CSE 522: Sublinear (and Streaming) Algorithms **Spring 2014**

Lecture 10: Johnson-Lindenstrauss Lemmas

April 30, 2014

Lecturer: Paul Beame *Scribe: Paul Beame*

Recall that in the Tug-of-War Sketch we began with a basic sketch y computed using a 4-wise independent $h : [M] \to \{-1, 1\}$ where y^2 is an unbiased estimator of F_2. The final sketch averages these sketches in groups of $k = \lceil 6/\varepsilon^2 \rceil$ such sketches and then outputs the median of $O(\log(/1\delta))$ of these averages.

As we observed earlier, the sketch for each group of k estimators is given by a $k \times M$ matrix $A \in \{-1, 1\}^{k \times M}$ and the average $y = A \cdot f$. The averaged estimator is $\sum_{i=1}^{k} y_i^2/k = ||y||_2^2/k$. The Tug-of-War Sketch chooses A with independent rows and within each row the entries are 4-wise independent. With those properties we showed that with probability at least 2/3.

$$(1-\varepsilon)F_2 = (1-\varepsilon)||f||_2^2 \leq ||y||_2^2/k \leq (1+\varepsilon)||f||_2^2 = (1-\varepsilon)F_2.$$

If we write $A' = \frac{1}{\sqrt{k}}A$ then with probability at least 2/3 we have

$$(1-\varepsilon)||f||_2^2 \leq ||A' \cdot f||_2^2 \leq (1+\varepsilon)||f||_2^2.$$

We would like to have this linear map succeed with much higher probability $1-\delta$ by, say, increasing the number of rows by an $O(\log(1/\delta))$ factor. However, the Chebyshev's inequality argument and the 4-wise independence in A do not yield such a decrease in failure probability.

What if the entries in A are uniformly independent? Achlioptas [] analyzed precisely this case.

Theorem 0.1 (Achlioptas)**.** *For $k \geq \frac{2\log_2(1/\delta)}{\varepsilon^2/2 - \varepsilon^3/3}$, if the elements of a $k \times M$ matrix A are independently to be ± 1 with probability $1/2$, then, for any fixed vector $x \in \mathbb{R}^M$, with probability at least $1-\delta$,*

$$(1-\varepsilon)||x||_2^2 \leq ||A \cdot x||_2^2/k \leq (1+\varepsilon)||x||_2^2.$$

Note that $1/(\varepsilon^2/2 - \varepsilon^2/3) \leq 6/\varepsilon^2$. We will discuss the ideas behind the proof of the theorem but first we discuss a simple application of the theorem to prove a lemma originally proved by Johnson and Lindenstrauss.

Lemma 0.2 (Johnson-Lindenstrauss Lemma)**.** *For any $\varepsilon > 0$ and integer n, and $k \geq k_0 = O(\varepsilon^{-2}\log n)$, for every set P of n points in $\mathbb{R}^d$ there is a (linear) map $f : \mathbb{R}^d \to \mathbb{R}^k$ such that for every $u, v \in P$,*

$$(1-\varepsilon)||u-v||_2^2 \leq ||f(u)-f(v)||_2^2 \leq (1+\varepsilon)||u-v||_2^2.$$

Proof. This follows immediately from the above theorem using the probabilistic method. Let $M = d$, and $\delta = 1/n^{2+c}$ for $c > 0$. Let

$$k = \lceil \frac{2\log_2(1/\delta)}{\varepsilon^2/2 - \varepsilon^3/3} = \lceil \frac{(4+2c)\log_2 n}{\varepsilon^2/2 - \varepsilon^3/3}.$$

For $A \in \{-1, 1\}^{k \times n}$ define

$$f_A(x) = \frac{A \cdot x}{\sqrt{k}}.$$

We show that for uniformly random A, f_A satisfies the properties of the J-L Lemma with probability at least $1 - n^{-c}/2$.

For each $u \neq v \in P$, write $x = u - v$, then $f_A(x) = f_A(u) - f_A(v) = \frac{A}{\sqrt{k}}(u - v) = \frac{A}{\sqrt{k}}(x)$. By the above theorem, except with probability at most δ for A chosen at random,

$$(1-\varepsilon)||u-v||_2^2 = (1-\varepsilon)||x||_2^2 \leq ||f_A(u)-f_A(v)||_2^2 = ||f_A(x) \leq (1+\varepsilon)||x||_2^2 = (1+\varepsilon)||u-v||_2^2.$$

Since there are only $\binom{n}{2}$ pairs $u \neq v \in P$, by a union bound, the probability that there is some pair $u, v \in P$ where the f_A fails to satisfy the required preservation of ℓ_2 norm is at most $\binom{n}{2}\delta \leq n^{-c}/2$ by definition of δ. Thus, not only does such an f with these properties exist, almost all functions f_A have the property. □

The Johnson-Lindenstrauss Lemma is very useful because it allows one to project high-dimensional data to a very lower dimensionsal space while approximately preserving all of its metric properties under the ℓ_2 metric. Many algorithms have runtimes that are exponential in the dimension and with only logarithmic dimension these algorithms become polynomial. It is also useful to note that the reduced dimension depends only on the number of points and not on the original dimension.

The theorem is sometimes called a Distributional Johnson-Lindenstrauss Lemma because of the connection to the J-L Lemma. Johnson and Lindenstrauss originally proved their result by beginning with a random rotation chosen from a spherically symmetric distribution followed by a projection on the first k coordinates. Vectors on the sphere can be chosen by selecting independent random entries according to the Gaussian distribution $N(0, 1)$ and then normalizing so that the vector has ℓ_2 norm 1.

It was later shown by Indyk and Motwani that the same properties hold by simply choosing each entry of A independently from $N(0, 1)$. Achlioptas, who described his version using ± 1 as a "database-friendly" version of the Johnson-Lindenstrauss Lemma because of the ease of computing with it, also showed that the same bound for the theorem holds if one independently sets each entry to 0 with probability 2/3 and scales everything up by a $\sqrt{3}$ factor to compensate for the fact the expected squared length would otherwise be only $k/3$. The 0s reduce the number of entries in the vector that need to be updated.

The sparsest matrices known for which one can prove a Distributional Johnson-Lindenstrauss Lemma are due to Kane and Nelson [] who show that a $k \times M$ matrix in which there are s blocks

of rows and precisely one randomly chosen ± 1 entry in each column of each block (with all other entries 0) will also work for $k = O(\varepsilon^{-2} \log n)$ provided $s = \Omega(\varepsilon^{-1} \log n)$ (and therefore there is a $\Theta(\varepsilon)$ fraction of non-zero entries). Note that this matrix is remarkably like the matrix for the Count Sketch. This is nearly tight since it is known that $\Omega((\varepsilon \log(1/\varepsilon))^{-1} \log n)$ non-zero entries per column are required for a distributional Johnson-Lindenstrauss Lemma to hold.

The general idea of the proofs is fairly similar. First, one begins with a distribution on each entry a_{ij} of A such that $\mathbb{E}(a_{ij}) = 0$ and $Var(a_{ij}) = \mathbb{E}(a_{ij}^2) = 1$. Then

$$\begin{aligned}
\mathbb{E}((A \cdot x)_i^2) &= \mathbb{E}((\sum_j a_{ij} x_j)^2) \\
&= \sum_j \sum_{j'} x_j x_{j'} \mathbb{E}(a_{ij} a_{ij'}) \\
&= \sum_j x_j^2 \mathbb{E}(a_{ij}^2) + \sum_{j' \neq j} x_j x_{j'} \mathbb{E}(a_{ij} a_{ij'}) \\
&= \sum_j x_j^2 \mathbb{E}(a_{ij}^2) \qquad \text{since } E(a_{ij})\mathbb{E}(a_{ij'}) = 0 \text{ for } j \neq j' \\
&= \sum_j x_j^2 \qquad \text{since } Var(a_{ij}) = 1 \\
&= ||x||_2^2
\end{aligned}$$

Therefore the square of each coordinate is an unbiased estimator of $||x||_2^2$. Because of independence, the proof for the deviation being small follows from methods similar to those used to prove Chernoff bounds and yields similar probability of error. In the case of random ± 1 elements, the method requires bounding all even moments of the distribution. We omit the details.

Alon showed that the dimension at which the Johnson-Lindenstrauss approximately preserves ℓ_2 distances is nearly optimal. As a consequence this shows that the dependence on the error ε in the Tug-of-War sketch is nearly optimal.

Theorem 0.3 (Alon)**.** *There is a set of $n + 1$ points in $\mathbb{R}^n$ such that for $1/2 \geq \varepsilon > 1/\sqrt{n}$, any mapping $f : \mathbb{R}^n \to \mathbb{R}^k$ that preserves the squares of ℓ_2^2 distance within $(1 \pm \varepsilon)$ factor requires that $k = \Omega(\frac{\log n}{\varepsilon^2 \log(1/\varepsilon)})$.*

Proof. Let the $n + 1$ points be $0 = (0, \ldots, 0)$ and $e_1 = (1, 0, \ldots, 0)$ through $e_n = (0, \ldots, 0, 1)$. Without loss of generality we can assume that $f(0) = 0$.

Let $v_i = f(e_i)$ for $i = 1, \ldots, n$. Since $||e_i - 0||_2^2 = ||e_i||_2^2 = 1$ must be approximately preserved for each i,

$$1 - \varepsilon \leq ||v_i||_2^2 \leq 1 + \varepsilon.$$

Similarly for $i \neq j$, $||e_i - e_j||_2^2 = 2$ so

$$2(1 - \varepsilon) \leq ||v_i - v_j||_2^2 \leq 2(1 + \varepsilon).$$

But, by bilinearity of the inner product and its symmetry over $\mathbb{R}$,

$$||v_i - v_j||_2^2 = \langle v_i - v_j \rangle^2 = \langle v_i, v_i \rangle - 2\langle v_i, v_j \rangle + \langle v_j, v_j \rangle,$$

from which it follows that $|\langle v_i, v_j \rangle| \leq 4\varepsilon$.

Consider the $n \times n$ matrix $\tilde{B}$ whose ij-th entry is $\langle v_i, v_j \rangle$. $\tilde{B}$ has rank at mose k since $\tilde{B} = V^T V$ where V is an $n \times k$ matrix whose i-th column is v_i.

The matrix $\tilde{B}$ is very close to an identity matrix. It is symmetric, its diagonal entries are within ε of 1, and its off-diagonal entries are within 4ε of 0. We will prove a lower bound on the rank of any such matrix in order to prove a lower bound on k.

To make this calculation convenient, we first normalize $\tilde{B}$ by dividing the i-th row of $\tilde{B}$ by $||v_i||_2^2$ which ensures that the diagonal is all 1's. This can increase the off-diagonal entries by a $1/(1-\varepsilon)$ factor, which still leaves them $O(\varepsilon)$ and does not change the rank. This can mess up the symmetry of $\tilde{B}$ in the resulting matrix B', so we set $B = (B' + (B')^T)/2$ which is symmetric. This matrix has 1 in all diagonal entries and $O(\varepsilon)$ off the diagonal.

Lemma 0.4. *Let $B = (b_{ij})_{ij}$, be an $n \times n$ symmetric real matrix with $b_{ii} = 1$ for all i and $|b_{ij}| \leq \varepsilon$ for all $i \neq j$. Then the rank r of B is at least*

$$\frac{n}{1 + (n-1)\varepsilon^2}.$$

Proof. Symmetric real matrices have all real eigenvalues and can be diagonalized. That is, there are real values $\lambda_1, \ldots, \lambda_n$ such that $B = ADA^{-1}$ where

$$D = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix},$$

and $\lambda_{r+1}, \ldots, \lambda_n = 0$ since B has rank r. Recall that the trace of a matrix is the sum of its diagonal entries and the trace of a matrix product is independent of the order of the matrices. By definition,

$$n = Trace(B) = Trace(ADA^{-1}) = Trace(DAA^{-1}) = Trace(D) = \sum_{i=1}^{n} \lambda_i = \sum_{i=1}^{r} \lambda_i.$$

Furthermore $\sum_{i=1}^n \lambda_i^2 = Trace(B^2) = Trace(B^T B) = \sum_{ij} b_{ij}^2$ since B is symmetric. But $b_{ij}^2 = 1$ for $i = j$ and $b_{ij}^2 \leq \varepsilon^2$ for $i \neq j$ so $\sum_{i=1}^n \lambda_i^2 \leq n + n(n-1)\varepsilon^2$. However, $\sum_{i=1}^n \lambda_i^2 = \sum_{i=1}^r \lambda_i^2 \geq r(n/r)^2 = n^2/r$ since $\sum_{i=1}^r \lambda_i = n$. Putting these two together we get $n^2/r \leq n + n(n-1)\varepsilon^2$ which yields the claimed bound on r by rearranging the inequality. □

we will finish the rest of the proof in the next class. □