega\left(\frac{1}{\varepsilon^2} \log n\right)$ (only against linear maps $f(x) = \Pi x$) [Larsen, Nelson ’14]
Lower bounds over time

- **Volume argument.** $m = \Omega(\log n)$ [Johnson, Lindenstrauss '84]
- **Incoherence + tensor trick.** $m = \Omega\left(\frac{1}{\varepsilon^2} \frac{\log n}{\log(1/\varepsilon)}\right)$ [Alon '03]
- **Net argument + probabilistic method.** $m = \Omega\left(\frac{1}{\varepsilon^2} \log n\right)$ (only against linear maps $f(x) = \Pi x$) [Larsen, Nelson '14]
- **Encoding argument.** $m = \Omega\left(\frac{1}{\varepsilon^2} \log n\right)$ [Larsen, Nelson '16]
Encoding argument.

[Larsen, Nelson '16]
JL is optimal even against non-linear maps

- The previous two lower bounds had different hard sets but . . .
  . . . they were both incoherent!

- Incoherent collection: \( n \) unit vectors \( x_1, \ldots, x_n \in \mathbb{R}^m \) pairwise dot products \( \leq \varepsilon \) in magnitude

- Embedding of simplex: \( e_i \mapsto x_i \) 
  \[ \|x_i - x_j\|_2^2 = \|x_i\|_2^2 + \|x_j\|_2^2 - 2\langle x_i, x_j \rangle = 2(1 \pm \varepsilon) \]

- But JL isn't optimal for incoherent sets! (for small \( \varepsilon \)) (can get incoherence with smaller \( m \) via codes [Alon '03])

- [Larsen, N. '16]: doesn't give explicit hard \( X \); shows one exists (compression argument / pigeonhole principle)

- Defines a large collection \( X \) of \( n \)-sized sets \( X \subset \mathbb{R}^d \) s.t. if all \( X \in X \) can be embedded into dimension \( \leq 10 - 10 \cdot \varepsilon - 2 \log_2 n \), then there is an encoding of elements of \( X \) into \( < \log_2|X| \) bits (i.e. a surjection from \( X \) to \( \{0, 1\}^t \) for \( t < \log_2|X| \)). Contradiction.
JL is optimal even against non-linear maps

- The previous two lower bounds had different hard sets but . . .
  . . . they were both incoherent!

- **Incoherent collection:** $n$ unit vectors $x_1, \ldots, x_n \in \mathbb{R}^m$
pairwise dot products $\leq \varepsilon$ in magnitude

- **Embedding of simplex:** $e_i \mapsto x_i$
  $$\|x_i - x_j\|_2^2 = \|x_i\|_2^2 + \|x_j\|_2^2 - 2 \langle x_i, x_j \rangle = 2 \cdot (1 \pm \varepsilon)$$

- But JL isn't optimal for incoherent sets! (for small $\varepsilon$)
  (can get incoherence with smaller $m$ via codes [Alon '03])

- [Larsen, N. '16]: doesn't give explicit hard $X$; shows one exists
  (compression argument / pigeonhole principle)

- Defines a large collection $X$ of $n$-sized sets $X \subset \mathbb{R}^d$
s.t. if all $X \in X$ can be embedded into dimension $\leq 10 - 10 \cdot \varepsilon - 2 \log_2 n$, then there is an encoding of elements of $X$ into $\log_2 |X|$ bits (i.e. a surjection from $X$ to $\{0, 1\}^t$ for $t < \log_2 |X|$).
  Contradiction.
JL is optimal even against non-linear maps

- The previous two lower bounds had different hard sets but ... ... they were both incoherent!

- **Incoherent collection:** $n$ unit vectors $x_1, \ldots, x_n \in \mathbb{R}^m$
pairwise dot products $\leq \varepsilon$ in magnitude

- **Embedding of simplex:** $e_i \mapsto x_i$

  $\|x_i - x_j\|_2^2 = \|x_i\|_2^2 + \|x_j\|_2^2 - 2 \langle x_i, x_j \rangle = 2 \cdot (1 \pm \varepsilon)$

- But JL isn’t optimal for incoherent sets! (for small $\varepsilon$)(can get incoherence with smaller $m$ via codes [Alon ’03])
JL is optimal even against non-linear maps

- The previous two lower bounds had different hard sets but ... 
  ... they were both incoherent!

- **Incoherent collection**: $n$ unit vectors $x_1, \ldots, x_n \in \mathbb{R}^m$
  pairwise dot products $\leq \varepsilon$ in magnitude

- **Embedding of simplex**: $e_i \mapsto x_i$
  $\|x_i - x_j\|_2^2 = \|x_i\|_2^2 + \|x_j\|_2^2 - 2 \langle x_i, x_j \rangle = 2 \cdot (1 \pm \varepsilon)$

- But JL isn’t optimal for incoherent sets! (for small $\varepsilon$)
  (can get incoherence with smaller $m$ via codes [Alon '03])

- [Larsen, N. '16]: doesn’t give explicit hard $X$; shows one exists
  (compression argument / pigeonhole principle)
JL is optimal even against non-linear maps

- The previous two lower bounds had different hard sets but . . .
  . . . they were both incoherent!

- **Incoherent collection**: $n$ unit vectors $x_1, \ldots, x_n \in \mathbb{R}^m$
  pairwise dot products $\leq \varepsilon$ in magnitude

- **Embedding of simplex**: $e_i \mapsto x_i$
  $\|x_i - x_j\|_2^2 = \|x_i\|_2^2 + \|x_j\|_2^2 - 2 \langle x_i, x_j \rangle = 2 \cdot (1 \pm \varepsilon)$

- But JL isn’t optimal for incoherent sets! (for small $\varepsilon$)
  (can get incoherence with smaller $m$ via codes [Alon ’03])

- [Larsen, N. ’16]: doesn’t give explicit hard $X$; shows one exists
  (compression argument / pigeonhole principle)

- Defines a 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. a surjection from $\mathcal{X}$ to $\{0, 1\}^t$ for $t < \log_2 |\mathcal{X}|$). **Contradiction.**
Encoding argument.

[Kasper, Nelson '16]
Encoding argument.

[Kasper, Nelson '16]

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

- Preserving distances implies preserving dot products. Say $\|x\|_2 = \|y\|_2 = 1$.

\[
\|x - y\|_2^2 = \|x\|_2^2 + \|y\|_2^2 - 2\langle x, y \rangle \quad (*)
\]

\[
\|f(x) - f(y)\|_2^2 = \|f(x)\|_2^2 + \|f(y)\|_2^2 - 2\langle f(x), f(y) \rangle
\]

Now subtract $(*)$ from $(**)$:

\[
\langle f(x), f(y) \rangle = \langle x, y \rangle \pm O(\epsilon)
\]
Observation

- Preserving distances implies preserving dot products. Say \(\|x\|_2 = \|y\|_2 = 1\).

\[
\|x - y\|_2^2 = \|x\|_2^2 + \|y\|_2^2 - 2\langle x, y \rangle \quad (*)
\]

\[
\|f(x) - f(y)\|_2^2 = \|f(x)\|_2^2 + \|f(y)\|_2^2 - 2\langle f(x), f(y) \rangle
\]

\[
\Rightarrow (1 \pm \varepsilon)\|x - y\|_2^2 = (1 \pm \varepsilon)\|x\|_2^2 + (1 \pm \varepsilon)\|y\|_2^2 - 2\langle f(x), f(y) \rangle \quad (**) \]

- Now subtract (*) from (**) : \(\langle f(x), f(y) \rangle = \langle x, y \rangle \pm O(\varepsilon)\)
JL lower bound outline

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

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

- **Idea:** low-distortion embedding preserves dot products up to $\pm \varepsilon$, which is enough to distinguish the two cases.
JL lower bound outline

- Pick $k = \frac{1}{100\varepsilon^2}$.
- For $S \subset [d]$ of size $k$, define vector $y_S = \frac{1}{\sqrt{k}} \sum_{j \in S} e_j$. Note
  \[
  \langle y_S, e_i \rangle = \begin{cases} 
  10\varepsilon, & i \in S \\
  0, & \text{otherwise}
  \end{cases}
  \]
- **Idea:** low-distortion embedding preserves dot products up to $\pm \varepsilon$, which is enough to distinguish the two cases
- $\mathcal{X}$ 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 vector $y_S = \frac{1}{\sqrt{k}} \sum_{j \in S} e_j$. Note
  \[
  \langle y_S, e_i \rangle = \begin{cases} 
  10\varepsilon, & i \in S \\
  0, & \text{otherwise}
  \end{cases}
  \]
- **Idea:** low-distortion embedding preserves dot products up to $\pm \varepsilon$, which is enough to distinguish the two cases
- $\mathcal{X}$ 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
  \[
  \geq (n - d - 1) \lg \binom{d}{k} = (1 - o(\varepsilon))nk \lg(d/k)
  \] bits.
JL lower bound outline

- Pick \( k = \frac{1}{100\varepsilon^2} \).
- For \( S \subset [d] \) of size \( k \), define vector \( y_S = \frac{1}{\sqrt{k}} \sum_{j \in S} e_j \). Note \[
\langle y_S, e_i \rangle = \begin{cases} 10\varepsilon, & i \in S \\ 0, & \text{otherwise} \end{cases}
\]
- **Idea:** low-distortion embedding preserves dot products up to \( \pm \varepsilon \), which is enough to distinguish the two cases
- \( X \) 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 \( S_i \in \binom{[d]}{k} \).
- \(|X| = \binom{d}{k}^{n-d-1} \), thus any encoding of \( X \in X \)
  requires \( \geq (n - d - 1) \log \binom{d}{k} = (1 - o_\varepsilon(1))nk \log(d/k) \) bits.
- Will show any \((1 + \varepsilon)\)-distortion embedding into \( \ell_2^m \) 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 \( \varepsilon \) 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?**

Want to violate pigeonhole principle, so range of the encoding must be of size $< \lg |X|$. But $f(x)$ has real entries, so the range is infinite!

The fix: Round each $f(x)$ to a point $\tilde{f}(x)$ in a finite set (a net).
**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 $< \log |\mathcal{X}|$. 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!

**The fix:** Round each $f(x)$ to a point $\tilde{f}(x)$ in a finite set (a net).
A warmup lower bound

Recall $X = (0, e_1, \ldots, e_d, y_{S_1}, \ldots, y_{S_{n-d-1}})$. For $(1 + \varepsilon)$-distortion embedding $f : X \rightarrow \ell_2^m$, wlog $f(0) = 0$ (by translating).
A warmup lower bound

Recall $X = (0, e_1, \ldots, e_d, y_{s_1}, \ldots, y_{s_{n-d-1}})$. For $(1 + \varepsilon)$-distortion embedding $f : X \to \ell^m_2$, wlog $f(0) = 0$ (by translating).

- Since distances to 0 preserved, $\|f(x)\|_2^2 \leq 1 + \varepsilon$ for $x \in X$
  i.e. $\forall x \in X, f(x) \in (1 + \varepsilon)B_{\ell^m_2}$
A warmup lower bound

Recall $X = (0, e_1, \ldots, e_d, y_{S_1}, \ldots, y_{S_{n-d-1}})$. For $(1 + \varepsilon)$-distortion embedding $f : X \to \ell^m_2$, wlog $f(0) = 0$ (by translating).

- Since distances to 0 preserved, $\|f(x)\|^2_2 \leq 1 + \varepsilon$ for $x \in X$
i.e. $\forall x \in X$, $f(x) \in (1 + \varepsilon)B_{\ell^m_2}$
- Pick $c\varepsilon$-net $T$ of $(1 + \varepsilon)B_{\ell^m_2}$ in $\ell_2$; has size $N = O(1/\varepsilon)^m$. 

▶ Remember: $\langle e_i, y_S \rangle \in \{0, \pm 10\varepsilon\}$ (depends on whether $i \in S$)
▶ Low-distortion embedding preserves dot products, so $\langle f(e_i), f(y_S) \rangle \in \{\pm \varepsilon, \pm 2\varepsilon\}$
▶ Mapping to $c\varepsilon$-net points again preserves dot products, so $\langle f(e_i), f(y_S) \rangle \in \{\pm 2\varepsilon, \pm 4\varepsilon\}$
▶ Thus from encodings can recover $\langle e_i, y_S \rangle$ to know which $i \in S$ (dot product either $< 2\varepsilon$ in magnitude, or $> 8\varepsilon$)
▶ Can decode $X$, implies $nm \log(1/\varepsilon) = \Omega(n\varepsilon^{-2}\log(\varepsilon^{-2}d))$
A warmup lower bound

Recall $X = (0, e_1, \ldots, e_d, y_{S_1}, \ldots, y_{S_{n-d-1}})$. For $(1 + \varepsilon)$-distortion embedding $f : X \to \ell_2^m$, wlog $f(0) = 0$ (by translating).

- Since distances to 0 preserved, $\|f(x)\|_2^2 \leq 1 + \varepsilon$ for $x \in X$
  
  i.e. $\forall x \in X, f(x) \in (1 + \varepsilon)B_{\ell_2^m}$

- Pick $c\varepsilon$-net $T$ of $(1 + \varepsilon)B_{\ell_2^m}$ in $\ell_2$; has size $N = O(1/\varepsilon)^m$.

- Encode $f(x)$ as $\hat{f}(x) \in T$: $|X| \cdot \lg N = nm\lg(1/\varepsilon)$ bits
A warmup lower bound

Recall $X = (0, e_1, \ldots, e_d, y_{S_1}, \ldots, y_{S_{n-d-1}})$. For $(1 + \varepsilon)$-distortion embedding $f : X \to \ell_2^m$, wlog $f(0) = 0$ (by translating).

- Since distances to 0 preserved, $\|f(x)\|_2^2 \leq 1 + \varepsilon$ for $x \in X$
  i.e. $\forall x \in X$, $f(x) \in (1 + \varepsilon)B_{\ell_2^m}$
- Pick $c\varepsilon$-net $T$ of $(1 + \varepsilon)B_{\ell_2^m}$ in $\ell_2$; has size $N = O(1/\varepsilon)^m$.
- Encode $f(x)$ as $\tilde{f}(x) \in T$: $|X| \cdot \lg N = nm \lg (1/\varepsilon)$ bits
- **Remember**: $\langle e_i, y_S \rangle \in \{0, 10\varepsilon\}$ (depends on whether $i \in S$)
A warmup lower bound

Recall \( X = (0, e_1, \ldots, e_d, y_{s_1}, \ldots, y_{s_{n-d-1}}) \). For \((1 + \varepsilon)\)-distortion embedding \( f : X \to \ell_m^2 \), wlog \( f(0) = 0 \) (by translating).

- Since distances to 0 preserved, \( \|f(x)\|_2^2 \leq 1 + \varepsilon \) for \( x \in X \)
  i.e. \( \forall x \in X, \ f(x) \in (1 + \varepsilon)B_{\ell_2^m} \)
- Pick \( c\varepsilon \)-net \( T \) of \((1 + \varepsilon)B_{\ell_2^m} \) in \( \ell_2 \); has size \( N = O(1/\varepsilon)^m \).
- Encode \( f(x) \) as \( \widehat{f(x)} \in T: |X| \cdot \lg N = nm \lg(1/\varepsilon) \) bits
- \textbf{Remember:} \( \langle e_i, y_S \rangle \in \{0, 10\varepsilon\} \) (depends on whether \( i \in S \))
- Low-distortion embedding preserves dot products, so \( \langle f(e_i), f(y_S) \rangle \in \{\pm \varepsilon, 10\varepsilon \pm \varepsilon\} \)
A warmup lower bound

Recall $X = (0, e_1, \ldots, e_d, y_{s_1}, \ldots, y_{s_{n-d-1}})$. For $(1 + \varepsilon)$-distortion embedding $f : X \rightarrow \ell_2^m$, wlog $f(0) = 0$ (by translating).

- Since distances to 0 preserved, $\|f(x)\|_2^2 \leq 1 + \varepsilon$ for $x \in X$
  i.e. $\forall x \in X$, $f(x) \in (1 + \varepsilon)B_{\ell_2^m}$
- Pick $c\varepsilon$-net $T$ of $(1 + \varepsilon)B_{\ell_2^m}$ in $\ell_2$; has size $N = O(1/\varepsilon)^m$.
- Encode $f(x)$ as $\hat{f}(x) \in T$: $|X| \cdot \lg N = nm \lg(1/\varepsilon)$ bits
- **Remember:** $\langle e_i, y_S \rangle \in \{0, 10\varepsilon\}$ (depends on whether $i \in S$)
- Low-distortion embedding preserves dot products, so $\langle f(e_i), f(y_S) \rangle \in \{-\varepsilon, 10\varepsilon \pm \varepsilon\}$
- Mapping to $c\varepsilon$-net points again preserves dot products, so $\langle \hat{f}(e_i), \hat{f}(y_S) \rangle \in \{-2\varepsilon, 10\varepsilon \pm 2\varepsilon\}$
A warmup lower bound

Recall $X = (0, e_1, \ldots, e_d, y_{S_1}, \ldots, y_{S_{n-d-1}})$. For $(1 + \varepsilon)$-distortion embedding $f : X \rightarrow \ell_2^m$, wlog $f(0) = 0$ (by translating).

- Since distances to 0 preserved, $\|f(x)\|_2^2 \leq 1 + \varepsilon$ for $x \in X$
  i.e. $\forall x \in X, f(x) \in (1 + \varepsilon)B_{\ell_2^m}$
- Pick $c\varepsilon$-net $T$ of $(1 + \varepsilon)B_{\ell_2^m}$ in $\ell_2$; has size $N = O(1/\varepsilon)^m$.
- Encode $f(x)$ as $\tilde{f}(x) \in T$: $|X| \cdot \lg N = nm \lg(1/\varepsilon)$ bits
- **Remember**: $\langle e_i, y_S \rangle \in \{0, 10\varepsilon\}$ (depends on whether $i \in S$)
- Low-distortion embedding preserves dot products, so $\langle f(e_i), f(y_S) \rangle \in \{\pm \varepsilon, 10\varepsilon \pm \varepsilon\}$
- Mapping to $c\varepsilon$-net points again preserves dot products, so $\langle \tilde{f}(e_i), \tilde{f}(y_S) \rangle \in \{\pm 2\varepsilon, 10\varepsilon \pm 2\varepsilon\}$
- Thus from encodings can recover $\langle e_i, y_S \rangle$ to know which $i \in S$
  (dot product either $< 2\varepsilon$ in magnitude, or $> 8\varepsilon$)
A warmup lower bound

Recall \( X = (0, e_1, \ldots, e_d, y_{S_1}, \ldots, y_{S_{n-d-1}}) \). For \((1 + \varepsilon)\)-distortion embedding \( f : X \to \ell^m_2 \), wlog \( f(0) = 0 \) (by translating).

Since distances to 0 preserved, \( \|f(x)\|_2^2 \leq 1 + \varepsilon \) for \( x \in X \)
i.e. \( \forall x \in X, \ f(x) \in (1 + \varepsilon)B_{\ell_2^m} \)

Pick \( c\varepsilon \)-net \( T \) of \((1 + \varepsilon)B_{\ell_2^m} \) in \( \ell_2 \); has size \( N = O(1/\varepsilon)^m \).

Encode \( f(x) \) as \( \overline{f(x)} \in T: \ |X| \cdot \lg N = nm \lg(1/\varepsilon) \) bits

**Remember:** \( \langle e_i, y_S \rangle \in \{0, 10\varepsilon\} \) (depends on whether \( i \in S \))

Low-distortion embedding preserves dot products, so \( \langle f(e_i), f(y_S) \rangle \in \{\pm \varepsilon, 10\varepsilon \pm \varepsilon\} \)

Mapping to \( c\varepsilon \)-net points again preserves dot products, so \( \langle \overline{f(e_i)}, \overline{f(y_S)} \rangle \in \{\pm 2\varepsilon, 10\varepsilon \pm 2\varepsilon\} \)

Thus from encodings can recover \( \langle e_i, y_S \rangle \) to know which \( i \in S \)
(dot product either \( < 2\varepsilon \) in magnitude, or \( > 8\varepsilon \))

Can decode \( X \), implies \( nm \lg(1/\varepsilon) = \Omega(n\varepsilon^{-2} \log(\varepsilon^2 d)) \)
A warmup lower bound

Recall $X = (0, e_1, \ldots, e_d, y_{S_1}, \ldots, y_{S_{n-d-1}})$. For $(1 + \varepsilon)$-distortion embedding $f : X \rightarrow \ell^m_2$, wlog $f(0) = 0$ (by translating).

- Can decode $X$, implies $nm \lg(1/\varepsilon) = \Omega(n\varepsilon^{-2} \log(\varepsilon^2 d))$
A warmup lower bound

Recall $X = (0, e_1, \ldots, e_d, y_{S_1}, \ldots, y_{S_{n-d-1}})$. For $(1 + \varepsilon)$-distortion embedding $f : X \to \ell^m_2$, wlog $f(0) = 0$ (by translating).

- Can decode $X$, implies $nm \log(1/\varepsilon) = \Omega(n \varepsilon^{-2} \log(\varepsilon^2 d))$
- Thus $m = \Omega(\varepsilon^{-2} \log(\varepsilon^2 d) / \log(1/\varepsilon))$.
  
  for $\varepsilon$ not too small, this is $m = \Omega(\varepsilon^{-2} \frac{\log n}{\log(1/\varepsilon)})$
A warmup lower bound

Recall $X = (0, e_1, \ldots, e_d, y_{s_1}, \ldots, y_{s_{n-d-1}})$. For $(1 + \varepsilon)$-distortion embedding $f : X \to \ell_2^m$, wlog $f(0) = 0$ (by translating).

- Can decode $X$, implies $nm \log(1/\varepsilon) = \Omega(n \varepsilon^{-2} \log(\varepsilon^2 d))$
- Thus $m = \Omega(\varepsilon^{-2} \log(\varepsilon^2 d) / \log(1/\varepsilon))$. for $\varepsilon$ not too small, this is $m = \Omega(\varepsilon^{-2} \frac{\log n}{\log(1/\varepsilon)})$
- Same lower bound as [Alon '03], but very different argument.

... but not what I promised you!
A warmup lower bound

Recall $X = (0, e_1, \ldots, e_d, y_{s_1}, \ldots, y_{s_{n-d-1}})$. For $(1 + \varepsilon)$-distortion embedding $f : X \rightarrow \ell^m_2$, wlog $f(0) = 0$ (by translating).

- Can decode $X$, implies $n m \lg(1/\varepsilon) = \Omega(n \varepsilon^{-2} \log(\varepsilon^2 d))$
- Thus $m = \Omega(\varepsilon^{-2} \log(\varepsilon^2 d)/\log(1/\varepsilon))$.

  for $\varepsilon$ not too small, this is $m = \Omega(\varepsilon^{-2} \frac{\log n}{\log(1/\varepsilon)})$
- Same lower bound as [Alon ’03], but very different argument.


... but not what I promised you!
A warmup lower bound

Recall $X = (0, e_1, \ldots, e_d, y_{s_1}, \ldots, y_{s_{n-d-1}})$. For $(1 + \varepsilon)$-distortion embedding $f : X \to \ell^m_2$, wlog $f(0) = 0$ (by translating).

- Can decode $X$, implies $nmlg(1/\varepsilon) = \Omega(n\varepsilon^{-2} \log(\varepsilon^2 d))$
- Thus $m = \Omega(\varepsilon^{-2} \log(\varepsilon^2 d) / \log(1/\varepsilon))$.
  
  for $\varepsilon$ not too small, this is $m = \Omega(\varepsilon^{-2} \frac{\log n}{\log(1/\varepsilon)})$

- Same lower bound as [Alon ’03], but very different argument.
  
  … but not what I promised you!

- Will now show a better encoding.

remember, we are for now assuming $d = n/ \lfloor \mathbf{lg}(1/\varepsilon) \rfloor$
An encoding of $X$ into $O(nm)$ bits

Sufficed for decoding $X$: knowing $\langle f(e_i), f(y_{S_j}) \rangle$ for each $i, j$.
An encoding of $X$ into $O(nm)$ bits

Sufficed for decoding $X$: knowing $\langle \tilde{f}(e_i), \tilde{f}(y_{Sj}) \rangle$ 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 $\langle \tilde{f}(e_i), \tilde{f}(y_{Sj}) \rangle$ 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 $A$ with $\dim(E) \leq m$.

- $A \in \mathbb{R}^{d \times m}$
- $f(y_{S_j}) \in \mathbb{R}^m$
- $v_j \in \mathbb{R}^d$
An encoding of $X$ into $O(nm)$ bits

Let $E$ denote the column space of $A$

$$\text{dim}(E) \leq m.$$  

Define $K = E \cap (13\varepsilon B_{\ell_\infty}^d)$, $\forall j \ v_j \in K$

- $A \in \mathbb{R}^{d \times m}$
- $f(y_{S_j}) \in \mathbb{R}^m$
- $v_j \in \mathbb{R}^d$
An encoding of $X$ into $O(nm)$ bits

Let $E$ denote the column space of $A$

$\text{dim}(E) \leq m$.

Define $K = E \cap (13\varepsilon B_{\ell_\infty}^d)$, $\forall j \ v_j \in K$

$\implies \mathcal{N}(K, \frac{1}{13}K) \leq 2^{O(m)}$

$A \in \mathbb{R}^{d \times m}$

$f(y_{S_j}) \in \mathbb{R}^m$

$v_j \in \mathbb{R}^d$
An encoding of $X$ into $O(nm)$ bits

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

Define $K = E \cap (13\varepsilon B_{\ell_\infty}^d)$, \( \forall j \ \tilde{v}_j \in K \)

\[ \Rightarrow \mathcal{N}(K, \frac{1}{13}K) \leq 2^{O(m)} \]

Define $\tilde{v}_j$ as center of translate which contains $v_j$. \( O(m) \) bits.

\[ \Rightarrow ||v_j - \tilde{v}_j||_{\infty} < \varepsilon \]

- $A \in \mathbb{R}^{d \times m}$
- $f(y_{S_j}) \in \mathbb{R}^m$
- $\tilde{v}_j \in \mathbb{R}^d$
An encoding of $X$ into $O(nm)$ bits

Let $E$ denote the column space of $A$

$$\text{dim}(E) \leq m.$$  

Define $K = E \cap (13\varepsilon B_{\ell_\infty}^d)$, $\forall j \ v_j \in K$

$$\Rightarrow \mathcal{N}(K, \frac{1}{13} K) \leq 2^{O(m)}$$

Define $\tilde{v}_j$ as center of translate which contains $v_j$. $O(m)$ bits.

$$\Rightarrow \|v_j - \tilde{v}_j\|_\infty < \varepsilon$$

Encoding needs to specify $E$ (i.e. $A$).

Encode $f(e_i)$ using $O(m \log(1/\varepsilon))$ bits

For the $e_i$: $O(dm \lg(1/\varepsilon)) = O(nm)$ bits

$A \in \mathbb{R}^{d \times m}$

$f(y_{s_j}) \in \mathbb{R}^m$

$v_j \in \mathbb{R}^d$
An encoding of $X$ into $O(nm)$ bits

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

Define $K = E \cap (13\varepsilon B_{\ell_\infty}^d)$, $\forall j \ v_j \in K$

$\implies N(K, \frac{1}{13}K) \leq 2^{O(m)}$

Define $\tilde{v}_j$ as center of translate which contains $v_j$. $O(m)$ bits.

$\implies \|v_j - \tilde{v}_j\|_{\infty} < \varepsilon$

Encoding needs to specify $E$ (i.e. $A$).

Encode $\tilde{f}(e_i)$ 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
What about when $d \neq n/\log(1/\varepsilon)$?
Extending to arbitrary $d, n$

- Our setup: “dictionary atoms” $e_1, \ldots, e_d$ with vectors $y_S = \frac{1}{\sqrt{k}} \sum_{i \in S} e_i$ recovered from dot products with atoms
Extending to arbitrary \( d, n \)

- **Our setup:** “dictionary atoms” \( e_1, \ldots, e_d \) with vectors
  \[ y_S = \frac{1}{\sqrt{k}} \sum_{i \in S} e_i \] recovered from dot products with atoms

- **Modification:** use different dictionary \( x_1, \ldots, x_\Delta \)
  \[ y_S = \frac{1}{\sqrt{k}} \sum_{i \in S} x_i \]

- **want property that for most** \( S \in \binom{[\Delta]}{k} \), \( \langle y_S, x_i \rangle \pm \varepsilon \) indicates whether \( i \in S \)
Extending to arbitrary $d, n$

- Our setup: “dictionary atoms” $e_1, \ldots, e_d$ with vectors $y_S = \frac{1}{\sqrt{k}} \sum_{i \in S} e_i$ recovered from dot products with atoms
- Modification: use different dictionary $x_1, \ldots, x_\Delta$
  
  $y_S = \frac{1}{\sqrt{k}} \sum_{i \in S} x_i$

- Want property that for most $S \in \binom{[\Delta]}{k}$, $\langle y_S, x_i \rangle \pm \varepsilon$ indicates whether $i \in S$

- Can show such dictionary exists via probabilistic method when $\Delta \leq \max\{d, e^{O(\varepsilon^2 d)}\} \leq e^{O(\varepsilon^2 d)}$ (2nd inequality for $\varepsilon \gg \frac{1}{\sqrt{d}}$)

**Proof:** pick $x_i$ as independent gaussian vectors then do some computation
Extending to arbitrary $d, n$

- Our setup: “dictionary atoms” $e_1, \ldots, e_d$ with vectors
  \[ y_S = \frac{1}{\sqrt{k}} \sum_{i \in S} e_i \] recovered from dot products with atoms
- Modification: use different dictionary $x_1, \ldots, x_{\Delta}$
  \[ y_S = \frac{1}{\sqrt{k}} \sum_{i \in S} x_i \]
- want property that for most $S \in \binom{[\Delta]}{k}$, $\langle y_S, x_i \rangle \pm \varepsilon$ indicates whether $i \in S$
- Can show such dictionary exists via probabilistic method when
  $\Delta \leq \max\{d, e^{O(\varepsilon^2d)}\} \leq e^{O(\varepsilon^2d)}$ (2nd inequality for $\varepsilon \gg \frac{1}{\sqrt{d}}$)

**Proof:** pick $x_i$ as independent gaussian vectors then do some computation

**Summary:** Hard point sets for JL exist for $n$ up to $e^{O(\varepsilon^2d)}$ (beyond that point $\varepsilon^{-2} \log n \gg d$, so JL isn’t optimal)
What next?
Static approximate dot product

Two days after [Larsen, N. ’16]

▶ Noga Alon: “Hi Jelani, Kasper, I wonder ... if you can get a tight estimate for the number of possibilities for the \( \binom{n}{2} \) distances among n vectors of length at most 1 ...”
Static approximate dot product

Two days after [Larsen, N. ’16]

- Noga Alon: “Hi Jelani, Kasper, I wonder . . . if you can get a tight estimate for the number of possibilities for the \( \binom{n}{2} \) distances among \( n \) vectors of length at most 1 . . .”

- 4 later: problem solved! (for knowing up to additive \( \varepsilon \))

[Alon, Klartag ’16]: Given \( X \subset S^{d-1}, |X| = n \), can create a data structure consuming \( f(n, d, \varepsilon) \) bits such that can answer \( \text{query}(i, j) = \langle x_i, x_j \rangle \pm \varepsilon \) for any \( x_i, x_j \in X \).
Static approximate dot product

Two days after [Larsen, N. ’16]

- Noga Alon: “Hi Jelani, Kasper, I wonder ... if you can get a tight estimate for the number of possibilities for the \( \binom{n}{2} \) distances among \( n \) vectors of length at most 1 ...”

- 4 later: problem solved! (for knowing up to additive \( \varepsilon \))

[Alon, Klartag ’16]: Given \( X \subset S^{d-1} \), \( |X| = n \), can create a data structure consuming \( f(n, d, \varepsilon) \) bits such that can answer query \( (i, j) = \langle x_i, x_j \rangle \pm \varepsilon \) for any \( x_i, x_j \in X \).

- Here \( f(n, d, \varepsilon) \) is a bound they prove optimal for this problem

\[
f(n, d, \varepsilon) = \begin{cases} 
\frac{n \log n}{\varepsilon^2}, & \frac{\log n}{\varepsilon^2} \leq d \leq n \\
nd \log(2 + \frac{\log n}{\varepsilon^2 d}), & \log n \leq d \leq \frac{\log n}{\varepsilon^2} \\
nd \log(1/\varepsilon), & 1 \leq d \leq \log n
\end{cases}
\]
Static approximate dot product

Two days after [Larsen, N. '16]

▶ Noga Alon: “Hi Jelani, Kasper, I wonder . . . if you can get a tight estimate for the number of possibilities for the $\binom{n}{2}$ distances among n vectors of length at most 1 . . .”

▶ 4 later: problem solved! (for knowing up to additive $\varepsilon$)
[Alon, Klartag '16]: Given $X \subset S^{d-1}$, $|X| = n$, can create a data structure consuming $f(n, d, \varepsilon)$ bits such that can answer query $(i, j) = \langle x_i, x_j \rangle \pm \varepsilon$ for any $x_i, x_j \in X$.

▶ Here $f(n, d, \varepsilon)$ is a bound they prove optimal for this problem

\[
f(n, d, \varepsilon) = \begin{cases} 
\frac{n \log n}{\varepsilon^2}, & \frac{\log n}{\varepsilon^2} \leq d \leq n \\
nd \log(2 + \frac{\log n}{\varepsilon^2 d}), & \log n \leq d \leq \frac{\log n}{\varepsilon^2} \\
nd \log(1/\varepsilon), & 1 \leq d \leq \log n
\end{cases}
\]

▶ First case for $d$, upper bound for this data structural problem achieved earlier by [Kushilevitz, Ostrovsky, Rabani '98]
Static approximate dot product

[Alon, Klartag '16]: Given $X \subset S^{d-1}$, $|X| = n$, can create a data structure consuming $f(n, d, \varepsilon)$ bits such that can answer query $(i, j) = \langle x_i, x_j \rangle + O(\varepsilon)$ for any $x_i, x_j \in X$.  

Proof also via encoding argument. Essentially the problem is equivalent to the following: let $G$ be the set of all $n \times n$ Gram matrices of rank $d$. What is the logarithm of the size of the smallest $\varepsilon$-net of $G$ under entrywise $\ell_\infty$-norm?

Encode $X$ as name of closest net point to its Gram matrix. Also implies optimal JL lower bound!

$f(n, n, 2\varepsilon) \leq f(n, m, \varepsilon)$ if low-distortion embedding into $\ell_m^2$ existed (first embed points then build data structure). But [AK'16] gave upper bound on $f(n, m, \varepsilon)$, so $m$ can't be too small lest their lower bound on $f(n, n, 2\varepsilon)$ be violated.
Static approximate dot product

[Alon, Klartag '16]: Given $X \subset S^{d-1}$, $|X| = n$, can create a data structure consuming $f(n, d, \varepsilon)$ bits such that can answer $\text{query}(i, j) = \langle x_i, x_j \rangle + O(\varepsilon)$ for any $x_i, x_j \in X$.

- Proof also via encoding argument.

Essentially the problem is equivalent to the following: let $G$ be the set of all $n \times n$ Gram matrices of rank $d$. What is the logarithm of the size of the smallest $\varepsilon$-net of $G$ under entrywise $\ell_\infty$-norm?

Encode $X$ as name of closest net point to its Gram matrix.
Static approximate dot product

[Alon, Klartag '16]: Given $X \subset S^{d-1}$, $|X| = n$, can create a data structure consuming $f(n, d, \varepsilon)$ bits such that can answer query $\langle i, j \rangle = \langle x_i, x_j \rangle + O(\varepsilon)$ for any $x_i, x_j \in X$.

▶ Proof also via encoding argument.

Essentially the problem is equivalent to the following: let $G$ be the set of all $n \times n$ Gram matrices of rank $d$. What is the logarithm of the size of the smallest $\varepsilon$-net of $G$ under entrywise $\ell_\infty$-norm?

Encode $X$ as name of closest net point to its Gram matrix.

▶ Also implies optimal JL lower bound!

$f(n, n, 2\varepsilon) \leq f(n, m, \varepsilon)$ if low-distortion embedding into $\ell_2^m$ existed (first embed points then build data structure)
[Alon, Klartag ‘16]: Given $X \subset S^{d-1}$, $|X| = n$, can create a data structure consuming $f(n, d, \varepsilon)$ bits such that can answer query $(i, j) = \langle x_i, x_j \rangle + O(\varepsilon)$ for any $x_i, x_j \in X$.

Proof also via encoding argument.

Essentially the problem is equivalent to the following: let $G$ be the set of all $n \times n$ Gram matrices of rank $d$. What is the logarithm of the size of the smallest $\varepsilon$-net of $G$ under entrywise $\ell_\infty$-norm?

Encode $X$ as name of closest net point to its Gram matrix.

Also implies optimal JL lower bound!

$f(n, n, 2\varepsilon) \leq f(n, m, \varepsilon)$ if low-distortion embedding into $\ell_2^m$ existed (first embed points then build data structure)

But [AK’16] gave upper bound on $f(n, m, \varepsilon)$, so $m$ can’t be too small lest their lower bound on $f(n, n, 2\varepsilon)$ be violated.
Static approximate dot product

[Alon, Klartag ’16]: Given $X \subset S^{d-1}$, $|X| = n$, can create a data structure consuming $f(n, d, \varepsilon)$ bits such that can answer query $(i, j) = \langle x_i, x_j \rangle + O(\varepsilon)$ for any $x_i, x_j \in X$.

OPEN:

- dynamic approx. dot product with fast update/query?
- approximate distance query with relative $1 + \varepsilon$ error? (see [Indyk, Wagner ’17]; potential gap of $\lg(1/\varepsilon)$ remains)
And yet there’s more
**Conjecture:** ([Larsen, Nelson '16]) If $s(n, d, \varepsilon)$ is the optimal $m$ for distortion $1 + \varepsilon$ for $n$-point subsets of $\ell_2^d$, then

$$s(n, d, \varepsilon) = \Theta(\min\{n, d, \varepsilon^{-2} \log(2 + \varepsilon^2 n)\})$$

for all $\varepsilon, n, d$. (i.e. JL is suboptimal for $\varepsilon$ approaching $1/\sqrt{n}$)
**Conjecture:** ([Larsen, Nelson ’16]) If $s(n, d, \varepsilon)$ is the optimal $m$ for distortion $1 + \varepsilon$ for $n$-point subsets of $\ell_2^d$, then $s(n, d, \varepsilon) = \Theta(\min\{n, d, \varepsilon^{-2} \log(2 + \varepsilon^2 n)\})$ for all $\varepsilon, n, d$. (i.e. JL is suboptimal for $\varepsilon$ approaching $1/\sqrt{n}$)

**[Alon, Klartag ’16]:** some progress toward conjecture. Proved lower bound. As for upper bound ...
**Conjecture:** (Larsen, Nelson '16) If \( s(n, d, \varepsilon) \) is the optimal \( m \) for distortion \( 1 + \varepsilon \) for \( n \)-point subsets of \( \ell_2^d \), then

\[
s(n, d, \varepsilon) = \Theta(\min\{ n, d, \varepsilon^{-2} \log(2 + \varepsilon^2 n) \})
\]

for all \( \varepsilon, n, d \).

(i.e. JL is suboptimal for \( \varepsilon \) approaching \( 1/\sqrt{n} \))

[Alon, Klartag '16]: some progress toward conjecture. Proved lower bound. As for upper bound ...

for bipartite version of problem with \( x_1, \ldots, x_n, y_1, \ldots, y_n \) of unit norm, can show there exist \( a_1, \ldots, a_n, b_1, \ldots, b_n \in \mathbb{R}^m \) for

\[
m = O(\varepsilon^{-2} \log(2 + \varepsilon^2 n))
\]

with

\[
\forall i, j \ | \langle x_i, y_j \rangle - \langle a_i, b_j \rangle | \leq \varepsilon
\]
**Conjecture:** ([Larsen, Nelson ’16]) If \( s(n, d, \varepsilon) \) is the optimal \( m \) for distortion \( 1 + \varepsilon \) for \( n \)-point subsets of \( \ell^d_2 \), then
\[
s(n, d, \varepsilon) = \Theta\left(\min\{n, d, \varepsilon^{-2} \log(2 + \varepsilon^2 n)\}\right)
\]
for all \( \varepsilon, n, d \).
(i.e. JL is suboptimal for \( \varepsilon \) approaching \( 1/\sqrt{n} \))

**[Alon, Klartag ’16]:** some progress toward conjecture. Proved lower bound. As for upper bound ...  

for *bipartite* version of problem with \( x_1, \ldots, x_n, y_1, \ldots, y_n \) of unit norm, can show there exist \( a_1, \ldots, a_n, b_1, \ldots, b_n \in \mathbb{R}^m \) for \( m = O(\varepsilon^{-2} \log(2 + \varepsilon^2 n)) \) with
\[
\forall i, j \mid |\langle x_i, y_j \rangle - \langle a_i, b_j \rangle| \leq \varepsilon
\]

**[AK’16] results for small \( \varepsilon \) using tools from convex geometry:** low \( M^* \)-estimate [Pajor, Tomczak-Jaegermann ’86] and special cases of Gaussian correlation inequality (Khatri-Sidak lemma [Khatri ’67], [Sidak ’67] and Hargé’s inequality [Hargé ’99]).
A taste of [Alon, Klartag ’16]

Theorem
Suppose \( x_1, \ldots, x_n \in S^{d-1} \) with \( \frac{\log(2+\varepsilon^2 n)}{8\varepsilon^2} \leq d \leq n \) and \( \varepsilon \geq \frac{2}{\sqrt{n}} \).

Then for \( y \in S^{d-1} \), the number of possibilities for the vector

\[
\left( \left\lfloor \frac{\langle x_1, y \rangle}{\varepsilon} \right\rfloor, \left\lfloor \frac{\langle x_2, y \rangle}{\varepsilon} \right\rfloor, \ldots, \left\lfloor \frac{\langle x_n, y \rangle}{\varepsilon} \right\rfloor \right)
\]

is \( (\varepsilon^2 n)^\Theta(\frac{1}{\epsilon^2}) \) (so \( \varepsilon^{-2} \log(\varepsilon^2 n) \) bits).
Theorem
Suppose $x_1, \ldots, x_n \in S^{d-1}$ with $\frac{\log(2+\varepsilon^2 n)}{8\varepsilon^2} \leq d \leq n$ and $\varepsilon \geq \frac{2}{\sqrt{n}}$. Then for $y \in S^{d-1}$, the number of possibilities for the vector

\[
\left( \left\lfloor \frac{\langle x_1, y \rangle}{\varepsilon} \right\rfloor, \left\lfloor \frac{\langle x_2, y \rangle}{\varepsilon} \right\rfloor, \ldots, \left\lfloor \frac{\langle x_n, y \rangle}{\varepsilon} \right\rfloor \right)
\]

is $(\varepsilon^2 n)^\Theta\left(\frac{1}{\varepsilon^2}\right)$ (so $\varepsilon^{-2} \log(\varepsilon^2 n)$ bits).

Note: implies data structure for bipartite static approximate dot product using $O(n\varepsilon^{-2} \log(\varepsilon^2 n))$ bits of memory.
A taste of [Alon, Klartag ’16]

**Theorem**

Suppose $x_1, \ldots, x_n, y \in S^{d-1}$. Then the logarithm of the number of possibilities for $(\langle x_i, y \rangle \pm \varepsilon)_{i=1}^n$ is $\Theta(\varepsilon^{-2} \log(\varepsilon^2 n))$.

**Proof of theorem (will show upper bound):**
A taste of [Alon, Klartag '16]

Theorem

Suppose \( x_1, \ldots, x_n, y \in S^{d-1} \). Then the logarithm of the number of possibilities for \((\langle x_i, y \rangle \pm \varepsilon)_{i=1}^n\) is \( \Theta(\varepsilon^{-2} \lg(\varepsilon^2 n)) \).

Proof of theorem (will show upper bound):

- Define \( t \) s.t. \( \varepsilon^2 t = 2 \ln(2n/t) \), so \( t = \Theta(\varepsilon^{-2} \lg(\varepsilon^2 n)) \).
A taste of [Alon, Klartag ’16]

Theorem
Suppose $x_1, \ldots, x_n, y \in S^{d-1}$. Then the logarithm of the number of possibilities for $(\langle x_i, y \rangle \pm \varepsilon)_{i=1}^n$ is $\Theta(\varepsilon^{-2} \lg(\varepsilon^2 n))$.

Proof of theorem (will show upper bound):

- Define $t$ s.t. $\varepsilon^2 t = 2 \ln(2n/t)$, so $t = \Theta(\varepsilon^{-2} \lg(\varepsilon^2 n))$.
- Define convex body $K = \{ z \in \mathbb{R}^n : \forall i \in [n], \ |\langle \frac{z}{\sqrt{t}}, x_i \rangle| \leq \varepsilon \}$.
A taste of [Alon, Klartag '16]

**Theorem**

Suppose $x_1, \ldots, x_n, y \in S^{d-1}$. Then the logarithm of the number of possibilities for $(\langle x_i, y \rangle \pm \varepsilon)_{i=1}^n$ is $\Theta(\varepsilon^{-2} \lg(\varepsilon^2 n))$.

**Proof of theorem (will show upper bound):**

- Define $t$ s.t. $\varepsilon^2 t = 2 \ln(2n/t)$, so $t = \Theta(\varepsilon^{-2} \lg(\varepsilon^2 n))$.
- Define convex body $K = \{ z \in \mathbb{R}^n : \forall i \in [n], |\langle z/\sqrt{t}, x_i \rangle| \leq \varepsilon \}$.
- Khatri-Sidak says $\gamma_n(K) \geq (1 - 2e^{-\varepsilon^2 t/2})^n = (1 - \frac{t}{n})^n \approx e^{-t}$ where $\gamma_n$ is gaussian volume.
A taste of [Alon, Klartag '16]

Theorem

Suppose $x_1, \ldots, x_n, y \in S^{d-1}$. Then the logarithm of the number of possibilities for $(\langle x_i, y \rangle \pm \varepsilon)_{i=1}^n$ is $\Theta(\varepsilon^{-2} \log(\varepsilon^2 n))$.

Proof of theorem (will show upper bound):

- Define $t$ s.t. $\varepsilon^2 t = 2 \ln(2n/t)$, so $t = \Theta(\varepsilon^{-2} \log(\varepsilon^2 n))$.
- Define convex body $K = \{ z \in \mathbb{R}^n : \forall i \in [n], |\langle \frac{z}{\sqrt{t}}, x_i \rangle| \leq \varepsilon \}$.
- Khatri-Sidak says $\gamma_n(K) \geq (1 - 2e^{-\varepsilon^2 t/2})^n = (1 - \frac{t}{n})^n \approx e^{-t}$ where $\gamma_n$ is gaussian volume.
- can show $\gamma_n(z + A) \geq e^{-\|z\|^2/2} \gamma_n(A)$ for any $z, A$, so $\gamma_n(\sqrt{t}b + K) \geq e^{-t/2} \gamma_n(K) > e^{-2t}$ for any $\|b\|_2 \leq 1$.
A taste of [Alon, Klartag '16]

Theorem

Suppose $x_1, \ldots, x_n, y \in S^{d-1}$. Then the logarithm of the number of possibilities for $(\langle x_i, y \rangle \pm \varepsilon)_{i=1}^n$ is $\Theta(\varepsilon^{-2} \log(\varepsilon^2 n))$.

Proof of theorem (will show upper bound):

- Define $t$ s.t. $\varepsilon^2 t = 2 \ln(2n/t)$, so $t = \Theta(\varepsilon^{-2} \log(\varepsilon^2 n))$.
- Define convex body $K = \{ z \in \mathbb{R}^n : \forall i \in [n], |\langle \frac{z}{\sqrt{t}}, x_i \rangle | \leq \varepsilon \}$.
- Khatri-Sidak says $\gamma_n(K) \geq (1 - 2e^{-\varepsilon^2 t/2})^n = (1 - \frac{t}{n})^n \approx e^{-t}$ where $\gamma_n$ is gaussian volume.
- can show $\gamma_n(z + A) \geq e^{-\|z\|_2^2/2} \gamma_n(A)$ for any $z, A$, so $\gamma_n(\sqrt{t}b + K) \geq e^{-t/2} \gamma_n(K) > e^{-2t}$ for any $\|b\|_2 \leq 1$.
- $\mathbb{P}(E_b) \overset{\text{def}}{=} \mathbb{P}(|\langle \frac{g}{\sqrt{t}}, x_i \rangle - \langle b, x_i \rangle | \leq \varepsilon) = \mathbb{P}(g - \sqrt{tb} \in K) > e^{-2t}$
A taste of [Alon, Klartag ’16]

Theorem

Suppose \( x_1, \ldots, x_n, y \in S^{d-1} \). Then the logarithm of the number of possibilities for \( (\langle x_i, y \rangle \pm \varepsilon)_{i=1}^n \) is \( \Theta(\varepsilon^{-2} \log(\varepsilon^2 n)) \).

Proof of theorem (will show upper bound):

- Define \( t \) s.t. \( \varepsilon^2 t = 2 \ln(2n/t) \), so \( t = \Theta(\varepsilon^{-2} \log(\varepsilon^2 n)) \).
- Define convex body \( K = \{ z \in \mathbb{R}^n : \forall i \in [n], \| \frac{z}{\sqrt{t}}, x_i \| \leq \varepsilon \} \).
- Khatri-Sidak says \( \gamma_n(K) \geq (1 - 2e^{-\varepsilon^2 t/2})^n = (1 - \frac{t}{n})^n \approx e^{-t} \) where \( \gamma_n \) is gaussian volume
- can show \( \gamma_n(z + A) \geq e^{-\|z\|^2/2} \gamma_n(A) \) for any \( z, A \), so \( \gamma_n(\sqrt{t}b + K) \geq e^{-t/2} \gamma_n(K) > e^{-2t} \) for any \( \| b \|_2 \leq 1 \)
- \( \mathbb{P}(E_b) \overset{\text{def}}{=} \mathbb{P}(\| \frac{g}{\sqrt{t}}, x_i \| - \langle b, x_i \rangle \| \leq \varepsilon) = \mathbb{P}(g - \sqrt{t}b \in K) > e^{-2t} \)
- Define \( B \subset S^{d-1} \) to be maximal s.t. for all \( b \neq b' \in B, \exists i \) s.t. \( \| \langle b, x_i \rangle - \langle b', x_i \rangle \| > 2\varepsilon \).
A taste of [Alon, Klartag '16]

**Theorem**

Suppose $x_1, \ldots, x_n, y \in S^{d-1}$. Then the logarithm of the number of possibilities for $(\langle x_i, y \rangle \pm \varepsilon)^n_{i=1}$ is $\Theta(\varepsilon^{-2} \lg(\varepsilon^2 n))$.

**Proof of theorem (will show upper bound):**

- Define $t$ s.t. $\varepsilon^2 t = 2 \ln(2n/t)$, so $t = \Theta(\varepsilon^{-2} \lg(\varepsilon^2 n))$.
- Define convex body $K = \{ z \in \mathbb{R}^n : \forall i \in [n], \ |\langle z/\sqrt{t}, x_i \rangle \leq \varepsilon \}$.
- Khatri-Sidak says $\gamma_n(K) \geq (1 - 2e^{-\varepsilon^2 t/2})^n = (1 - \frac{t}{n})^n \approx e^{-t}$ where $\gamma_n$ is gaussian volume
- can show $\gamma_n(z + A) \geq e^{-\|z\|_2^2/2} \gamma_n(A)$ for any $z, A$, so $\gamma_n(\sqrt{t}b + K) \geq e^{-t/2} \gamma_n(K) > e^{-2t}$ for any $\|b\|_2 \leq 1$
- $\mathbb{P}(E_b) \overset{\text{def}}{=} \mathbb{P}(|\langle g/\sqrt{t}, x_i \rangle - \langle b, x_i \rangle| \leq \varepsilon) = \mathbb{P}(g - \sqrt{t}b \in K) > e^{-2t}$
- Define $B \subset S^{d-1}$ to be maximal s.t. for all $b \neq b' \in B$, $\exists i$ s.t. $|\langle b, x_i \rangle - \langle b', x_i \rangle| > 2\varepsilon$. Then $E_b$ are disjoint events, so $1 \geq \mathbb{P}(\bigcup_{b \in B} E_b) = \sum_b \mathbb{P}(E_b) > |B| \cdot e^{-2t}$, so $\lg |B| = O(t)$.
More open problems
Open problems

- Improved upper bound for constructing incoherent vectors? Maybe [Alon '03] sharp and **GV bound always suboptimal!**?
- Instance-wise optimality for $\ell_2$ dimensionality reduction? What’s the right $m$ in terms of $X$ itself? Bicriteria results?
- JL map that can be applied to $x$ in time $\tilde{O}(m + \|x\|_0)$? \(\|\cdot\|_0\) denotes support size
- Explicit DJL distribution with seed length $O(\log \frac{d}{\delta})$?
- **Rasmus Pagh:** Las Vegas algorithm for computing a JL map for set of $n$ points in time $o(n^2)$?