Random linear maps

Daniel Hsu

COMS 4772

JL lemma

JL lemma

Johnson and Lindenstrauss (1984) theorem. There is a constant $C > 0$ such that the following holds. For any $\varepsilon \in (0, 1/2)$, point set $S \subset \mathbb{R}^d$ of cardinality $|S| = n$, and $k \in \mathbb{N}$ such that $k \geq \frac{C \log n}{\varepsilon^2}$, there exists a linear map $f : \mathbb{R}^d \to \mathbb{R}^k$ such that

$$(1-\varepsilon)\|\boldsymbol{x}-\boldsymbol{y}\|_2^2 \;\leq\; \|f(\boldsymbol{x})-f(\boldsymbol{y})\|_2^2 \;\leq\; (1+\varepsilon)\|\boldsymbol{x}-\boldsymbol{y}\|_2^2 \quad \text{for all } \boldsymbol{x}, \boldsymbol{y} \in S\,.$$

- There is a randomized procedure to efficiently construct f.
- Target dimension k need not depend on original dimension d.
- Any data analysis based on Euclidean distances among n points can be approximately carried out in dimension $O(\log n)$.
  - E.g., nearest-neighbor computations, many clustering procedures

Proofs of JL lemma

Many ways to (randomly) construct f that proves the lemma.

1. Original construction:

$$f(\boldsymbol{x}) \;=\; \sqrt{\frac{d}{k}}\boldsymbol{A}\boldsymbol{x}$$

   where rows of $\boldsymbol{A}$ are orthonormal basis (ONB) for k-dimensional subspace chosen uniformly at random.

2. Simpler construction (Indyk & Motwani, 1998):

$$f(\boldsymbol{x}) \;=\; \frac{1}{\sqrt{k}}\boldsymbol{A}\boldsymbol{x}$$

   where $\boldsymbol{A}$ is a random matrix whose entries are iid $\mathrm{N}(0, 1)$.
   - Can replace $\mathrm{N}(0, 1)$ with any subgaussian distribution with mean zero and unit variance.

Uniformly random unit vector

Pick $Z_1, Z_2, \ldots, Z_d$ iid $\mathrm{N}(0,1)$, and set

$$\boldsymbol{U} \;:=\; \frac{(Z_1, Z_2, \ldots, Z_d)}{\sqrt{Z_1^2 + Z_2^2 + \cdots + Z_d^2}}\,.$$

Aside: if $\boldsymbol{U}$ and $W_d \sim \chi^2(d)$ are independent, then

$$\sqrt{W_d}\,\boldsymbol{U} \;\sim\; \mathrm{N}(\boldsymbol{0}, \boldsymbol{I})\,.$$

ONB for uniformly random k-dimensional subspace

- Pick $\boldsymbol{U}_1$ uniformly at random from S^{d-1}.
  - Let columns of $\boldsymbol{V}_1$ be ONB for subspace orthogonal to $\operatorname{span}\{\boldsymbol{U}_1\}$.
- Pick $\boldsymbol{U}_2$ uniformly at random from $\boldsymbol{V}_1 S^{d-2}$.
  - Let columns of $\boldsymbol{V}_2$ be ONB for subspace orthogonal to $\operatorname{span}\{\boldsymbol{U}_1, \boldsymbol{U}_2\}$.
- Pick $\boldsymbol{U}_3$ uniformly at random from $\boldsymbol{V}_2 S^{d-3}$.
  - Let columns of $\boldsymbol{V}_3$ be ONB for subspace orthogonal to $\operatorname{span}\{\boldsymbol{U}_1, \boldsymbol{U}_2, \boldsymbol{U}_3\}$.
- ...
- Mapping is

$$f(\boldsymbol{x}) \;=\; \sqrt{\frac{d}{k}} \begin{bmatrix} \langle \boldsymbol{U}_1, \boldsymbol{x} \rangle \\ \langle \boldsymbol{U}_2, \boldsymbol{x} \rangle \\ \vdots \\ \langle \boldsymbol{U}_k, \boldsymbol{x} \rangle \end{bmatrix}.$$

ONB for uniformly random k-dimensional subspace

Easier method:

- Pick $k \times d$ random matrix $\boldsymbol{A}$ with all entries iid $\mathrm{N}(0,1)$.
- Run *Gram-Schmidt orthogonalization* on the rows.

Requirements of the randomized construction

- f is a linear map, so $f(\boldsymbol{x}) - f(\boldsymbol{y}) = f(\boldsymbol{x} - \boldsymbol{y})$.
- f "works" for all $\binom{n}{2}$ squared lengths $\|f(\boldsymbol{x} - \boldsymbol{y})\|_2^2$:

$$(1-\varepsilon)\|\boldsymbol{x}-\boldsymbol{y}\|_2^2 \;\leq\; \|f(\boldsymbol{x}-\boldsymbol{y})\|_2^2 \;\leq\; (1+\varepsilon)\|\boldsymbol{x}-\boldsymbol{y}\|_2^2 \,.$$

- Equivalently, ensure for each of $\binom{n}{2}$ unit vectors $\boldsymbol{v} := \frac{\boldsymbol{x}-\boldsymbol{y}}{\|\boldsymbol{x}-\boldsymbol{y}\|_2}$,

$$1-\varepsilon \;\leq\; \|f(\boldsymbol{v})\|_2^2 \;\leq\; 1+\varepsilon \,.$$

- **Proof strategy**: prove that, for any such unit vector $\boldsymbol{v}$,

$$\mathbb{P}\Big(\|f(\boldsymbol{v})\|_2^2 \notin [1-\varepsilon, 1+\varepsilon]\Big) \;\leq\; \frac{2}{n^2} \,.$$

- By a union bound over all $\binom{n}{2}$ choices of $\boldsymbol{v}$, we achieve the required properties with probability at least $1/n$.

Key lemma

Key lemma: for any fixed $\boldsymbol{v} \in S^{d-1}$,

$$\mathbb{P}\Big(\|f(\boldsymbol{v})\|_2^2 \notin [1-\varepsilon, 1+\varepsilon] \Big) \ \leq\ \frac{2}{n^2}\,.$$

- **Simple construction**: $f(\boldsymbol{v}) = \frac{1}{\sqrt{k}} \boldsymbol{A}\boldsymbol{v}$, where $\boldsymbol{A}$ is $k \times d$ random matrix with iid $\mathrm{N}(0,1)$ entries.
- Each entry of $\boldsymbol{A}\boldsymbol{v}$ is a linear combination of iid $\mathrm{N}(0,1)$ random variables: for $Z \sim \mathrm{N}(0,1)$,

$$\sum_{j=1}^{d} A_{i,j} v_j \ \stackrel{\text{dist}}{=}\ \left(\sum_{j=1}^{d} v_j^2 \right)^{1/2} Z \ =\ Z\,.$$

- So distribution of $\|\boldsymbol{A}\boldsymbol{v}\|_2^2$ is same as that of $\sum_{i=1}^{k} Z_i^2$, where $Z_1, Z_2, \dots, Z_k$ are iid $\mathrm{N}(0,1)$.
- I.e., $Y := \|\boldsymbol{A}\boldsymbol{v}\|_2^2 \sim \chi^2(k)$.

Proof of key lemma

To prove: for $Y \sim \chi^2(k)$,

$$\mathbb{P}\big(Y \notin k\,[1-\varepsilon, 1+\varepsilon] \big) \ \leq\ \frac{2}{n^2}\,.$$

- Recall: Y is $(4k, 4)$-subexponential, so

$$\mathbb{P}(Y \geq k + t) \ \leq\ \exp\Big(-\min\Big\{ t^2/k,\ t \Big\}/8 \Big)\,.$$

- Also can show that $-Y$ is $2k$-subgaussian, so

$$\mathbb{P}(Y \leq k - t) \ =\ \mathbb{P}(-Y \geq -k + t) \ \leq\ \exp\Big(-t^2/(4k) \Big)\,.$$

- For $t := k\varepsilon$, each bound is at most $\exp(-k\varepsilon^2/8)$.
- Proof follows by using assumption $k \geq \frac{16 \ln(n)}{\varepsilon^2}$. $\square$

Finishing the proof of JL lemma

- For any pair of distinct points $\boldsymbol{x}, \boldsymbol{y} \in S$,

$$\mathbb{P}\left(\frac{\|f(\boldsymbol{x}) - f(\boldsymbol{y})\|_2^2}{\|\boldsymbol{x} - \boldsymbol{y}\|_2^2} \notin [1-\varepsilon, 1+\varepsilon]\right) \leq 2\exp\left(-k\varepsilon^2/8\right) \leq \frac{2}{n^2}.$$

- Union bound over all $\binom{n}{2}$ pairs:

$$\mathbb{P}\left(\exists \boldsymbol{x}, \boldsymbol{y} \in S \,\textbf{.}\, \frac{\|f(\boldsymbol{x}) - f(\boldsymbol{y})\|_2^2}{\|\boldsymbol{x} - \boldsymbol{y}\|_2^2} \notin [1-\varepsilon, 1+\varepsilon]\right) \leq \binom{n}{2}\frac{2}{n^2}.$$

- Therefore, with probability at least $1/n$,

$$\frac{\|f(\boldsymbol{x}) - f(\boldsymbol{y})\|_2^2}{\|\boldsymbol{x} - \boldsymbol{y}\|_2^2} \in [1-\varepsilon, 1+\varepsilon] \quad \text{for all } \boldsymbol{x}, \boldsymbol{y} \in S. \quad \square$$

- *Note*: success probability is $1-\delta$ if $k \geq \frac{16\ln(n)+8\ln(1/\delta)}{\varepsilon^2}$.

Original construction

Original construction:

$$f(\boldsymbol{x}) = \sqrt{\frac{d}{k}}\boldsymbol{A}\boldsymbol{x}$$

where rows of $\boldsymbol{A}$ are ONB for k-dimensional subspace chosen uniformly at random.

- Elementary proof by Dasgupta and Gupta (2002) also reduces to similar key lemma: for any *fixed* $\boldsymbol{v} \in S^{d-1}$,

$$\mathbb{P}\left(\|f(\boldsymbol{v})\|_2^2 \notin [1-\varepsilon, 1+\varepsilon]\right) \leq 2\exp\left(-\Omega(k\varepsilon^2)\right).$$

- *Key insight*: Distribution of $\|\boldsymbol{A}\boldsymbol{v}\|_2^2$ is the same as $\|\boldsymbol{R}\boldsymbol{U}\|_2^2$, where $\boldsymbol{R}$'s rows are ONB for *fixed* k-dimensional subspace, and $\boldsymbol{U}$ is a uniformly random unit vector in S^{d-1}.

Fast JL transform

Computational issues

- d = original dimension; k = target dimension.
- Time to apply $f \colon \mathbb{R}^d \to \mathbb{R}^k$ is $O(kd)$.
  - Due to matrix-vector multiplication.
  - Not obvious how to speed-up this up because matrix is mostly unstructured.

Using a structured random matrix

- **Simple idea**: suppose $\boldsymbol{M}$ is *sparse*, i.e., $\mathrm{nnz}(\boldsymbol{M}) \ll kd$.
  - Can multiply vector by $\boldsymbol{M}$ in time $O(\mathrm{nnz}(\boldsymbol{M}))$.
  - Still want $\boldsymbol{M}$ to satisfy "JL property": for any *fixed* $\boldsymbol{x} \in S^{d-1}$,

$$\mathbb{P}\Big(\|\boldsymbol{M}\boldsymbol{x}\|_2^2 \notin [1-\varepsilon, 1+\varepsilon]\Big) \;\leq\; 2\exp\Big(-\Omega(k\varepsilon^2)\Big)\,.$$

Sparse random matrix

Define $\boldsymbol{M}$ to be $k \times d$ random matrix with iid entries

$$M_{i,j} \;:=\; \frac{1}{\sqrt{\theta k}} A_{i,j} B_{i,j}\,,$$

where $A_{i,j} \sim \mathrm{N}(0,1)$ and $B_{i,j} \sim \mathrm{Bern}(\theta)$, which are also independent of each other.

- Write as $\boldsymbol{M} = \frac{1}{\sqrt{\theta k}}(\boldsymbol{A} \odot \boldsymbol{B})$.
- Scaling ensures $\mathbb{E}\,\|\boldsymbol{M}\boldsymbol{x}\|_2^2 = 1$ for every $\boldsymbol{x} \in S^{d-1}$.
- $\mathbb{E}(\mathrm{nnz}(\boldsymbol{M})) = \theta k d$.
- Great if we can use $\theta = O(1/d + 1/k)$, which would give $\mathbb{E}(\mathrm{nnz}(\boldsymbol{M})) = O(k+d)$.
- But does it satisfy JL property?
  - Depends on $\boldsymbol{x}$ ...

JL property for sparse random matrix

$$\|\boldsymbol{Mx}\|_2^2 = \sum_{i=1}^{k}\left(\sum_{j=1}^{d}\frac{1}{\sqrt{\theta k}}A_{i,j}B_{i,j}x_j\right)^2 \overset{\text{dist}}{=} \frac{1}{\theta k}\sum_{i=1}^{k}\left(\sum_{j=1}^{d}B_{i,j}x_j^2\right)Z_i^2$$

where $Z_1, Z_2, \ldots, Z_k$ are iid $\mathrm{N}(0,1)$.

- Suppose $\boldsymbol{x} = (1, 0, \ldots, 0)$.
  - $\|\boldsymbol{Mx}\|_2^2$ depends only on first column of $\boldsymbol{M}$:

$$\|\boldsymbol{Mx}\|_2^2 \overset{\text{dist}}{=} \frac{1}{\theta k}\sum_{i=1}^{k}B_{i,1}Z_i^2 .$$

  - Variance is $\approx 3/(\theta k)$, which is $O(\varepsilon^2)$ only if $\theta = \Omega(1/(k\varepsilon^2))$.

JL property for sparse random matrix

$$\|\boldsymbol{Mx}\|_2^2 = \sum_{i=1}^{k}\left(\sum_{j=1}^{d}\frac{1}{\sqrt{\theta k}}A_{i,j}B_{i,j}x_j\right)^2 \overset{\text{dist}}{=} \frac{1}{\theta k}\sum_{i=1}^{k}\left(\sum_{j=1}^{d}B_{i,j}x_j^2\right)Z_i^2$$

where $Z_1, Z_2, \ldots, Z_k$ are iid $\mathrm{N}(0,1)$.

- Suppose instead $\boldsymbol{x} = (d^{-1/2}, d^{-1/2}, \ldots, d^{-1/2})$.
  - Averaging effect: with high probability,

$$\sum_{j=1}^{d}B_{i,j}x_j^2 = \frac{1}{d}\sum_{j=1}^{d}B_{i,j} = \theta \pm O\left(\sqrt{\frac{\theta}{d}} + \frac{1}{d}\right).$$

  - Just need $\theta = \Omega(1/d)$. In general, just need $\theta = \Omega(\|\boldsymbol{x}\|_\infty^2)$.

Densification

- Sparse random matrix not great for *sparse unit vectors*, but great for *dense unit vectors*, which have

$$\|\boldsymbol{x}\|_\infty^2 \;=\; \max_{i\in[d]} x_i^2 \;\approx\; \frac{1}{d}\,.$$

- **Idea**: compose two linear maps.
  1. "Densifying" orthogonal transformation:

$$\text{(maybe sparse)}\;\; \boldsymbol{x} \;\mapsto\; \boldsymbol{Qx} \;\; \text{(likely dense)}\,.$$

  2. Sparse linear map:

$$\boldsymbol{Qx} \;\mapsto\; \frac{1}{\sqrt{\theta k}}(\boldsymbol{A}\odot\boldsymbol{B})(\boldsymbol{Qx})\,.$$

Simple densification (picture)

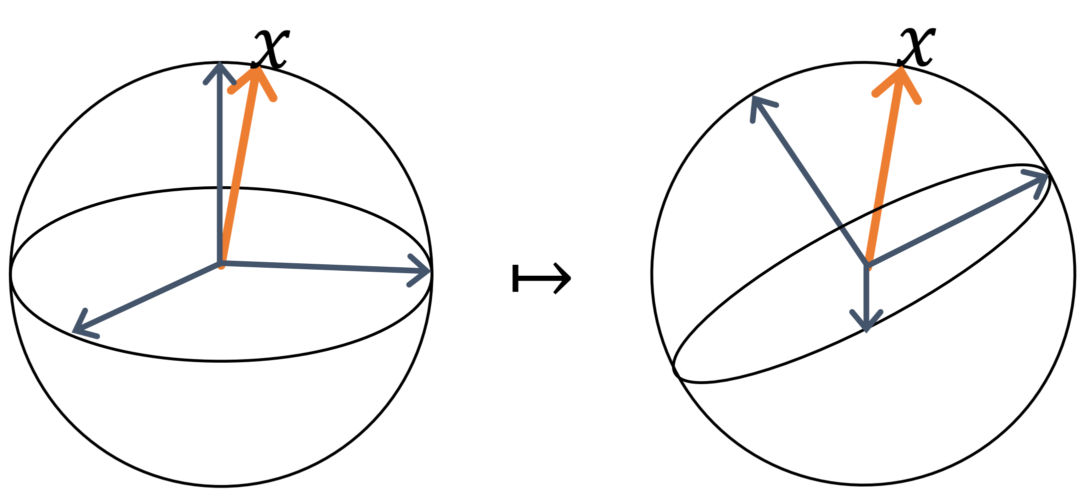

Figure 1: Densifying orthogonal transformation

Simple densification

- Let $\boldsymbol{Q}$ be uniformly random $d \times d$ orthogonal matrix.
  - i-th row $\boldsymbol{Q}_i^\top$ of $\boldsymbol{Q}$ is a uniformly random unit vector.
  - i-th entry of $\boldsymbol{Qx}$ is $\langle \boldsymbol{Q}_i, \boldsymbol{x} \rangle$.
- Can show that

$$\mathbb{P}(|\langle \boldsymbol{Q}_i, \boldsymbol{x} \rangle| \geq \varepsilon) \;\leq\; 2e^{-\varepsilon^2(d-1)/2}\,.$$

- Union bound $\Rightarrow$ with high probability,

$$\langle \boldsymbol{Q}_i, \boldsymbol{x} \rangle^2 \;\leq\; O\left(\frac{\log d}{d}\right) \quad \text{for all } i = 1, 2, \ldots, d\,.$$

Faster densification

- Unfortunately, uniformly random orthogonal matrix also mostly unstructured; time to apply is $O(d^2)$.
- **Insight of (Ailon and Chazelle, 2006)**: can use highly structured "densifying" orthogonal matrix:

$$\boldsymbol{x} \;\mapsto\; \frac{1}{\sqrt{d}}\boldsymbol{HDx}\,.$$

  - $\boldsymbol{H} = \boldsymbol{H}_d$ is the $d \times d$ *Hadamard matrix* (not random).
  - $\boldsymbol{D}$ is random diagonal matrix where diagonal entries are iid Rademacher.

Hadamard matrices

- Recursive definition (for d a power of two):

$$\boldsymbol{H}_1 := +1, \qquad \boldsymbol{H}_d := \begin{bmatrix} +\boldsymbol{H}_{d/2} & +\boldsymbol{H}_{d/2} \\ +\boldsymbol{H}_{d/2} & -\boldsymbol{H}_{d/2} \end{bmatrix}.$$

- Example: $d = 4$

$$\boldsymbol{H}_4 = \begin{bmatrix} +1 & +1 & +1 & +1 \\ +1 & -1 & +1 & -1 \\ +1 & +1 & -1 & -1 \\ +1 & -1 & -1 & +1 \end{bmatrix}.$$

- **Fact 1**: $\frac{1}{\sqrt{d}}\boldsymbol{H}_d$ is orthogonal, and so is $\frac{1}{\sqrt{d}}\boldsymbol{H}_d\boldsymbol{D}$.
- **Fact 2**: Multiplication by $\boldsymbol{D}$ requires $O(d)$ time.
- **Fact 3**: Multiplication by $\boldsymbol{H}_d$ requires $O(d \log d)$ time!

Hadamard transform via divide-and-conquer

- To compute $\boldsymbol{H}_d\boldsymbol{x}$:
  - Partition $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2)$, so $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^{d/2}$.
  - Recursively compute $\boldsymbol{H}_{d/2}\boldsymbol{x}_1$ and $\boldsymbol{H}_{d/2}\boldsymbol{x}_2$.
  - Compute $\boldsymbol{H}_{d/2}\boldsymbol{x}_1 + \boldsymbol{H}_{d/2}\boldsymbol{x}_2$ and $\boldsymbol{H}_{d/2}\boldsymbol{x}_1 - \boldsymbol{H}_{d/2}\boldsymbol{x}_2$.
  - Return $\boldsymbol{H}_d\boldsymbol{x} = \begin{bmatrix} \boldsymbol{H}_{d/2}\boldsymbol{x}_1 + \boldsymbol{H}_{d/2}\boldsymbol{x}_2 \\ \boldsymbol{H}_{d/2}\boldsymbol{x}_1 - \boldsymbol{H}_{d/2}\boldsymbol{x}_2 \end{bmatrix}$.
- Total time: $O(d \log d)$.

Analysis of randomized Hadamard transform

- Let $\boldsymbol{Y} := \frac{1}{\sqrt{d}} \boldsymbol{H}\boldsymbol{D}\boldsymbol{x}$ for *fixed* unit vector $\boldsymbol{x} \in S^{d-1}$.
- Want to show that $\|\boldsymbol{Y}\|_\infty^2 = O\left(\frac{\log d}{d}\right)$ with high probability.
- For each $i = 1, 2, \ldots, d$,

$$Y_i \;=\; \frac{1}{\sqrt{d}} \sum_{j=1}^{d} H_{i,j} \sigma_j x_j \;\stackrel{\text{dist}}{=}\; \frac{1}{\sqrt{d}} \sum_{j=1}^{d} x_j \sigma_j \,,$$

  where $\sigma_1, \sigma_2, \ldots, \sigma_d$ are iid Rademacher.
- Each Y_i has mean zero and is 1-subgaussian, so with high probability,

$$Y_i^2 \;\leq\; O\left(\frac{\log d}{d}\right) \quad \text{for all } i = 1, 2, \ldots, d \,.$$

Overall random linear map (picture)

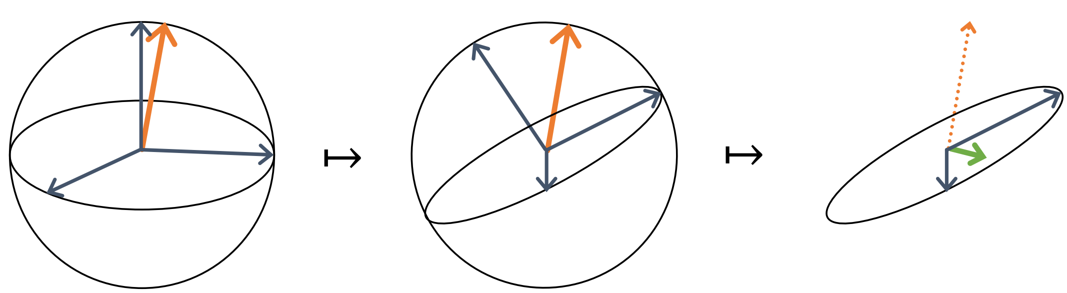

Figure 2: Randomized Hadamard transform + sparse random linear map

Overall random linear map

- Overall linear map from $\mathbb{R}^d$ to $\mathbb{R}^k$:
  1. Densification (randomized Hadamard transform):
  $$\boldsymbol{x} \;\mapsto\; \boldsymbol{y} \;:=\; \frac{1}{\sqrt{d}}\boldsymbol{H}\boldsymbol{D}\boldsymbol{x}\,.$$
  2. Dimension reduction (sparse random linear map):
  $$\boldsymbol{y} \;\mapsto\; \frac{1}{\sqrt{\theta k}}(\boldsymbol{A} \odot \boldsymbol{B})\boldsymbol{y}\,.$$
- Overall running time: $O(d \log d + \theta k d)$.
- Can use $\theta \approx \frac{\log d}{d}$, so running time is $O((d + k) \log d)$.