njb2154
Noah J.X. Bergam July 24, 2023

Probably Approximate Johnson-Lindenstrauss Lemma

The Johnson-Lindenstrauss Lemma asserts that for any point cloud in $\mathbb{R}^d$, there exists a relatively low-dimensional subspace of $\mathbb{R}^d$ which preserves pairwise distances up to a given tolerance level $\epsilon \in (0,1)$. In this writeup, we would like to reframe the result with an independent confidence parameter $\delta \in (0,1)$, which characterizes the likelihood with which a randomly sampled projection leads to such a subspace. In other words, we will prove a probably approximately correct formalation of Johnson-Lindenstrauss.

Formulation

Given some $S = \{x_1, ..., x_n\} \subset \mathbb{R}^d$, we call $f : \mathbb{R}^d \mapsto \mathbb{R}^k$ an ϵ-**JL map** of S if we have the following guarantee for all distinct $u, v \in S$:

$$\frac{||f(u) - f(v)||^2}{||u - v||^2} \in [1 - \epsilon, 1 + \epsilon] \tag{1}$$

The usual Johnson-Lindenstrauss Lemma then states:

Theorem 1. *For any $S = \{x_1, ..., x_n\} \subset \mathbb{R}^d$ and any tolerance level $\epsilon \in (0,1)$, an ϵ-JL map of S, denoted $f : \mathbb{R}^d \mapsto \mathbb{R}^k$, exists if $k \geq \frac{4\ln(n)}{\epsilon^2/2 - \epsilon^3/3}$, i.e. $k = O(\ln(n)/\epsilon^2)$.*

Most proofs of the above statement uncover an efficient probabilistic algorithm for finding the map f. In (Dasgupta and Gupta 2002), this algorithm involves a random projection whose probability of success is at least $1 - \frac{1}{n}$. In this writeup, we explore a slightly different approach: we would like the probability of success to depend on an independently chosen variable δ. This motivates the following result:

Theorem 2. *If A is a random $k \times d$ matrix whose entries are random normal variables $A_{ij} \sim \mathcal{N}(0,1)$ and $k \geq \frac{8\ln(2/\delta)}{\epsilon^2}$, then for any $x \in \mathbb{R}^d$, we have, with probability $1 - \delta$:*

$$\frac{||Ax||^2}{k||x||^2} \in [1 - \epsilon, 1 + \epsilon]$$

With this property, we can specify a probabilistic algorithm for generating a JL map.

Theorem 3. *For any $\epsilon, \delta \in (0,1)$, and for any $S = \{x_1, ..., x_n\} \subset \mathbb{R}^d$, take $A \in M_{k\times n}(\mathbb{R})$, $A_{ij} \sim \mathcal{N}(0,1)$ and $k \geq \frac{8\ln(2\binom{n}{2}/\delta)}{\epsilon^2}$. Then, with probability $(1 - \delta)$, $f(x) = \frac{1}{\sqrt{k}}Ax$ is an ϵ-JL map of S from $\mathbb{R}^d \mapsto \mathbb{R}^k$.*

Proofs

For the proof of Theorem 1, we refer to the reader to (Dasgupta and Gupta 2002)

Theorem 2

$||Ax||^2$ is a random variable. We want to show that, for a big enough k, it approximates $||x||^2$ with precision $(1 \pm \epsilon)$. Like most proofs of approximate algorithms, we show that the mean of the distribution of $||Ax||^2$ is in the right place, and then we show that it is concentrated enough.

First we observe that $\mathbb{E}[||Ax||^2] = k||x||^2$ for any positive integer d. The expectation is over the randomness in the choice of the matrix A.

$$||Ax||^2 = \sum_{i=1}^{k} (Ax)_i^2 = \sum_{i=1}^{k} \Big[\sum_{j=1}^{d} A_{ij} x_j \Big]^2 = \sum_{i=1}^{k} \Big[\sum_{j=1}^{d} A_{ij}^2 x_j^2 + \sum_{k=1, k \neq j}^{d} A_{ij} x_j A_{ik} x_k \Big]$$

Apply the expectation to the whole expression, and apply linearity of expectation (and the fact that, for independent random variables X and Y, we have $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$).

$$\mathbb{E}[||Ax||^2] = \sum_{i=1}^{k} \Big[\sum_{j=1}^{d} \mathbb{E}[A_{ij}^2] x_j^2 + \sum_{k=1, k \neq j}^{d} \mathbb{E}[A_{ij}] x_j \mathbb{E}[A_{ik}] x_k \Big]$$

Recall that the expectation of $X \sim \mathcal{N}(0, 1)$ is clearly 0, and $\mathbb{E}[X^2] = \text{Var}[X] + \mathbb{E}^2[X] = 1$.

$$= \sum_{i=1}^{k} \Big[\sum_{j=1}^{d} x_j^2 \Big] = k \Big[\sum_{j=1}^{d} x_j^2 \Big] = k||x||^2$$

We furthermore observe that $||Ax||^2$ follows a chi-squared distribution. This follows from the following observations.

- Gaussian random variables are stable, i.e. given two independent Gaussian random variables centered at the same point, their sum is Gaussian and the variances add.
- This means $(Ax)_i = \sum_{j=1}^{d} A_{ij} x_j$ is Gaussian, because each $A_{ij} x_j$ is Gaussian with mean zero and variance x_j^2. So the total variance of $(Ax)_i$ is $||x||^2 = \sum_j x_j^2$.
- Thus, $\frac{||Ax||^2}{||x||^2}$ follows a chi-squared distribution with k degrees of freedom, since each $(Ax)_i/||x|| \sim \mathcal{N}(0, 1)$.
$$\frac{||Ax||^2}{k||x||^2} = \frac{1}{k} \sum_{i=1}^{k} \Big[\frac{(Ax)_i}{||x||} \Big]^2$$

Lemma: A chi-squared distribution with k-degrees of freedom $X \sim \chi^2(k)$ obeys the following tail bound for $\epsilon \in (0,1)$.

$$P[|X/k - 1| < \epsilon] \leq 2e^{-k\epsilon^2/8}$$

Thus, if we take $k \geq \frac{8\ln(2/\delta)}{\epsilon^2}$, we win.

$$\begin{aligned} P[|||Ax||^2/k||x||^2 - 1| > \epsilon] &\leq 2\exp(-\epsilon^2(8\ln(2/\delta)/\epsilon^2)/8) \\ &= 2\exp(-\ln(2/\delta)) \\ &= 2\exp(\ln(\delta/2)) = \delta \quad \square \end{aligned}$$

Thus, the probability of failure is less than δ, so the probability of success is at least $1-\delta$. This concludes the proof. $\square$

Proof of Lemma: We apply the Chernoff-Hoeffding bounding method, to $X = \sum_k X_i^2$ with $X_i \sim \mathcal{N}(0,1)$ i.i.d. Let $Y_i = X_i^2 - 1$.

$$\begin{aligned} P(X/k - 1 > \epsilon) &= P(\sum_{i=1}^{k}(X_i^2 - 1) > k\epsilon) \\ &= P(\sum_{i=1}^{k}(X_i^2 - 1) > k\epsilon) \\ &= P(\sum_{i=1}^{k} Y_i > k\epsilon) \end{aligned}$$

Let $t > 0$. Apply Markov's inequality with the Chernoff twist. Then apply the i.i.d property of Y_i.

$$\begin{aligned} &\leq \frac{\mathbb{E}(\exp(t\sum_{i=1}^{k} Y_i))}{\exp(tk\epsilon)} \\ &= \prod_{i=1}^{k} \frac{\mathbb{E}(\exp(tY_i))}{\exp(tk\epsilon)} \end{aligned}$$

The game here is to find a bound on $\mathbb{E}[\exp(t(X_i^2 - 1))]$. At first glance, this looks like a difficult expectation to get a handle on. Thus, we apply the "Law of the Unconscious Statistician" (a tongue-in-cheek call-out to the fact that this is often treated as the definition of expectation, when it is in fact a nontrivial theorem) which allows to calculate the expectation of a random variable Y, transformed by some function g (it is helpful when we know the distribution of Y but not the distribution of $g(Y)$). We state the law in general below (f_Y is the pdf of Y).

$$\mathbb{E}[g(Y)] = \int_{-\infty}^{\infty} g(z)f_Y(z)dz$$

In this case, let $g(z) = \exp(t(z^2 - 1))$ and f_Y be the standard normal. Using a standard change-of-variables, we can evaluate the integral and find a nice bound.

$$\mathbb{E}[e^{tY_i}] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{t(z^2-1)} e^{-z^2/2} dz = \frac{e^{-t}}{\sqrt{1-2t}} \leq e^{2t^2} \text{ for } 0 \leq t \leq 1/4$$

We will now apply the above information to the main inequality. Take $t \leq \epsilon/4 \leq 1/4$, so the above applies. In the second step below, we apply monotonicity of the exponential with the understanding that $tk^2\epsilon > 0$.

$$P(X/k - 1 > \epsilon) \leq \exp(2t^2k - tk^2\epsilon) \leq \exp(2t^2k) \leq \exp(k\epsilon^2/8)$$

We can apply roughly the same method to the other tail and we find the same bound. The only difference in the proof would be that $Y_i = 1 - X_i^2$.

Thus, we have the following tail bound:

$$P(|X/k - 1| > \epsilon) \leq 2\exp(k\epsilon^2/8) \quad \square$$

Theorem 2

Fix $\epsilon > 0$, $\delta > 0$, and $S = \{x_1, ..., x_n\} \subset \mathbb{R}^d$ as given.

Pick $\tilde{\delta} < \frac{\delta}{\binom{n}{2}}$, and $k \geq \frac{8 \ln(2/\tilde{\delta})}{\epsilon^2}$.

Let $f(x) = Ax/\sqrt{k}$, where A is as specified in Theorem 2. Since f is linear, then for any pair of points x and y, we observe the following with probability $1 - \delta$.

$$\frac{||f(x-y)||^2}{||x-y||^2} = \frac{||f(x) - f(y)||^2}{||x-y||^2} \in [1-\epsilon, 1+\epsilon]$$

For each pair of points, the probability of failure is $\tilde{\delta}$. There are $\binom{n}{2}$ points. Let $L(u, v)$ denote the event in which the Johnson-Lindenstrauss property fails for distinct points $u, v \in S$. Apply the union bound.

$$P\Big(\bigcup_{u,v \in S, u \neq v} L(u,v)\Big) \leq \sum_{u,v \in S, u \neq v} L(u,v) = \binom{n}{2}\tilde{\delta} < \delta$$

Since the probability of failure for f is less than δ, the probability of success (i.e. the probability that f is an ϵ-JL map for the set of points S) is at least $1 - \delta$. This completes the proof. $\square$

Works Cited

https://www.cs.princeton.edu/ smattw/Teaching/Fa19Lectures/lec9/lec9.pdf

https://cseweb.ucsd.edu/ dasgupta/papers/jl.pdf

https://www.stat.berkeley.edu/ mjwain/stat210b/Chap2$_TailBounds_Jan22_2015.pdf$