Near-equivalence of the Restricted Isometry Property and Johnson-Lindenstrauss Lemma

Rachel Ward

University of Texas at Austin

September 20, 2011

Joint work with Felix Krahmer (Hausdorff Center, Bonn, Germany)

Theorem (Johnson-Lindenstrauss (1984))

Let $\varepsilon \in (0,1)$ *and let* $x_1, \ldots, x_p \in \mathbb{R}^N$ *be arbitrary points. Let* $m = O\big(\varepsilon^{-2}\log(p)\big)$ *be a natural number. Then there exists a Lipschitz map* $f : \mathbb{R}^N \to \mathbb{R}^m$ *such that*

$$(1-\varepsilon)\|x_i - x_j\|^2 \leq \|f(x_i) - f(x_j)\|^2 \leq (1+\varepsilon)\|x_i - x_j\|^2$$

for all $i, j \in \{1, 2, \ldots, p\}$*.*

Theorem (Johnson-Lindenstrauss (1984))

Let $\varepsilon \in (0,1)$ and let $x_1, \ldots, x_p \in \mathbb{R}^N$ be arbitrary points. Let $m = O\big(\varepsilon^{-2}\log(p)\big)$ be a natural number. Then there exists a Lipschitz map $f : \mathbb{R}^N \to \mathbb{R}^m$ such that

$$(1-\varepsilon)\|x_i - x_j\|^2 \leq \|f(x_i) - f(x_j)\|^2 \leq (1+\varepsilon)\|x_i - x_j\|^2$$

for all $i, j \in \{1, 2, \ldots, p\}$.

Original proof: f is taken as a random orthogonal projection

Theorem (Johnson-Lindenstrauss (1984))

Let $\varepsilon \in (0,1)$ and let $x_1, \dots, x_p \in \mathbb{R}^N$ be arbitrary points. Let $m = O\big(\varepsilon^{-2}\log(p)\big)$ be a natural number. Then there exists a Lipschitz map $f : \mathbb{R}^N \to \mathbb{R}^m$ such that

$$(1-\varepsilon)\|x_i - x_j\|^2 \le \|f(x_i) - f(x_j)\|^2 \le (1+\varepsilon)\|x_i - x_j\|^2$$

for all $i, j \in \{1, 2, \dots, p\}$.

Original proof: f is taken as a random orthogonal projection

[Alon 2003] m-dependence on p and ε optimal up to $\log(1/\varepsilon)$

Theorem (Johnson-Lindenstrauss (1984))

Let $\varepsilon \in (0,1)$ *and let* $x_1, \ldots, x_p \in \mathbb{R}^N$ *be arbitrary points. Let* $m = O\big(\varepsilon^{-2}\log(p)\big)$ *be a natural number. Then there exists a Lipschitz map* $f : \mathbb{R}^N \to \mathbb{R}^m$ *such that*

$$(1-\varepsilon)\|x_i - x_j\|^2 \le \|f(x_i) - f(x_j)\|^2 \le (1+\varepsilon)\|x_i - x_j\|^2$$

for all $i, j \in \{1, 2, \ldots, p\}$*.*

Original proof: f is taken as a random orthogonal projection

[Alon 2003] m-dependence on p and ε optimal up to $\log(1/\varepsilon)$

(Even with suboptimal dependence we call such f "JL embeddings" or "distance-preserving embeddings")

Probabilistic distance-preserving embeddings

We want a linear map $\Phi : \mathbb{R}^N \to \mathbb{R}^m$ such that

$$\Big| \|\Phi(x_i - x_j)\| - \|x_i - x_j\| \Big| \leq \varepsilon \|x_i - x_j\| \text{ for } \binom{p}{2} \text{ vectors } x_i - x_j.$$

Probabilistic distance-preserving embeddings

We want a linear map $\Phi : \mathbb{R}^N \to \mathbb{R}^m$ such that

$$\Big| \|\Phi(x_i - x_j)\| - \|x_i - x_j\| \Big| \le \varepsilon \|x_i - x_j\| \text{ for } \binom{p}{2} \text{ vectors } x_i - x_j.$$

- For any fixed vector $v \in \mathbb{R}^N$, and for a matrix $\Phi : \mathbb{R}^N \to \mathbb{R}^m$ with i.i.d. Gaussian entries,

$$\mathbb{P}\Big(\big| \|\Phi v\|^2 - \|v\|^2 \big| \ge \varepsilon \|v\|^2 \Big) \le \exp(-c\varepsilon^2 m).$$

- Take union bound over $\binom{p}{2}$ vectors $x_i - x_j$;
- If $m \ge c' \varepsilon^{-2} \log(p)$, then Φ is optimal embedding with probability $\ge 1/2$.

Practical distance-preserving embeddings

For computational efficiency, $\Phi : \mathbb{R}^N \to \mathbb{R}^m$ should

- allow fast matrix-vector multiplies: $O(N \log N)$ flops per matrix-vector multiply is optimal
- not involve too much randomness

Practical distance-preserving embeddings

- [Ailon, Chazelle '06] : "Fast Johnson-Lindenstrauss Transform"

$$\Phi = \mathcal{G}\mathcal{F}\mathcal{D};$$

  - $\mathcal{D} : \mathbb{R}^N \to \mathbb{R}^N$ is diagonal matrix with random ± 1 entries.
  - $\mathcal{F} : \mathbb{R}^N \to \mathbb{R}^N$ is discrete Fourier matrix,
  - $\mathcal{G} : \mathbb{R}^N \to \mathbb{R}^m$ is sparse Gaussian matrix.

$\mathcal{O}(N \log N)$ multiplication when $p < e^{N^{1/2}}$

Practical distance-preserving embeddings

- [Ailon, Chazelle ′06] : “Fast Johnson-Lindenstrauss Transform”

$$\Phi = \mathcal{G}\mathcal{F}\mathcal{D};$$

  - $\mathcal{D} : \mathbb{R}^N \to \mathbb{R}^N$ is diagonal matrix with random ± 1 entries.
  - $\mathcal{F} : \mathbb{R}^N \to \mathbb{R}^N$ is discrete Fourier matrix,
  - $\mathcal{G} : \mathbb{R}^N \to \mathbb{R}^m$ is sparse Gaussian matrix.

  $\mathcal{O}(N \log N)$ multiplication when $p < e^{N^{1/2}}$

- Many more constructions ...

Practical Johnson-Lindenstrauss embeddings

- [Ailon, Liberty '10]: $\Phi = \mathcal{F}_{rand}\mathcal{D}$,
  - $\mathcal{D} : \mathbb{R}^N \to \mathbb{R}^N$ is diagonal matrix with random ± 1 entries.
  - $\mathcal{F}_{rand} : \mathbb{R}^N \to \mathbb{R}^m$ consists of m randomly-chosen rows from the discrete Fourier matrix
  - $\mathcal{O}(N \log(N))$ multiplication, but suboptimal embedding dimension for distance-preservation:

$$m = \mathcal{O}\big(\varepsilon^{-4} \log(p) \log^4(N)\big)$$

Proof relies on (nontrivial) estimates for $\mathcal{F}_{rand}$ from [Rudelson, Vershynin '08] (operator LLN, Dudley's inequality, ...)- these estimates are used in *compressed sensing* for sparse recovery guarantees.

Practical Johnson-Lindenstrauss embeddings

[Krahmer, W '10]: Improved embedding dimension for $\Phi = \mathcal{F}_{rand}\mathcal{D}$ to $m = \mathcal{O}\left(\varepsilon^{-2}\log(p)\log^4(N)\right)$.

Practical Johnson-Lindenstrauss embeddings

[Krahmer, W '10]: Improved embedding dimension for $\Phi = \mathcal{F}_{rand}\mathcal{D}$ to $m = \mathcal{O}\Big(\varepsilon^{-2}\log(p)\log^4(N)\Big)$.

Proof relies only on a certain *restricted isometry property* of $\mathcal{F}_{rand}$ introduced in context of sparse recovery. Many random matrix constructions share this property...

The Restricted Isometry Property (RIP)

The Restricted Isometry Property

A vector $x \in \mathbb{R}^N$ with at most k nonzero coordinates is $k-$sparse.

Definition (Candès/Romberg/Tao (2006))

A matrix $\Phi : \mathbb{R}^N \to \mathbb{R}^m$ is said to have the *restricted isometry property* of order k and level δ if

$$(1-\delta)\|x\|_2^2 \leq \|\Phi x\|_2^2 \leq (1+\delta)\|x\|_2^2$$

for all k-sparse $x \in \mathbb{R}^N$.

The Restricted Isometry Property

A vector $x \in \mathbb{R}^N$ with at most k nonzero coordinates is $k-$sparse.

Definition (Candès/Romberg/Tao (2006))

A matrix $\Phi : \mathbb{R}^N \to \mathbb{R}^m$ is said to have the *restricted isometry property* of order k and level δ if

$$(1-\delta)\|x\|_2^2 \leq \|\Phi x\|_2^2 \leq (1+\delta)\|x\|_2^2$$

for all k-sparse $x \in \mathbb{R}^N$.

Usual context: If $\Phi : \mathbb{R}^N \to \mathbb{R}^m$ has (k, δ)-RIP with δ sufficiently small, and if $x^{\#}$ is a k-sparse solution to the system $y = \Phi x$, then

$$x^{\#} = \underset{\Phi z = y}{\operatorname{argmin}} \|z\|_1.$$

RIP through concentration of measure

Recall the concentration inequality for distance-preserving embeddings (i.e. when Φ is Gaussian):

$$\mathbb{P}\left(\left|\|\Phi v\|^2 - \|v\|^2\right| \geq \varepsilon\|v\|^2\right) \leq \exp(-c\varepsilon^2 m) \qquad (1)$$

RIP through concentration of measure

Recall the concentration inequality for distance-preserving embeddings (i.e. when Φ is Gaussian):

$$\mathbb{P}\Big(\big|\|\Phi v\|^2 - \|v\|^2\big| \geq \varepsilon \|v\|^2\Big) \leq \exp(-c\varepsilon^2 m) \qquad (1)$$

[Baraniuk et al 2008]: If $\Phi : \mathbb{R}^N \to \mathbb{R}^m$ satisfies the concentration inequality, then with high probability a particular realization of Φ satisfies (k, ε)-RIP for $m \geq c' \varepsilon^2 k \log N$

- Implies RIP with optimally small m for Gaussian (and more generally subgaussian) matrices

Known RIP bounds

The following random matrices satisfy (k, δ)-RIP with high probability (proved via other methods):

- [Rudelson/Vershynin '08]: Partial Fourier matrix $\mathcal{F}_{rand}$; $m \gtrsim \delta^{-2} k \log^4(N)$
- [Adamczak et al '09]: Matrices whose columns are i.i.d. from log-concave distribution - $m \gtrsim \delta^{-2} k \log^2(N)$
- ...

- The best known deterministic constructions require $m \gtrsim k^{2-\mu}$ for some small μ (Bourgain et al (2011)).

Main results

Theorem (Krahmer, W. 2010)

Fix $\eta > 0$ and $\varepsilon > 0$. Let $\{x_j\}_{j=1}^{p} \subset \mathbb{R}^N$ be arbitrary. Set $k \geq 40 \log \frac{4p}{\eta}$, and suppose that $\Phi : \mathbb{R}^N \to \mathbb{R}^m$ has the $(k, \varepsilon/4)$-restricted isometry property. Let $\mathcal{D}$ be a diagonal matrix of random signs. Then with probability $\geq 1 - \eta$,

$$(1-\varepsilon)\|x_j\|_2^2 \leq \|\Phi \mathcal{D} x_j\|_2^2 \leq (1+\varepsilon)\|x_j\|_2^2$$

uniformly for all x_j.

- $\mathcal{F}_{rand}$ has (k, δ)-RIP with $m \geq c\varepsilon^{-2} k \log^4(N) \Rightarrow$ $\mathcal{F}_{rand}\mathcal{D}$ is a distance-preserving embedding if $m \geq c'\varepsilon^{-2} \log(p) \log^4(N)$.

A Geometric Observation

- A matrix that acts as an approximate isometry on sparse vectors (an RIP matrix) also acts as an approximate isometry on most vertices of the Hamming cube $\{-1, 1\}^N$).
  - Apply our result to the vector $x = (1, \ldots, 1)$.

Idea of Proof:

- Assume w.l.o.g. x is in decreasing arrangement.
- Partition x in $R = \frac{2N}{k}$ blocks of length $s = \frac{k}{2}$:

$$x = (x_1, \dots, x_N) = (x_{(1)}, x_{(2)}, \dots, x_{(R)}) = (x_{(1)}, x_{(\flat)})$$

- Need to bound

$$\begin{aligned}\|\Phi D_\xi x\|_2^2 &= \|\Phi D_x \xi\|_2^2 = \|\sum_{j=1}^{R} \Phi_{(J)} D_{x_{(J)}} \xi_{(J)}\|_2^2 \\ &= \sum_{J=1}^{R} \|\Phi_{(J)} D_{x_{(J)}} \xi_{(J)}\|_2^2 \; + \; 2\xi_{(1)}^* D_{x_{(1)}} \Phi_{(1)}^* \Phi_{(\flat)} D_{x_{(\flat)}} \xi_{(\flat)} \\ &+ \sum_{\substack{J,L=2 \\ J \neq L}}^{R} \left\langle \Phi_{(J)} D_{x_{(J)}} \xi_{(J)}, \Phi_{(L)} D_{x_{(L)}} \xi_{(L)} \right\rangle\end{aligned}$$

- Estimate each term separately.

First term

- Φ has (k, δ)-RIP, hence also has (s, δ)-RIP, and each $\Phi_{(J)}$ is almost an isometry.
- Noting that $\|D_{x_{(J)}} \xi_{(J)}\|_2 = \|D_{\xi_{(J)}} x_{(J)}\|_2 = \|x_{(J)}\|_2$, we estimate

$$(1-\delta)\|x\|_2^2 \leq \sum_{J=1}^{R} \|\Phi_{(J)} D_{x_{(J)}} \xi_{(J)}\|_2^2 \leq (1+\delta)\|x\|_2^2.$$

- Conclude with $\delta \leq \frac{\varepsilon}{4}$ that

$$\left(1-\frac{\varepsilon}{4}\right)\|x\|_2^2 \leq \sum_{J=1}^{R} \|\Phi_{(J)} D_{x_{(J)}} \xi_{(J)}\|_2^2 \leq \left(1+\frac{\varepsilon}{4}\right)\|x\|_2^2.$$

Second term

$$2\xi^*_{(1)} D_{x_{(1)}} \Phi^*_{(1)} \Phi_{(\flat)} D_{x_{(\flat)}} \xi_{(\flat)}$$

- Keep $\xi_{(1)} = b$ fixed, then use Hoeffding's inequality.

Proposition (Hoeffding (1963))

Let $v \in \mathbb{R}^N$, and let $\xi = (\xi_j)_{j=1}^N$ be a Rademacher sequence. Then, for any $t > 0$,

$$\mathbb{P}\Big(|\sum_j \xi_j v_j| > t\Big) \leq 2\exp\Big(-\frac{t^2}{2\|v\|_2^2}\Big).$$

- Need to estimate $\|v\|_2$ for $v = D_{x_{(\flat)}} \Phi^*_{(\flat)} \Phi_{(1)} D_{x_{(1)}} b$.

Key estimate

Proposition

Let $R = \lceil N/s \rceil$. Let $\Phi = (\Phi_j) = (\Phi_{(1)}, \Phi_{(\flat)}) \in \mathbb{R}^{m \times N}$ have the $(2s, \delta)$-RIP, let $x = (x_{(1)}, x_{(\flat)}) \in \mathbb{R}^N$ be in decreasing arrangement with $\|x\|_2 \leq 1$, fix $b \in \{-1, 1\}^s$, and consider the vector

$$v \in \mathbb{R}^N, \quad v = D_{x_{(\flat)}} \Phi^*_{(\flat)} \Phi_{(1)} D_{x_{(1)}} b.$$

Then $\|v\|_2 \leq \frac{\delta}{\sqrt{s}}$.

Key ingredients for the proof of the proposition

- $\|x_{(J)}\|_\infty \leq \frac{1}{\sqrt{k}}\|x_{(J-1)}\|_2$ for $J > 1$ (decreasing arrangement).
- Off-diagonal RIP estimate: $\|\Phi^*_{(J)}\Phi_{(L)}\| \leq \delta$ for $J \neq L$.

Third term

$$\sum_{\substack{J,L=2\\ J\neq L}}^{R} \left\langle \Phi_{(J)} D_{x_{(J)}} \xi_{(J)}, \Phi_{(L)} D_{x_{(L)}} \xi_{(L)} \right\rangle$$

- Use concentration inequality for Rademacher Chaos:

Proposition (Hanson/Wright '71, Boucheron et al '03)

Let X be the $N \times N$ matrix with entries $x_{i,j}$ and assume that $x_{i,i} = 0$ for all $i \in [N]$. Let $\xi = (\xi_j)_{j=1}^N$ be a Rademacher sequence. Then, for any $t > 0$,

$$\mathbb{P}\Big(\big|\sum_{i,j} \xi_i \xi_j x_{i,j}\big| > t \Big) \leq 2 \exp\left(-\frac{1}{64} \min\left(\frac{\frac{96}{65} t}{\|X\|}, \frac{t^2}{\|X\|_{\mathcal{F}}^2} \right)\right).$$

- Need $\|C\|$ and $\|C\|_{\mathcal{F}}$ for

$$C \in \mathbb{R}^{N\times N}, \quad C_{j,\ell} = \begin{cases} x_j \Phi_j^* \Phi_\ell x_\ell, & j,\ell > s \text{ in different blocks} \\ 0, & \text{else.} \end{cases}$$

Summary and discussion

Novel connection: An RIP matrix with randomized column signs is a distance-preserving (Johnson-Lindenstrauss) embedding.

- Yields "near-equivalence" between RIP and JL-Lemma
- Allows to transfer the theoretical results developed in compressed sensing to the setting of distance-preserving embeddings
- Yields improved bounds for embedding dimension of several classes of random matrices, and optimal dependence on distortion ε for a fast embedding.

Thanks!