Matousek's Proof of the Johnson-Lindenstrauss Lemma

The reference for this note is "On Variants of the Johnson-Lindenstrauss Lemma" by J. Matousek.

Theorem 0.1. (Johnson-Lindenstrauss Lemma.) *Given any n points $(v_1, \ldots, v_n)$ in $\mathbf{R}^d$ and any $\varepsilon \in (0, 1/2)$, there exists a mapping $\Phi : \mathbf{R}^d \to \mathbf{R}^k$ where $k \leq \frac{200 \log n}{\varepsilon^2}$ such that*

$$\forall i, j \quad (1-\varepsilon)||v_i - v_j||_2 \leq ||\Phi(v_i) - \Phi(v_j)||_2 \leq (1+\varepsilon)||v_i - v_j||_2$$

The mapping is indeed a "random linear transformation", that is, $\Phi(x) = Ax$ where each entry of A is a suitable random variable. Before going into this, let us review what are called subgaussian random variables, and the proof will then follow in less than half a page.

0.1 Subgaussian Random Variables

Let's start by calculating the moment generating function of a gaussian $Z \sim N(0, \sigma^2)$.

$$\begin{aligned} \mathbf{E}[e^{tZ}] &= \frac{1}{\sigma\sqrt{2\pi}} \int_{\infty}^{\infty} e^{tz} \cdot e^{-z^2/2} dz \\ &= e^{\frac{t^2\sigma^2}{2}} \end{aligned}$$

Motivated by this, here's the standard definition of subgaussian random variables.

Definition 1. *A random variable Z is said to be σ-subgaussian, if for all $t \in \mathbf{R}$*

$$\mathbf{E}[e^{tZ}] \leq e^{\frac{t^2\sigma^2}{2}} \tag{1}$$

A random variable Z is said to be σ-subgaussian up to t_0 *if* (1) *holds for all $|t| \leq t_0$.*

Lemma 1. (Subgaussian Random Variables and Concentration.) *If Z is σ-subgaussian up to t_0, then $\Pr[|Z| > u] \leq 2e^{-\frac{u^2}{2\sigma^2}}$ for $0 < u < \sigma^2 t_0$.*

Proof. This is the usual Chernoff step:

$$\begin{aligned} \Pr[Z > u] &= \Pr[e^{tZ} > e^{tu}] && \text{for positive } t \\ &\leq e^{-tu}\mathbf{E}[e^{tZ}] && \text{as long as } t \leq t_0 \\ &\leq e^{-tu} e^{\frac{t^2\sigma^2}{2}} \end{aligned}$$

Setting $t := u/\sigma^2 \leq t_0$ if $u < \sigma^2 t_0$, we get the upper tail, and the lower tail is similar. □

Examples of Subgaussian RVs. The Gaussian $Z \sim N(0, \sigma^2)$ is by definition subgaussian. Bounded random variables are subgaussian. More generally, if $|Z| \leq b$, then Z is b-subgaussian.

Moments of Subgaussian RVs.

- **Higher Even Moments.** Let Z be σ-subgaussian.

$$
\begin{aligned}
\mathbf{E}[Z^{2k}] = \mathbf{E}[|Z|^{2k}] &= \int_0^\infty (2k)t^{2k-1}\Pr[|Z| \geq t]dt \\
&\leq 4k \int_0^\infty t^{2k-1} e^{-t^2/2\sigma^2} dt \\
&= 2k\left(2\sigma^2\right)^k \int_0^\infty x^{k-1}e^{-x}dx \quad \text{change of variable: } x = t^2/2\sigma^2 \\
&= 2\left(2\sigma^2\right)^k k! \qquad (2)
\end{aligned}
$$

In the first equality we use the following formula: if Z is a non-negative rv, and h is a differentiable function, then

$$\mathbf{E}[h(Z)] = \int_0^\infty h(x)f(x)dx = h(0) + \int_0^\infty h'(x)\Pr[X \geq x]dx$$

One can see by integrating the second term in the RHS by parts.

Lemma 2. **(Linear combination of subgaussian random variables.)** *Let $X_1, \ldots, X_k$ be a collection of independent random variables, such that each X_i is σ_i-subgaussian up to t_0. Let $Y = \sum_{i=1}^n w_i X_i$. Then Y is a $\bar{\sigma}$-subgaussian random variable up to t_0/w_{max}, where $w_{max} = \max_i w_i$ and $\bar{\sigma} := \sqrt{\sum_{i=1}^k w_i^2 \sigma_i^2}$.*

Proof.

$$\mathbf{E}[e^{tY}] = \prod_{i=1}^n \mathbf{E}[e^{tw_iX_i}] \leq \prod_{i=1}^n e^{\frac{\sigma_i^2 t^2 w_i^2}{2}} = e^{\frac{t^2\sum_{i=1}^k w_i^2\sigma_i^2}{2}}$$

as long as $tw_i \leq t_0$, that is, $t \leq t_0/w_i$ for all i. ☐

Corollary 0.2. **(Average of Subgaussian RVs.)** *Let $X_1, \ldots, X_k$ be mutually independent, σ-subgaussian random variable up to t_0. Then $X := \frac{1}{k}\sum_{i=1}^k X_i$ is a $\frac{\sigma}{\sqrt{k}}$-subgaussian random variable up to kt_0.*

Lemma 3. **(Square of a subgaussian random variable.)** *Let X be a σ-subgaussian random variable. Then the centred random variable $Z := X^2 - \mathbf{E}[X^2]$ is a $\sqrt{32}\sigma^2$-subgaussian random variable up to $\frac{1}{4\sigma^2}$.*

Proof. This uses the higher even moments of X which we computed earlier.

$$
\begin{aligned}
\mathbf{E}[e^{tX^2}] &= \sum_{k=0}^\infty \frac{t^k \mathbf{E}[X^{2k}]}{k!} \\
&= 1 + t\mathbf{E}[X^2] + \sum_{k\geq 2} \frac{t^k\mathbf{E}[X^{2k}]}{k!} \\
&\leq 1 + t\mathbf{E}[X^2] + 2\sum_{k\geq 2}\frac{t^k}{k!}\cdot(2\sigma^2)^k k! \quad \text{see higher moments previously bounded} \\
&= 1 + t\mathbf{E}[X^2] + 2\sum_{k\geq 2}(2t\sigma^2)^k = 1 + t\mathbf{E}[X^2] + 8t^2\sigma^4\sum_{k\geq 0}\left(2t\sigma^2\right)^k \\
&\leq 1 + t\mathbf{E}[X^2] + 16t^2\sigma^4 \quad \text{if } t < t_0 = \frac{1}{4\sigma^2} \\
&\leq e^{-t\mathbf{E}[X^2] + 16t^2\sigma^4}
\end{aligned}
$$

Thus,

$$\mathbf{E}[e^{tZ}] = e^{-t\mathbf{E}[X^2]}\mathbf{E}[e^{tX^2}] \le e^{-\frac{32t^2\sigma^4}{2}} \quad \forall |t| \le \frac{1}{4\sigma^2}$$

implying Z is $\sqrt{32}\sigma^2$-subgaussian up tp $\frac{1}{4\sigma^2}$. $\square$

0.2 Proof of Johnson Lindenstrauss.

Now we have all the tools. Let R be a $k \times n$ random matrix where each entry R_{ij} is a σ-subgaussian random variable with variance $= 1 \le \sigma^2$. In particular, we can choose $R_{ij} \sim N(0,1)$ which would give $\sigma = 1$. Let $A := \frac{1}{\sqrt{k}} \cdot R$ be the scaled version of it, and define $\Phi(x) = Ax$. We show the following

$$\forall x, ||x|| = 1, \text{ we have with probability} \ge 1 - 1/n^3, \quad (1-\varepsilon)||x||_2 \le ||Ax||_2 \le (1+\varepsilon)||x||_2 \qquad (*)$$

By union bound, we ge that with probability $> 1 - 1/n$, the above holds for $x = (v_i - v_j)/||v_i - v_j||$ for all pairs, which implies the JLT.

Fix an unit vector x. For $i = 1, \ldots, k$, let $Y_i = \sum_{j=1}^n R_{ij}x_j$. Observe that $\mathbf{E}[Y_i^2] = \sum_{i=1}^n x_i^2 \mathbf{E}[R_{ij}^2] = 1$ where we used that the variance of each R_{ij} is exactly 1 and $||x||_2 = 1$. Note that $(Ax)_i := \frac{1}{\sqrt{k}}Y_i$, and so $||Ax||_2^2 = \frac{1}{k}\sum_{i=1}^k Y_i^2$.

We need to show that the following event occurs wp $\ge 1 - 2/n^3$.

$$\mathcal{E} := \{(1-2\varepsilon) \le \frac{1}{k}\sum_{i=1}^k Y_i^2 \le (1+2\varepsilon)\} \text{ or equivalently } \{\left|\frac{1}{k}\sum_{i=1}^k (Y_i^2 - 1)\right| \le 2\varepsilon\}$$

By Lemma 2, each Y_i is also a σ-subgaussian random variable because $||x||_2 = 1$. So, by Lemma 3, each $Z_i = Y_i^2 - \mathbf{E}[Y_i^2] = Y_i^2 - 1$ is a $\sqrt{32}\sigma^2$-subgaussian rv till $t_0 = \frac{1}{4\sigma^2}$. Therefore, by Corollary 0.2,

$$Z := \tfrac{1}{k}\textstyle\sum_{i=1}^k Z_i \text{ is a } \sqrt{32}\sigma^2/\sqrt{k}\text{-subgaussian random variable up to } \tfrac{k}{4\sigma^2}.$$

and therefore, by Lemma 1, we get

$$\Pr[|Z| > \varepsilon] \le 2e^{-\frac{k\varepsilon^2}{64\sigma^4}} \quad \forall \varepsilon \le \frac{32\sigma^4}{k} \cdot \frac{k}{4\sigma^2} = 8\sigma^2$$

Therefore, if $k \ge \frac{200\sigma^2 \cdot \log n}{\varepsilon^2}$, we get that $\Pr[|Z| > \varepsilon] < 2/n^3$, proving (*).