COMP 7070 Advanced Topics in AI and ML February 27, 2024

Lecture 6: Johnson-Lindenstrauss & Matrix Sketching

Instructor: Yifan Chen *Scribes: Hongduan Tian, Yi Ding* *Proof reader: Zhanke Zhou*

6.1 Johnson-Lindenstrauss Lemma

In this section, we are going to learn about Johnson-Lindenstrauss Lemma, which has been highly impactful in the design of algorithms for high-dimensional delta.

Theorem 6.1 (JL lemma)**.** *For any $\varepsilon \in (0,1)$ and any $X \in \mathbb{R}^d$ for $|X| = n$ finite, there exists an embedding $f : X \to \mathbb{R}^m$ for $m = O(\varepsilon^{-2}\log n)$ such that*

$$\forall x, y \in X, (1-\varepsilon)||x-y||_2^2 \le ||f(x)-f(y)||_2^2 \le (1+\varepsilon)||x-y||_2^2. \tag{6.1}$$

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. In other words, a set of data points in a high-dimensional space can be embedded into a much lower dimension in a way that the distance information is nearly preserved.

With the desirable property of the Johnson-Lindenstrauss lemma in Theorem 6.1, the algorithms that contain heavy matrix computation can be improved:

- **Approximate Matrix Multiplication (AMM).** Consider two matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ and $\boldsymbol{B} \in \mathbb{R}^{n\times n}$, the complexity of the multiplication $\boldsymbol{AB}$ is $O(n^3)$. According to JL lemma, by embedding $\boldsymbol{A}$ and $\boldsymbol{B}$ to lower dimension with the transformation $f : \mathbb{R}^n \to \mathbb{R}^m$, where $m < n$, we can approximate $\boldsymbol{AB}$ with $f(\boldsymbol{A})f(\boldsymbol{B})$. Then, the complexity is reduced to $O(n^2m)$.
- **Graph Convolutional Network (GCN).** In graph convolutional network, given an adjacency matrix $\boldsymbol{A} \in \mathbb{R}^{N\times N}$, the hidden state of the last layer $\boldsymbol{H}^{t-1} \in \mathbb{R}^{N\times dim}$ and a set of weights $\boldsymbol{W} \in \mathbb{R}^{dim\times dim}$, the output of current layer can be calculated as: $\boldsymbol{AH}^{t-1}\boldsymbol{W}$. In this case, matrices with lower dimensions can also be applied for efficient computation: $\boldsymbol{ASS}^\top\boldsymbol{H}^{t-1}\boldsymbol{W}$, where $\boldsymbol{S} \in \mathbb{R}^{d\times d'}$, $d' < d$.
- **Attention Calculation.** Attention mechanism is also computationally dense. Given query matrix $\boldsymbol{Q}$, key matrix $\boldsymbol{K}$ and value matrix $\boldsymbol{V}$, the attention features $\boldsymbol{H} \in \mathbb{R}^{d\times d}$ can be calculated in the way:
$$\boldsymbol{H} = \text{softmax}\left(\frac{\boldsymbol{Q}\cdot\boldsymbol{K}^\top}{\sqrt{d}}\right)\cdot\boldsymbol{V}.$$
By introducing $\boldsymbol{S} \in \mathbb{R}^{d\times d'}$, $d' < d$, $\boldsymbol{H}$ can be further calculated as:
$$\boldsymbol{H} = \text{softmax}\left(\frac{\boldsymbol{Q}\cdot\boldsymbol{K}^\top}{\sqrt{d}}\right)\boldsymbol{SS}^\top\boldsymbol{V}.$$

6.1.1 Distributed Johnson-Lindenstrauss Lemma (DJL).

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

$$\mathbb{P}_{\Pi\sim\mathcal{D}_{\varepsilon,\delta}}(|||\Pi_z||_2^2 - 1| > \varepsilon) < \delta.$$

Based on Lemma. 6.2, assume that $\delta \leq \frac{1}{n^2}$, $Z = \frac{X-Y}{||X-Y||}$, with other mild assumptions, we have:

$$\begin{aligned}\mathbb{E}_{X,Y} &= \left\{\left|||\Pi Z||^2 - 1\right| > \varepsilon\right\} \\ &= \left\{\left|||\Pi(X-Y)||^2 - ||X-Y||^2\right| > \varepsilon||X-Y||^2\right\}.\end{aligned}$$

Since $|X| = |Y| = n$, we then have $\frac{n(n-1)}{2}$ pairs. Then, we have:

$$\mathbb{P}\left(\bigcup_{X,Y} \mathbb{E}_{X,Y}\right) \leq \sum_{X,Y} \mathbb{P}\left(\mathbb{E}_{X,Y}\right) = \frac{n(n-1)}{2}\delta.$$

Thus, with the probability of least $1-c$, the JL lemma would hold.

6.2 Sketching Method

6.2.1 Sketch

Let replace a vector/matrix x/X by its sketch $\Pi x / X\Pi^\top$, then we have:

$$||\Pi z||^2 \to 1 \iff z^\top \Pi^\top \Pi z \to 1.$$

When $z^\top z = 1$ holds, we have to ensure that $\mathbb{E}\Pi^\top\Pi = I$, where I denotes the identity matrix. Here, we introduce two applications:

- **Coordinate.** Consider a special case where $\Pi \in \mathbb{R}^{1\times m}$, then it will be uniformly distributed in the set $\{\sqrt{d}\boldsymbol{e}_i\}_{i=1}^d$. Then, we have:

$$\mathbb{E}\Pi^\top\Pi = \frac{1}{d}\sum_{i=1}^{d} d\boldsymbol{e}_i\boldsymbol{e}_i^\top = I.$$

- **AMM.** The other application is approximate matrix multiplication, such as

$$B^\top\Pi^\top\Pi C \approx B^\top C_{d\times n} \quad if \quad m << d.$$

6.3 Matrix Concentration Sketching

6.3.1 Sub-Gaussian Random Variable

Definition 6.3 (Moment Generating Function, MGF)**.** *Given a random variable $X \sim \mathsf{subG}(\sigma^2)$, where $\mathbb{E}(X) = 0$, for $\forall\lambda \in \mathbb{R}$, the moment generating function satisfies:*

$$\mathbb{E}\exp\left(\lambda(X-\mu)\right) \leq \exp(\frac{\sigma^2\lambda^2}{2}).$$

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

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

In particular,

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

Proof. We first calculate $\mathbb{E}[|X|^p]$:

$$\begin{aligned}
\mathbb{E}[|X|^p] &= \int_0^\infty \mathbb{P}(|X|^p > t)dt \\
&= \int_0^\infty \mathbb{P}(|X| > t^{1/p})dt \\
&\leq 2\int_0^\infty e^{-\frac{t^{2/p}}{2\sigma^2}}dt \\
&= (2\sigma^2)^{2/p}p\int_0^\infty e^{-u}u^{p/2-1}du, \qquad u = -\frac{t^{2/p}}{2\sigma^2} \\
&= (2\sigma^2)^{p/2}p\Gamma(p/2), \qquad \Gamma(n) = \int_0^\infty e^{-u}u^{n-1}du = (n-1)!.
\end{aligned}$$

With the conditions: $\Gamma(p/2) \leq (p/2)^{p/2}$ and $p^{1/p} \leq e^{1/e}$ for any $p \geq 2$, we further have:

$$\begin{aligned}
((2\sigma^2)^{p/2}p\Gamma(p/2))^{1/p} &= ((2\sigma^2)^{p/2})^{1/p}p^{1/p}(\Gamma(p/2))^{1/p} \\
&= \sqrt{2}\sigma \cdot p^{1/p}(\Gamma(p/2))^{1/p} \\
&\leq \sqrt{2}\sqrt{p/2} \\
&= \sigma e^{1/e}\sqrt{p}.
\end{aligned}$$

□

Where K_2 is the sub-Gaussian norm and $K_2 \sim \sigma$. For a large K_2, X is not very sub-Gaussian.

$$||X||_{\Psi_2} = K_2.$$

6.3.2 Sub-Gaussian Random Vector X.

Definition 6.5. *For $\forall x \in \mathbb{R}^d$, $< X, x >$ is sub-Gaussian. Then, we have*

$$||X||_{\Psi_2} := \sup_{x\in\mathcal{S}^{d-1}} || < X, x > ||_{\Psi_2}.$$

For coordinate $< X, x >= \sqrt{d}\sum_{i=1}^d x_i\mathbb{I}_{\{\omega=i\}}$, we then have $||X||_{\Psi_2} \sim \sqrt{d}$.

6.4

Given a matrix Π with independent sub-Gaussian rows, where $\frac{1}{m}\mathbb{E}\Pi^\top\Pi = I$, we then have:

$$\sqrt{m} - C\sqrt{d} - t \leq \mathcal{S}_{\min}(\Pi) \leq \mathbb{S}_{\max}(\Pi) \leq \sqrt{m} + C\sqrt{d} + t,$$

with the probability at least $1 - 2\exp(-ct^2)$.
1 ε-net to approximate $\mathcal{S}^{d-1}$

Definition 6.6 (ε-net)**.** *Given $\mathcal{N} \subseteq \mathcal{S}^{d-1}$ and $\forall x \in \mathcal{S}^{d-1}$, we can have $d(x, y) \leq \varepsilon$, where $y \in \mathcal{N}^\varepsilon$ and $d(\cdot, \cdot)$ is a distance measure.*

A special case of ε-net is *covering number* $\mathcal{N}(\mathcal{S}^{d-1}, \varepsilon)$, which is the minimal cardinality of $\mathcal{N}_\varepsilon$.

Bound of $\mathcal{N}(\mathcal{S}^{d-1}, \varepsilon)$ Here, we explore the bound of $\mathcal{N}(\mathcal{S}^{d-1}, \varepsilon)$.

- Consider the maximal ε-separated subset of $\mathcal{S}^{d-1}$, then for $\forall y_1, y_2, d(y_1, y_2) > \varepsilon$.
- Such a subset aforementioned is a ε-net, o.w. $\exists x \in \mathcal{S}^{d-1}$, $d(x, y_i) > \varepsilon$. Then, the point can be added to the subset.

- Based on the content above, we can formulate the relationship as follows:

$$\sum \frac{\varepsilon}{2}\mathcal{B} \leq (1+\frac{\varepsilon}{2})\mathcal{B},$$

where $\mathcal{B}$ denotes the volume of a unit ℓ_2 ball of $\mathbb{R}^d$. Then, we have:

$$|\mathcal{N}_\varepsilon|(\frac{\varepsilon}{2})^d \leq (1+\frac{\varepsilon}{2})^d$$

Prove $1-\delta \leq \mathcal{S}_{\min}(\Pi) \leq \mathcal{S}_{\max}(\Pi) \leq 1+\delta$. The proof can be equivalently transformed to prove $||\Pi^\top\Pi - I|| \leq \max(\delta, \delta^2)$.

Proof.

$$|||\Pi x|| - 1| \equiv |Z-1| \leq \max\left\{|Z-1|, |Z-1|^2\right\}.$$

When $Z \in [0,2]$, then

$$\max\left\{|Z-1|, |Z-1|^2\right\} = |Z-1| \leq |Z^2-1|.$$

If $\delta > 1$, then $|Z-1| \leq 1 < \delta$; otherwise,

$$\begin{aligned}
|Z-1| &\leq |Z^2-1| \\
&= \left|x^\top\Pi^\top\Pi x - x^\top x\right| \\
&= \left|x^\top(\Pi^\top\Pi - I)x\right| \\
&\leq ||\Pi^\top\Pi - I|| \qquad \leq \max(\delta,\delta^2) \\
&= \delta.
\end{aligned}$$

When $Z > 2$, then $|Z-1|^2 \leq |Z^2-1| \leq \max(\delta,\delta^2) = \delta$.
Thus, we can say that $1-\delta \leq \mathcal{S}_{\min}(\Pi) \leq \mathcal{S}_{\max}(\Pi) \leq 1+\delta$ holds for $\forall x, |||\Pi x|| - 1| \leq \delta$. □

The bound can be given up to a constant factor with a $\frac{1}{4}$-net. Consider $\exists x_1 \in \mathcal{S}^{d-1}, ||A|| = x_1^\top A x_1$, $\exists y \in \mathcal{N}_{\frac{1}{4}}, d(x_1, y) \leq \frac{1}{4}$, we then have

$$\begin{aligned}
|< Ax_1, x_1 > - < Ay, y >| &= |< Ax_1, x_1 - y > + < A(x_1 - y), y >| \\
&\leq ||A|| \cdot ||x_1|| \cdot ||x_1 - y||| + ||A|| \cdot ||x_1 - y|| \cdot ||y|| \\
&= 2||A|| \cdot ||x_1 - y|| \\
&= 2||A|| \cdot d(x_1, y) \\
&\leq \frac{1}{2}||A||.
\end{aligned}$$

$\Rightarrow < Ay, y > \leq < Ax_1, x_1 > - \frac{1}{2}||A|| = \frac{1}{2}||A||$.
$\Rightarrow ||A|| \leq 2 \cdot y^\top Ay \leq 2 \cdot \max_{y \in \mathcal{N}_{\frac{1}{4}}} y^\top Ay$.
The above conclusion is also equivalent to:

$$2 \cdot \max_{x \in \mathcal{N}_{\frac{1}{4}}} \left|\frac{1}{m}||\Pi x||^2 - 1\right| \leq \varepsilon.$$

② Concentration: $||\Pi x||^2 = \sum_{i=1} m < \Pi_i, x >^2$, where $Z_1 =< \Pi_i, x >$ and $\mathbb{E}Z_i^2 = 1$.
③ Union bound:

$$\mathbb{P}\left(\max_{x \in \mathcal{N}_{\frac{1}{4}}} \left|\frac{1}{m}\sum_{i=1}^{m} Z_i^2(x) - 1\right| > \frac{\varepsilon}{2}\right) \leq \left|\mathcal{N}_{\frac{1}{4}}\right| \cdot 2 \cdot \exp\left(-\frac{\varepsilon^2}{128\sigma^2}\right).$$