CS787: Advanced Algorithms	**Scribe:** Tim Becker
Lecture 23: Dimension Reduction and Sparsification	**Date:** April 23, 2019

23.1 Dimension Reduction

The goal of Dimension reduction is to map points in $\mathbb{R}^d$ space to some smaller space $\mathbb{R}^k$ while preserving euclidean distance. In this lecture we will construct a random map $\mathbb{R}^d \to \mathbb{R}^k$ which satisfies the statement of the following lemmas.

23.1.1 Johnson-Lindenstrauss Lemmas

Lemma 23.1.1 (Distributional Johnson-Lindenstrauss Lemma) *For every $d > 0$, $\epsilon > 0$, $\delta > 0$, there exists a distribution on matrices $M \in \mathbb{R}^{k\times d}$ with $k = O(\frac{1}{\epsilon^2}\log\frac{1}{\delta})$ such that for all $x \in \mathbb{R}^d$, with probability at least $1-\delta$,*

$$||Mx||_2 \in (1 \pm \epsilon)||x||_2. \tag{23.1.1}$$

Dimension reduction is typically used to map a finite collection of n data points in $\mathbb{R}^d$ into $\mathbb{R}^k$, while preserving the distances between them. The following lemma is more useful in this context, and it follows from the previous lemma by a union bound argument: since there are in total $\binom{n}{2} \approx n^2$ pairs of points, we need to preserve all the pairwise distances, so we apply the distributional JL lemma with probability $\frac{\delta}{n^2}$ to get the following JL lemma:

Lemma 23.1.2 (Johnson-Lindenstrauss Lemma) *Given n points in $\mathbb{R}^d$ and $\epsilon \in [0,1]$, there exists a distribution on matrices $M \in \mathbb{R}^{k\times d}$ with $k = O(\frac{1}{\epsilon^2}\log\frac{n}{\delta})$, such that with probability at least $1-\delta$, for all pairs of data points x_1, x_2,*

$$||Mx_1 - Mx_2||_2 \in (1 \pm \epsilon)||x_1 - x_2||_2.$$

23.1.2 Construction

We will now construct the distribution on matrices $M \in \mathbb{R}^{k\times d}$ which satisfies these lemmas. Draw each M_{ij} independently from $N(0,1)$, i.e. the normal distribution with mean 0 and variance 1. Recall that $N(0,1)$ has density function $f(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$. Now observe that for any row M_i, the density of M_i in the $N(0,1)^d$ distribution is

$$\prod_j f(M_{ij}) = \frac{1}{\sqrt{2\pi}^d} \cdot \exp\left(\sum_j M_{ij}^2\right)^{\frac{1}{2}} \propto \exp\left(||M_i||_2^2\right)^{\frac{1}{2}},$$

and thus the density depends only on the length of M_i (this says that the d-dimensional normal distribution is "spherically symmetric").

Now, it suffices to only consider x such that $||x||_2 = 1$. This is because M is a linear map, so

$$\left|\left|M\left(\frac{x}{||x||_2}\right)\right|\right|_2 = \left|\left|\frac{1}{||x||_2}Mx\right|\right|_2 = \frac{1}{||x||_2}||Mx||_2$$

and thus Equation 23.1.1 holds for x if and only if it holds for the rescaled version of x. So fix some x with $||x||_2 = 1$. Then

$$\mathbf{E}[||M_i \cdot x||_2] = \mathbf{E}\left[\left(\sum_j M_{ij}x_j\right)^2\right] = \sum_j x_j^2 \mathbf{E}\left[M_{ij}^2\right] + 2\sum_{j \neq j'} x_j x_{j'} \mathbf{E}\left[M_{ij}M_{ij'}\right].$$

Now since $M_{ij} \sim N(0,1)$, $\mathbf{E}\left[M_{ij}^2\right] = 1$. Also since each M_{ij} is drawn independently, $\mathbf{E}\left[M_{ij} \cdot M_{ij'}\right] = \mathbf{E}[M_{ij}] \cdot \mathbf{E}\left[M_{ij'}\right] = 0 \cdot 0$, so the entire second term vanishes. Thus we're left with

$$\mathbf{E}[||M_i \cdot x||_2] = \sum_j x_j^2 = ||x||_2^2 = 1.$$

This implies

$$\mathbf{E}\left[||Mx||_2^2\right] = \mathbf{E}\left[\sum_i (M_i \cdot x)^2\right] = \sum_i \mathbf{E}[M_i \cdot x]^2 = k$$

So in expectation, M grows vectors by a factor of k. We want M to preserve lengths, so we can rescale by selecting $M_{ij} \sim \frac{1}{\sqrt{k}}N(0,1)$. Repeating the above analysis will then show $\mathbf{E}\left[||Mx||_2^2\right] = 1 = ||x||_2$.

Now consider rotating our axes so that $x_1 = 1$ and $x_i = 0$ for all $i > 1$. Then we have $M_i \cdot x = M_{i1}$. Thus $||Mx||_2^2 = \sum_i M_{i1}^2$, i.e. the squared length of the first column of M. Thus the probability that $||Mx||_2 \in (1 \pm \epsilon)||x||$ boils down to how concentrated the lengths of the columns are. Note that $\|Mx\|_2$ follows a χ^2 distribution, and it concentrates well around its mean. Apply a Bernstein type inequality gives the desired JL embedding.

23.2 Sparsification

Given a G with n vertices and m edges, we say G is *sparse* if $m = O(n \operatorname{polylog} n)$. Sparsification refers to replacing G with a sparse subgraph H of G such that some graph property is preserved. Examples of such graph properties are distances between nodes and cuts / capacities of cuts in G. We will focus on the latter property.

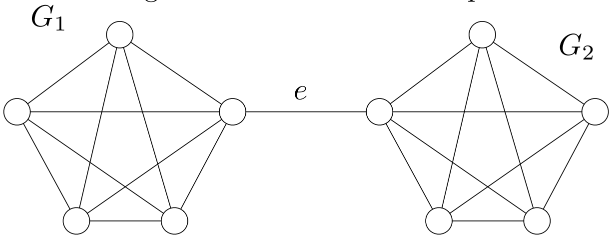
Definition 23.2.1 (Cut Sparsifier) *An ϵ-cut sparsifier of a graph $G = (V, E)$ is a weighted subgraph $H = (V, E')$ such that for all partitions (U_1, U_2) of V,*

$$W_H(U_1 \times U_2) \in (1 \pm \epsilon)W_G(U_1 \times U_2)$$

where $W_G(U_1 \times U_2) = |(U_1 \times U_2) \cap E|$ and

$$W_H(U_1 \times U_2) = \sum_{e \in (U_1 \times U_2) \cap E'} w_H(e).$$

Figure 23.2.1: Second Example



For the purpose of sparsification, we will want an ϵ-cut sparsifier with $m = O(n \operatorname{polylog} n)$. We want to give a randomized process to find an ϵ-cut sparsifier of G.

Example: Consider the following process: Take $G = K_n$, the complete graph on n vertices. For every edge e, w.p. p put e in E' and set $w_H(e) = \frac{1}{p}$. Otherwise don't put e in E'. ■

Example: Let G consist of two K_ns (call them G_1, G_2) joined together by a single edge e. Note that it is critical that we add e to E'. Otherwise, the cut (G_1, G_2) has weight 0 in H but weight 1 in G, and thus can never satisfy the inequality required of a cut sparsifier. See Figure 23.2.1. ■

The previous example suggests that we consider edges which are in cuts of small size more important than edges which are only in large cuts. For every $e \in E$, define k_e to be the size of the smallest cut containing e in G. Our strategy will be the following: with probability $\frac{1}{k_e}$ put e in E' and set $w_H(e) = k_e$. We will analyze this strategy in the next lecture.