Fast Random Projections

Edo Liberty

Joint work with Nir Ailon.
Dimensionality reduction

Original space

\[ x_i, x_j \in \mathbb{R}^d \]

\[ \| x_i - x_j \|_2 \]

Target space

\[ \Psi : \mathbb{R}^d \rightarrow \mathbb{R}^k \]

\[ \| \Psi(x_i) - \Psi(x_j) \|_2 \approx \| x_i - x_j \|_2 \]

\[ (1 - \varepsilon) \| x_i - x_j \|_2 \leq \| \Psi(x_i) - \Psi(x_j) \|_2 \leq (1 + \varepsilon) \| x_i - x_j \|_2 \]

- \( \binom{n}{2} \) distances are \( \varepsilon \) preserved
- Target dimension \( k \) smaller than original dimension \( d \)
**Simple task:** search through your library of 10,000 images for near duplicates (on your PC).

**Problem:** your images are 5 Mega-pixels each. Your library occupies 22 Gigabytes of disk space and does not fit in memory.

**Possible solution:** Embed each image in a lower dimension (say 500). Then, search for close neighbors in the embedded points.

This can be done in memory on a moderately strong computer.
A distribution $\mathcal{D}$ over $k \times d$ matrices $\Psi$ s.t.

$$\forall x \in \mathbb{R}^{d-1} \quad \Pr_{\Psi \sim \mathcal{D}} [\|\Psi x\|_2 - 1 > \varepsilon] \leq 1/n^2$$

All $\binom{n}{2}$ pairwise distances are preserved w.p. at least $1/2$. 

Johnson Lindenstrauss Lemma

Lemma (Johnson Lindenstrauss 84)

\( \Psi = \text{uniformly chosen } k \text{ dimensional subspace (projection)} \)

\[
\Pr[|\|\Psi x\|_2 - 1| > \varepsilon] \leq c_1 e^{-c_2 \varepsilon^2 k}
\]

\[
k = \Theta(\log(n)/\varepsilon^2) \quad \rightarrow \quad \Pr \leq \frac{1}{n^2}
\]

Definition

Such distributions are said to exhibit the JL property.
What is this good for?

We get:
- Target dimension $k$ independent of $d$
- Target dimension $k$ logarithmic in $n$
- $\Psi$ chosen independently of input points

These make random projection extremely useful in:
- Linear Embedding / Dimensionality reduction
- Approximate-nearest-neighbor algorithms
- Rank $k$ approximation
- $\ell_1$ and $\ell_2$ regression
- Compressed sensing
- Learning

...
The distribution over the choice of $\Psi$ is rotation invariant, thus:

$$\Pr[|\|\Psi x\|_2 - 1| > \varepsilon] = \Pr_{x \sim U(S^{d-1})}[|\|l_k x\|_2 - 1| > \varepsilon]$$

Informally: projecting a **fixed vector** on a **random subspace** is equivalent to projecting a **random vector** on a **fixed subspace**.

From an isoperimetric inequality on the sphere, the norm of the first $k$ coordinates of a random unit vector is strongly concentrated around its mean.
Lemma (Frankl Meahara 87)

\[ \psi(i, j) \sim N(0, \frac{1}{\sqrt{k}}) \quad \rightarrow \quad \text{JL property.} \]

Proof.

Due to the rotational invariance of the Gaussian distribution:

\[ \| \psi x \|_2 \sim \sqrt{\frac{1}{k} x_k^2} \approx N(1, \frac{1}{\sqrt{k}}) \]

Which gives the JL property
Dense i.i.d. distributions

Lemma (Achlioptas 03, Matousek 06)

\[ \Psi(i, j) \in \{+1, -1\} \text{ uniformly} \rightarrow \text{JL property.} \]
\[ \Psi(i, j) \sim \text{any subgaussian distribution} \rightarrow \text{JL property.} \]

Proof.

\[ \| \Psi x \|_2^2 = \sum_{i=1}^{k} \langle \Psi(i), x \rangle^2 = \sum_{i=1}^{k} y_i^2 \]

The random variables \( y_i \) are i.i.d. and sub-Gaussian (Due to Hoeffding).

The proof above is due to Matousek.
The need for speed

All of the above distributions are such that:

- $\Psi$ requires $O(kd)$ space to store.
- Mapping $x \mapsto \Psi x$ requires $O(kd)$ operations.

Example: projecting a 5 Megapixel image to dimension 500:

- $\Psi$ takes up roughly 10 Gigabytes of memory.
- It takes roughly 5 hours to compute $x \mapsto \Psi x$.
  (very optimistic estimate for a 2Ghz CPU)
Sparse i.i.d. distributions

Can the projecting matrix be made sparser?

- Dasgupta, Kumar, Sarlos 09
- Kane, Nelson 10
- Braverman, Ostrovsky, Rabani 10

**Lemma (Kane, Nelson 10)**

*Number of non zeros in $\Psi$ can be $O(d \log(n) / \varepsilon)$, factor $\varepsilon$ better than naive.*

**Lemma (Dasgupta, Kumar, Sarlos 09)**

*This cannot be improved much.*

Proof: Consider input vectors like $[0, 0, 1, 0, 0, \ldots, 0, 1, 0]^T$

Can the projection be sparser if the input vectors are not sparse?
Sparse i.i.d. distributions

If the vectors are dense, the projection can be sparse!

**Lemma (Ailon Chazelle 06, Matousek 06)**

For some $q \in O(\eta^2 k) \leq 1$:

$$\psi(i, j) = \begin{cases} 
1/\sqrt{q} & \text{w.p. } q/2 \\
-1/\sqrt{q} & \text{w.p. } q/2 \\
0 & \text{w.p. } 1-q.
\end{cases}$$

for $x$ such that $\|x\|_\infty/\|x\|_2 \leq \eta$ (i.e. not sparse). **→ JL property**
FJLT: random-sign Fourier + sparse projection

Lemma (Ailon, Chazelle 06)

Let $\Phi$ be HD:

- $H$ is a Hadamard transform
- $D$ is a random $\pm 1$ diagonal matrix

$$\forall x \in S^{d-1} \quad \text{w.h.p.} \quad \|\Phi x\|_\infty \leq \sqrt{k/d}$$
FJLT: random-sign Fourier + sparse projection

\[ x \in S^{d-1} \]

\[ \| \Phi x \|_2 = O\left( \sqrt{\frac{k}{d}} \right) \]

\[ (S \Phi x) \in \mathbb{R}^k \]

\[ \| S \Phi x \|_2 \approx \| x \|_2 \]

Preprocess: Random-sign Fourier

Requires \( O(d \log(d)) \) operations

Project: Sparse projection matrix

Contains \( O(k^3) \) non zeros in expectation

Lemma (Ailon, Chazelle 06)

After the rotation, an expected number of \( O(k^3) \) nonzeros in \( S \) is sufficient for the JL property to hold.
FJLT: random-sign Fourier + sparse projection

\[ x \in S^{d-1} \]

Preprocess: Random-sign Fourier
Requires \( O(d \log(d)) \) operations

\[ \Phi x \|
\]

Project: Sparse projection matrix
contains \( O(k^3) \) non zeros in expectation

Lemma (Ailon, Chazelle 06)

\( S\Phi \) exhibits the JL property

Computing \( x \mapsto S\Phi x \) requires \( O(d \log(d) + k^3) \) operations

This is \( O(d \log(d)) \) if \( k \lesssim d^{1/3} \)
The belief is that \( O(d \log(d)) \) time is possible for JL property for all \( k \).
Can we remove this constraint by derandomizing the projection matrix?

Consider the distribution $\Psi = AD$:
- $A$ is a fixed $k \times d$ matrix.
- $D$ is a diagonal matrix, $D(i, i) = s(i)$ (Rademacher).

We have that:

$$\|ADx\|_2 = \left\| \sum_{i=1}^{d} A^{(i)} D(i, i) x(i) \right\|_2 = \left\| \sum_{i=1}^{d} A^{(i)} x(i) s(i) \right\|_2 = \|Ms\|_2$$

where $M^{(i)} = A^{(i)} x(i)$. 
Lemma ((L, Ailon, Singer 09) derived from Ledoux, Talagrand 91)

For any matrix $M$:

$$\Pr \left[ |\| Ms \|_2 - \| M \|_{Fro} | \geq \varepsilon \right] \leq 16e^{-\varepsilon^2/32\|M\|_2^2}$$

- Since $Ms = ADx$
- if $\|M\|_{Fro} = 1$ (true if $A$ is column normalized).
- and $\|M\|_2 = O(k^{-1/2})$.

$$\Pr \left[ |\| ADx \|_2 - 1 | \geq \varepsilon \right] \leq c_1 e^{-c_2\varepsilon^2 k}$$

We get the JL property
FJLT using dual BCH codes

**Holder’s inequality**

\[ \|M\|_{2\rightarrow 2} \in O \left( \|A^T\|_{2\rightarrow 4} \|x\|_4 \right) \]

**Lemma**

\( A \leftarrow \text{four-wise independent code matrix (concatenated code matrices)} \)

- \[ \|A^T\|_{2\rightarrow 4} \in O(d^{1/4}k^{-1/2}). \]
- Computing \( z \mapsto Az \) requires \( O(d \log(k)) \) operations.

**Lemma**

\( \Phi \leftarrow \text{concatenated random-sign Fourier transforms} \)

- \[ \|\Phi x\|_4 = O(d^{-1/4}) \text{ w.h.p.} \]
- Computing \( z \mapsto \Phi z \) requires \( O(d \log(d)) \) operations.
Lemma (Ailon, Liberty 08)

*Exhibits JL property and applicable in time $O(d \log d)$
*Construction exists for $k \lesssim d^{1/2}$.

The constraint on $k$ is weaker but still there...
Motivation from compressed sensing...

We want to get rid of the constraint on $k$ altogether.

On the one hand:
Preprocessing becomes a bottleneck for $k \in \Omega(\sqrt{d})$. We need to avoid it.

On the other hand:
Sparse vectors seem to require it.

There is hope:
Sparse Reconstruction (Compressed Sensing) constructions naturally deal with reconstructing sparse signals...
Motivation from compressed sensing...

Definition (Restricted Isometry Property (RIP))

For all $r$-sparse vectors $x$:

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

Lemma (Rudelson, Vershynin 08, Candes, Romberg, Tau 06)

$$\Psi \leftarrow \frac{r \log^4(d)}{\varepsilon^2} \text{ random rows (frequencies) from Hadamard matrix, then w.p. } \Psi \text{ is RIP.}$$

- The same approximate isometric condition as random projections
- Deals with sparse vectors without preprocessing
- No constraint (e.g. $\sqrt{d}$ upper bound) on $r$
- Very simple construction
Almost optimal JL transform

\[ k = O \left( \log(n) \text{ polylog}(d) / \varepsilon^4 \right) \]

Hadamard Matrix

\[
\begin{pmatrix}
H
\end{pmatrix}
\]

RIP

\[
\begin{pmatrix}
\Phi
\end{pmatrix}
\]

JL property

\[
\begin{pmatrix}
+1 & -1 \\
-1 & \ddots & -1
\end{pmatrix}
\]

Lemma

For any set \( X \) of cardinality \( n \), with constant probability:

\[
\forall x \in X \quad (1 - \varepsilon) \| x \|_2^2 \leq \left\| \frac{1}{\sqrt{k}} \Phi Dx \right\|_2^2 \leq (1 + \varepsilon) \| x \|_2^2.
\]

- Fast for all \( k \).
- Very simple construction (application time is \( O(d \log(d)) \))
Almost optimal JL transform

\[ r = O\left(\frac{\log(n)}{\epsilon^2}\right) \]

\[ x = \hat{x} + \tilde{x} \]

- \( \hat{x} \) is the \( r \)-sparse vector containing the \( r \) largest entries in \( x \).
- \( \tilde{x} \) contains the rest. \( \|\tilde{x}\|_\infty \leq 1/\sqrt{r} \).
Almost optimal JL transform

\[ r = O \left( \log \left( \frac{n}{k^{1/2}} \right) \right) \]

\[ k = O \left( \log(n) \log^4(d) / \epsilon^4 \right) \]

Lemma (Rudelson, Vershynin 08)

w.p. \( \forall \ x \in X \)

\[ \left\| \frac{1}{\sqrt{k}} \Phi D \hat{x} \right\|^2 = \| \hat{x} \|^2 + O(\epsilon) \]

Using the RIP property as black box.
Almost optimal JL transform

\[ 2 \begin{pmatrix} k^{-1/2} \Phi D \end{pmatrix} \begin{pmatrix} \hat{x} \end{pmatrix}^T = O\left(\epsilon\right) \]

Lemma

\[ w.p. \quad \forall \ x \in X \quad \frac{2}{k} (\Phi D\hat{x})^T \Phi D\hat{x} = O(\epsilon) \]

Not hard to show using Hoeffding’s inequality. (Note that this function is linear in random bits supporting \( \hat{x} \))
Almost optimal JL transform

\[
\left\| k^{-1/2} \Phi D \right\|_2^2 = \| \tilde{x} \|_2^2 + O(\epsilon)
\]

Main technical lemma:

Lemma (Extension of Rudelson and Vershynin, and Talagrand.)

\[
w.p. \quad \forall \ x \in X \quad \left\| \frac{1}{\sqrt{k}} \Phi D \tilde{x} \right\|^2 = \| \tilde{x} \|^2 + O(\epsilon)
\]
Almost optimal JL transform

From Talagrand: \[ \left\| \frac{1}{\sqrt{k}} \Phi D\tilde{x} \right\| = \|\tilde{x}\| + O(\varepsilon) \] if:

\[ \left\| \frac{1}{\sqrt{k}} \Phi D\tilde{x} \right\|^2_2 \in O\left(\frac{\varepsilon^2}{\log(n)}\right) \]

where \( D\tilde{x} \) is diagonal matrix with \( \tilde{x} \) on its diagonal.

By triangle inequality:

\[ \left\| \frac{1}{\sqrt{k}} \Phi D\tilde{x} \right\|^2_2 = \left\| \frac{1}{k} D\tilde{x} \Phi^t \Phi D\tilde{x} \right\|_2 \leq \left\| \frac{1}{k} D\tilde{x} \Phi^t \Phi D\tilde{x} - D^2_{\tilde{x}} \right\|_2 + \left\| D^2_{\tilde{x}} \right\|_2 \]

By the choice of \( \tilde{x} \):

\[ \left\| D^2_{\tilde{x}} \right\|_2 = \|\tilde{x}\|_\infty^2 \leq 1/r = \varepsilon^2 / \log(n) \]

To conclude the proof we need a similar bound for

\[ \left\| \frac{1}{k} D\tilde{x} \Phi^t \Phi D\tilde{x} - D^2_{\tilde{x}} \right\|_2. \]
Lemma (Rudelson, Vershynin + careful modifications)

\[ E_\Phi \left[ \sup_{\|z\|_2 \leq 1, \|z\|_\infty \leq \alpha} \left\| D_z^2 - \frac{1}{k} D_z \phi^t \phi D_z \right\| \right] \in O \left( \frac{\alpha \log^2(d)}{\sqrt{k}} \right). \]

Substituting our choice of \( \alpha^2 = 1/r = \frac{\varepsilon^2}{\log(n)} \) and

\[ k \in \Omega \left( \frac{\log(n) \log^4(d)}{\varepsilon^4} \right) \]

Satisfies the required bound and concludes the proof.
This approach seems to actually give dependence $\varepsilon^{-3}$ instead of $\varepsilon^{-4}$ as presented.

Krahmer and Ward 10 show that any RIP construction becomes a JL construction if you add a random sign matrix. This fixes the dependence on $\varepsilon$ to the correct $\varepsilon^{-2}$. It also uses RIP constructions as a black box.

Future work:

- Eliminating the $\text{polylog}(d)$ factor for JL with no restriction on $k$. This will also give an improved RIP construction.
- Improving our understanding of random projections for sparse input vectors, e.g. bag of words models of text documents.
Fin