CS 229r: Algorithms for Big Data Fall 2013

Lecture 9 — October 1, 2013

Prof. Jelani Nelson *Scribe: Colin Lu*

1 Overview

In the last lecture we proved several space lower bounds for streaming algorithms using the communication complexity model, and some ideas from information theory.

In this lecture we will move onto the next topic: dimensionality reduction.

2 Dimensionality Reduction

Dimensionality reduction is useful when solving high-dimensional computational geometry problems, such as:

- clustering
- nearest neighbors search
- numerical linear algebra (on big matrices)

The **main idea** of dimensionality reduction is to reduce the dimensionality of the input while preserving the geometric structure of the input.

Reducing the dimensionality of the input enables our algorithms to run faster, but since we preserve the geometric structure of the input, our algorithms are still approximately correct

2.1 Distortion

Definition 1. *Suppose we have two metric spaces,* (X, d_X)*, and* (Y, d_Y)*, and a function* $f : X \to Y$*. Then* f *has distortion* D_f *if* $\forall x, x' \in X$*,* $C_1 \cdot d_X(x, x') \leq d_Y(f(x), f(x')) \leq C_2 \cdot d_X(x, x')$*, where* $\frac{C_2}{C_1} = D_f$*.*

We will focus on spaces in which $d_X(x, x') = \|x - x'\|_X$ (ie. normed spaces).

2.2 Limitations of Dimensionality Reduction

If $\|\cdot\|_X$ is the l_1 norm, then $D_f \leq C \implies$ in worst case, target dimension is $n^{\Omega(\frac{1}{C^2})}$. That is, there exists a set of n points X, such that for all functions $f : (X, l_1) \to (X', l_1^m)$, with distortion $\leq C$, then m must be at least $n^{\Omega(\frac{1}{C^2})}$[1].

3 Johnson-Lindenstrauss lemma

The Johnson-Lindenstrauss (JL) lemma [2] states that for all $\epsilon \in (0, \frac{1}{2})$, $\forall x_1, ..., x_n \in l_2$, there exists $\Pi \in R^{m\times n}$, $m = O(\frac{1}{\epsilon^2}\log(n))$ such that for all i, j , $(1-\epsilon)\|x_i - x_j\|_2 \le \|\Pi x_i - \Pi x_j\|_2 \le (1+\epsilon)\|x_i - x_j\|_2$

$f : (x, l_2) \to (x, l_2^m)$, $f(x) = \Pi x$

Theorem 2 (Johnson and Naor [3])**.** *: Suppose X is a Banach space (complete normed space), such that for any n point subset, there exists a linear map Π into a linear subspace $F \subseteq X$ of dimension $O(\log n)$ with $O(1)$ distortion. Then for every positive integer k, every k-dimensional linear subspace of X can be embedded into ℓ_2 with distortion at most $2^{2^{O(\log^* n)}}$.*

In a sense, this theorem states that any complete normed space that enjoys Johnson-Lindenstrauss type dimensionality reduction is similar to l_2.

4 Distributional Johnson-Lindenstrauss lemma

Proofs of the JL lemma typically first prove the distributional JL lemma, which states that for all $0 < \epsilon, \delta < \frac{1}{2}$, there exists a distribution $D_{\epsilon,\delta}$ on matrices $\Pi \in \mathbb{R}^{m\times n}$, $m = O(\frac{1}{\epsilon^2}\log(\frac{1}{\delta}))$ such that for all $x \in \mathbb{R}^n$, and Π drawn from the distribution $D_{\epsilon,\delta}$,

$$\mathbb{P}(\|\Pi x\|_2 \notin [(1-\epsilon)\|x\|_2, (1+\epsilon)\|x\|_2] < \delta$$

This is equivalent to saying that for all x of unit Euclidean norm,

$$\mathbb{P}(|\|\Pi x\|_2^2 - 1)| > \epsilon) < \delta \text{ (up to changing } \epsilon \text{by a factor of 2, from the squaring)}$$

Claim 3. *Distributional JL lemma implies JL lemma*

Proof. Set $\delta < \frac{1}{\binom{n}{2}}$, and union bound over the $\binom{n}{2}$ points, with $x = \frac{x_i - x_j}{\|x_i - x_j\|_2}$ □

There are various ways to prove the distributional JL lemma

- Johnson and Lindenstrauss' approach: consider a random rotation, and project onto the first m coordinates
- other proofs: choose $\Pi_{i,j}$ independent, mean 0, variance $\frac{1}{m}$, and subgaussian (ie. decays at a rate beneath a constant factor of the Gaussian distribution).
- more proofs: coming in future lectures

5 Proof of the distributional JL lemma

We will now prove the distributional JL lemma using Π with random sign entries

$$\Pi_{i,j} = \frac{\sigma_{i,j}}{\sqrt{m}}, \ \sigma_{i,j} \in \{-1, 1\}$$

The "usual proof" that this matrix satisfies the conditions of the lemma involves expanding the expression:

$$\|\Pi x\|_2^2 - 1 = \frac{1}{m} \sum_{r=1}^{m} \sum_{i \neq j} \sigma_{r,i} \sigma_{r,j} x_i x_j := z$$

Then by some computations,

$$\mathbb{P}(z > \epsilon) = \mathbb{P}(e^{tz} > e^{t\epsilon}) < \frac{\mathbb{E}\, e^{tz}}{e^{t\epsilon}} \text{ by Markov}$$

and also

$$\mathbb{P}(z < -\epsilon) = \mathbb{P}(-z > \epsilon) = \mathbb{P}(e^{-tz} > e^{t\epsilon}) < \frac{\mathbb{E}\, e^{-tz}}{e^{t\epsilon}}$$

This approach is not particularly insightful, so instead we will instead present a proof of distributional JL using the Hanson-Wright inequality.

5.1 Hanson-Wright inequality

Theorem 4 (Hanson-Wright inequality [4])**.** *Suppose $\sigma = (\sigma_1, ..., \sigma_n)$, a vector of i.i.d. uniform random signs, $A = (a_{ij}) \in \mathbb{R}^{n \times n}$*

$$\mathbb{P}(|\sigma^T A \sigma - \mathbb{E}(\sigma^T A \sigma)| > \lambda) \lesssim e^{-min\{\frac{c \cdot \lambda^2}{\|A\|_F^2}, \frac{c\lambda}{\|A\|}\}}$$

$\|A\|_F^2 = \sum a_{ij}^2$, $\|A\| = \sqrt{\lambda_{max}(A^T A)}$ *=largest magnitude of eigenvalue of A if A is symmetric.*

Claim 5. *The Hanson-Wright Inequality implies the distributional JL lemma*

Proof. Set $\lambda = \epsilon$, $\|\Pi x\|_2^2 - 1 = \sigma^T A_x \sigma$, where

$$A_x = \frac{1}{m} \begin{bmatrix} xx^T & 0 & ... & 0 & 0 \\ 0 & xx^T & ... & 0 & 0 \\ ... & ... & ... & ... & ... \\ 0 & 0 & ... & xx^T & 0 \\ 0 & 0 & ... & 0 & xx^T \end{bmatrix} \text{ (m blocks each n×n)}$$

$\|A_x\|_F^2 = \frac{1}{m^2}\sum_{r=1}^m \sum_{i,j} x_i^2 x_j^2 = \frac{1}{m}$

$\|A_x\| = \frac{1}{m}(1) = \frac{1}{m}$ since xx^T has eigenvalue $\|x\|_2^2 = 1$ with eigenvector x. Then Hanson-Wright directly gives the distributional JL lemma. □

5.1.1 Useful Theorems and Lemmas

We will prove the Hanson-Wright inequality using the following theorems and lemmas (they have other applications as well)

Definition 6. *For X a random variable, define $\|X\|_p = (\mathbb{E}\,|X|^p)^{\frac{1}{p}}$*

Theorem 7. *Minkowski's inequality: $\|\cdot\|_p$ is a norm for $p \geq 1$. In other words, the triangle inequality holds.*

Theorem 8 (Jensen's inequality)**.** *If Φ is a convex function (ie. $\Phi(tx+(1-t)y) \leq t\Phi(x)+(1-t)\Phi(y)$ for all $t \in [0,1]$, x, y)), then $\Phi(\mathbb{E}\,X) \leq \mathbb{E}\,\Phi(X)$*

Claim 9. *If $1 \leq p < q$, then $\|X\|_p \leq \|X\|_q$*

Proof. Use Jensen's inequality on $\Phi(z) = |z|^{q/p}$, giving $(\mathbb{E}\,|X|^p)^{q/p} \leq \mathbb{E}\,|X|^q$, then take $1/p$th powers on both sides. □

Definition 10. *The standard normal distribution $N(0,1)$ is a distribution over $\mathbb{R}$ with probability density function $f(t) = \frac{1}{\sqrt{2\pi}}e^{-t^2/2}$*

Fact 11. *If $g \sim N(0,1)$, then $\mathbb{E}\,g^p$ is 0 if p is odd, and $\mathbb{E}\,g^p = \frac{p!}{2^{p/2}\cdot(p/2)!})$ when p is even. By Stirling's approximation, the latter is $O(\sqrt{p})^p$.*

Lemma 12 (Concentration of Lipschitz functions of Gaussians [5])**.** *Suppose $f : \mathbb{R}^n \to \mathbb{R}$. $g = (g_1, ..., g_n)$ is a vector of independent Gaussians $N(0,1)$. Then*

$$\mathbb{P}(|f(g) - \mathbb{E}(f(g))| > \lambda) \leq 2 \cdot e^{-c\lambda^2/\|f\|_{lip}^2} \tag{1}$$

where $\|f\|_{lip} = sup_{x,y}\frac{|f(x)-f(y)|}{\|x-y\|_2}$

(1) $\iff \forall p \geq 1$, $\|f(g) - \mathbb{E}(f(g))\|_p \leq \sqrt{p} \cdot \|f\|_{lip}$. This implies (1) by Markov's inequality on the pth moment.

Lemma 13 (Decoupling)**.**

$$\sum_{i \neq j} \|a_{ij}\sigma_i\sigma_j\|_p \leq 4 \cdot \|\sum_{i,j} a_{ij}\sigma_i\sigma'_j\|_p$$

$\sigma = (\sigma_1, ..., \sigma_n)$, $\sigma' = (\sigma'_1, ..., \sigma'_n)$, all independent random signs.

5.2 Proof of Hanson-Wright inequality

The following proof was shown to me by Gilles Pisier on a whiteboard in October 2011. It suffices to show $\|\sum_{i\neq j} a_{ij}\sigma_i\sigma_j\|_p \leq \sqrt{p}\|A\|_F + p\cdot\|A\|$, since subtracting the expectation removes the trace, and this moment bound can be plugged into Markov's inequality with the appropriate pth moment to get the desired exponential tail bound. Then

$$\begin{aligned}
\|\sum_{i\neq j} a_{ij}\sigma_i\sigma_j\|_p &\leq 4\cdot\|\sum_{i,j} a_{ij}\sigma_i\sigma_j'\|_p \text{ (by decoupling)}\\
&= \frac{4}{(\mathbb{E}\,|g|)^2}\cdot\|\sum_{i,j} a_{ij}\sigma_i\sigma_j'\,\mathbb{E}\,|g_i|\cdot|g_j'|\|_p\\
&\lesssim \|\sum_{i,j} a_{ij}\sigma_i\sigma_j'\cdot|g_i|\cdot|g_j'|\|_p\\
&= \|\sum_{i,j} a_{ij}g_ig_j'\|_p\\
&= \|\langle Ag', g\rangle\|_p\\
&\lesssim \sqrt{p}\cdot\|\|Ag'\|_2\|_p\\
&\leq \sqrt{p}\left(\|\mathbb{E}\,\|Ag'\|_2\|_p + \|\|Ag'\|_2 - \mathbb{E}\,\|Ag'\|_2\|_p\right) (L^p \text{ is a norm, Minkowski})
\end{aligned}$$

But we have that by Jensen's inequality, $\|\mathbb{E}\,\|Ag'\|_2\|_p \leq (\mathbb{E}\,\|Ag'\|_2^2)^{\frac{1}{2}} = \|A\|_F$.

The other term is bounded by the lemma on Lipschitz functions of Gaussians, on $f(g') = \|Ag'\|_2$, $\|f\|_{lip} = \|A\|$, so
$||Ag'\|_2 - \mathbb{E}\,\|Ag'\|_2\|_p \leq \sqrt{p}\|A\|$, and

$$\sqrt{p}\left(\|\mathbb{E}\,\|Ag'\|_2\|_p + ||Ag'\|_2 - \mathbb{E}\,\|Ag'\|_2\|_p\right) \leq \sqrt{p}\|A\|_F + p\cdot\|A\|$$

completing the proof of the Hanson-Wright inequality.

References

[1] B. Brinkman and M. Charikar, On the impossibility of dimension reduction in l_1, J. ACM, vol. 52, no. 5, pp. 766–788, 2005.

[2] William B. Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.

[3] William B. Johnson and Assaf Naor, The Johnson-Lindenstrauss Lemma Almost Characterizes Hilbert Space, But Not Quite. *Discrete and Computational Geometry*, vol. 43, no. 3, pp. 542–553, 2010.

[4] David Lee Hanson and Farroll Tim Wright. A bound on tail probabilities for quadratic forms in independent random variables. *Ann. Math. Statist.*, 42(3):1079–1083, 1971.

[5] Gilles Pisier. Probabilistic methods in the geometry of Banach spaces. *Probability and Analysis. Lecture Notes in Mathematics.* Vol. 1206, 167–241, 1986.