Deck 9: Random embeddings

Math 7870: Topics in Randomized Numerical Linear Algebra

Spring 2026

Akil Narayan

(Linear) Embeddings

We now consider the task of compressing high-dimensional vectors into lower-dimensional space

- We've already seen this through the Johnson-Lindenstrauss Lemma, though that operated on finite sets of points.
- Accomplishing this compression in more general scenarios will reveal new capabilites for us: randomized least squares and the randomized SVD.

(Linear) Embeddings

We now consider the task of compressing high-dimensional vectors into lower-dimensional space

- We've already seen this through the Johnson-Lindenstrauss Lemma, though that operated on finite sets of points.
- Accomplishing this compression in more general scenarios will reveal new capabilites for us: randomized least squares and the randomized SVD.

Let $V \subset \mathbb{R}^n$ be an arbitrary set. A linear map (matrix) $\boldsymbol{S} : \mathbb{R}^n \to \mathbb{R}^d$ is called an ℓ^2 **embedding of distortion** $\epsilon \in (0, 1)$ on V if,

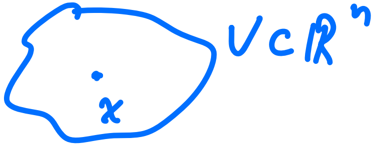
$S \in \mathbb{R}^{d \times n}$

$$(1 - \epsilon)\|\boldsymbol{x}\|_2 \leqslant \|\boldsymbol{S}\boldsymbol{x}\|_2 \leqslant (1 + \epsilon)\|\boldsymbol{x}\|_2, \qquad \boldsymbol{x} \in V.$$

Of particular interest:

- $d \ll n$
- How large must d be to achieve a specific ϵ?
- Generate $\boldsymbol{S}$ randomly.

Random embeddings



Randomness is a conceptually reasonable strategy to generate $\boldsymbol{S}$ since if $\boldsymbol{S}$ is random, then for every deterministic $\boldsymbol{x}$:

$$\mathbb{E}\|\boldsymbol{S}\boldsymbol{x}\|_2^2 = \boldsymbol{x}^* \mathbb{E}\left[\boldsymbol{S}^*\boldsymbol{S}\right]\boldsymbol{x}.$$

Therefore, if $\boldsymbol{S}$ is isotropic: $\mathbb{E}\boldsymbol{S}^*\boldsymbol{S} = \boldsymbol{I}$, then this *on average* is an embedding with zero distortion.

By now, we know that we can make low-distrortion a reality with high probability if we take enough rows of $\boldsymbol{S}$.

≈ "large enough d"

E.g., if $\boldsymbol{x} \in \mathbb{R}^n$ is an isotropic random vector, then filling the rows of $\boldsymbol{S}$ with sufficiently many iid copies of $\boldsymbol{x}$ should do the trick.

$$S = \begin{bmatrix} - & x_1 & - \\ - & x_2 & - \\ & \vdots & \\ - & x_d & - \end{bmatrix}$$

Restricted singular values

Before discussing randomness, let's probe some machinery for embeddings a bit more. Recall that $\boldsymbol{S}$ is an ϵ-distortion of V if,

$$(1-\epsilon)\|\boldsymbol{x}\|_2 \leqslant \|\boldsymbol{S}\boldsymbol{x}\|_2 \leqslant (1+\epsilon)\|\boldsymbol{x}\|_2, \qquad \boldsymbol{x} \in V.$$

The (extremal) V-**restricted singular values of $\boldsymbol{S}$** are,

$$\sigma_{\min}(\boldsymbol{S}; V) := \inf_{\boldsymbol{x}\in V\backslash\{0\}} \frac{\|\boldsymbol{S}\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2}, \qquad \sigma_{\max}(\boldsymbol{S}; V) := \sup_{\boldsymbol{x}\in V\backslash\{0\}} \frac{\|\boldsymbol{S}\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2}.$$

Note $V = \mathbb{R}^n$ reocvers the standard definition of singular values.

What's really important: $\sigma_{\min}(\boldsymbol{S}; V) \neq 0$.
It's ok if $\sigma_{\max}(\boldsymbol{S}; V)$ is large. (This is in general an injection, which works just fine.)

Note then that $\boldsymbol{S}$ is an ϵ-distortion embedding of V iff $\sigma_{\min}(\boldsymbol{S}; V) \geqslant 1-\epsilon$ and $\sigma_{\max}(\boldsymbol{S}; V) \leqslant 1+\epsilon$.

Gaussian embeddings and width

To study randomness, let's keep things simple to start, and consider Gaussian embeddings.

E.g., if $\boldsymbol{S} \in \mathbb{R}^{d \times n}$ is a matrix with iid $\mathcal{N}\left(0, \frac{1}{d}\right)$ (Gaussian) entries, then it's isotropic.

$$\mathbb{E}\, S^* S = I$$

diagonal terms: $\sum_{j \in [d]} \mathbb{E} (S)_{j,1}^2 = \sum_j \frac{1}{d}$

$$= 1$$

Gaussian embeddings and width

To study randomness, let's keep things simple to start, and consider Gaussian embeddings.

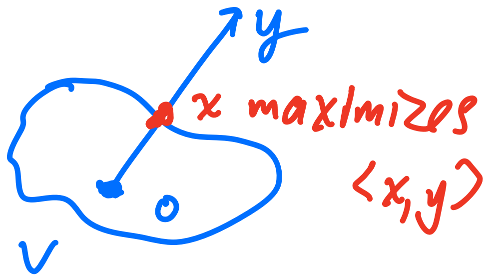
E.g., if $\boldsymbol{S} \in \mathbb{R}^{d \times n}$ is a matrix with iid $\mathcal{N}\left(0, \frac{1}{d}\right)$ (Gaussian) entries, then it's isotropic.

A key tool to understand random Gaussian embeddings is the following concept: Given a set $V \subset \mathbb{R}^n$, the quantity

$$w(V) := \mathbb{E} \sup_{\boldsymbol{x} \in V} \langle \boldsymbol{x}, \boldsymbol{y} \rangle, \qquad \boldsymbol{y} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}),$$

is called the **Gaussian width** of V.

Informally: the Gaussian width measures the size or complexity of V. The quantity $\sup_{\boldsymbol{x} \in V} \langle \boldsymbol{x}, \boldsymbol{y} \rangle$ measures the maximum reach of V in the (unnormalized) direction $\boldsymbol{y}$.
The Gaussian width averages these maximum reaches.



Gaussian width properties

Here are some important properties of Gaussian width:

- $w(V) < \infty$ iff V is a bounded set.
- $\frac{1}{\sqrt{2\pi}}\text{diam}(V) \leqslant w(V) \leqslant \frac{\sqrt{n}}{2}\text{diam}(V)$, where V is the diameter of V.
- The Gaussian width is invariant under taking convex hulls: $w(V) = w(\text{conv}(V))$.
- The Gaussian width is monotonic: if $V_1 \subset V_2$, then $w(V_1) \leqslant w(V_2)$.
- For any deterministic unitary matrix ~~$\boldsymbol{A}$~~ $\boldsymbol{U}$ and any vector $\boldsymbol{b}$, we have $w(V) = w(\boldsymbol{U}V + \boldsymbol{b})$.
- For any $c \in \mathbb{R}$, $w(cV) = |c|w(V)$.
- For any sets V_1 and V_2, $w(V_1 + V_2) = w(V_1) + w(V_2)$, where $V_1 + V_2$ is defined as Minkowski (set) addition.

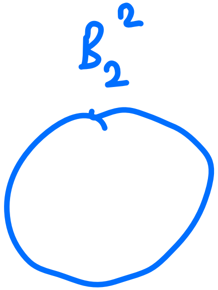
Gaussian width examples

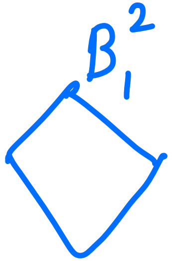
Let B^n_p denote the unit ball in $\mathbb{R}^n$ in the $\ell^p(\mathbb{R}^n)$ norm. Then for absolute constants c_j:

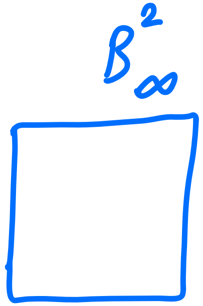
$$w(B^n_2) \in [\sqrt{n} - c_1, \sqrt{n} + c_1], \qquad w(B^n_1) \in [c_2\sqrt{\log n}, c_3\sqrt{\log n}], \qquad w(B^n_\infty) = n\frac{\sqrt{2\pi}}{\;}.$$

In fact, in general $w(B^n_p) \leqslant c_4\sqrt{q}n^{1/q}$, where (p, q) are conjugate exponents: $\frac{1}{p} + \frac{1}{q} = 1$.

If V is a finite point set, $|V| < \infty$, then $w(V) \leqslant c_5 \mathrm{diam}(V)\sqrt{\log |V|}$.







Gaussian width examples

Let B_p^n denote the unit ball in $\mathbb{R}^n$ in the $\ell^p(\mathbb{R}^n)$ norm. Then for absolute constants c_j:

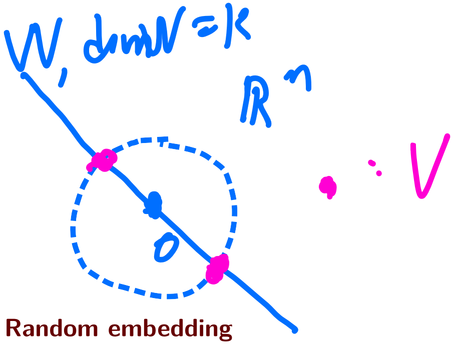
$$w(B_2^n) \in [\sqrt{n} - c_1, \sqrt{n} + c_1], \qquad w(B_1^n) \in [c_2\sqrt{\log n}, c_3\sqrt{\log n}], \qquad w(B_\infty^n) = n\frac{\sqrt{2\pi}}{.}$$

In fact, in general $w(B_p^n) \leqslant c_4\sqrt{q}n^{1/q}$, where (p, q) are conjugate exponents: $\frac{1}{p} + \frac{1}{q} = 1$.

If V is a finite point set, $|V| < \infty$, then $w(V) \leqslant c_5 \mathrm{diam}(V)\sqrt{\log |V|}$.

If V is a subpsace of dimension k, restricted to the unit sphere, then

$$w(V) \in [\sqrt{k-1}, \sqrt{k}].$$



Gaussian widths to embeddings

We can tie many of these results together: The Gaussian width controls concentration of a Gaussian embedding.

Theorem 1. *Let V be a subset of the unit sphere in $\mathbb{R}^n$. Let $\boldsymbol{S} \in \mathbb{R}^{d \times n}$ be a Gaussian embedding. (It has iid $\mathcal{N}\left(0, \frac{1}{d}\right)$ entries.) Then the restricted singular values of $\boldsymbol{S}$ satisfy,*

For all $t > 0$

$$\Pr\left(\sigma_{\min}(\boldsymbol{S}; V) \geqslant 1 + \frac{w(V)}{\sqrt{d}} + t\right) \leqslant \exp(-dt^2/2)$$

$$\Pr\left(\sigma_{\max}(\boldsymbol{S}; V) \leqslant 1 - \frac{1 + w(V)}{\sqrt{d}} - t\right) \leqslant \exp(-dt^2/2).$$

The punch line: picking $t = \epsilon$ and $d \gtrsim w(V)^2/\epsilon^2$ achieves an embedding with high probability.

NB: While Gaussian widths care about sizes of vectors, *subspace* embeddings do not: Given $\sigma_{\min}(V; \boldsymbol{S})$ for a spherical set V, the set cV for any nontrivial constant c has the same restricted singular values.

$$(1-\epsilon) \leq \frac{\|Sx\|^2}{\|x\|^2} \leq (1+\epsilon) \quad x \in V$$

Quick-and-dirty application: Johnson-Lindenstrauss

Let's revisit the Johnson-Lindenstrauss Lemma: Let $\{\boldsymbol{x}_j\}_{j \in [N]} \subset \mathbb{R}^n$ be given. The conclusion of the Johnson-Lindenstrauss Lemma is that there's a linear embedding $\boldsymbol{S} \in \mathbb{R}^{d \times n}$ such that,

$$1 - \epsilon \leqslant \frac{\|\boldsymbol{S}(\boldsymbol{x}_i - \boldsymbol{x}_j)\|_2}{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2} \leqslant 1 + \epsilon.$$

Define the set of points,

$$V := \left\{ \frac{\boldsymbol{x}_i - \boldsymbol{x}_j}{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2} \right\}_{i<j}, \qquad |V| \sim N^2.$$

This is a finite point set, so $w(V) \sim \sqrt{\log |V|} \sim$ ~~$\sqrt{N}$~~. $\sqrt{\log N}$

Then from the previous slide, we know choosing

$$d \sim \frac{w(V)^2}{\epsilon^2} \sim (\log N)/\epsilon^2$$

achieves an ϵ-embedding with high probability.

Application 2: Subspace embeddings

We can now immediately craft subspace embeddings: Let V be a k-dimensional subspace of $\mathbb{R}^n$. We wish to identify $\boldsymbol{S} \in \mathbb{R}^{d \times n}$ such that,

$$1 - \epsilon \leqslant \frac{\|\boldsymbol{S}\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2} \leqslant 1 + \epsilon$$

$$\Updownarrow$$

$$\sigma_{\min}(\boldsymbol{S}; V) \leqslant 1 - \epsilon, \sigma_{\max}(\boldsymbol{S}; V) \geqslant 1 + \epsilon.$$

We know that the Guassian width $w(V \cap S^{n-1}) \leqslant \sqrt{k}$, so taking

$$d \gtrsim \frac{w(V \cap S^{n-1})^2}{\epsilon^2} = k\epsilon^2,$$

and choosing $\boldsymbol{S}$ as a Gaussian embedding does the trick.

Structured Embeddings

The Gaussian strategy for constructing embeddings is useful as a benchmark. It has some weaknesses:

- it's a dense matrix
- naive matrix-vector multiplications cost $\mathcal{O}(dn)$ complexity per vector.

Of course, Gaussian embeddings are a great barometer as they provide essentially optimal complexity (in terms of d vs ϵ), and they are *oblivious*, i.e., they don't depend on the set V.

Structured Embeddings

The Gaussian strategy for constructing embeddings is useful as a benchmark. It has some weaknesses:

- it's a dense matrix
- naive matrix-vector multiplications cost $\mathcal{O}(dn)$ complexity per vector.

Of course, Gaussian embeddings are a great barometer as they provide essentially optimal complexity (in terms of d vs ϵ), and they are *oblivious*, i.e., they don't depend on the set V.

There are two classes of structured embeddings that directly address some of the weaknesses mentioned above:

- randomized sign matrices
- subsampled randomized Fourier transforms

Analyzing subspace embeddings, I

We will be interested in subspace embeddings, so we specialize to that situation now. With V is a k-dimensional subspace of $\mathbb{R}^n$, let $\boldsymbol{U} \in \mathbb{R}^{n \times k}$ be any semi-unitary matrix spanning V:

$$\boldsymbol{U}^* \boldsymbol{U} = \boldsymbol{I}_k, \qquad \text{range}(\boldsymbol{U}) = V.$$

In order to accomplish an ϵ-embedding, we want,

$$1 - \epsilon \leqslant \frac{\|\boldsymbol{S}\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2} \leqslant 1 + \epsilon \quad \forall\ \boldsymbol{x} \in V \backslash \{\boldsymbol{0}\}.$$

This is equivalent to requiring,

$$1 - \epsilon \leqslant \frac{\|\boldsymbol{S}\boldsymbol{U}\boldsymbol{y}\|_2}{\|\boldsymbol{y}\|_2} \leqslant 1 + \epsilon, \quad \forall\ \boldsymbol{y} \in \mathbb{R}^k.$$

Hence, we seek to study the extremal singular values of $\boldsymbol{S}\boldsymbol{U}$.

Analyzing subspace embeddings, II

The above is equivalent to studying the extremal eigenvalues of,

$$\boldsymbol{Y} := (\boldsymbol{SU})^* \boldsymbol{SU} = \boldsymbol{U}^* \boldsymbol{S}^* \boldsymbol{SU}.$$

We wish to design $\boldsymbol{S}$ so that,

$$\lambda_{\text{min}}(\boldsymbol{Y}) \geqslant (1-\epsilon)^2, \qquad \lambda_{\text{max}}(\boldsymbol{Y}) \leqslant (1+\epsilon)^2.$$

Analyzing subspace embeddings, II

The above is equivalent to studying the extremal eigenvalues of,

$$\boldsymbol{Y} := (\boldsymbol{SU})^* \boldsymbol{SU} = \boldsymbol{U}^* \boldsymbol{S}^* \boldsymbol{SU}.$$

We wish to design $\boldsymbol{S}$ so that,

$$\lambda_{\min}(\boldsymbol{Y}) \geqslant (1-\epsilon)^2, \qquad \lambda_{\max}(\boldsymbol{Y}) \leqslant (1+\epsilon)^2.$$

Suppose that we can write $\boldsymbol{S}^*\boldsymbol{S}$ as a sum of iid matrices:

$$\boldsymbol{S}^*\boldsymbol{S} = \sum_{j \in [d]} \boldsymbol{X}_j, \qquad \boldsymbol{X}_j \succeq \mathbf{0}.$$

If this were the case, we'd be able to use our matrix concentration inequalities, in particular Matrix Chernoff....

Sparse sign matrices

Consider the following *sparse sign matrix* model: $\boldsymbol{S} \in \mathbb{R}^{n \times d}$ is a matrix with iid entries, having distribution:

$$(S)_{i,j} = \begin{cases} \alpha, & \text{with probability } p/2 \\ -\alpha, & \text{with probability } p/2 \\ 0, & \text{otherwise} \end{cases}$$

where $p \in (0,1)$ is a free parameter, and α is chosen to maintain isotropy: $pd\alpha^2 = 1$.

The value of p controls sparsity: larger p yields denser matrices.

Sparse sign matrices

Consider the following *sparse sign matrix* model: $\boldsymbol{S} \in \mathbb{R}^{n \times d}$ is a matrix with iid entries, having distribution:

$$(S)_{i,j} = \begin{cases} \alpha, & \text{with probability } p/2 \\ -\alpha, & \text{with probability } p/2 \\ 0, & \text{otherwise} \end{cases}$$

where $p \in (0,1)$ is a free parameter, and α is chosen to maintain isotropy: $pd\alpha^2 = 1$.

The value of p controls sparsity: larger p yields denser matrices.

Since each row of $\boldsymbol{S}$ is independent, then letting $\boldsymbol{s}_j$ denote the jth row of $\boldsymbol{S}$, we have,

$$\boldsymbol{Y} = \boldsymbol{U}^* \boldsymbol{S}^* \boldsymbol{S} \boldsymbol{U} = \sum_{j \in [d]} \boldsymbol{U}^* \boldsymbol{s}_j \boldsymbol{s}_j^* \boldsymbol{U} =: \sum_{j \in [d]} \boldsymbol{X}_j,$$

where each $\boldsymbol{X}_j$ is independent.

Sparse sign matrices

Consider the following *sparse sign matrix* model: $\boldsymbol{S} \in \mathbb{R}^{n \times d}$ is a matrix with iid entries, having distribution:

$$(S)_{i,j} = \begin{cases} \alpha, & \text{with probability } p/2 \\ -\alpha, & \text{with probability } p/2 \\ 0, & \text{otherwise} \end{cases}$$

where $p \in (0,1)$ is a free parameter, and α is chosen to maintain isotropy: $pd\alpha^2 = 1$.

The value of p controls sparsity: larger p yields denser matrices.

Since each row of $\boldsymbol{S}$ is independent, then letting $\boldsymbol{s}_j$ denote the jth row of $\boldsymbol{S}$, we have,

$$\boldsymbol{Y} = \boldsymbol{U}^* \boldsymbol{S}^* \boldsymbol{S} \boldsymbol{U} = \sum_{j \in [d]} \boldsymbol{U}^* \boldsymbol{s}_j \boldsymbol{s}_j^* \boldsymbol{U} =: \sum_{j \in [d]} \boldsymbol{X}_j,$$

where each $\boldsymbol{X}_j$ is independent.

It's possible to compute somewhat sophisticated bounds on $\|\boldsymbol{X}_j\|_2$, yielding to an application of Matrix Chernoff:

If $d \gtrsim (k + \log n) \log k$, and $p \gtrsim (\log k)/d$, then the sparse sign matrix model is a $\frac{1}{2}$-distortion subspace embedding with high probability.

Subsampled Randomized Fourier transforms

Another "structured" random matrix yields similar guarantees. Let $\boldsymbol{F} \in \mathbb{C}^{n \times n}$ denote the discrete Fourier transform matrix, which is unitary.

We construct $\boldsymbol{S}$ as,

$$\boldsymbol{S} = \frac{1}{\sqrt{d}} \boldsymbol{R} \boldsymbol{F} \boldsymbol{\Sigma},$$

where:

- $\boldsymbol{\Sigma}$ is a square, diagonal matrix with iid Bernoulli$(1/2)$ entries.
- $\boldsymbol{R} \in \mathbb{R}^{d \times n}$ is a random restriction matrix (its rows are $\boldsymbol{e}_{q_j}$, where q_j, $j \in [d]$ are iid uniform random variables on $[n]$).

The advantage/motivation: applying $\boldsymbol{S}$ is quite cheap, requiring $\mathcal{O}(n \log d)$ complexity.

This random transform $\boldsymbol{F}\boldsymbol{\Sigma}$ attempts to construct output vectors whose entries have similar magnitudes.

Subsampled Randomized Fourier transforms

Like the sparse sign model, the subsampled randomized Fourier transform can be written as,

$$\boldsymbol{Y} = \boldsymbol{U}^* \boldsymbol{S}^* \boldsymbol{S} \boldsymbol{U} = \frac{1}{d} \sum_{j \in [d]} \boldsymbol{w}_j \boldsymbol{w}_j^*,$$

The following bound also holds:

$$\|\boldsymbol{X}_j\| \leqslant \frac{1}{d} \|\boldsymbol{w}_j\|_2^2 = \frac{1}{d} \|\boldsymbol{e}_{q_j}^* \boldsymbol{F} \boldsymbol{\Sigma} \boldsymbol{U}\|_2^2 \leqslant \frac{k + c \log n}{d}.$$

This last step requires proving a bound on the last vector norm, which requires Hoeffding's inequality.

With all of this, we have that if $d \gtrsim (k + \log n) \log k$, then the subsampled randomized Fourier transform is a subspace embedding with distortion $1/2$ with high probability.