CSE 392: Matrix and Tensor Algorithms for Data

Instructor: Shashanka Ubaru

University of Texas, Austin
Spring 2024

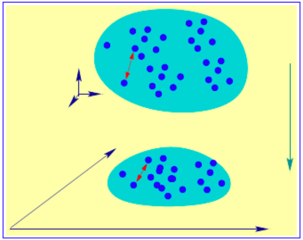
Lecture 7: JL Lemma and subspace embedding

Outline

1. Near orthogonal vectors and ϵ-Net
2. Gaussian matrix properties
3. Johnson-Lindenstrauss Lemma
4. Subspace embedding

High-dimensional vectors

- Often we deal with data vectors that are high-dimensional.
- **Dimensionality reduction:** One popular approach is to embed these vectors on a low-dimensional space.
- What criteria should we use to compute this low-dimensional embedding? What properties of the data do we wish to preserve?



Near-orthogonal vectors

Given a d-dimensional space, what is the largest set of *mutually orthogonal* unit vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t$ we can have? I.e. with the inner products

$$|\boldsymbol{x}_i^\top \boldsymbol{x}_j| = 0 \quad \forall i, j$$

Near-orthogonal vectors

Given a d-dimensional space, what is the largest set of *mutually orthogonal* unit vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t$ we can have? I.e. with the inner products

$$|\boldsymbol{x}_i^\top \boldsymbol{x}_j| = 0 \quad \forall i, j$$

Answer: d

Given a d-dimensional space, what is the largest set of nearly orthogonal unit vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t$? I.e. with the inner products

$$|\boldsymbol{x}_i^\top \boldsymbol{x}_j| \leq \epsilon \quad \forall i, j$$

Suppose ϵ is a constant. E.g. $\epsilon = 1/10$.

Near-orthogonal vectors

Given a d-dimensional space, what is the largest set of *mutually orthogonal* unit vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t$ we can have? I.e. with the inner products

$$|\boldsymbol{x}_i^\top \boldsymbol{x}_j| = 0 \quad \forall i, j$$

Answer: d

Given a d-dimensional space, what is the largest set of nearly orthogonal unit vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t$? I.e. with the inner products

$$|\boldsymbol{x}_i^\top \boldsymbol{x}_j| \leq \epsilon \quad \forall i, j$$

Suppose ϵ is a constant. E.g. $\epsilon = 1/10$.
Answer: $2^{\Theta(d)}$

Near-orthogonal vectors

Claim: There is an exponential number of nearly orthogonal unit vectors in d-dimensional space ($\sim 2^d$).
Proof approach: One approach is to use *Probabilistic Argument.* For $t = 2^{\Theta(d)}$, define a random process which generates random vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t$ that are unlikely to have large inner product

- Show that, with high probability, $|\boldsymbol{x}_i^\top \boldsymbol{x}_j| \leq \epsilon \quad \forall i, j$.
- Hence, there must exists some set of unit vectors with all pairwise inner-products bounded by ϵ.

Proof: Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t$ be normalized Radmacher vectors, i.e., have independent random entries, each set to $\pm 1/\sqrt{d}$ with equal probability.

$$\mathbb{E}[\boldsymbol{x}_i^\top \boldsymbol{x}_j] = ?$$

Let $S = \boldsymbol{x}_i^\top \boldsymbol{x}_j = \sum_{i=1}^{d} \mathrm{c}_i$, where c_i is random $\pm 1/d$.
S is sum of i.i.d random variables. Lets use Hoeffding's inequality:

Hoeffding Inequality

Let $\mathrm{c}_1, \ldots, \mathrm{c}_d$ be independent random variables with each $\mathrm{c}_i \in [a_i, b_i]$. Let $\mathbb{E}[\mathrm{c}_i] = \mu_i$ and $\mathrm{Var}[\mathrm{c}_i] = \sigma_i^2$. Let $\mu = \sum_i \mu_i$ and $\sigma^2 = \sum_i \sigma_i^2$. Then, for and $\alpha > 0$, $S = \sum_i \mathrm{c}_i$ satisfies

$$\Pr[|S - \mu| \geq \alpha] \leq 2e^{-\frac{2\alpha^2}{\sum_i (a_i - b_i)^2}}.$$

Hoeffding Inequality

Let $\mathrm{c}_1, \ldots, \mathrm{c}_d$ be independent random variables with each $\mathrm{c}_i \in [a_i, b_i]$. Let $\mathbb{E}[\mathrm{c}_i] = \mu_i$ and $\mathrm{Var}[\mathrm{c}_i] = \sigma_i^2$. Let $\mu = \sum_i \mu_i$ and $\sigma^2 = \sum_i \sigma_i^2$. Then, for and $\alpha > 0, S = \sum_i \mathrm{c}_i$ satisfies

$$\Pr[|S - \mu| \geq \alpha] \leq 2e^{-\frac{2\alpha^2}{\sum_i (a_i - b_i)^2}}.$$

Here, $a_i = -1/d, b_i = 1/d$. $\mu_i =$?

Hoeffding Inequality

Let $c_1, \ldots, c_d$ be independent random variables with each $c_i \in [a_i, b_i]$. Let $\mathbb{E}[c_i] = \mu_i$ and $\mathrm{Var}[c_i] = \sigma_i^2$. Let $\mu = \sum_i \mu_i$ and $\sigma^2 = \sum_i \sigma_i^2$. Then, for and $\alpha > 0, S = \sum_i c_i$ satisfies

$$\Pr[|S - \mu| \geq \alpha] \leq 2e^{-\frac{2\alpha^2}{\sum_i (a_i - b_i)^2}}.$$

Here, $a_i = -1/d, b_i = 1/d$. $\mu_i =$?

We have

$$\Pr[|\boldsymbol{x}_i^\top \boldsymbol{x}_j| \geq \epsilon] \leq 2e^{-\epsilon^2 d/2}$$

For any pair i, j, we have $\Pr[|\boldsymbol{x}_i^\top \boldsymbol{x}_j| < \epsilon] > 1 - 2e^{-\epsilon^2 d/2}$. Taking union bound over all possible pairs, we get

$$\Pr[|\boldsymbol{x}_i^\top \boldsymbol{x}_j| < \epsilon] > 1 - \binom{t}{2} 2e^{-\epsilon^2 d/2}$$

Near-orthogonal vectors

- **Result:** In d-dimensional space, there are $t = 2^{\Theta(\epsilon^2 d)}$unit vectors with all pairwise inner products $\leq \epsilon$.
- **Alternate point of view :** Random vectors tend to be far apart (and roughly equidistant) in high-dimensions.
- **Curse of dimensionality:** If our data distribution is truly random, suppose we want to use say k-nearest neighbors to learn a function or classify points in $\mathbb{R}^d$, we typically need an exponential amount of data.
- Hope is that there exists low dimensional structure is our data.

Alternate approach: ϵ-Nets

Some definitions:

- **Unit sphere:** Let $\mathcal{S}_p^{d-1} \equiv \{\boldsymbol{x} \in \mathbb{R}^d \mid \|\boldsymbol{x}\|_p = 1\}$.
  We will omit p, when $p = 2$, and d when in context.
- **Semi-norms from sets:** For symmetric matrix $\boldsymbol{W} \in \mathbb{R}^{d\times d}$ and non-empty $\mathcal{N} \subset \mathbb{R}^d$, let
  $$\|\boldsymbol{W}\|_{\mathcal{N}} \equiv \sup\{|\boldsymbol{x}^\top \boldsymbol{W}\boldsymbol{x}|/\|\boldsymbol{x}\|^2 \mid \boldsymbol{x} \in \mathcal{N}, \boldsymbol{x} \neq 0\}$$
  so when $\mathcal{N} \subset \mathcal{S}$, $\|\boldsymbol{W}\|_{\mathcal{N}} \equiv \sup_{\boldsymbol{x}\in\mathcal{N}} |\boldsymbol{x}^\top \boldsymbol{W}\boldsymbol{x}|$.
- **Embedding of $\mathcal{N}$:** For $\mathcal{N} \subset \mathbb{R}^d$, $\boldsymbol{B} \in \mathbb{R}^{m\times d}$, and $\beta \in (0, 1]$,
  $\|\boldsymbol{B}^\top \boldsymbol{B} - \boldsymbol{I}\|_{\mathcal{N}} \leq \beta \implies \boldsymbol{B}$ is a β-embedding of $\mathcal{N}$.
- $\boldsymbol{B}^\top \boldsymbol{B} - \boldsymbol{I}$ is called the centered Grammian of $\boldsymbol{B}$.
- If $\|\boldsymbol{B}^\top \boldsymbol{B} - \boldsymbol{I}\|_{\mathcal{S}} \leq \beta$, then $\boldsymbol{B}$ is a β-embedding of $\mathbb{R}^d$.

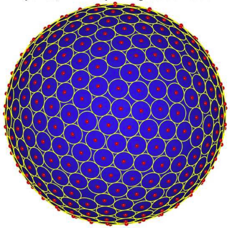
ϵ-Nets

- $\mathcal{N} = \mathcal{N}(\epsilon)$ is an ϵ-net of set $\mathcal{P}$ if it is both:
  - ϵ-*packing*: all $p \in \mathcal{N}$ at least ϵ from $\mathcal{N}$

$$d(p, \mathcal{N} \setminus \{p\}) \geq \epsilon \text{ for } p \in \mathcal{N}$$

  - ϵ-*covering*: all $p \in \mathcal{P}$ at most ϵ from $\mathcal{N}$

$$d(p, \mathcal{N}) \leq \epsilon \text{ for } p \in \mathcal{P}$$



Sphere covering number

The unit sphere $\mathcal{S}$ in $\mathbb{R}^d$ has an ϵ-net of size at most $(1+2/\epsilon)^d$.

Proof is through a volume argument. Since the points in $\mathcal{N}(\epsilon)$ are ϵ-separated, the balls of radii $\epsilon/2$ centered at the points in $\mathcal{N}(\epsilon)$ are disjoint. Also, all such balls lie in $(1+\epsilon/2)B_2^d$ where B_2^d denotes the unit Euclidean ball centered at the origin. So, we have

$$vol(\frac{\epsilon}{2}B_2^d) \cdot |\mathcal{N}(\epsilon)| \le vol((1+\frac{\epsilon}{2})B_2^d)$$

Since, $vol(rB_2^d) = r^d vol(B_2^d)$, we get

$$|\mathcal{N}(\epsilon)| \le (1+\frac{\epsilon}{2})^d/(\frac{\epsilon}{2})^d = (1+\frac{2}{\epsilon})^d.$$

ϵ-Net bound

For $\mathcal{N}_\epsilon$ an ϵ-net of unit sphere $\mathcal{S}$ in $\mathbb{R}^d$ and $\epsilon < 1$, if matrix $\boldsymbol{W}$ is symmetric, then

$$(1 - 2\epsilon)\|\boldsymbol{W}\|_2 \leq \|\boldsymbol{W}\|_{\mathcal{N}_\epsilon} \leq \|\boldsymbol{W}\|_{\mathcal{S}} = \|\boldsymbol{W}\|_2$$

and so if $\boldsymbol{B}$ is a β-embedding of $\mathcal{N}_\epsilon$, then
it is a $\beta/(1 - 2\epsilon)$- embedding of $\mathcal{S}$, and so of $\mathbb{R}^d$.

ϵ-Net bound

For $\mathcal{N}_\epsilon$ an ϵ-net of unit sphere $\mathcal{S}$ in $\mathbb{R}^d$ and $\epsilon < 1$, if matrix $\boldsymbol{W}$ is symmetric, then

$$(1-2\epsilon)\|\boldsymbol{W}\|_2 \leq \|\boldsymbol{W}\|_{\mathcal{N}_\epsilon} \leq \|\boldsymbol{W}\|_{\mathcal{S}} = \|\boldsymbol{W}\|_2$$

and so if $\boldsymbol{B}$ is a β-embedding of $\mathcal{N}_\epsilon$, then
it is a $\beta/(1-2\epsilon)$- embedding of $\mathcal{S}$, and so of $\mathbb{R}^d$.

Proof: Let unit $\boldsymbol{y}$ be such that $|\boldsymbol{y}^\top \boldsymbol{W} \boldsymbol{y}| = \|\boldsymbol{W}\|_2 = \|\boldsymbol{W}\|_{\mathcal{S}}$.
Since $\mathcal{N}_\epsilon$ is an ϵ-net, there is $\boldsymbol{z}$ with $\|\boldsymbol{z}\| \leq \epsilon$ and $(\boldsymbol{y}-\boldsymbol{z}) \in \mathcal{N}_\epsilon$.
Next,

$$\begin{aligned}
\|\boldsymbol{W}\|_2 = |\boldsymbol{y}^\top \boldsymbol{W} \boldsymbol{y}| &= |(\boldsymbol{y}-\boldsymbol{z})^\top \boldsymbol{W}(\boldsymbol{y}-\boldsymbol{z}) + \boldsymbol{z}^\top \boldsymbol{W}\boldsymbol{y} + \boldsymbol{z}^\top \boldsymbol{W}(\boldsymbol{y}-\boldsymbol{z})| \\
&\leq |(\boldsymbol{y}-\boldsymbol{z})^\top \boldsymbol{W}(\boldsymbol{y}-\boldsymbol{z})| + |\boldsymbol{z}^\top \boldsymbol{W}\boldsymbol{y}| + |\boldsymbol{z}^\top \boldsymbol{W}(\boldsymbol{y}-\boldsymbol{z})| \\
&\leq \|\boldsymbol{W}\|_{\mathcal{N}_\epsilon} + \|\boldsymbol{z}\| \cdot \|\boldsymbol{W}\boldsymbol{y}\| + \|\boldsymbol{z}\| \cdot \|\boldsymbol{W}(\boldsymbol{y}-\boldsymbol{z})\| \\
&\leq \|\boldsymbol{W}\|_{\mathcal{N}_\epsilon} + 2\epsilon\|\boldsymbol{W}\|_2.
\end{aligned}$$

Independent Gaussians

Recall the norm estimation random vectors.

- **Gaussians are stable:** Given $\boldsymbol{y} \in \mathbb{R}^d$, if $\boldsymbol{g} \in \mathbb{R}^d$ has entries i.i.d $\mathcal{N}(0, 1)$, then

$$\boldsymbol{g}^\top \boldsymbol{y} \sim \mathcal{N}(0, \|\boldsymbol{y}\|^2)$$

- A sum of independent Gaussians is Gaussian, and a scalar multiple of a Gaussian is Gaussian.

Independent Gaussians

Recall the norm estimation random vectors.

- **Gaussians are stable:** Given $\boldsymbol{y} \in \mathbb{R}^d$, if $\boldsymbol{g} \in \mathbb{R}^d$ has entries i.i.d $\mathcal{N}(0, 1)$, then

$$\boldsymbol{g}^\top \boldsymbol{y} \sim \mathcal{N}(0, \|\boldsymbol{y}\|^2)$$

- A sum of independent Gaussians is Gaussian, and a scalar multiple of a Gaussian is Gaussian.
- **Vector embedding:** Given a unit vector $\boldsymbol{y} \in \mathbb{R}^d, \epsilon \in (0, 1]$. If $\boldsymbol{G} \in \mathbb{R}^{m \times d}$ has independent entries $g_{ij} \sim \mathcal{N}(0, 1/m)$, then

$$\Pr\{|\|\boldsymbol{G}\boldsymbol{y}\|_2^2 - 1| \geq \epsilon\} \leq 2\exp(-\epsilon^2 m/16).$$

  We know $\sqrt{m}\boldsymbol{G}\boldsymbol{y} \sim \mathcal{N}(0, 1)$ and squared norm is a χ_m^2 distribution. Using the standard bounds for concentration of a χ_m^2, we get the above.
- With high probability, $\boldsymbol{G}$ ϵ-embeds unit vectors $\boldsymbol{y} \in \mathbb{R}^d$. Also, for any fixed $\boldsymbol{y} \in \mathbb{R}^d$.

Gaussian width

- **Gaussian width:** Given $\mathcal{R} \subset \mathbb{R}^d$, the Gaussian width of $\mathcal{R}$ is

$$w(\mathcal{R}) \equiv \mathbb{E}_{\boldsymbol{g} \sim \mathcal{N}(0, \boldsymbol{I})}\left[\sup_{\boldsymbol{y}, \boldsymbol{x} \in \mathcal{R}} \boldsymbol{g}^\top (\boldsymbol{y} - \boldsymbol{x})\right].$$

- Alternatively, the Gaussian width of $\mathcal{R}$ is

$$w(\mathcal{R}) \equiv \mathbb{E}_{\boldsymbol{g} \sim \mathcal{N}(0, \boldsymbol{I})}\left[\sup_{\boldsymbol{y} \in \mathcal{R}} \boldsymbol{g}^\top \boldsymbol{y} / \|\boldsymbol{y}\|\right].$$

- Gaussian widths:
  - $w(\mathbb{R}^d) \leq \sqrt{d}$
  - $w(\mathcal{L}) \leq \sqrt{k}$ for $\mathcal{L}$ a k-dimensional subspace.
  - $w(\mathcal{R}) \leq \sqrt{2 \log |\mathcal{R}|}$ for finte $\mathcal{R}$.

Gordon's theorem

Gordon's theorem [G88]

For given $\mathcal{R} \subset \mathbb{R}^d$, if $\boldsymbol{G} \in \mathbb{R}^{m \times d}$ has independent entries $g_{ij} \sim \mathcal{N}(0, 1/m)$, then

$$\Pr\{\|\boldsymbol{G}^\top \boldsymbol{G} - \boldsymbol{I}\|_{\mathcal{R}} \geq 2\beta + \beta^2\} \leq 2\exp(-t^2/2),$$

where $\beta \equiv \frac{w(\mathcal{R})+1+t}{\sqrt{m}}$.

Euclidean dimensionality reduction

Johnson-Lindenstrauss, 1984

For any set of n data points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$ there exists a *linear map* $\Pi : \mathbb{R}^d \to \mathbb{R}^m$ where $m = O(\frac{\log n}{\epsilon^2})$ such that for all i, j,

$$(1-\epsilon)\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2 \le \|\Pi\boldsymbol{x}_i - \Pi\boldsymbol{x}_j\|_2 \le (1+\epsilon)\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2$$

Johnson-Lindenstrauss, 1984

For any set of n data points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$ there exists a *linear map* $\Pi : \mathbb{R}^d \to \mathbb{R}^m$ where $m = O(\frac{\log n}{\epsilon^2})$ such that for all i, j,

$$(1-\epsilon)\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2 \leq \|\Pi\boldsymbol{x}_i - \Pi\boldsymbol{x}_j\|_2 \leq (1+\epsilon)\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2$$

Proof:

- We show that for a Gaussian matrix $\boldsymbol{G} \in \mathbb{R}^{m\times d}$ has independent entries $g_{ij} \sim \mathcal{N}(0, 1/m)$, the result holds.
- Use the vector embedding result from before (squared norm $\|\boldsymbol{G}(\boldsymbol{x}_i - \boldsymbol{x}_j)\|^2$ is χ^2_m distribution with mean $\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2$).
- Set the probability to $1/n^2$. Since we have $< n^2$ possible pairs i, j, using union bound, we get the result.
- For vectors in finite $\mathcal{R} \subset \mathbb{R}^d$, we can use Gordon's theorem to prove similar result.

Original result used rows of a random orthogonal matrix. Random sign matrix, where rows are Radamacher vectors, is an example.

Oblivious subspace embedding

- For real $\boldsymbol{x}, \boldsymbol{y}$ and ϵ, by $\boldsymbol{x} = (1 \pm \epsilon)\boldsymbol{y}$ we mean that $|\boldsymbol{x} - \boldsymbol{y}| \leq \epsilon |\boldsymbol{y}|$.
- **Embedding:** A matrix $\boldsymbol{S} \in \mathbb{R}^{m \times n}$ is an ϵ-embedding of set $\mathcal{P} \subset \mathbb{R}^n$ if, for all $\boldsymbol{y} \in \mathcal{P}$,

$$\|\boldsymbol{S}\boldsymbol{y}\|_2 = (1 \pm \epsilon)\|\boldsymbol{y}\|_2.$$

  We will call $\boldsymbol{S}$ a "sketching matrix".

Subspace embedding

For $\boldsymbol{A} \in \mathbb{R}^{n \times d}$, a matrix $\boldsymbol{S} \in \mathbb{R}^{m \times n}$ is a subspace ϵ-embedding for $\boldsymbol{A}$ if $\boldsymbol{S}$ is an ϵ-embedding for $span(\boldsymbol{A}) = \{\boldsymbol{A}\boldsymbol{x} \mid \boldsymbol{x} \in \mathbb{R}^d\}$. I.e., for all $\boldsymbol{x} \in \mathbb{R}^d$,

$$\|\boldsymbol{S}\boldsymbol{A}\boldsymbol{x}\|_2 = (1 \pm \epsilon)\|\boldsymbol{A}\boldsymbol{x}\|_2.$$

We will call $\boldsymbol{S}\boldsymbol{A}$ a "sketch".

Obliviousness

An *Oblivious* subspace embedding is:

- A probability distribution $\mathcal{D}$ over matrices $\boldsymbol{S} \in \mathbb{R}^{m \times n}$, so that
- For any unknown but fixed matrix $\boldsymbol{A}$, $\boldsymbol{S}$ is a subspace ϵ-embedding for $\boldsymbol{A}$ with high probability.

Advantages:

- Distribution $\mathcal{D}$ does not depend on input data. Construct $\boldsymbol{S}$ without knowing $\boldsymbol{A}$.
- *Streaming:* when entries of $\boldsymbol{A}$ change, $\boldsymbol{SA}$ is easy to update.
- *Distributed:* If each p processor has matrix $\boldsymbol{A}^{(p)}$ and $\boldsymbol{A} = \sum_p \boldsymbol{A}^{(p)}$, compute sketch at each processor.
- *Analysis:* If $\boldsymbol{U}$ has $span(\boldsymbol{U}) = span(\boldsymbol{A})$, then the embedding condition holds for $span(\boldsymbol{A})$ iff it holds for $span(\boldsymbol{U})$. So, we can assume $\boldsymbol{A}$ is orthonormal.

Subspace embedding

Given $\epsilon, \delta > 0$, $\boldsymbol{A} \in \mathbb{R}^{n \times d}$, and unit vector $\boldsymbol{y} \in \mathbb{R}^n$. There is $m = O(\frac{d \log(1/\delta)}{\epsilon^2})$ so that if $\boldsymbol{S} \in \mathbb{R}^{m \times n}$ is randomly chosen from a fixed (oblivious to $\boldsymbol{A}$) distribution with the property that $\boldsymbol{S}$ is an $\epsilon/6$-*embedding* of $\boldsymbol{y}$ (JL property) with failure probability $\delta' = K_1 \exp(-K_2 \epsilon^2 m)$, for some $K_1, K_2 > 0$, then $\boldsymbol{S}$ is a *subspace* ϵ-*embedding* for $\boldsymbol{A}$ with failure probability δ.

Subspace embedding

Given $\epsilon, \delta > 0$, $\boldsymbol{A} \in \mathbb{R}^{n \times d}$, and unit vector $\boldsymbol{y} \in \mathbb{R}^n$. There is $m = O(\frac{d \log(1/\delta)}{\epsilon^2})$ so that if $\boldsymbol{S} \in \mathbb{R}^{m \times n}$ is randomly chosen from a fixed (oblivious to $\boldsymbol{A}$) distribution with the property that $\boldsymbol{S}$ is an $\epsilon/6$-*embedding* of $\boldsymbol{y}$ (JL property) with failure probability $\delta' = K_1 \exp(-K_2 \epsilon^2 m)$, for some $K_1, K_2 > 0$, then $\boldsymbol{S}$ is a *subspace* ϵ-*embedding* for $\boldsymbol{A}$ with failure probability δ.

Proof: We will use the ϵ-net argument with the ϵ-embedding (JL) property.

- Since $\boldsymbol{S}$ is oblivious, assume $\boldsymbol{A}$ has orthonormal columns.
- For some $\epsilon_0 > 0$ (to be determined), we pick an ϵ_0-net $\mathcal{N}_{\epsilon_0} \subset \mathcal{S}$.
- For $\boldsymbol{x} \in \mathcal{N}_{\epsilon_0}, \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} \in span(\boldsymbol{A})$ is a unit vector.
- Let $\boldsymbol{W} := \boldsymbol{A}^\top \boldsymbol{S}^\top \boldsymbol{S} \boldsymbol{A} - \boldsymbol{I}$.
- Note that, for any $\beta \in (0, 1], (1 + \beta)^2 \leq (1 + 3\beta)$ and $(1 - \beta)^2 \geq (1 - 3\beta)$.

So, we have $|\|\boldsymbol{S}\boldsymbol{y}\|_2^2 - 1| \leq \epsilon/2$. Also,

$$|\|\boldsymbol{S}\boldsymbol{y}\|_2^2 - 1| = |\boldsymbol{y}^\top \boldsymbol{S}^\top \boldsymbol{S}\boldsymbol{y} - \boldsymbol{y}^\top \boldsymbol{y}| = |\boldsymbol{x}^\top \boldsymbol{A}^\top \boldsymbol{S}^\top \boldsymbol{S}\boldsymbol{A}\boldsymbol{x} - \boldsymbol{x}^\top \boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{x}| = |\boldsymbol{x}^\top \boldsymbol{W}\boldsymbol{x}| \leq \epsilon/2$$

with failure probability δ'.
Applying this to all vectors in $\mathcal{N}_{\epsilon_0}$, and union bound,

$$\|\boldsymbol{W}\|_{\mathcal{N}_{\epsilon_0}} \leq \epsilon/2 \text{ with failure probability } \leq \delta' |\mathcal{N}_{\epsilon_0}|$$

Using the relation between $\|\boldsymbol{W}\|_{\mathcal{S}}$ and $\|\boldsymbol{W}\|_{\mathcal{N}_{\epsilon_0}}$ and the bound on net size $|\mathcal{N}_{\epsilon_0}|$,

$$\|\boldsymbol{W}\|_{\mathcal{S}} \leq \epsilon/2/(1-\epsilon_0) \text{ with failure probability } \leq \delta' |\mathcal{N}_{\epsilon_0}| \leq (1 + \frac{2}{\epsilon_0})^d K_1 \exp(-K_2 \epsilon^2 m).$$

For fixed ϵ_0, there is $m = O(\frac{d \log(1/\delta)}{\epsilon^2})$, so that this is at most δ.
For $\epsilon_0 \leq 1/2$, we have $\|\boldsymbol{W}\|_{\mathcal{S}} \leq \epsilon$.

Further Reading

- Woodruff, David P. "Sketching as a tool for numerical linear algebra." Foundations and Trends® in Theoretical Computer Science 10.1–2 (2014): 1-157.
- Martinsson, P. G., and Tropp, J. "Randomized numerical linear algebra: foundations and algorithms". arXiv preprint arXiv:2002.01387 (2020).

Questions?