Dimension reduction via random projections

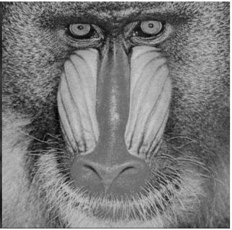
Rachel Ward

University of Texas at Austin

May 9, 2014

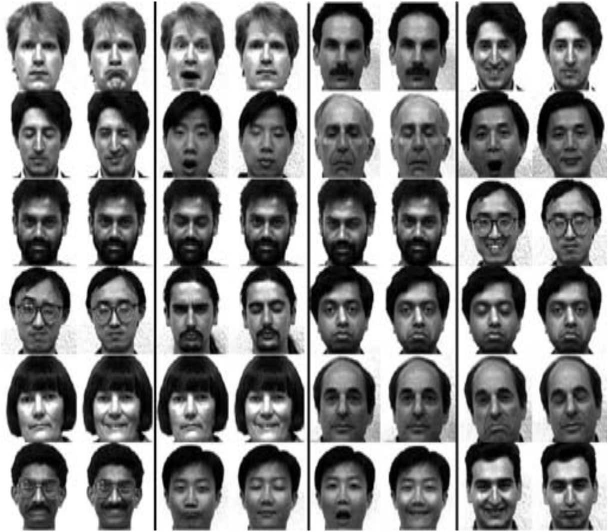
High dimensional data with low intrinsic dimension is everywhere





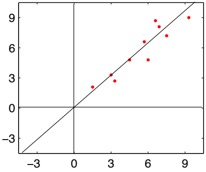
300 by 300 pixel images $= 90,000$ dimensions

High dimensional data with low intrinsic dimension is everywhere



Principal component analysis (PCA)

Standard tool for dimension reduction if data approximately lies on a linear subspace



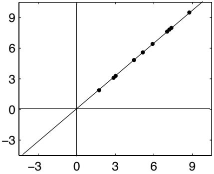


Figure: Original data and projection onto first principal component

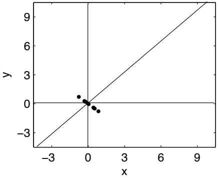
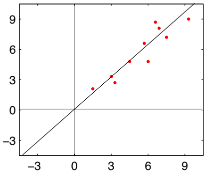
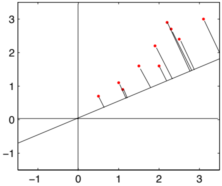


Figure: Residual

Random projections vs PCA





Principal components:
Directions of projection are data-dependent

Random projections:
Directions of projection are *independent* of the data

Wen random projections can be better:

1. Data is so high dimensional that it is too expensive to compute principal components directly
2. You do not have access to all the data at once, as in data streaming
3. Data is approximately low-dimensional, but not near a linear subspace

In this talk:

To what extent can information in a high dimensional data set be preserved if we acquire it through random projections?

- The Johnson-Lindenstrauss Lemma / concentration of measure
- Connections to sparse recovery
- Preserving non-Euclidean distances (especially ℓ_1)

Set-up

- Data as vectors $\mathbf{x}_j \in \mathbb{R}^n, \quad j = 1, 2, \ldots, n$

Set-up

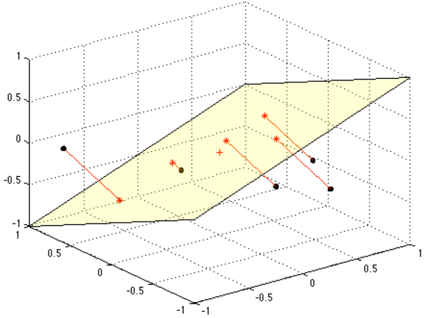
- Data as vectors $\mathbf{x}_j \in \mathbb{R}^n, \quad j = 1, 2, \ldots, n$
- Recall that
  $\|\mathbf{x}\|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}, 1 \le p < \infty, \quad \|\mathbf{x}\|_\infty = \max_i |x_i|$
- Data is processed through small number of *linear sketches*
  $y_k = \langle \mathbf{a}_k, \mathbf{x}_j \rangle, \ldots k \in \{1, 2, \ldots, m\}$ and $m \ll n$

  E.g.
  - Computing the mean: $\mu = \frac{1}{n} \langle \mathbf{1}, \mathbf{x} \rangle$
  - *Random* sketches: $y = \langle \mathbf{a}, \mathbf{x}_j \rangle, \quad \mathbf{a} = (\pm 1, \pm 1, \pm 1, \ldots)$
- In matrix-vector form: $\mathbf{y} = \mathbf{A}\mathbf{x}$

$$\begin{bmatrix} \mathbf{y} \end{bmatrix} = \begin{bmatrix} \quad \mathbf{A} \quad \end{bmatrix} \begin{bmatrix} \mathbf{x} \end{bmatrix}$$

Linear Dimensionality Reduction

- **The Johnson-Lindenstrauss Lemma:** *"A set of p points in high-dimensional Euclidean space can be linearly embedded in* $m > 9\varepsilon^{-2}\log p$ *dimensions without distorting the distance between any two points by more than a factor of* $(1 \pm \varepsilon)$*"*



The Johnson-Lindenstrauss Lemma

More precisely,

Theorem (Johnson/Lindenstrauss (1984))

Let $\varepsilon \in (0,1)$ *and let* $\mathcal{X} = \{\mathbf{x}_1, ..., \mathbf{x}_n\} \subset \mathbb{R}^n$.
Let $m \geq 9\varepsilon^{-2} \log n$ *be a natural number. Then there exists a linear map* $\Phi : \mathbb{R}^n \to \mathbb{R}^m$ *such that*

$$(1-\varepsilon)\|\mathbf{x}_i - \mathbf{x}_j\|^2 \leq \|\Phi\mathbf{x}_i - \Phi\mathbf{x}_j\|^2 \leq (1+\varepsilon)\|\mathbf{x}_i - \mathbf{x}_j\|^2 \quad \forall i,j \in \mathcal{X}$$

The Johnson-Lindenstrauss Lemma

More precisely,

Theorem (Johnson/Lindenstrauss (1984))

Let $\varepsilon \in (0,1)$ *and let* $\mathcal{X} = \{\mathbf{x}_1, ..., \mathbf{x}_n\} \subset \mathbb{R}^n$.
Let $m \geq 9\varepsilon^{-2} \log n$ *be a natural number. Then there exists a linear map* $\Phi : \mathbb{R}^n \to \mathbb{R}^m$ *such that*

$$(1-\varepsilon)\|\mathbf{x}_i - \mathbf{x}_j\|^2 \leq \|\Phi\mathbf{x}_i - \Phi\mathbf{x}_j\|^2 \leq (1+\varepsilon)\|\mathbf{x}_i - \mathbf{x}_j\|^2 \quad \forall i, j \in \mathcal{X}$$

- [Alon '03] m dependence on n and ε optimal up to $\log(1/\varepsilon)$ factor.
- With high probability, a random projection Φ works.

Probabilistic JL constructions

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

$$\|\Phi(\mathbf{x}_i - \mathbf{x}_j)\| \approx \|\mathbf{x}_i - \mathbf{x}_j\| \quad \text{for } \binom{n}{2} \text{ vectors } \mathbf{x}_i - \mathbf{x}_j.$$

- For any **fixed** vector $\mathbf{v} \in \mathbb{R}^n$, and for a matrix $\Phi : \mathbb{R}^n \to \mathbb{R}^m$ with i.i.d. Gaussian entries, $\mathbb{E}\|\Phi\mathbf{v}\|^2 = \|\mathbf{v}\|^2$ and

$$\mathbb{P}\Big((1-\varepsilon)\|\mathbf{v}\|^2 \le \|\Phi\mathbf{v}\|^2 \le (1+\varepsilon)\|\mathbf{v}\|^2\Big) \ge 1 - 2e^{-c\varepsilon^2 m}.$$

  This is *concentration of measure* for Gaussian random matrices.

Probabilistic JL constructions

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

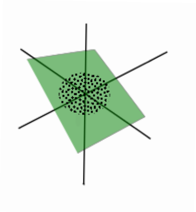
$$\|\Phi(\mathbf{x}_i - \mathbf{x}_j)\| \approx \|\mathbf{x}_i - \mathbf{x}_j\| \quad \text{for } \binom{n}{2} \text{ vectors } \mathbf{x}_i - \mathbf{x}_j.$$

- For any **fixed** vector $\mathbf{v} \in \mathbb{R}^n$, and for a matrix $\Phi : \mathbb{R}^n \to \mathbb{R}^m$ with i.i.d. Gaussian entries, $\mathbb{E}\|\Phi\mathbf{v}\|^2 = \|\mathbf{v}\|^2$ and

  $$\mathbb{P}\Big((1-\varepsilon)\|\mathbf{v}\|^2 \leq \|\Phi\mathbf{v}\|^2 \leq (1+\varepsilon)\|\mathbf{v}\|^2\Big) \geq 1 - 2e^{-c\varepsilon^2 m}.$$

  This is *concentration of measure* for Gaussian random matrices.
- Take union bound over $\binom{n}{2}$ vectors $\mathbf{x}_i - \mathbf{x}_j \quad \Rightarrow$
  Φ works with probability $\geq 1/2$ if $m = \mathcal{O}(\varepsilon^{-2}\log(n))$

From finite sets to continuous subsets

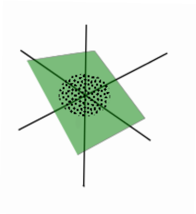


- Suppose K is a bounded subset of $\mathbb{R}^n$ and $\varepsilon > 0$ is fixed.
- A finite subset $Q \subset K$ is called an ε-net of K if for every $\mathbf{x} \in K$ one can find $\mathbf{y} \in Q$ such that

$$\|\mathbf{x} - \mathbf{y}\|_2 \leq \varepsilon$$

- The minimal possible size $\#Q$ is the ε-*covering number* $N(K, \varepsilon)$.

From finite sets to continuous subsets



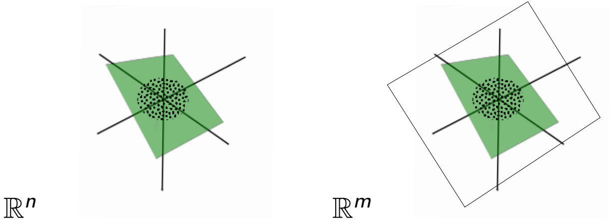
- Suppose K is a bounded subset of $\mathbb{R}^n$ and $\varepsilon > 0$ is fixed.
- A finite subset $Q \subset K$ is called an ε-net of K if for every $\mathbf{x} \in K$ one can find $\mathbf{y} \in Q$ such that

$$\|\mathbf{x} - \mathbf{y}\|_2 \leq \varepsilon$$

- The minimal possible size $\#Q$ is the ε-*covering number* $N(K, \varepsilon)$.

Example: For B_2^k the Euclidean ball, $N(B_2^k, \varepsilon) \leq (3/\varepsilon)^k$.

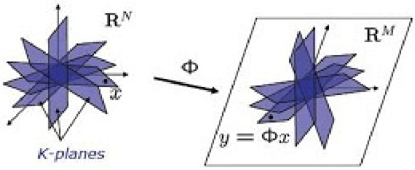
Random projections preserve information



If $\mathcal{S}$ is a k-dimensional subspace of high-dimensional Euclidean space, a Gaussian random matrix $\Phi : \mathbb{R}^n \to \mathbb{R}^m$ with $m \geq C\varepsilon^{-2} \log(\#Q) = C\varepsilon^{-2} k \log(1/\varepsilon)$ will, with high probability, preserve all pairwise distances between points in the subspace:

$$(1-\varepsilon)\|\mathbf{x}-\mathbf{y}\|^2 \leq \|\Phi(\mathbf{x}-\mathbf{y})\|^2 \leq (1+\varepsilon)\|\mathbf{x}-\mathbf{y}\|^2, \quad \forall \mathbf{x}, \mathbf{y} \in \mathcal{S}$$

Random projections preserve information

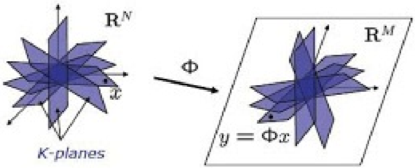


A Gaussian random matrix $\Phi : \mathbb{R}^n \to \mathbb{R}^m$ will also preserve the geometry of a *union* of low-dimensional subspaces.

[Baraniuk/Davenport/DeVore/Wakin 06] Consider the subset of *k-sparse signals*

$$\mathcal{S}_k = \{\mathbf{x} \in \mathbb{R}^n : \#\{i : |x_i| > 0\} \leq k\}.$$

Random projections preserve information



A Gaussian random matrix $\Phi : \mathbb{R}^n \to \mathbb{R}^m$ will also preserve the geometry of a *union* of low-dimensional subspaces.

[Baraniuk/Davenport/DeVore/Wakin 06] Consider the subset of *k-sparse signals*

$$\mathcal{S}_k = \{\mathbf{x} \in \mathbb{R}^n : \#\{i : |x_i| > 0\} \leq k\}.$$

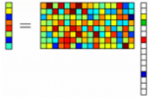
ε-covering number this set $\leq \binom{n}{k}(3/\varepsilon)^k \leq (\frac{n}{k})^k(3/\varepsilon)^k$

$\Rightarrow$ If $m = \mathcal{O}(\varepsilon^{-2} k \log(n/k))$ then with high probability,

$$(1-\varepsilon)\|\mathbf{x}-\mathbf{y}\|^2 \leq \|\Phi(\mathbf{x}-\mathbf{y})\|^2 \leq (1+\varepsilon)\|\mathbf{x}-\mathbf{y}\|^2, \quad \forall \mathbf{x}, \mathbf{y} \in \mathcal{S}_k$$

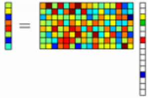
- The Johnson-Lindenstrauss Lemma / concentration of measure
- Connections to sparse recovery
- Non-Euclidean metrics (especially ℓ_1)

Sparse recovery



Sparse recovery concerns the "inverse problem": Can we recover a given $\mathbf{x} \in \mathbb{R}^n$ which is k-sparse from lower-dimensional projection $\Phi\mathbf{x} \in \mathbb{R}^m, \quad m \ll n$.

Sparse recovery

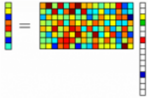


Sparse recovery concerns the "inverse problem": Can we recover a given $\mathbf{x} \in \mathbb{R}^n$ which is k-sparse from lower-dimensional projection $\Phi\mathbf{x} \in \mathbb{R}^m, \quad m \ll n$.

- Definition [Candès/Romberg/Tao (2006)]: $\Phi : \mathbb{R}^n \to \mathbb{R}^m$ has the *restricted isometry property* (RIP) of order k and level $\varepsilon \in (0,1)$ if

$$(1-\varepsilon)\|\mathbf{x}\|_2^2 \leq \|\Phi\mathbf{x}\|_2^2 \leq (1+\varepsilon)\|\mathbf{x}\|_2^2 \quad \forall\ k\text{-sparse } \mathbf{x} \in \mathbb{R}^n$$

Sparse recovery



Sparse recovery concerns the "inverse problem": Can we recover a given $\mathbf{x} \in \mathbb{R}^n$ which is k-sparse from lower-dimensional projection $\Phi\mathbf{x} \in \mathbb{R}^m, \quad m \ll n$.

- Definition [Candès/Romberg/Tao (2006)]: $\Phi : \mathbb{R}^n \to \mathbb{R}^m$ has the *restricted isometry property* (RIP) of order k and level $\varepsilon \in (0,1)$ if

$$(1-\varepsilon)\|\mathbf{x}\|_2^2 \leq \|\Phi\mathbf{x}\|_2^2 \leq (1+\varepsilon)\|\mathbf{x}\|_2^2 \quad \forall\ k\text{-sparse } \mathbf{x} \in \mathbb{R}^n$$

We have seen that with high probability, a Gaussian random matrix $\Phi \in \mathbb{R}^{m\times n}$ has RIP if $m \geq \varepsilon^{-2} k \log(n/k)$

RIP of order $2k$ and small ε implies that Φ is *invertible* and *well-conditioned* over the subset of k-sparse signals:

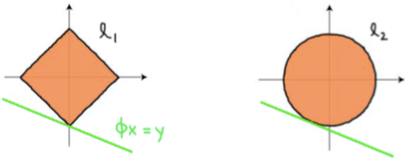
$$\|\Phi(\mathbf{x}_1 - \mathbf{x}_2)\|_2^2 \geq (1-\varepsilon)\|\mathbf{x}_1 - \mathbf{x}_2\|_2^2, \qquad \mathbf{x}_1, \mathbf{x}_2 \text{ } k\text{-sparse} .$$

This implies that if $\mathbf{x}$ is k-sparse and Φ has RIP of order $2k$,

$$\mathbf{x} = \arg\min_{\mathbf{z} \in \mathbb{R}^n} \|\mathbf{z}\|_0 \quad \text{subject to } \Phi\mathbf{z} = \Phi\mathbf{x}$$

This is not a tractable optimization algorithm (NP hard in general).

Sparse recovery through ℓ_1 minimization



[Candès/Romberg/Tao, Donoho (2006)] RIP of order $2k$ also implies:

- If $\mathbf{x}$ is k-sparse, then

$$\mathbf{x} = \arg\min_{\mathbf{z}\in\mathbb{R}^n} \|\mathbf{z}\|_1 \quad \text{subject to } \Phi\mathbf{z} = \Phi\mathbf{x}$$

- More generally, if $\mathbf{x}$ is "close to" k-sparse, then

$$\mathbf{x}^{\#} = \arg\min_{\mathbf{z}\in\mathbb{R}^n} \|\mathbf{z}\|_1 \quad \text{subject to } \Phi\mathbf{z} = \Phi\mathbf{x}$$

is close to $\mathbf{x}$.

Sparse recovery and linear dimension reduction

Recall the crucial concentration inequality for a (properly normalized) Gaussian random matrix: For a fixed $\mathbf{x} \in \mathbb{R}^n$,

$$\mathbb{P}\Big((1-\varepsilon)\|\mathbf{x}\|_2^2 \le \|\Phi\mathbf{x}\|_2^2 \le (1+\varepsilon)\|\mathbf{x}\|_2^2\Big) \ge 1 - 2e^{-c\varepsilon^2 m}.$$

Sparse recovery and linear dimension reduction

Recall the crucial concentration inequality for a (properly normalized) Gaussian random matrix: For a fixed $\mathbf{x} \in \mathbb{R}^n$,

$$\mathbb{P}\Big((1-\varepsilon)\|\mathbf{x}\|_2^2 \le \|\Phi\mathbf{x}\|_2^2 \le (1+\varepsilon)\|\mathbf{x}\|_2^2\Big) \ge 1 - 2e^{-c\varepsilon^2 m}.$$

- We have seen that Φ having this concentration for an *arbitrary* $\mathbf{x} \Rightarrow \Phi$ has the Restricted Isometry Property with high probability, once $m \ge Ck\log(n/k)$.
- The RIP has also been shown for many *structured random matrix constructions*, via more complicated arguments, such as *random partial discrete Fourier matrices.*
- Is there a converse to this result? Does RIP for a matrix Φ imply that Φ satisfies the concentration inequality for an arbitrary $\mathbf{x}$? Not quite, but ...

We can recover a "near" converse result:

Theorem (Krahmer, W. '11)

Suppose $\Phi : \mathbb{R}^n \to \mathbb{R}^m$ *satisfies*

$$(1-\varepsilon)\|\mathbf{x}\|_2^2 \leq \|\Phi\mathbf{x}\|^2 \leq (1+\varepsilon)\|\mathbf{x}\|^2 \quad \forall \mathbf{x} \text{ } k\text{-sparse.}$$

Fix $\mathbf{x} \in \mathbb{R}^n$ *arbitrary and suppose* $\mathcal{D}_\xi \in \mathbb{R}^{n\times n}$ *is a diagonal matrix with* $\xi = \pm 1$ *on diagonal. Then*

$$\mathbb{P}\Big((1-\varepsilon)\|\mathbf{x}\|_2^2 \leq \|\Phi\mathcal{D}_\xi\mathbf{x}\|_2^2 \leq (1+\varepsilon)\|\mathbf{x}\|_2^2\Big) \geq 1 - 2e^{\left(-\frac{c\varepsilon^2 m}{\log(n)}\right)}$$

We can recover a "near" converse result:

Theorem (Krahmer, W. '11)

Suppose $\Phi : \mathbb{R}^n \to \mathbb{R}^m$ *satisfies*

$$(1-\varepsilon)\|\mathbf{x}\|_2^2 \le \|\Phi\mathbf{x}\|^2 \le (1+\varepsilon)\|\mathbf{x}\|^2 \quad \forall \mathbf{x} \text{ } k\text{-sparse.}$$

Fix $\mathbf{x} \in \mathbb{R}^n$ *arbitrary and suppose* $\mathcal{D}_\xi \in \mathbb{R}^{n\times n}$ *is a diagonal matrix with* $\xi = \pm 1$ *on diagonal. Then*

$$\mathbb{P}\Big((1-\varepsilon)\|\mathbf{x}\|_2^2 \le \|\Phi\mathcal{D}_\xi\mathbf{x}\|_2^2 \le (1+\varepsilon)\|\mathbf{x}\|_2^2\Big) \ge 1 - 2e^{\left(-\frac{c\varepsilon^2 m}{\log(n)}\right)}$$

Informally, RIP + random column sign flips implies Johnson-Lindenstrauss concentration for Φ up to a $\log(n)$ factor in the embedding dimension m.

A Geometric Observation

- A matrix Φ that acts as an approximate isometry on sparse vectors (an RIP matrix) also acts as an approximate isometry on most maximally flat vectors (i.e., in the Hamming cube $\{-1,1\}^N$).
  - Follows from $\|\Phi \mathcal{D}_\xi \mathbf{x}\|_2 \approx \|\mathbf{x}\|_2$ with $\mathbf{x} = (1,\ldots,1)$.

- The Johnson-Lindenstrauss Lemma / concentration of measure
- Connections to sparse recovery
- Non-Euclidean metrics (especially ℓ_1)

Probabilistic JL embeddings: ℓ_2^n to ℓ_1^m

A random Gaussian matrix can also be used to embed finite subsets of ℓ_2^n into ℓ_1^m :

Proposition

[1] *Fix* $\mathbf{x} \in \mathbb{R}^n$*. Let* $\Phi : \mathbb{R}^n \to \mathbb{R}^m$ *with standard i.i.d. Gaussian entries. Then*

$$\mathbb{P}\left((1-\varepsilon)\|\mathbf{x}\|_2 \leq \sqrt{\frac{\pi}{2}}\frac{1}{m}\sum_{i=1}^{m}|(\Phi\mathbf{x})_i| \leq (1+\varepsilon)\|\mathbf{x}\|_2\right) \geq 1 - Ce^{-c\varepsilon^2 m}$$

[1] Plan, Vershynin, *One-bit compressed sensing by linear programming*, 2012.

Probabilistic JL embeddings: ℓ_2^n to ℓ_1^m

A random Gaussian matrix can also be used to embed finite subsets of ℓ_2^n into ℓ_1^m :

Proposition

[1] *Fix* $\mathbf{x} \in \mathbb{R}^n$. *Let* $\Phi : \mathbb{R}^n \to \mathbb{R}^m$ *with standard i.i.d. Gaussian entries. Then*

$$\mathbb{P}\left((1-\varepsilon)\|\mathbf{x}\|_2 \leq \sqrt{\frac{\pi}{2}}\frac{1}{m}\sum_{i=1}^{m} |(\Phi\mathbf{x})_i| \leq (1+\varepsilon)\|\mathbf{x}\|_2 \right) \geq 1 - Ce^{-c\varepsilon^2 m}$$

What if $\|\mathbf{x}\|_2$ above is replaced by $\|\mathbf{x}\|_1$?

That is, given an arbitrary set of n points in $\mathbb{R}^n$, does there exist a linear map $T : \mathbb{R}^n \to \mathbb{R}^{c\log(n)}$ which preserves pairwise ℓ_1 distances between points in the set?

[1] Plan, Vershynin, *One-bit compressed sensing by linear programming*, 2012.

Dimension reduction in ℓ_1

In high dimensions, ℓ_1 norm is more meaningful than ℓ_2 for nearest neighbor comparisons.

- Consider d points in $\mathbb{R}^n$, each coordinate of each point drawn i.i.d. from some underlying distribution.
- Let dmax_p^n be farthest point from origin and dmin_p^n be closest point to origin with respect to ℓ_p^n metric. Then[2]

$$\lim_{n\to\infty} \mathbb{E}\left[\mathrm{dmax}_p^n - \mathrm{dmin}_p^n\right] \asymp n^{1/p-1/2}.$$

[2] Hinneburg, Aggarwal, and Keim. *What is the nearest neighbor in high dimensional spaces?*, 2000.

Dimension reduction in ℓ_1

In high dimensions, ℓ_1 norm is more meaningful than ℓ_2 for nearest neighbor comparisons.

- Consider d points in $\mathbb{R}^n$, each coordinate of each point drawn i.i.d. from some underlying distribution.
- Let dmax_p^n be farthest point from origin and dmin_p^n be closest point to origin with respect to ℓ_p^n metric. Then[2]

$$\lim_{n\to\infty} \mathbb{E}\left[\mathrm{dmax}_p^n - \mathrm{dmin}_p^n\right] \asymp n^{1/p-1/2}.$$

- For ℓ_2, all points become equidistant up to a constant.
- For ℓ_p with $p > 2$, all points become completely equidistant
- ℓ_1 is only “simple” metric where the difference between nearest and farthest neighbor increases with dimension

[2] Hinneburg, Aggarwal, and Keim. *What is the nearest neighbor in high dimensional spaces?*, 2000.

The curse of non-Euclideanity

- Hardness result for ℓ_p to ℓ_p embedding:[3] for each $1 \leq p \leq \infty$, there are arbitrarily large n-point subsets X such that any linear mapping $T : \mathbb{R}^n \to \mathbb{R}^m$ satisfies

$$\left(\frac{m}{n}\right)^{|1/p-1/2|} \|\mathbf{x}-\mathbf{y}\|_p \leq \|T(\mathbf{x}-\mathbf{y})\|_p \leq \left(\frac{n}{m}\right)^{|1/p-1/2|} \|\mathbf{x}-\mathbf{y}\|_p$$

for some $\mathbf{x}, \mathbf{y} \in X$.

[3] Charikar, Sahai '02, Lee, Mendel, Naor '05

The curse of non-Euclideanity

- Hardness result for ℓ_p to ℓ_p embedding:[3] for each $1 \leq p \leq \infty$, there are arbitrarily large n-point subsets X such that any linear mapping $T : \mathbb{R}^n \to \mathbb{R}^m$ satisfies

$$\left(\frac{m}{n}\right)^{|1/p-1/2|} \|\mathbf{x}-\mathbf{y}\|_p \leq \|T(\mathbf{x}-\mathbf{y})\|_p \leq \left(\frac{n}{m}\right)^{|1/p-1/2|} \|\mathbf{x}-\mathbf{y}\|_p$$

  for some $\mathbf{x}, \mathbf{y} \in X$.
  - For $p = 2$, everything is nice!

[3] Charikar, Sahai '02, Lee, Mendel, Naor '05

The curse of non-Euclideanity

- Hardness result for ℓ_p to ℓ_p embedding:[3] for each $1 \leq p \leq \infty$, there are arbitrarily large n-point subsets X such that any linear mapping $T : \mathbb{R}^n \to \mathbb{R}^m$ satisfies

$$\left(\frac{m}{n}\right)^{|1/p-1/2|} \|\mathbf{x}-\mathbf{y}\|_p \leq \|T(\mathbf{x}-\mathbf{y})\|_p \leq \left(\frac{n}{m}\right)^{|1/p-1/2|} \|\mathbf{x}-\mathbf{y}\|_p$$

  for some $\mathbf{x}, \mathbf{y} \in X$.
  - For $p = 2$, everything is nice!
  - For $p = 1$, linear dimensionality reduction with constant distortion is not possible in general.

[3] Charikar, Sahai '02, Lee, Mendel, Naor '05

Dimension reduction in ℓ_1 for *sparse* vectors

The negative result for dimension reduction in ℓ_1 is a *worst case bound* over arbitrary sets of n points

Dimension reduction in ℓ_1 for *sparse* vectors

The negative result for dimension reduction in ℓ_1 is a *worst case bound* over arbitrary sets of n points

If $\mathbf{x} \in \mathbb{R}^n$ is *s-sparse*, the situation is much better:

Proposition (Berinde, Gilbert, Indyk, Karloff, Strauss '08)

There exists a linear map $T : \mathbb{R}^n \to \mathbb{R}^m$ *with* $m \geq C\varepsilon^{-2} s \log(n)$ *such that the following holds uniformly over all s-sparse* $\mathbf{x} \in \mathbb{R}^n$*:*

$$(1 - 2\varepsilon)\|\mathbf{x}\|_1 \leq \|T\mathbf{x}\|_1 \leq \|\mathbf{x}\|_1.$$

Such a matrix is said to have the $1-$*restricted isometry property* (1-RIP).

Dimension reduction in ℓ_1 for *sparse* vectors

The negative result for dimension reduction in ℓ_1 is a *worst case bound* over arbitrary sets of n points

If $\mathbf{x} \in \mathbb{R}^n$ is *s-sparse*, the situation is much better:

Proposition (Berinde, Gilbert, Indyk, Karloff, Strauss '08)

There exists a linear map $T : \mathbb{R}^n \to \mathbb{R}^m$ *with* $m \geq C\varepsilon^{-2} s \log(n)$ *such that the following holds uniformly over all s-sparse* $\mathbf{x} \in \mathbb{R}^n$*:*

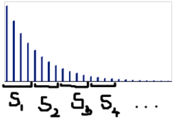
$$(1 - 2\varepsilon)\|\mathbf{x}\|_1 \leq \|T\mathbf{x}\|_1 \leq \|\mathbf{x}\|_1.$$

Such a matrix is said to have the $1-$*restricted isometry property* (1-RIP).

Probabilistic construction of such a T: sparse binary random matrix with $d = c\varepsilon^{-1} \log(n)$ ones per column. Corresponds to adjacency matrix of an (s, d, ε) lossless expander graph

Question: can we say anything about dimension reduction in ℓ_1 in between the worst-case setting where dimension reduction is not possible, and the setting of sparse vectors, where very strong dimension reduction is possible?

An interpolation norm

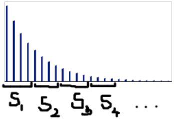


Given $\mathbf{x} \in \mathbb{R}^n$, partition its support into disjoint subsets $S_1, S_2, S_3 \ldots$ of size s according to the decreasing rearrangement of $\mathbf{x}$. The following is a norm:

$$\|\mathbf{x}\|_{1,2,s} := \sqrt{\sum_{\ell=1}^{\lceil n/s \rceil} \|\mathbf{x}_{S_\ell}\|_1^2}$$

1. When $s = 1$, $\|\cdot\|_{1,2,s} \equiv \|\cdot\|_2$.
2. When $s = n$, $\|\cdot\|_{1,2,s} \equiv \|\cdot\|_1$.
3. For any s, $\|\mathbf{x}\|_{1,2,s} = \|\mathbf{x}\|_1$ if $\mathbf{x}$ is s-sparse.

An interpolation norm

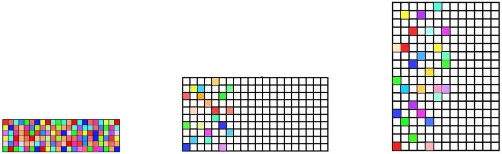


Given $\mathbf{x} \in \mathbb{R}^n$, partition its support into disjoint subsets $S_1, S_2, S_3 \dots$ of size s according to the decreasing rearrangement of $\mathbf{x}$. The following is a norm:

$$\|\mathbf{x}\|_{1,2,s} := \sqrt{\sum_{\ell=1}^{\lceil n/s \rceil} \|\mathbf{x}_{S_\ell}\|_1^2}$$

1. When $s = 1$, $\|\cdot\|_{1,2,s} \equiv \|\cdot\|_2$.
2. When $s = n$, $\|\cdot\|_{1,2,s} \equiv \|\cdot\|_1$.
3. For any s, $\|\mathbf{x}\|_{1,2,s} = \|\mathbf{x}\|_1$ if $\mathbf{x}$ is s-sparse.

Related to classical interpolation norms appearing in Banach space literature

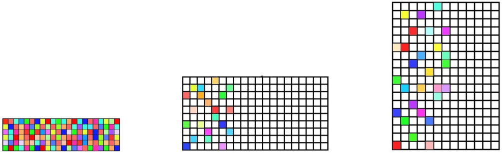


Theorem (W., 2014)

Fix $\mathbf{x} \in \mathbb{R}^n$. *Fix* $s < m \in \mathbb{N}$. *There is a distribution on linear maps* $\Psi_s : \mathbb{R}^n \to \mathbb{R}^m$ *such that, with probability exceeding* $1 - 2ne^{-\frac{\varepsilon^2 m}{s}}$,

$$(.63 - \varepsilon)\|\mathbf{x}\|_{1,2,s} \leq \|\Psi_s \mathbf{x}\|_1 \leq (1.77 + \varepsilon)\|\mathbf{x}\|_{1,2,s}$$

1. When $s = 1$ and $\|\cdot\|_{1,2,s} = \|\cdot\|_2$, Ψ is a Gaussian matrix, and we recover ℓ_2 to ℓ_1 JL embedding result up to factors .63 and 1.77

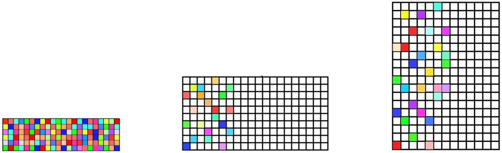


Theorem (W., 2014)

Fix $\mathbf{x} \in \mathbb{R}^n$. *Fix* $s < m \in \mathbb{N}$. *There is a distribution on linear maps* $\Psi_s : \mathbb{R}^n \to \mathbb{R}^m$ *such that, with probability exceeding* $1 - 2ne^{-\frac{\varepsilon^2 m}{s}}$,

$$(.63 - \varepsilon)\|\mathbf{x}\|_{1,2,s} \leq \|\Psi_s \mathbf{x}\|_1 \leq (1.77 + \varepsilon)\|\mathbf{x}\|_{1,2,s}$$

2. When $s = n$ and $\|\cdot\|_{1,2,s} = \|\cdot\|_1$, we find that $m \geq cn\log(n)$ - can't hope for dimension reduction in ℓ_1 in general



Theorem (W., 2014)

Fix $\mathbf{x} \in \mathbb{R}^n$. *Fix* $s < m \in \mathbb{N}$. *There is a distribution on linear maps* $\Psi_s : \mathbb{R}^n \to \mathbb{R}^m$ *such that, with probability exceeding* $1 - 2ne^{-\frac{\varepsilon^2 m}{s}}$,

$$(.63 - \varepsilon)\|\mathbf{x}\|_{1,2,s} \leq \|\Psi_s \mathbf{x}\|_1 \leq (1.77 + \varepsilon)\|\mathbf{x}\|_{1,2,s}$$

3. If $\mathbf{x}$ is s-sparse, $\|\mathbf{x}\|_{1,2,s} = \|\mathbf{x}\|_1$ and we recover that s-sparse vectors in ℓ_1^n embed into ℓ_1^m, with $m = \mathcal{O}(s \log(n))$

Summary

- The Johnson-Lindenstrauss Lemma says that a set of n points in high-dimensional Euclidean space can be mapped down to $m = \mathcal{O}(\varepsilon^{-2} log(n))$ dimensions while preserving pairwise ℓ_2 distances up to $1 \pm \varepsilon$, and a Gaussian random matrix can be used for such an embedding.
- The Johnson-Lindenstrauss embedding property implies the Restricted Isometry Property (RIP), and has applications to sparse recovery. A near-converse result is also true: any matrix with the RIP, with column signs randomly flipped, will be a Johnson-Lindenstrauss embedding.
- In many cases, ℓ_1 distance preservation is more meaningful than ℓ_2 distances. Although there is no analog of the Johnson-Lindenstrauss for ℓ_1, we may consider a block norm which interpolates between ℓ_1 and ℓ_2, and derive near-ℓ_1 embedding results for approximately sparse vectors through this interpolation.

Thank you!

References

Ward, "A unified framework for linear dimensionality reduction in ℓ_1". arXiv preprint arXiv:1405.1332 (2014).

Krahmer, Ward. "New and improved Johnson-Lindenstrauss embeddings via the restricted isometry property." *SIAM Journal on Mathematical Analysis* 43.3 (2011): 1269-1281.