New and improved Johnson-Lindenstrauss embeddings via the Restricted Isometry Property

Felix Krahmer

Hausdorff Center for Mathematics, Universität Bonn

3/10/2011

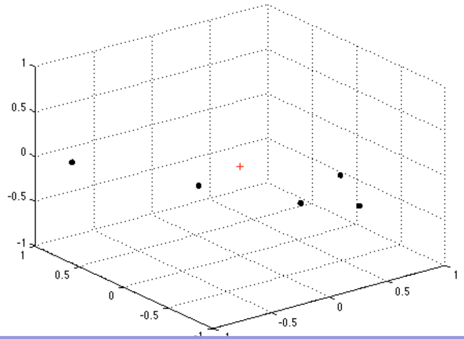
Joint work with Rachel Ward (Courant Institute, NYU)

Linear Dimensionality reduction

- Set up: We have many data vectors $\vec{x}_j \in \mathbb{R}^N$ for N large

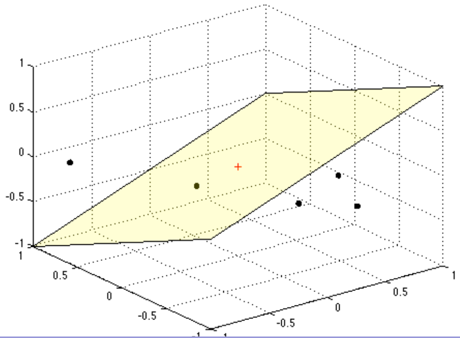
Linear Dimensionality reduction

- Set up: We have many data vectors $\vec{x}_j \in \mathbb{R}^N$ for N large
- We would like a linear map $\Phi \in \mathbb{R}^{m \times N}$, with $m \ll N$, such that the geometry of the set $\{\vec{x}_j\}_{j=1}^p$ is preserved under the embedding $\vec{x}_j \mapsto \Phi \vec{x}_j$



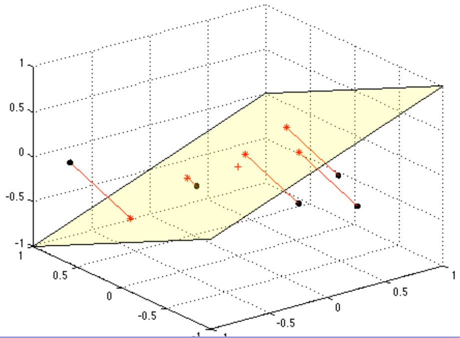
Linear Dimensionality reduction

- Set up: We have many data vectors $\vec{x}_j \in \mathbb{R}^N$ for N large
- We would like a linear map $\Phi \in \mathbb{R}^{m \times N}$, with $m \ll N$, such that the geometry of the set $\{\vec{x}_j\}_{j=1}^{p}$ is preserved under the embedding $\vec{x}_j \mapsto \Phi \vec{x}_j$



Linear Dimensionality reduction

- Set up: We have many data vectors $\vec{x}_j \in \mathbb{R}^N$, and N is large
- We would like a linear map $\Phi \in \mathbb{R}^{m \times N}$, with $m \ll N$, such that the geometry of the set $\{\vec{x}_j\}_{j=1}^{p}$ is preserved under the embedding $\vec{x}_j \mapsto \Phi \vec{x}_j$



The Johnson-Lindenstrauss (JL) Lemma

Theorem (Johnson-Lindenstrauss (1984))

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

$$(1-\varepsilon)\|x_i - x_j\|_2^2 \leq \|f(x_i) - f(x_j)\|_2^2 \leq (1+\varepsilon)\|x_i - x_j\|_2^2 \tag{1}$$

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

Original proof: Random orthogonal projections

Applications

Dimension reduction for

- Computer science
- Numerical linear algebra
- Manifold Learning
- ...

Applications

Dimension reduction for

- Computer science
- Numerical linear algebra
- Manifold Learning
- ...

To use JL Lemma in practice, f should

- be efficiently computable
- not involve too much randomness

Linear JL embeddings

- In practice: Linear JL embeddings, represented by $\Phi \in \mathbb{R}^{m \times N}$.
- Consider set of differences. $E = \{x_i - x_j\}$. Then Φ should satisfy:

$$(1-\varepsilon)\|y\|_2^2 \leq \|\Phi y\|_2^2 \leq (1+\varepsilon)\|y\|_2^2, \quad \text{for all } y \in E.$$

- For a random matrix Φ, we need for an arbitrary fixed $x \in \mathbb{R}^N$

$$\mathbb{P}\big((1-\varepsilon)\|x\|_2^2 \leq \|\Phi x\|_2^2 \leq (1+\varepsilon)\|x\|_2^2\big) \geq 1 - 2\exp(-c_0\varepsilon^2 m).$$

  - c_0 constant (possibly mildly dependent on N)
  - Then Φ is a JL embedding with high probability (union bound).

Previous work

- [Ailon, Chazelle '06] "Fast Johnson-Lindenstrauss transform": $\Phi = \mathcal{P}W\mathcal{D}$ is fast if $p \leq e^{N^{1/2}}$, slow if $e^{N^{1/2}} < p < e^{N}$:
  - $\mathcal{D} \in \mathbb{R}^{N \times N}$ is diagonal matrix of random signs,
  - $W \in \mathbb{R}^{N \times N}$ is discrete Fourier matrix,
  - $\mathcal{P} \in \mathbb{R}^{m \times N}$ is sparse Gaussian matrix.
- [Vybiral '10]: $\Phi = \mathcal{C}_{part}\mathcal{D}$; $\mathcal{C}_{part}$ is partial circulant matrix
  - Fast, but suboptimal embedding bound of $m = \mathcal{O}(\varepsilon^{-2} \log^2(p))$.
- [Ailon, Liberty '10]: Random partial Fourier matrix $W_{rand}\mathcal{D}$:
  - Fast, but suboptimal embedding dimension $m = \mathcal{O}(\varepsilon^{-4} \log(p) \log^4(N))$.

The Restricted Isometry Property

Definition (Candès/Romberg/Tao (2006))

A matrix $\Phi \in \mathbb{R}^{m \times N}$ is said to have the Restricted Isometry Property of order k and level $\delta \in (0,1)$ (equivalently, (k,δ)-RIP) if

$$(1-\delta)\|x\|_2^2 \leq \|\Phi x\|_2^2 \leq (1+\delta)\|x\|_2^2 \qquad \text{for all } k\text{-sparse } x \in \mathbb{R}^N.$$

Usual context: If Φ satisfies $(2k,\delta)$-RIP with $\delta \leq .46$, and if $y = \Phi x$ admits a k-sparse solution $x^\#$, then $x^\# = \underset{\Phi z = y}{\operatorname{argmin}} \|z\|_1$.

Known RIP bounds

The following random matrices have RIP with high probability :

- Gaussian and Bernoulli matrices if $m \gtrsim \delta^{-2} k \log(N)$
- Partial Fourier/Hadamard if $m \gtrsim \delta^{-2} k \log^4(N)$
- Partial Circulant Matrices (based on a Rademacher vector) if $m \gtrsim max(\delta^{-2} k \log(N), \delta^{-1} k^{3/2} \log^{3/2}(N))$
- ...

Contributors: Baraniuk, Candès, Davenport, DeVore, Pfander, Rauhut, Romberg, Rudelson, Tao, Tropp, Vershynin, Wakin, Ward, ...

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

Proof of RIP through the JL Lemma

Recall the crucial concentration inequality for the JL Lemma:

$$\mathbb{P}\big((1-\varepsilon)\|x\|_2^2 \leq \|\Phi x\|_2^2 \leq (1+\varepsilon)\|x\|_2^2\big) \geq 1 - 2\exp(-c_0\varepsilon^2 m) \quad (2)$$

Baraniuk, Davenport, DeVore, Wakin (2008) establish a connection between this inequality and RIP:

Theorem (Baraniuk et al.)

Suppose that m, N, and $0 < \delta < 1$ are given. If the $m \times N$ random matrix Φ satisfies the concentration inequality (2) with $\varepsilon = \delta$ and absolute constant c_0, then there exist constants c_1, c_2 such that with probability $\geq 1 - 2e^{-c_2\delta^2 m}$, the (k, δ)-RIP holds for Φ with any $k \leq c_1\delta^2 m/\log(N/k)$.

- In this sense, the JL Lemma implies the RIP.

RIP implies the JL Lemma

Theorem (K., Ward (2010))

Fix $\eta > 0$ *and* $\varepsilon > 0$, *let* $E \subset \mathbb{R}^N$ *with* $|E| = p$. *Set* $k \geq 40 \log \frac{4p}{\eta}$, *and suppose that* $\Phi \in \mathbb{R}^{m \times N}$ *has the* (k, δ)*-RIP with* $\delta \leq \frac{\varepsilon}{4}$. *Let* $\xi \in \mathbb{R}^N$ *be a Rademacher sequence. Then with probability* $\geq 1 - \eta$,

$$(1 - \varepsilon)\|x\|_2^2 \leq \|\Phi D_\xi x\|_2^2 \leq (1 + \varepsilon)\|x\|_2^2$$

uniformly for all $x \in E$.

- Rademacher sequence: Uniformly distributed on $\{-1, 1\}^N$
- Notation: $D_\xi =$ diagonal matrix with ξ on the diagonal.

A converse to the result by Baraniuk et al.

Proposition (K., Ward (2010))

Fix $\varepsilon > 0$, and suppose that for some c_3 and all pairs (k, m) with $k \leq c_3\delta^2 m/\log(N/k)$, $\Phi = \Phi(m) \in \mathbb{R}^{m\times N}$ has the (k, δ)-RIP with $\delta \leq \frac{\varepsilon}{4}$. Fix $x \in \mathbb{R}^N$ and let $\xi \in \mathbb{R}^N$ be a Rademacher sequence. Then there exists a constant c_4 such that for all m, ΦD_ξ satisfies the concentration inequality (2) for $c_0 = c_4 \log^{-1}\left(\frac{N}{k}\right)$.

- This converse is optimal up to a factor of $\log(N)$

	RIP bounds	Previous JL Bound	JL Bound from our result
Partial Fourier	$\delta^{-2} k \log^3(k) \log(N)$ [1,2]	$\varepsilon^{-4} \log(\frac{p}{\eta}) \log^3(\log(\frac{p}{\eta})) \log(N)$ [3]	$\varepsilon^{-2} \log(\frac{p}{\eta}) \log^3(\log(\frac{p}{\eta})) \log(N)$
Partial Circulant	$\max\left(\delta^{-1} k^{\frac{3}{2}} \log^{\frac{3}{2}}(N), \delta^{-2} k \log^2(k) \log^2(N)\right)$ [4]	$\varepsilon^{-2} \log^2(\frac{p}{\eta})$ [5]	$\max\left(\varepsilon^{-1} \log^{\frac{3}{2}}(\frac{p}{\eta}) \log^{\frac{3}{2}}(N), \varepsilon^{-2} \log(\frac{p}{\eta}) \log^2(\log(\frac{p}{\eta})) \log^2(N)\right)$
Deterministic (DeVore, Iwen)	$\delta^{-2} k^2 \log^2(N)$ [6,7]		$\varepsilon^{-2} \log^2(\frac{p}{\eta}) \log^2(N)$

References

[1] Candès/Tao (2006)
[2] Rudelson/Vershynin (2008)
[3] Ailon/Liberty (2010)
[4] Rauhut/Romberg/Tropp (2010)
[5] Vybíral (2010)
[6] DeVore (2007)
[7] Iwen (2010)

Idea of Proof:

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

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

- Need to bound

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

- Estimate each term separately.
- Union bound over x.

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

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

Proposition (Hoeffding (1963))

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

Estimate of $\|v\|_2$

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

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

- Use concentration inequality for Rademacher Chaos:

Proposition (Hanson/Wright (1971))

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 $t > 0$,

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

- 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: RIP implies JL Lemma.
- Yields best-known bounds for embedding dimension for many random matrices, optimal dependence on distortion ε.

- Important balance: log-factors in N and log factors in p.
- Structured matrices also reduce randomness. Can randomness be reduced further?