Fast Random Projections using Lean Walsh Transforms

Edo Liberty [1] Nir Ailon [2] Amit Singer [3]

[1]Yale University, Department of Computer Science.
[2]Google Research
[3]Yale University, Program in Applied Mathematics

What are random projections?

A $k \times d$ matrix Ψ is a Random Projection if:

$$\Pr[|\|\Psi x\|_2 - 1| > \varepsilon] \leq \frac{1}{n} \tag{1}$$

for some $0 < \varepsilon < 1/2$ and arbitrary large n.

Johnson and Lindenstrauss showed that there exist such a matrix Ψ if $k > O(\log(n)/\varepsilon^2)$

What are they good for?

For different choices of ε and n Random projections found exciting applications in theory and in practice:

- Approximate nearest neighbors searches
- Linear Embedding / Dimensionality reduction
- Rank k approximation
- ℓ_1 and ℓ_p regressions
- Compressed sensing

and the list continues...

Known constructions

Classic random projection constructions:

- The original proof and construction, W.B.Johnson and J.Lindenstrauss (1984). They used k rows from random orthogonal matrix (random projection matrix).
- P.Frankl and H.Meahara (1987), P.Indyk and R.Motowani (1998), and S.DasGupta and A.Gupta (1999) used a random Gaussian distribution, $\Psi(i,j) \sim N(0,1)$. The rotational invariance of the normal distribution makes the proof much easier.
- Dimitris Achlioptas (2003) showed that Ψ can be a dense $\Psi(i,j) \in \{0,-1,1\}$ matrix.
- Jiri Matousek (2006) extended the result of Achlioptas to any i.i.d sub-gaussian entrees.

Fast Johnson Lindenstrauss Transforms

Lemma (Ailon, Chazelle (2006))

The matrix $\Psi = PHD_s$ is a random projection.

- *Let $P \in \mathbb{R}^{k \times d}$ be a sparse matrix. Let $q = \Theta(\frac{\log^2(n)}{d})$, set $P(i,j) \sim N(0, q^{-1})$ w.p q and $P(i,j) = 0$ else.*
- *Let H denote the $d \times d$ Walsh Hadamard matrix.*
- *Let D_s denote a $d \times d$ diagonal matrix such that $D(i,i)$ are i.i.d random ± 1.*

Notice that P contains only $O(k^3)$ entrees (in expectancy) and applying $x \mapsto HD_s x$ takes $d \log(d)$ operations to compute.

Computing $x \mapsto HD_s x$ is $O(d \log(d) + k^3)$ operations.

Fast Johnson Lindenstrauss Transforms

Lemma (Ailon, Liberty (2008))

The matrix $\Psi = BD_s\Phi$ is a random projection.

- *Let $B \in \mathbb{R}^{k \times d}$ be a 4-wise independent ± 1 matrix.*
- *and let D_s denote a $d \times d$ diagonal random ± 1 matrix.*
- *Let Φ be block-wise composition of randomized Hadamard matrices (details omitted)*

Running time is $d \log(k)$ for all $k < d^{1/2-\delta}$ for arbitrary small δ.

Fast JL idea

The idea for both Fast JL algorithms is the same.
Choose $\Psi = AD_s\Phi$ where:

- A is a fast applicable $k \times d$ matrix.
- D_s is a random diagonal.
- Φ is an isometric preprocessing matrix.

Schematically

$$\begin{array}{ccccc} \mathbb{R}^k \leftarrow \chi & & \chi \leftarrow \mathbb{R}^d & & \mathbb{R}^d \\ & \Leftarrow & & \Leftarrow & \\ AD_s & & \Phi & & x \end{array} \tag{2}$$

χ for A

χ is the "good" set of x for which AD_s is a good random projection.

We say that $x \in \chi(A, \varepsilon, n)$ if:

$$\Pr(|\|AD_s x\|_2 - 1| > \varepsilon) \leq \frac{1}{n} \tag{3}$$

What is the relationship between χ and A?

How fast can A be applied?

Previous results

	The rectangular $k \times d$ matrix A	Application time	$x \in \chi$ if:
Johnson, Lindenstrauss	Top k rows of a random unitary matrix	$O(kd)$	$x \in \mathbb{R}^d$
Various, Authors	i.i.d random entries	$O(dk)$	$x \in \mathbb{R}^d$
Ailon, Chazelle	Sparse Gaussian entrees	$O(k^3)$	$\\|x\\|_\infty \leq O((d/k)^{-1/2})$
Ailon, Liberty	4-wise independent Code matrix	$O(d \log k)$	$\\|x\\|_4 \leq O(d^{-1/4})$
This work	Any matrix	?	$\\|x\\|_A \leq O(k^{-1/2})$
This work	Lean Walsh	$O(d)$	$\\|x\\|_\infty \leq O(k^{-1/2} d^{-\delta})$

Probability for distortion

Let us compute

$$\Pr(|\|AD_s x\|_2 - 1| > \varepsilon) \tag{4}$$

Notice that $AD_s x = AD_x s \equiv Ms$.

Lemma (Talagrand)

Consider a matrix M and a random vector s, $s(i)$ are i.i.d ± 1 w.p $1/2$. Define the random variable $Y = \|Ms\|_2$. Denote by μ the median of Y, and $\sigma = \|M\|_{2\to 2}$ the spectral norm of M. Then

$$\Pr[\,|Y - \mu| > t] \leq 4e^{-t^2/8\sigma^2} \tag{5}$$

Success condition for σ

We substitute the following terms:

- $E[Y^2] = 1$ if the columns of A are normalized to 1 (or normalized in expectancy).
- $|1 - \mu| \leq \sqrt{32}\sigma$ (By computing $E[(Y - \mu)^2]$)

And we get:

$$\Pr(|\|AD_s x\|_2 - 1| > \varepsilon) \leq \frac{1}{n} \tag{6}$$

if $\sigma = \|AD_x\|_{2\to 2} \leq k^{-1/2}$ where $k = O(\log(n)/\varepsilon^2)$.

Definition of, and condition on $\|x\|_A$

Define $\|x\|_A = \|AD_x\|_{2\to 2}$.

$$\|x\|_A \leq k^{-1/2} \Rightarrow x \in \chi(A, \varepsilon, n) \tag{7}$$

Lemma

For any column normalized matrix A and x, such that $\|x\|_2 = 1$ and $\|x\|_A \leq k^{-1/2}$:

$$\Pr(|\|AD_s x\|_2 - 1| > \varepsilon) \leq \frac{1}{n} \tag{8}$$

Lean Walsh Transforms, Definition by example

Lean Walsh matrices are Kronecker powers of a *seed* matrix. (Much like Walsh Hadamard matrices).

$$A_1 = \frac{1}{\sqrt{3}} \begin{pmatrix} 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix} \tag{9}$$

$$A_\ell = \frac{1}{\sqrt{3}} \begin{pmatrix} A_{\ell-1} & A_{\ell-1} & -A_{\ell-1} & -A_{\ell-1} \\ A_{\ell-1} & -A_{\ell-1} & A_{\ell-1} & -A_{\ell-1} \\ A_{\ell-1} & -A_{\ell-1} & -A_{\ell-1} & A_{\ell-1} \end{pmatrix} \tag{10}$$

Applying A_ℓ

The size of A is $3^\ell \times 4^\ell$.
Since $d = 4^\ell$ we have $\ell = \log d / \log 4$.

So, the size of A is $d^\alpha \times d$ for $\alpha < 1$.

Applying the lean Walsh matrix of any size is $O(d)$.

$$A'_\ell z = A'_\ell \begin{pmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{pmatrix} = \frac{1}{\sqrt{3}} \begin{pmatrix} A'_{\ell-1}(z_1 + z_2 - z_3 - z_4) \\ A'_{\ell-1}(z_1 - z_2 + z_3 - z_4) \\ A'_{\ell-1}(z_1 - z_2 - z_3 + z_4) \end{pmatrix} \tag{11}$$

We see that $T(d) = 3T(d/4) + 3d$
and so $T(d) = O(d)$.

Connection between $\|x\|_A$ and $\|x\|_\infty$

What is the condition on $\|x\|_\infty$ such that $\|x\|_A \leq k^{-1/2}$.

$$\begin{aligned}
\|x\|_A^2 &= \|AD_x\|_{2\to 2}^2 = \max_{y,\|y\|_2=1} \|y^T AD_x\|_2^2 && (12)\\
&= \max_{y,\|y\|_2=1} \sum_{i=1}^{d} x^2(i)(y^T A^{(i)})^2 && (13)\\
&\leq \left(\sum_{i=1}^{d} x^{2p}(i)\right)^{1/p} \left(\max_{y,\|y\|_2=1} \sum_{i=1}^{d} (y^T A^{(i)})^{2q}\right)^{1/q} && (14)\\
&= \|x\|_\infty^2 \|A^T\|_{2\to 2}^2 && (15)
\end{aligned}$$

And so:

$$\|x\|_A \leq \|x\|_\infty \|A\|_{2\to 2} \qquad (16)$$

Lean Walsh $\|A_\ell\|_{2\to 2}$

From the recursive structure of A_ℓ we have:

$$\|A_\ell\|_{2\to 2} = \|A_1\|_{2\to 2}^{\ell} \tag{17}$$

From $d = 4^\ell$ and $\|A_1\|_{2\to 2} = \sqrt{4/3}$:

$$\|A_\ell\|_{2\to 2} = d^\delta \tag{18}$$

For $\delta = (1 - \log_4(3))/2 \sim 0.1$.

Using other choices of seeds δ can be arbitrarily small.

Putting it all together

We have that:

- $\|x\|_A \leq k^{-1/2} \Rightarrow x \in \chi(A, \varepsilon, n)$.
- $\|x\|_A \leq \|x\|_\infty \|A\|_{2\to 2}$.
- $\|A\|_{2\to 2} \leq d^\delta$

Combining all the above we conclude:

$$\|x\|_\infty \leq k^{-1/2} d^{-\delta} \Rightarrow x \in \chi(A, \varepsilon, n) \tag{19}$$

And so

$$\Pr(|\|AD_s x\|_2 - 1| > \varepsilon) \leq \frac{1}{n} \tag{20}$$

Conclusion

What did we see?

- Any (column normalized) matrix A is a good random projection for a subset χ of $\mathbb{R}^d$.
- For any matrix A we have that $\|x\|_A \leq k^{-1/2} \Rightarrow x \in \chi$.
- We saw the lean Walsh matrices.
- For the lean Walsh A, we have $\|x\|_\infty \leq k^{-1/2} d^{-\delta}$ implies $x \in \chi$.

What did we not see?

- Other exciting properties of lean Walsh matrices.
- A proof that there exists also an ℓ_∞ bound on x which does not depend on d.
- Experimental results showing that this method is extremely fast in practice.

Thank you