Lecture 6

Transcribed from notes dated 9/11/2021

Johnson-Lindenstrauss Lemma and Dimensionality Reduction

A fundamental yet simple result from convex geometry that has found many applications in data analysis and algorithms.

Let $\overline{v}_1, \overline{v}_2, \ldots, \overline{v}_n$ be n vectors (points) in $\mathbb{R}^d$ where d is very large. The lemma says that one can project these vectors to a lower dimensional space $\mathbb{R}^k$ to create new vectors $\overline{u}_1, \overline{u}_2, \ldots, \overline{u}_n$ such that for all $i, j \in [n]$:

$$(1-\epsilon)\|\overline{v}_i - \overline{v}_j\|_2 \leq \|\overline{u}_i - \overline{u}_j\|_2 \leq (1+\epsilon)\|\overline{v}_i - \overline{v}_j\|_2$$

i.e. the pairwise Euclidean distance is approximately preserved. And:

$$k = O\left(\frac{\log n}{\epsilon^2}\right)$$

Note that d does not show up.

Clearly this is useful only if d was originally larger than k.

Further the projection is via a linear map $\mathbb{R}^{k \times d}$ that is randomized and oblivious to the data.

The core of the result is about a single vector and we then use a union bound.

Distributional JL Lemma

Fix a vector $\overline{x} \in \mathbb{R}^d$. Let Π be a $k \times d$ matrix where $\Pi_{i,j}$ is chosen independently from a standard Gaussian distribution $N(0,1)$.

If $k = \Omega\left(\frac{1}{\epsilon^2}\log\frac{1}{\delta}\right)$ then with probability $\geq 1-\delta$:

$$(1-\epsilon)\|\overline{x}\|_2 \leq \left\|\frac{1}{\sqrt{k}}\Pi\overline{x}\right\|_2 \leq (1+\epsilon)\|\overline{x}\|_2$$

Assuming the DJL the dimensionality reduction is easy.

Choose $k = \Omega(\frac{1}{\epsilon^2}\log n)$. Then for $i \neq j$ look at vector $\overline{x} = \overline{v}_i - \overline{v}_j$. With probability $1 - \frac{1}{n^3}$, $\left\|\frac{1}{\sqrt{k}}(\Pi\overline{v}_i - \Pi\overline{v}_j)\right\|_2$ is within $(1 \pm \epsilon)\|\overline{v}_i - \overline{v}_j\|_2$. Let $\overline{u}_i = \frac{1}{\sqrt{k}}\Pi\overline{v}_i$.

So by union bound all pairs are preserved with probability $1 - \binom{n}{2}\frac{1}{n^3} \geq 1 - \frac{1}{n}$.

The projection is oblivious since Π was chosen without considering the data $\overline{v}_1, \ldots, \overline{v}_n$. This implies it can be used for "future" data.

Proof of the DJL Lemma

The proof relies on some nice properties of the Gaussian distribution. One can use other distributions, and there are some advantages to do so, but the analysis is more involved.

Sum of Independent Normally Distributed Random Variables

Recall $N(\mu, \sigma^2)$ is the normal distribution with mean μ and variance σ^2. Its probability density function is:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Lemma 1. *If $Z_1 \sim N(\mu_1, \sigma_1^2)$ and $Z_2 \sim N(\mu_2, \sigma_2^2)$ and Z_1, Z_2 are independent, then $Z_1 + Z_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.*

Corollary 1. *Suppose $\vec{z} = (z_1, z_2, \ldots, z_d)$ is a vector of independent $N(0,1)$ random variables. Let $\overline{x} \in \mathbb{R}^d$ be a unit vector, i.e., $\|\overline{x}\|_2 = 1$. Then $\sum_{i=1}^d x_i z_i$ is distributed as $N(0,1)$.*

Proof. The sum is a sum of independent Normal variables, as each $x_i z_i \sim N(0, x_i^2)$. The mean is $\sum_{i=1}^d E[x_i z_i] = \sum_{i=1}^d x_i E[z_i] = 0$. The variance is $\sum_{i=1}^d \mathrm{Var}(x_i z_i) = \sum_{i=1}^d x_i^2 \mathrm{Var}(z_i) = \sum_{i=1}^d x_i^2 = 1$, since $\overline{x}$ is a unit vector. □

Proof of Lemma (via Moment Generating Functions)

We will consider the moment generating function (MGF) of $X \sim N(\mu, \sigma^2)$, which is $E[e^{tX}]$.

$$E[e^{tX}] = \int_{-\infty}^{\infty} e^{tx} \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} dx$$

This evaluates to:

$$E[e^{tX}] = e^{\mu t + \frac{\sigma^2 t^2}{2}}$$

Suppose we have k independent random variables $X_1, X_2, \ldots, X_k$. If $M_1(t), \ldots, M_k(t)$ are their respective MGFs, then the MGF of their sum $X = \sum X_i$ is:

$$M(t) = M_1(t) M_2(t) \ldots M_k(t)$$

This is because $E[e^{tX}] = E[e^{t\sum X_i}] = E\left[\prod_{i=1}^k e^{tX_i}\right] = \prod_{i=1}^k E[e^{tX_i}]$ by independence. If $X_i \sim N(\mu_i, \sigma_i^2)$, then the MGF for the sum is:

$$M(t) = \prod_{i=1}^k e^{\mu_i t + \frac{\sigma_i^2 t^2}{2}} = e^{(\sum_i \mu_i)t + \frac{(\sum_i \sigma_i^2) t^2}{2}}$$

This is the MGF of $N(\sum \mu_i, \sum \sigma_i^2)$.

Main Proof of DJL

Let Π be a $k \times d$ Gaussian matrix and let $\overline{x}$ be a unit vector (wlog). Let $\overline{y} = \Pi\overline{x} = (Y_1, Y_2, \ldots, Y_k)$. From the corollary, each $Y_i = \sum_{j=1}^d \Pi_{ij} x_j$ is distributed as $Y_i \sim N(0,1)$, and the Y_i are independent. We are interested in the squared norm $\|\overline{y}\|_2^2 = \sum_{i=1}^k Y_i^2$. The expectation is $E[\|\overline{y}\|_2^2] = \sum_{i=1}^k E[Y_i^2] = k$, since $E[Y_i^2] = \mathrm{Var}(Y_i) + (E[Y_i])^2 = 1 + 0 = 1$. This suggests that $\frac{1}{\sqrt{k}}\|\Pi\overline{x}\|_2$ should be close to 1.

Our goal is to prove that $Y = \sum_{i=1}^k Y_i^2$ is concentrated around its mean. The distribution of the sum of squares of t independent $N(0,1)$ random variables is called the χ^2 distribution with parameter t, denoted $\chi^2(t)$.

Lemma 2 (Concentration for χ^2)**.** *Let $Y_1, \ldots, Y_k$ be independent $N(0,1)$ random variables and let $Y = \sum_{i=1}^k Y_i^2$. Then for $\epsilon \in (0, 1/2)$:*

$$Pr[(1-\epsilon)^2 k \leq Y \leq (1+\epsilon)^2 k] \geq 1 - 2e^{-c\epsilon^2 k}$$

where c is an absolute constant.

This concentration lemma implies the DJL lemma. By choosing $k = O(\frac{1}{\epsilon^2} \ln \frac{1}{\delta})$, we get $Pr[(1-\epsilon) \leq \left\| \frac{1}{\sqrt{k}} \Pi\overline{x} \right\|_2 \leq (1+\epsilon)] \geq 1 - \delta$.

Proof of the Concentration Lemma

We use the exponential moment method (Chernoff bound). For $t \geq 0$:

$$Pr[Y > \alpha] = Pr[e^{tY} > e^{t\alpha}] \leq \frac{E[e^{tY}]}{e^{t\alpha}}$$

The MGF for $Y \sim \chi^2(k)$ has a closed form for $t < 1/2$:

$$E[e^{tY}] = (1-2t)^{-k/2}$$

Plugging this in, we get $Pr[Y > \alpha] \leq (1-2t)^{-k/2}e^{-t\alpha}$. Optimizing for t and setting $\alpha = (1+\epsilon)^2 k$ gives a bound of the form $e^{-c'\epsilon^2 k}$. The lower tail is similar.

To derive the MGF for Y_i^2 where $Y_i \sim N(0,1)$:

$$M_{Y_i^2}(t) = E[e^{tY_i^2}] = \int_{-\infty}^{\infty} e^{ty^2} \frac{e^{-y^2/2}}{\sqrt{2\pi}} dy = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}(1-2t)} dy$$

This integral evaluates to $(1-2t)^{-1/2}$. Since the Y_i^2 are independent, the MGF of their sum $Y = \sum Y_i^2$ is the product of their individual MGFs:

$$MGF(Y) = \prod_{i=1}^{k} (1-2t)^{-1/2} = (1-2t)^{-k/2}$$

New Topic: Hashing

This topic requires some background in pseudo-randomness and derandomization. A key question is how to convert randomized algorithms into deterministic ones. Our goal today is to introduce the notion of limited independence.

Definition: A set of random variables $X_1, \ldots, X_n$ are **pairwise independent** (or 2-wise independent) if X_i and X_j are independent for any $i \neq j$.

Example: Let $X_1, X_2 \in \{0,1\}$ be independent random bits. Let $X_3 = X_1 \oplus X_2$ (XOR). Then X_1, X_2, X_3 are pairwise independent but not fully (mutually) independent.

Lemma 3. *One can construct $n = 2^k - 1$ pairwise independent random variables $X_1, \ldots, X_n$ (where each $X_i \in \{0,1\}$) from just k truly random bits $Y_1, \ldots, Y_k$.*

Construction: Let S be any non-empty subset of $\{1, 2, \ldots, k\}$. Define $X_S = \bigoplus_{i \in S} Y_i$. For any two distinct non-empty subsets S and T, the variables X_S and X_T are independent.

Application: Derandomizing the Max-Cut algorithm

Recall the simple randomized algorithm for Max-Cut on a graph $G = (V, E)$: for each vertex $v \in V$, assign it to a set S with probability $1/2$. The expected number of edges in the cut is $|E|/2$. The analysis for any given edge $e = (u, v)$ only requires that the random choices for u and v are independent. Therefore, the linearity of expectation argument holds even if the random variables for all vertices are only pairwise independent.

We can derandomize this as follows:

- We need $n = |V|$ pairwise independent random bits. We can generate these from a "seed" of $k = \lceil \log_2(n+1) \rceil$ true random bits using the construction above.
- The number of possible seeds is 2^k, which is $O(n)$.
- A deterministic algorithm can iterate through all $O(n)$ possible seeds. For each seed, it generates the n bits, defines the corresponding cut, and computes its size.
- The algorithm then outputs the largest cut found. Since the expected size is $|E|/2$, at least one of these deterministic outcomes must produce a cut of size at least $|E|/2$. This yields a deterministic polynomial-time approximation algorithm.