COMP 7070 Advanced Topics in AI and ML February 26, 2025

Lecture 6: Johnson-Lindenstrauss & Matrix Sketching

Instructor: Yifan Chen *Scribes: XIA Lei, YU Donghong* *Proof reader: Xiong Peng*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*
Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

6.1 Johnson-Lindenstrauss Lemma

In this section, we explore the Johnson-Lindenstrauss Lemma, a cornerstone in high-dimensional data analysis that has profoundly influenced algorithm design for processing such data.

Theorem 6.1 (Johnson-Lindenstrauss Lemma) *For any $\varepsilon \in (0,1)$ and any finite set $X \subseteq \mathbb{R}^d$ with $|X| = n$, there exists a linear map $f : \mathbb{R}^d \to \mathbb{R}^m$ with $m = O\left(\frac{\log n}{\varepsilon^2}\right)$ such that*

$$\forall x, y \in X, \quad (1-\varepsilon)\|x-y\|_2^2 \leq \|f(x) - f(y)\|_2^2 \leq (1+\varepsilon)\|x-y\|_2^2.$$

A simple intuition of Theorem 6.1 is that the distance between the embeddings of two data points, which are randomly sampled from space $\mathbb{R}^d$, is bounded, and the bound is related to the distance of the two data points in $\mathbb{R}^d$ space. It states that a finite set of points in a high-dimensional Euclidean space can be embedded into a lower-dimensional space via a linear map f, typically a random projection, while approximately preserving pairwise Euclidean distances. The target dimension m depends only on the number of points n and the distortion parameter ε, not the original dimension d, making it highly efficient for dimensionality reduction.

6.1.1 Applications

The JL Lemma enhances algorithms involving heavy matrix computations:

- **Approximate Matrix Multiplication (AMM):** For matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times n}$, direct multiplication $\mathbf{AB}$ has complexity $O(n^3)$. Using a projection $f : \mathbb{R}^n \to \mathbb{R}^m$ with $m \ll n$, we approximate $\mathbf{AB}$ as $f(\mathbf{A})f(\mathbf{B})$, reducing complexity to $O(n^2 m)$.

- **Graph Convolutional Networks (GCN):** Given an adjacency matrix $\mathbf{A} \in \mathbb{R}^{N\times N}$, hidden state $\mathbf{H}^{t-1} \in \mathbb{R}^{N\times d}$, and weights $\mathbf{W} \in \mathbb{R}^{d\times p}$, the layer output is $\mathbf{AH}^{t-1}\mathbf{W}$. Using a sketching matrix $\mathbf{\Pi} \in \mathbb{R}^{d\times d'}$ with $d' < d$, we approximate it as $\mathbf{A\Pi}(\mathbf{\Pi}^\top \mathbf{H}^{t-1}\mathbf{W})$, lowering computational cost.

- **Attention Mechanisms:** In attention models, the feature matrix $\mathbf{H} \in \mathbb{R}^{d\times d}$ is computed as:

  $$\mathbf{H} = \mathrm{softmax}\left(\frac{\mathbf{QK}^\top}{\sqrt{d}}\right)\mathbf{V},$$

  where $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ are query, key, and value matrices. With $\mathbf{S} \in \mathbb{R}^{d\times d'}$ and $d' < d$, we approximate:

  $$\mathbf{H} \approx \mathrm{softmax}\left(\frac{\mathbf{QK}^\top}{\sqrt{d}}\right)\mathbf{S}(\mathbf{S}^\top \mathbf{V}),$$

  reducing complexity.

6.1.2 Distributional Johnson-Lindenstrauss Lemma

The practical implementation of the JL Lemma relies on a probabilistic version:

Lemma 6.2 (Distributional JL Lemma) *For any $\varepsilon, \delta \in (0, 1/2)$ and integer $d > 1$, there exists a distribution $\mathcal{D}_{\varepsilon,\delta}$ over matrices $\Pi \in \mathbb{R}^{m\times d}$ with $m = O\left(\frac{\log(1/\delta)}{\varepsilon^2}\right)$ such that, for any fixed $z \in \mathbb{R}^d$ with $\|z\|_2 = 1$,*

$$\forall \|z\|_2 = 1, \quad \mathbb{P}_{\Pi\sim\mathcal{D}_{\varepsilon,\delta}}\left(\left|\|\Pi z\|_2^2 - 1\right| > \varepsilon\right) < \delta.$$

This lemma ensures that a random projection matrix Π preserves the norm of any unit vector within $(1 \pm \varepsilon)$ with high probability. To extend this to all pairs in X, consider $z = \frac{x-y}{\|x-y\|_2}$ for $x, y \in X$. The event that the distance is not preserved is:

$$\mathbb{E}_{x,y} = \left\{\left|\|\Pi z\|_2^2 - 1\right| > \varepsilon\right\} = \left\{\left|\|\Pi(x-y)\|_2^2 - \|x-y\|_2^2\right| > \varepsilon\|x-y\|_2^2\right\}.$$

With $\binom{n}{2} = \frac{n(n-1)}{2}$ pairs, the union bound gives:

$$\mathbb{P}\left(\bigcup_{x,y\in X} \mathbb{E}_{x,y}\right) \le \sum_{x,y\in X} \mathbb{P}\left(\mathbb{E}_{x,y}\right) \le \frac{n(n-1)}{2}\delta.$$

Choosing $\delta = \frac{2}{n^2}$, we get:

$$\mathbb{P}\left(\bigcup_{x,y\in X} \mathbb{E}_{x,y}\right) \le \frac{n(n-1)}{2} \cdot \frac{2}{n^2} < 1,$$

ensuring the JL condition holds with probability at least $1 - \frac{n(n-1)}{n^2} \approx 1 - 1/n$.

6.2 Randomized Sketching Method

6.2.1 Sketch

In randomized sketching, we replace a vector or matrix x/X with its sketch $\Pi x/X\Pi^\top$, where Π is a random projection matrix. For a vector z, we aim to preserve its norm approximately: The derivation involving expectation is as follows:

$$\mathbb{E}\left[z^\top \pi^\top \pi z\right] = 1 = z^\top z = z^\top\left(\mathbb{E}[\pi^\top \pi]\right) z \quad \Rightarrow \quad \mathbb{E}[\pi^\top\pi] = I,$$

To ensure this holds in expectation, we require:

$$\mathbb{E}[\Pi^\top \Pi] = I,$$

where I is the identity matrix. Below, we introduce two applications of sketching that satisfy this condition.

- **Coordinate Sketching:** Consider $\Pi \in \mathbb{R}^{1\times d}$, a row vector uniformly distributed over the set $\{\sqrt{d}e_i\}_{i=1}^d$, where e_i is the i-th standard basis vector in $\mathbb{R}^d$. Then:

  $$\mathbb{E}[\Pi^\top\Pi] = \frac{1}{d}\sum_{i=1}^d (\sqrt{d}e_i)(\sqrt{d}e_i^\top) = \frac{1}{d}\sum_{i=1}^d d e_i e_i^\top = \sum_{i=1}^d e_i e_i^\top = I.$$

  This confirms that the expectation condition is met.

- **Approximate Matrix Multiplication (AMM):** For matrices $B \in \mathbb{R}^{d\times n}$ and $C \in \mathbb{R}^{d\times n}$, computing $B^\top C$ is costly when d is large. Using a sketching matrix $\Pi \in \mathbb{R}^{m\times d}$ with $m \ll d$, we approximate:

  $$B^\top \Pi^\top \Pi C \approx B^\top C,$$

  leveraging $\mathbb{E}[\Pi^\top\Pi] = I$ to reduce the dimensionality and computational complexity.

6.3 Matrix Concentration Inequalities

6.3.1 Sub-Gaussian Random Variable

Definition 6.3 (Moment Generating Function, MGF) *Given a random variable $X \sim \mathrm{subG}(\sigma^2)$ with $\mathbb{E}(X) = 0$, the moment generating function satisfies:*

$$\mathbb{E}\left[\exp(\lambda X)\right] \leq \exp\left(\frac{\sigma^2\lambda^2}{2}\right) \quad \forall\lambda \in \mathbb{R}.$$

Lemma 6.4 *Let $X \sim \mathrm{subG}(\sigma^2)$. Then, for any $p \geq 1$,*

$$\mathbb{E}\left[|X|^p\right] \leq (2\sigma^2)^{p/2} p\Gamma\left(\frac{p}{2}\right).$$

In particular,

$$\mathbb{E}\left[|X|^p\right]^{1/p} \leq \sigma e^{1/e}\sqrt{p}.$$

Proof. We compute $\mathbb{E}\left[|X|^p\right]$ using the integral representation:

$$\begin{aligned}
\mathbb{E}\left[|X|^p\right] &= \int_0^\infty \mathbb{P}\left(|X|^p > t\right)\, dt \\
&= \int_0^\infty \mathbb{P}\left(|X| > t^{1/p}\right)\, dt \\
&\leq 2\int_0^\infty e^{-\frac{t^{2/p}}{2\sigma^2}}\, dt \quad \text{(by sub-Gaussian tail bound)} \\
&= 2\int_0^\infty e^{-\frac{u}{2\sigma^2}}\left(\frac{p}{2}u^{\frac{p}{2}-1}\right) du \quad \text{where } u = t^{2/p},\ dt = \frac{p}{2}u^{\frac{p}{2}-1}\, du \\
&= p\int_0^\infty e^{-\frac{u}{2\sigma^2}} u^{\frac{p}{2}-1}\, du \\
&= p(2\sigma^2)^{\frac{p}{2}}\int_0^\infty e^{-v} v^{\frac{p}{2}-1}\, dv \quad \text{where } v = \frac{u}{2\sigma^2} \\
&= p(2\sigma^2)^{\frac{p}{2}}\Gamma\left(\frac{p}{2}\right), \quad \Gamma(n) = \int_0^\infty e^{-u}u^{n-1}\, du = (n-1)!.
\end{aligned}$$

since $\Gamma\left(\frac{p}{2}\right) = \int_0^\infty e^{-v}v^{\frac{p}{2}-1}\, dv$.

For the second part, we bound $\mathbb{E}\left[|X|^p\right]^{1/p}$:

$$\mathbb{E}\left[|X|^p\right]^{1/p} \leq \left[p(2\sigma^2)^{p/2}\Gamma\left(\frac{p}{2}\right)\right]^{1/p}.$$

Using the approximation $\Gamma\left(\frac{p}{2}\right) \leq \left(\frac{p}{2}\right)^{\frac{p}{2}-1}\Gamma\left(\frac{p}{2}-1\right)$ and properties of the Gamma function, combined with $p^{1/p} \leq e^{1/e}$, we derive:

$$\mathbb{E}\left[|X|^p\right]^{1/p} \leq \sigma e^{1/e}\sqrt{p}.$$

The sub-Gaussian norm is defined as:

$$\|X\|_{\psi_2} = \inf\left\{K > 0 : \mathbb{E}\left[\exp\left(\frac{X^2}{K^2}\right)\right] \leq 2\right\},$$

and for $X \sim \mathrm{subG}(\sigma^2)$, $\|X\|_{\psi_2} \sim \sigma$. A larger $\|X\|_{\psi_2}$ implies X is less sub-Gaussian.

$$\|X\|_{\psi_2} = K_2.$$

Where K_2 is the sub-Gaussian norm and $K_2 \sim \sigma$.

6.3.2 Sub-Gaussian Random Vector X

Definition 6.5 *A random vector $X \in \mathbb{R}^d$ is sub-Gaussian if, for every $x \in \mathbb{R}^d$, the projection $\langle X, x\rangle$ is sub-Gaussian. The sub-Gaussian norm is:*

$$\|X\|_{\psi_2} := \sup_{x\in S^{d-1}} \|\langle X, x\rangle\|_{\psi_2},$$

where $S^{d-1} = \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$.

For example, let X be uniformly distributed over $\{\sqrt{d}e_i\}_{i=1}^d$, where e_i are standard basis vectors. Then:

$$\langle X, x\rangle = \sqrt{d}x_i \quad \text{with probability } \frac{1}{d} \text{ for each } i,$$

and $\|X\|_{\psi_2} \sim \sqrt{d}$, reflecting the dimensional scaling of sub-Gaussianity.

6.4 Π with Independent Sub-Gaussian Rows

Given a random matrix $\Pi \in \mathbb{R}^{m\times d}$ with independent sub-Gaussian rows Π_i, where $\frac{1}{m}\mathbb{E}[\Pi^\top\Pi] = I$, we have:

$$\sqrt{m} - C\sqrt{d} - t \le \sigma_{\min}(\Pi) \le \sigma_{\max}(\Pi) \le \sqrt{m} + C\sqrt{d} + t,$$

with probability at least $1 - 2\exp(-ct^2)$, where $\sigma_{\min}(\Pi)$ and $\sigma_{\max}(\Pi)$ denote the smallest and largest singular values of Π, respectively, and $C, c > 0$ are constants depending on the sub-Gaussian parameters.

6.4.1 ε-Net to Approximate S^{d-1}

To study the singular values of Π, we discretize the unit sphere $S^{d-1} = \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$ using an ε-net.

Definition 6.6 (ε-Net) *A set $\mathcal{N} \subseteq S^{d-1}$ is an ε-net for S^{d-1} if for every $x \in S^{d-1}$, there exists $y \in \mathcal{N}$ such that $\|x - y\|_2 \le \varepsilon$.*

The covering number $\mathcal{N}(S^{d-1}, \varepsilon)$ is the minimal cardinality of such an ε-net.
Bound on $\mathcal{N}(S^{d-1}, \varepsilon)$:

- Let $\mathcal{N}_\varepsilon \subseteq S^{d-1}$ be a maximal ε-separated subset, i.e., for all $y_1, y_2 \in \mathcal{N}_\varepsilon$, $\|y_1 - y_2\|_2 > \varepsilon$.
- This set is an ε-net: if there existed $x \in S^{d-1}$ with $\|x - y\|_2 > \varepsilon$ for all $y \in \mathcal{N}_\varepsilon$, we could add x to $\mathcal{N}_\varepsilon$, contradicting maximality.
- Using volume arguments, the number of $\varepsilon/2$-balls centered at points in $\mathcal{N}_\varepsilon$ that pack S^{d-1} satisfies:

  $$|\mathcal{N}_\varepsilon| \left(\frac{\varepsilon}{2}\right)^d \mathcal{B} \le \left(1 + \frac{\varepsilon}{2}\right)^d \mathcal{B},$$

  where $\mathcal{B}$ is the volume of the unit ℓ_2-ball in $\mathbb{R}^d$. Thus:

  $$|\mathcal{N}_\varepsilon| \le \left(\frac{2}{\varepsilon} + 1\right)^d.$$

6.4.2 Singular Value Bounds via ε-Net

We aim to prove:

$$1-\delta \leq S_{\min}(\Pi) \leq S_{\max}(\Pi) \leq 1+\delta,$$

which is equivalent to showing:

$$\left\|\frac{1}{m}\Pi^\top\Pi - I\right\| \leq \delta.$$

Proof:

1. **Norm Approximation via ε-Net:** For a $\frac{1}{4}$-net $\mathcal{N}_{\frac{1}{4}}$ of S^{d-1}, the operator norm can be approximated as:

   $$\left\|\frac{1}{m}\Pi^\top\Pi - I\right\| = \sup_{x\in S^{d-1}}\left|\frac{1}{m}\|\Pi x\|_2^2 - 1\right|.$$

   For any $x \in S^{d-1}$, choose $y \in \mathcal{N}_{\frac{1}{4}}$ such that $\|x-y\|_2 \leq \frac{1}{4}$. Then:

   $$\begin{aligned}\left|\frac{1}{m}\|\Pi x\|_2^2 - \frac{1}{m}\|\Pi y\|_2^2\right| &= \left|x^\top\left(\frac{1}{m}\Pi^\top\Pi\right)x - y^\top\left(\frac{1}{m}\Pi^\top\Pi\right)y\right| \\ &= \left|\langle\frac{1}{m}\Pi^\top\Pi x, x-y\rangle + \langle\frac{1}{m}\Pi^\top\Pi(x-y), y\rangle\right| \\ &\leq \left\|\frac{1}{m}\Pi^\top\Pi\right\|(\|x\|_2\|x-y\|_2 + \|x-y\|_2\|y\|_2) \\ &\leq 2\left\|\frac{1}{m}\Pi^\top\Pi\right\|\cdot\frac{1}{4} = \frac{1}{2}\left\|\frac{1}{m}\Pi^\top\Pi\right\|.\end{aligned}$$

   Since $\left\|\frac{1}{m}\Pi^\top\Pi\right\| \geq \left|\frac{1}{m}\|\Pi y\|_2^2\right|$, it follows that:

   $$\left\|\frac{1}{m}\Pi^\top\Pi - I\right\| \leq 2\max_{y\in\mathcal{N}_{\frac{1}{4}}}\left|\frac{1}{m}\|\Pi y\|_2^2 - 1\right|.$$

2. **Concentration for Fixed x:** For a fixed $x \in S^{d-1}$, define $Z_i = \langle\Pi_i, x\rangle$. Then $\|\Pi x\|_2^2 = \sum_{i=1}^m Z_i^2$, and since $\frac{1}{m}\mathbb{E}[\Pi^\top\Pi] = I$, we have $\mathbb{E}[Z_i^2] = 1$. As Π_i are sub-Gaussian, Z_i is sub-Gaussian, and $Z_i^2 - 1$ is sub-exponential. By concentration inequalities:

   $$\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^m Z_i^2 - 1\right| > \frac{\varepsilon}{2}\right) \leq 2\exp\left(-\frac{c\varepsilon^2 m}{\sigma^2}\right),$$

   where σ^2 is the sub-Gaussian parameter of Π_i.

3. **Union Bound Over ε-Net:** Apply the union bound over the $\frac{1}{4}$-net:

   $$\mathbb{P}\left(\max_{y\in\mathcal{N}_{\frac{1}{4}}}\left|\frac{1}{m}\|\Pi y\|_2^2 - 1\right| > \frac{\varepsilon}{2}\right) \leq |\mathcal{N}_{\frac{1}{4}}|\cdot 2\exp\left(-\frac{c\varepsilon^2 m}{\sigma^2}\right).$$

   Since $|\mathcal{N}_{\frac{1}{4}}| \leq \left(\frac{2}{\frac{1}{4}}+1\right)^d = 9^d$, we have:

   $$\mathbb{P}\left(\left\|\frac{1}{m}\Pi^\top\Pi - I\right\| > \varepsilon\right) \leq 2\cdot 9^d\exp\left(-\frac{c\varepsilon^2 m}{\sigma^2}\right).$$

   Setting $\varepsilon = \delta$ and ensuring $m \geq Cd/\delta^2$ for some constant C, the probability becomes exponentially small, proving the bound.