Recent Advances in Dimensionality Reduction with Provable Guarantees

Jelani Nelson
Harvard

July 12, 2018

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.

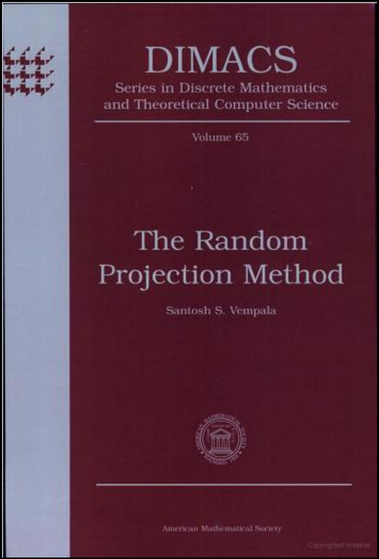
General setup

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 $x, y \in X$,

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

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 $z \in \mathbb{R}^d$

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

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.

Some natural questions

(**this talk:** and some recent answers)

Some natural questions

(**this talk:** and some recent answers)

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

Some natural questions

(**this talk:** and some recent answers)

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

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 $\|x_i - x_j\| \le \|L(x_i) - L(x_j)\| \le O(1) \cdot \|x_i - x_j\|$ for all $1 \le i, j \le n$. 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 $\|x_i - x_j\| \le \|L(x_i) - L(x_j)\| \le O(1) \cdot \|x_i - x_j\|$ for all $1 \le i, j \le n$. 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 $2^{\Omega(\alpha(k))}$).

Some natural questions

(**this talk:** and some recent answers)

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

The DJL distribution over Π

Older proofs

- [Johnson-Lindenstrauss, 1984], [Frankl-Maehara, 1988]: Random rotation, then projection onto first m coordinates.
- [Indyk-Motwani, 1998], [Dasgupta-Gupta, 2003]: Random matrix with independent Gaussian entries.
- [Achlioptas, 2001]: Independent ± 1 entries.
- [Clarkson-Woodruff, 2009]: $O(\log(1/\delta))$-wise independent ± 1 entries.
- [Arriaga-Vempala, 1999], [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]: Random rotation, then projection onto first m coordinates.
- [Indyk-Motwani, 1998], [Dasgupta-Gupta, 2003]: Random matrix with independent Gaussian entries.
- [Achlioptas, 2001]: Independent ± 1 entries.
- [Clarkson-Woodruff, 2009]: $O(\log(1/\delta))$-wise independent ± 1 entries.
- [Arriaga-Vempala, 1999], [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, 2006]: $x \mapsto PHDx$, $O(d \log d + m^3)$ time
  P is a random sparse matrix, H is Hadamard, D has random ± 1 on diagonal

Fast JL Transforms

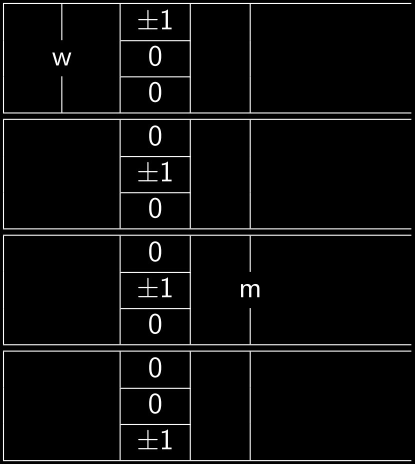
- [Ailon-Chazelle, 2006]: $x \mapsto PHDx$, $O(d \log d + m^3)$ time
  P is a random sparse matrix, H is Hadamard, D has random ± 1 on diagonal
- Several follow-up works with technical improvements: [Ailon-Liberty'08], [Ailon-Liberty'11], [Krahmer-Ward'11], [Rudelson-Vershynin'08], [Cheraghchi-Guruswami-Velingker'13], [N.-Price-Wootters'14], [Bourgain'14], [Haviv-Regev'16]

Fast JL Transforms

- [Ailon-Chazelle, 2006]: $x \mapsto PHDx$, $O(d \log d + m^3)$ time
  P is a random sparse matrix, H is Hadamard, D has random ± 1 on diagonal
- Several follow-up works with technical improvements: [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 $\Omega(\min\{m \cdot \|x\|_0, d \log d\})$.

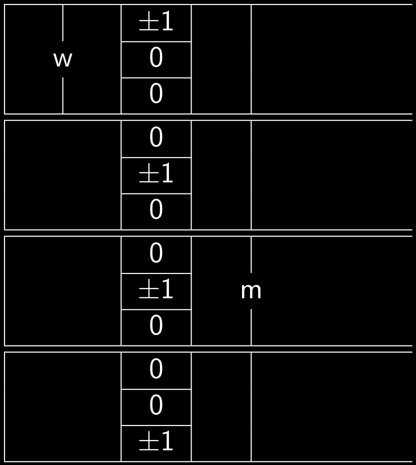
CountSketch [Charikar-Chen-FarachColton'02]

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

- 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], ...	$m \approx 4\varepsilon^{-2}\log(1/\delta)$	dense
[Achlioptas01]	$m/3$	sparse Bernoulli
[WDALS09]	no proof	hashing
[DKS10]	$\tilde{O}(\varepsilon^{-1}\log^3(1/\delta))$	hashing
[KN10a]*, [BOR10]*	$\tilde{O}(\varepsilon^{-1}\log^2(1/\delta))$	"
[KN14]	$O(\varepsilon^{-1}\log(1/\delta))$	CountSketch

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

Some natural questions

(**this talk:** and some recent answers)

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

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* $u \in S^{d-1}$

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

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* $u \in S^{d-1}$

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

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 $\mathbb{R}^{m\times d}$.

$$\begin{aligned}
&\forall x \in S^{d-1} \ \mathop{\mathbb{P}}_{\Pi\sim\mathcal{D}_{\varepsilon,\delta}}(|\|\Pi x\|_2^2 - 1| > \varepsilon) < \delta \\
\implies& \mathop{\mathbb{P}}_{x\sim\mathcal{F}} \ \mathop{\mathbb{P}}_{\Pi\sim\mathcal{D}_{\varepsilon,\delta}}(|\|\Pi x\|_2^2 - 1| > \varepsilon) < \delta \\
\implies& \mathop{\mathbb{P}}_{\Pi\sim\mathcal{D}_{\varepsilon,\delta}} \ \mathop{\mathbb{P}}_{x\sim\mathcal{F}}(|\|\Pi x\|_2^2 - 1| > \varepsilon) < \delta \\
\implies& \exists \Pi \in \mathbb{R}^{m\times d} \ \mathop{\mathbb{P}}_{x\sim\mathcal{F}}(|\|\Pi x\|_2^2 - 1| > \varepsilon) < \delta
\end{aligned}$$

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

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

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

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

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

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

(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 $\mathbb{R}^{m\times d}$.

$$\begin{aligned}
&\forall x \in S^{d-1} \quad \mathop{\mathbb{P}}_{\Pi\sim\mathcal{D}_{\varepsilon,\delta}}(|\|\Pi x\|_2^2 - 1| > \varepsilon) < \delta \\
\Longrightarrow &\mathop{\mathbb{P}}_{x\sim\mathcal{F}} \mathop{\mathbb{P}}_{\Pi\sim\mathcal{D}_{\varepsilon,\delta}}(|\|\Pi x\|_2^2 - 1| > \varepsilon) < \delta \\
\Longrightarrow &\mathop{\mathbb{P}}_{\Pi\sim\mathcal{D}_{\varepsilon,\delta}} \mathop{\mathbb{P}}_{x\sim\mathcal{F}}(|\|\Pi x\|_2^2 - 1| > \varepsilon) < \delta \\
\Longrightarrow &\exists \Pi \in \mathbb{R}^{m\times d} \quad \mathop{\mathbb{P}}_{x\sim\mathcal{F}}(|\|\Pi x\|_2^2 - 1| > \varepsilon) < \delta
\end{aligned}$$

(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 x is $\exp(-O(\varepsilon^2 m + 1)) \Longrightarrow m = \Omega(\varepsilon^{-2}\log(1/\delta))$ to succeed.

Some natural questions

(**this talk:** and some recent answers)

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

JL 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* $X \subset \ell_2^d$*,* $|X| = n$*, such that any embedding* $f : X \to \ell_2^m$ *with distortion at most* $1 + \varepsilon$ *must have*

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

JL 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 $X \subset \ell_2^d$, $|X| = n$, 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.

JL 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 $X \subset \ell_2^d$, $|X| = n$, 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.
- 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.** $m = \Omega(\frac{1}{\varepsilon^2} \frac{\log n}{\log(1/\varepsilon)})$ [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]

Encoding argument.

[Larsen, Nelson '17]

JL is optimal even against non-linear maps

We define 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. an injection from $\mathcal{X}$ to $\{0,1\}^t$ for $t < \log_2 |\mathcal{X}|$). **Contradiction.**

JL is optimal even against non-linear maps

We define 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 $\le 10^{-10} \cdot \varepsilon^{-2} \log_2 n$, 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 procedure based on very simple metric entropy / convex geometry argument.

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 $\varepsilon \to 1/\sqrt{n}$?

Some natural questions

(**this talk:** and some recent answers)

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

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

Some natural questions

(**this talk:** and some recent answers)

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

Application-specific

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

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

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

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

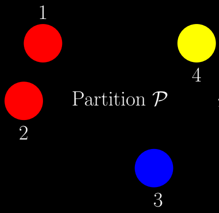
Clustering induces a k-partition $\mathcal{P}$ on $[n]$, so want to find best $\mathcal{P} = (P_1, \ldots, P_k)$. For fixed $\mathcal{P}$, best choice of y_j is centroid of P_j.

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

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

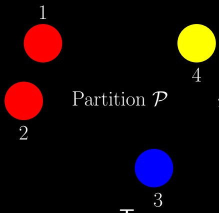
[Boutsidis-Zouzias-Mahoney-Drineas'11]: can reformulate k-means as a constrained low-rank approximation problem



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

Better dimension reduction for k-means

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

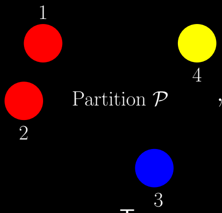


$$, \quad X_{\mathcal{P}} = \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & 0 \\ \frac{1}{\sqrt{2}} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, X_{\mathcal{P}} X_{\mathcal{P}}^{\mathsf{T}} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0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

$X_{\mathcal{P}} X_{\mathcal{P}}^{\mathsf{T}}$ is a rank-k orthogonal projection, and if we put points as rows of a matrix A, then $X_{\mathcal{P}} X_{\mathcal{P}}^{\mathsf{T}} A$ 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



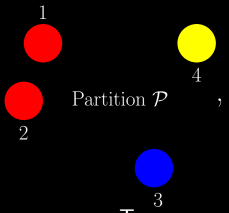
$$, \quad X_{\mathcal{P}} = \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & 0 \\ \frac{1}{\sqrt{2}} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, X_{\mathcal{P}} X_{\mathcal{P}}^{\mathsf{T}} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0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

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

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

Better dimension reduction for k-means

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



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

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

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

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

Better dimension reduction for k-means

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

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

Better dimension reduction for k-means

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

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

Dimensionality reduction for optimization:

$$\tilde{Q}_{opt} = \underset{Q \in \mathcal{Q}}{argmin} \|(I - Q)A\Pi^{\mathsf{T}}\|_F^2.$$

Better dimension reduction for k-means

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

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

Dimensionality reduction for optimization:

$$\tilde{Q}_{opt} = \underset{Q \in \mathcal{Q}}{argmin} \|(I - Q)A\Pi^{\mathsf{T}}\|_F^2.$$

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

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

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

Dimensionality reduction for optimization:

$$\tilde{Q}_{opt} = \underset{Q \in \mathcal{Q}}{argmin} \|(I - Q)A\Pi^{\mathsf{T}}\|_F^2.$$

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]).

Evidence "k" may be $\log k$: [CEMMP'15] shows $O(\log k)$ dimensions suffice for $O(1)$-approximation.
(just can't get down to $1 + \varepsilon$).

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

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

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

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

Equivalent to

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

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

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

- 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. $w(T) := \mathbb{E} \sup_{x \in T} \langle g, x \rangle$.

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

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

- 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. $w(T) := \mathbb{E} \sup_{x \in T} \langle g, x \rangle$.

  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.

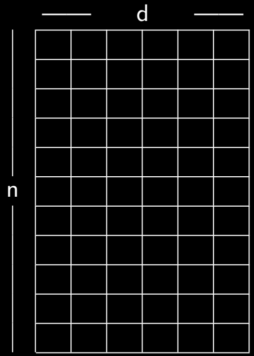
More dimensionality reduction: large matrices

Sketching for large matrix problems

Subspace embeddings [Sarlós'06]

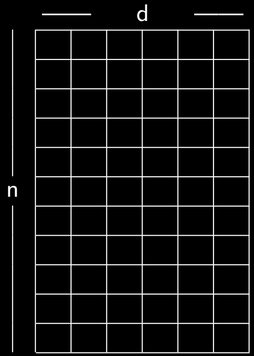
Following works by Drineas, Kannan, Mahoney, Mutukrishnan,

$A =$



Sketching for large matrix problems

$A =$

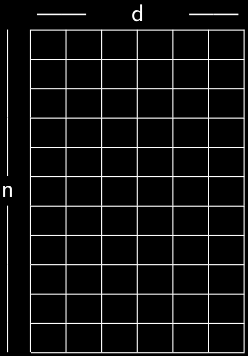


Subspace embeddings [Sarlós'06]

Following works by Drineas, Kannan, Mahoney, Mutukrishnan,

- **Subspace embedding:** Π is an ε-*subspace embedding* for a linear subspace E if $\forall x \in E$, $\|\Pi x\|_2^2 = (1 \pm \varepsilon)\|x\|_2^2$

Sketching for large matrix problems

$A =$ 

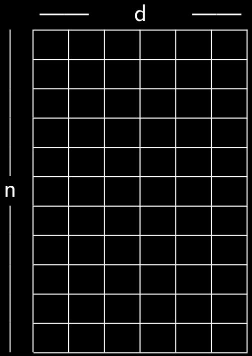
Subspace embeddings [Sarlós'06]

Following works by Drineas, Kannan, Mahoney, Mutukrishnan,

- **Subspace embedding:** Π is an *ε-subspace embedding* for a linear subspace E if $\forall x \in E$, $\|\Pi x\|_2^2 = (1 \pm \varepsilon)\|x\|_2^2$
- Given A, B with many rows n, approximate $A^\mathsf{T} B$ by $(\Pi A)^\mathsf{T} \Pi B$

Sketching for large matrix problems

$A =$



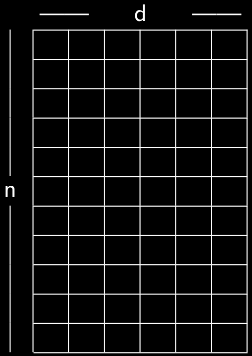
Subspace embeddings [Sarlós'06]

Following works by Drineas, Kannan, Mahoney, Mutukrishnan,

- **Subspace embedding:** Π is an *ε-subspace embedding* for a linear subspace E if $\forall x \in E$, $\|\Pi x\|_2^2 = (1 \pm \varepsilon)\|x\|_2^2$
- Given A, B with many rows n, approximate $A^\top B$ by $(\Pi A)^\top \Pi B$
- Given X, β, approximate least squares: $\tilde{\beta}^{LS} = argmin_\beta \|\Pi X\beta - \Pi y\|_2$

Sketching for large matrix problems

$A =$



Subspace embeddings [Sarlós'06]

Following works by Drineas, Kannan, Mahoney, Mutukrishnan,

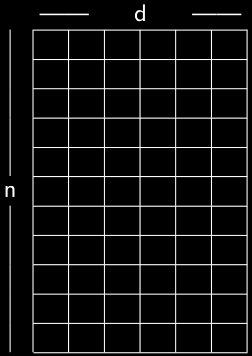
- **Subspace embedding:** Π is an ε-*subspace embedding* for a linear subspace E if $\forall x \in E$, $\|\Pi x\|_2^2 = (1 \pm \varepsilon)\|x\|_2^2$
- Given A, B with many rows n, approximate $A^{\mathsf{T}} B$ by $(\Pi A)^{\mathsf{T}} \Pi B$
- Given X, β, approximate least squares: $\tilde{\beta}^{LS} = argmin_{\beta} \|\Pi X \beta - \Pi y\|_2$
- Also applications to clustering [BZMD'11], [CEMMP'15], [CNW'16],

  PCA, and many other problems; see

  [Halko-Martinsson-Trop'11], [Mahoney'11], [Woodruff'14]

Sketching for large matrix problems

$A =$



Subspace embeddings [Sarlós'06]

Following works by Drineas, Kannan, Mahoney, Mutukrishnan,

- **Subspace embedding:** Π is an ε*-subspace embedding* for a linear subspace E if $\forall x \in E$, $\|\Pi x\|_2^2 = (1 \pm \varepsilon)\|x\|_2^2$
- Given A, B with many rows n, approximate $A^\top B$ by $(\Pi A)^\top \Pi B$
- Given X, β, approximate least squares: $\tilde{\beta}^{LS} = argmin_\beta \, \|\Pi X \beta - \Pi y\|_2$
- Also applications to clustering [BZMD'11], [CEMMP'15], [CNW'16],

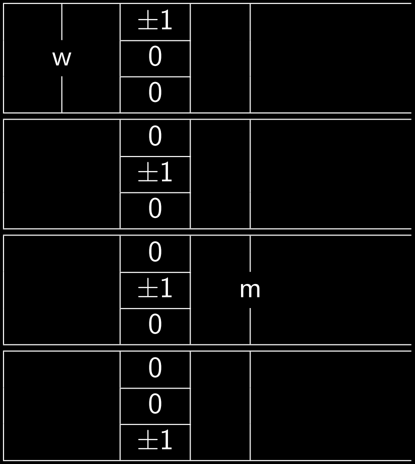
  PCA, and many other problems; see

  [Halko-Martinsson-Trop'11], [Mahoney'11], [Woodruff'14]

What Π to use?
E.g. regression want to compute ΠX quickly.

CountSketch [Charikar-Chen-FarachColton'02] (it's me again)

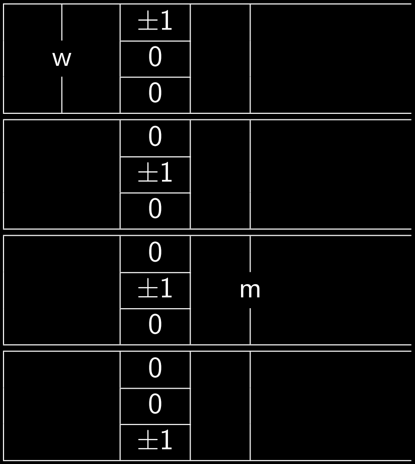
$$\Pi = \frac{1}{\sqrt{s}} \times$$



- Analyzed for approx. matrix mult, then regression and PCA, in [Kane-N.'12], [Clarkson-Woodruff'13], [Meng-Mahoney'13], [N.-Nguyễn'13], [BDN'15], [Cohen'16]

CountSketch [Charikar-Chen-FarachColton'02] (it's me again)

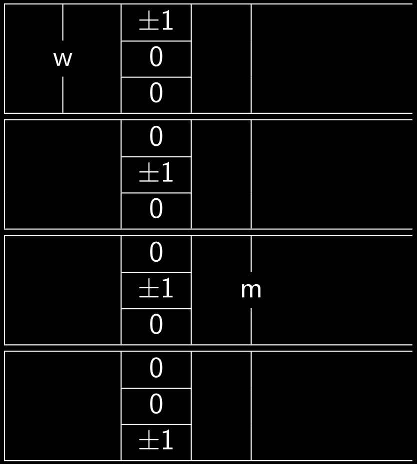
$$\Pi = \frac{1}{\sqrt{s}} \times$$



- Analyzed for approx. matrix mult, then regression and PCA, in [Kane-N.'12], [Clarkson-Woodruff'13], [Meng-Mahoney'13], [N.-Nguyễn'13], [BDN'15], [Cohen'16]
- For regression/PCA, $m = O(rank(X)/\varepsilon^2), s = 1$, or $m = \tilde{O}(rank(X)/\varepsilon^2), s = polylog(rank(X))/\varepsilon$ suffice

CountSketch [Charikar-Chen-FarachColton'02] (it's me again)

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



- Analyzed for approx. matrix mult, then regression and PCA, in [Kane-N.'12], [Clarkson-Woodruff'13], [Meng-Mahoney'13], [N.-Nguyễn'13], [BDN'15], [Cohen'16]
- For regression/PCA, $m = O(rank(X)/\varepsilon^2), s = 1$, or $m = \tilde{O}(rank(X)/\varepsilon^2), s = polylog(rank(X))/\varepsilon$ suffice
- Leads to algorithms with runtime $\approx$ the sparsity of X

Open problems

- Instance-wise optimality for ℓ_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 faster than repeated random projections then checking?