Randomized Algorithms 2021A – Lecture 5 (second part) Dimension Reduction in ℓ_2*

Robert Krauthgamer

1 The Johnson-Lindenstrauss (JL) Lemma

The Johnson-Lindenstrauss (JL) Lemma: Let $x_1, \ldots, x_n \in \mathbb{R}^d$ and fix $0 < \varepsilon < 1$. Then there exist $y_1, \ldots, y_n \in \mathbb{R}^k$ for $k = O(\varepsilon^{-2} \log n)$, such that

$$\forall i, j \in [n], \qquad \|y_i - y_j\|_2 \in (1 \pm \varepsilon)\|x_i - x_j\|_2.$$

Moreover, there is a randomized linear mapping $L : \mathbb{R}^d \to \mathbb{R}^k$ (oblivious to the given points), such that if we define $y_i = Lx_i$, then with probability at least $1 - 1/n$ all the above inequalities hold.

Throughout, all norms are ℓ_2, unless stated otherwise.

Remark: there is no assumption on the input points (e.g., that they lie in a low-dimensional space).

Idea: The map L is essentially (up to normalization) a matrix of standard Gaussians. In fact, random signs ± 1 work too!

Since L is linear, $Lx_i - Lx_j = L(x_i - x_j)$, and it suffices to verify that L preserves the norm of arbitrary vector WHP (instead of arbitrary pair of vectors).

Lemma 2 (Main): Fix $\delta \in (0, 1)$ and let $G \in \mathbb{R}^{k \times d}$ be a random matrix of standard Gaussians, for suitable $k = O(\varepsilon^{-2} \log \frac{1}{\delta})$. Then

$$\forall v \in \mathbb{R}^d, \qquad \Pr\left[\|Gv\| \notin (1 \pm \varepsilon)\sqrt{k}\|v\|\right] \leq \delta.$$

Using main lemma: Let $L = G/\sqrt{k}$, and recall we defined $y_i = Lx_i$. For every $i < j$, apply the lemma to $x_i - x_j$, then with probability at least $1 - \delta = 1 - 1/n^3$,

$$\|y_i - y_j\| = \|L(x_i - x_j)\| = \|G(x_i - x_j)\|/\sqrt{k} \in (1 \pm \varepsilon)\|x_i - x_j\|.$$

Now apply a union bound over $\binom{n}{2}$ pairs.

QED

*These notes summarize the material covered in class, usually skipping proofs, details, examples and so forth, and possibly adding some remarks, or pointers. The exercises are for self-practice and need not be handed in. In the interest of brevity, most references and credits were omitted.

It remains to prove the main lemma.

Fact 3 (the sum of Gaussians is Gaussian): Let $X \sim N(0, \sigma_X^2)$ and $Y \sim N(0, \sigma_Y^2)$ be independent Gaussian random variables. Then $X + Y \sim N(0, \sigma_X^2 + \sigma_Y^2)$.

The proof is by writing the CDF function (integration), recall that PDF is $\frac{1}{\sqrt{2\pi}} e^{-x^2/2}$.

Corollary 4 (Gaussians are 2-stable): Let $X_1, \ldots, X_n$ be independent standard Gaussians $N(0,1)$, and let $\sigma_1, \ldots, \sigma_n \in \mathbb{R}$. Then $\sum_i \sigma_i X_i \sim N(0, \sum_i \sigma_i^2)$.

Follows by induction.

Proof of main lemma: Was seen in class, using the next claim.

Claim 5: Let Y have chi-squared distribution with parameter k, i.e., $Y = \sum_{i=1}^{k} X_i^2$ for independent $X_1, \ldots, X_k \sim N(0,1)$. Then

$$\forall \varepsilon \in (0,1), \qquad \Pr[Y \geq (1+\varepsilon)^2 k] \leq e^{-\varepsilon^2 k/2}.$$

Remark: The claim and its proof are similar to Hoeffding bounds. Indeed, one may compare Claim 5 to another random variable $Y' \sim 2 \cdot B(k, 1/2)$ which has the same expectation.

It remains to prove Claim 5, which we will see in the next class.