0368-3248-01-Algorithms in Data Mining **Fall 2012**

Lecture 5: Random Projections

Lecturer: Edo Liberty

Warning: *This note may contain typos and other inaccuracies which are usually discussed during class. Please do not cite this note as a reliable source. If you find mistakes, please inform me.*

We will give a simple proof of the following, rather amazing, fact. Every set of n points in a Euclidian space (say in dimension d) can be embedded into the Euclidian space of dimension $k = O(\log(n)/\varepsilon^2)$ such that all pairwise distances are preserved up distortion $1 \pm \varepsilon$. We will follow ideas from [1] and [2] and later improve on running time using methods introduced in [3].

Random projection

We will argue that a certain distribution over the choice of a matrix $R \in \mathbb{R}^{k \times d}$ gives that:

$$\forall x \in \mathbb{R}^d \quad \Pr\left[\left| ||\frac{1}{\sqrt{k}} Rx|| - ||x|| \right| > \varepsilon ||x|| \right] \leq \frac{1}{n^2} \tag{1}$$

Before we pick this distribution and show that Equation 1 holds for it, let us first see that this gives the opening statement.

Consider a set of n points $x_1, \ldots, x_n$ in Euclidian space $\mathbb{R}^d$. Embedding these points into a lower dimension while preserving all distances between them up to distortion $1 \pm \varepsilon$ means approximately preserving the norms of all $\binom{n}{2}$ vectors $x_i - x_j$. Assuming Equation 1 holds and using the union bound, this property will fail to hold for at least one $x_i - x_j$ pair with probability at most $\binom{n}{2}\frac{1}{n^2} \leq 1/2$. Which means that all $\binom{n}{2}$ point distances are preserved up to distortion ε with probability at least $1/2$.

1 i.i.d. gaussian distribution

We consider the distribution of matrices R such that each $R(i,j)$ is drawn independently from a normal distribution with mean zero and variance 1, $R(i,j) \sim \mathcal{N}(0,1)$. We show that for this distribution Equation 1 holds for some $k \in O(\log(n)/\varepsilon^2)$.

First consider the random variable $z = \sum_{i=1}^{d} r(i)x(i)$ where $r(i) \sim \mathcal{N}(0,1)$. To understand how the variable z distributes we recall the two-stability of the normal distribution. Namely, if $z_3 = z_2 + z_1$ and $z_1 \sim \mathcal{N}(\mu_1, \sigma_1)$ and $z_2 \sim \mathcal{N}(\mu_2, \sigma_2)$ then,

$$z_3 \sim \mathcal{N}(\mu_1 + \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2}).$$

In our case, $r(i)x(i) \sim \mathcal{N}(0, x_i)$ and therefore, $z = \sum_{i=1}^{d} r(i)x(i) \sim \mathcal{N}(0, \sqrt{\sum_{i=1}^{d} x_i^2}) \sim \mathcal{N}(0, ||x||) \sim ||x||y$ where $y_i \sim \mathcal{N}(0,1)$. Now, note that each element in the vector Rx distributes exactly like z. Defining k identical copies of z, $z_1, \ldots, z_k$, We get that $||\frac{1}{\sqrt{k}} Rx||$ distributes exactly like:

$$||\frac{1}{\sqrt{k}} Rx|| \sim \sqrt{\frac{1}{k}\sum_{i=1}^{k} z_i^2} \sim ||x|| \sqrt{\frac{1}{k}\sum_{i=1}^{k} y_i^2}$$

Thus, proving Equation 1 reduces to showing that:

$$\Pr\left[\left|\sqrt{\frac{1}{k}\sum_{i=1}^{k} y_i^2} - 1\right| > \varepsilon\right] \leq \frac{1}{n^2} \tag{2}$$

The sum of k squared normal variables is a very known distribution called chi-square with k degrees of freedom, denoted by χ_k^2. It is exactly defined by $\chi_k^2 = \sum_{i=1}^{k} y_i^2$ where $y_i \sim \mathcal{N}(0,1)$. Since χ_k^2 is a sum of independent random variables, due to the central limit theorem, χ_k^2 converges to a normally distributed quantity as k grows. We will use here a slightly different property $\sqrt{\chi_k^2} \sim_{k\to\infty} \mathcal{N}(\sqrt{k}, 1/\sqrt{2})$. Somewhat sloppily, we will assume that k is large enough so that it is harmless to substitute:

$$\sqrt{\chi_k^2} \sim \mathcal{N}(\sqrt{k}, 1/\sqrt{2})$$

I that case we have$\sqrt{\frac{1}{k}\sum_{i=1}^{k} y_i^2} - 1 \sim \mathcal{N}(0, \frac{1}{\sqrt{2k}})$. Thus, we only need to show that for a random variable $Z \sim \sqrt{2k}\left[\sqrt{\frac{1}{k}\sum_{i=1}^{k} y_i^2} - 1\right] \sim \mathcal{N}(0,1)$ it holds that

$$\Pr\left[|Z| > \varepsilon\sqrt{2k}\right] \leq \frac{1}{n^2} \tag{3}$$

We now use a simple bound on the error function

$$\Pr[Z > t] = \int_t^\infty \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz < \int_t^\infty \frac{1}{\sqrt{2\pi}} \frac{z}{t} e^{-z^2/2} dz = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$$

Setting $t = \varepsilon\sqrt{2k}$ and noting that $\Pr[Z > t] = \Pr[Z < -t]$ we demand that $\frac{1}{\sqrt{2\pi}} e^{-\varepsilon^2 k} \leq \frac{1}{2n^2}$. This yields the bound $k \geq \frac{2\log(n)+O(1)}{\varepsilon^2}$ which completes the proof.

2 Sparse Random Projections

The goal of this section is The matrix R will contain a non zero only w.p. q. That is, $R(i,j) = N(0, 1/\sqrt{q})$ with probability q and zero else. Again, we define $y_i = \sum_j R(i,j)x_j = \sum_j b_{i,j} g_{i,j} x_j$ where $b_{i,j} = 1$ w.p. q and $g_{i,j} \sim \mathcal{N}(0, 1/\sqrt{q})$. For simplicity and w.o.l.g. we set $\|x\|_2 = 1$.

First notice that $\mathbb{E}[y_i^2] = 1$. Also, given the values of $b_{i,j}$ we have that y_i is Gaussian. More accurately, $\sigma_i^2 = \sum_j b_{i,j} x_j^2/q$

First, let us see that this is sufficient in some sense

$$\begin{aligned} \Pr[\frac{1}{k}\sum_i y_i^2 \geq (1+\varepsilon)] &= \Pr[e^{\lambda\sum_i y_i^2} \geq e^{\lambda k(1+\varepsilon)}] & (4)\\ &\leq e^{-\lambda k(1+\varepsilon)} \Pi_i \mathbb{E}[e^{\lambda y_i^2}] & (5) \end{aligned}$$

Given σ_i we have that y_i is Gaussian and so we can compute $\mathbb{E}[e^{y_i^2}]$ exaclty.

$$\begin{aligned} \mathbb{E}[e^{\lambda y_i^2}] &= \frac{1}{\sqrt{2\pi\sigma_i^2}} \int_\infty^\infty e^{-\frac{y_i^2}{2\sigma_i^2}} e^{\lambda y_i^2} dy & (6)\\ &= \frac{1}{\sqrt{2\pi\sigma_i^2}} \int_\infty^\infty e^{-(\frac{1}{\sigma_i^2} - 2\lambda)\frac{y_i^2}{2}} dy & (7)\\ &= \frac{1}{\sqrt{1-2\lambda\sigma_i^2}} & (8) \end{aligned}$$

Note that we must enforce now that $2\lambda\sigma_i^2 < 1$. Plugging this back into our formula we get

$$\begin{aligned} \Pr &\leq e^{-\lambda k(1+\varepsilon)}\Pi_i \frac{1}{\sqrt{1-2\lambda\sigma_i^2}} && (9)\\ &= e^{-\lambda k(1+\varepsilon)+\frac{1}{2}\sum_i \log(\frac{1}{1-2\lambda\sigma_i^2})} && (10)\end{aligned}$$

We now use the Tailor expansion by $log(\frac{1}{1-x}) \geq x + x^2$

$$\begin{aligned} \Pr &\leq e^{-\lambda k(1+\varepsilon)+\frac{1}{2}\sum_i 2\lambda\sigma_i^2+4\lambda^2\sigma_i^4} && (11)\\ &\leq e^{\lambda(\sum_i \sigma_i^2-k)-\lambda k\varepsilon+\sum_i 2\lambda^2\sigma_i^4} && (12)\\ & && (13)\end{aligned}$$

Now, assume that $\sigma_i^2 \leq 1 + \varepsilon/2$ (we will fix this soon) then

$$\begin{aligned} \Pr &\leq e^{-\lambda k\varepsilon/2+2\lambda^2\sum_i \sigma_i^4} && (14)\\ &\leq e^{-\frac{1}{32}\frac{k^2\varepsilon^2}{\sum_i \sigma_i^4}} \leq e^{-ck\varepsilon^2} && (15)\end{aligned}$$

for some constant c. As before, invoking the union bound completes the proof for some $k \in O(\log(n)/\varepsilon^2)$.

Alas, we are left to show that $\sigma_i^2 \leq 1+\varepsilon/2$. This is where the bounds on q come in. We will see that this is not true for every vector x and every value of q. Never the less, we'll be able to fix this and still gain on running time. Let us recap , $\sigma_i^2 = \sum_j b_{i,j}x_j^2/q$ where $b_{i,j} = 1$ w.p. q and zero otherwise. Take for example $x = [1, 0, \ldots, 0]$. In this case $\sigma_i^2 = b_{i,1}/q$ which is potentially $1/q$ which is significantly more that $1 + \varepsilon/2$. On the other hand, consider the vector $x = [\frac{1}{\sqrt{d}}, \ldots, \frac{1}{\sqrt{d}}]$. In this case $\sigma_i^2 = \frac{1}{d}\sum_j b_{i,j}/q$ whose expectation is 1 and which we expect from Chernoff's inequality to be less that $1 + \varepsilon/2$ w.h.p.

Let us restrict our selves to vectors such that $\|x\|_\infty \leq \eta$. I claim that the "worst" vectors we can have of this form contain $1/\eta^2$ entries of value η and the rest zeros. This is a result of the convexity of the moment generating functions of σ_i^2 with respect to x and the fact that the set of possible values for $\|x\|_\infty \leq \eta$ lies in a polytop. Hence, the maximal value is attained in an extreme point as above. Computing for this vector we have $\sigma_i^2 = \sum_{j=1}^{1/\eta^2} b_{i,j}\eta^2/q$. Bounding σ_i^2 by $1 + \varepsilon$ we get

$$\begin{aligned} \Pr[\sigma_i^2 \geq 1+\varepsilon] &= \Pr[\sum_{j=1}^{1/\eta^2} b_{i,j} - \frac{q}{\eta^2} \geq \frac{q\varepsilon}{\eta^2}] && (16)\\ &\leq e^{-\frac{q\varepsilon^2}{2\eta^2}} \leq \frac{1}{cnd} \quad \text{(fail w.p. at most 1/5)} && (17)\\ &\text{for} \quad q \geq \frac{3\log(n)\eta^2}{\varepsilon^2} && (18)\end{aligned}$$

Thus, if our vectors are "spread" such that $\|x\|_\infty \leq \eta < \frac{\varepsilon}{\sqrt{3\log(n)}}$ we can save one computation and storage by being able to set $q < 1$.

3 Fast Vector Spreading

The question is, can we actively make sure that $\|x\|_\infty$ is low. The answer is yes and a method for doing that was suggested in [3]. For this we will need to learn what Hadamard matrices are. Hadamard matrices are commonly used in coding theory and are conceptually close for Fourier matrices. We assume for convenience that d is a power of 2. The Walsh Hadamard transform of a vector $x \in \mathbb{R}^d$ is the result of the matrix-vector multiplication Hx where H is a $d \times d$ matrix whose entries are $H(i,j) = \frac{1}{\sqrt{d}}(-1)^{\langle i,j\rangle}$. Here $\langle i, j\rangle$ means the

dot product over F_2 of the bit representation of i and j as binary vectors of length $\log(d)$. Another way to view this is to define Hadamard Matrices recursively.

$$H_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \quad H_d = \frac{1}{\sqrt{2}} \left(\begin{array}{c:c} H_{d/2} & H_{d/2} \\ \hdashline H_{d/2} & -H_{d/2} \end{array}\right)$$

Here are a few interesting (and easy to show) facts about Hadamard matrices.

1. H_d is a unitary matrix $\|Hx\| = \|x\|$ for any vector $x \in \mathbb{R}^d$.
2. Computing $x \mapsto Hx$ requires $O(d \log(d))$ operations.

We also define a diagonal matrix D to be such that $D(i,i) \in \{1,-1\}$ uniformly. Clearly, we have that $\|HDx\|_2 = \|x\|_2$ since both H and D are isotropies. Let us now bound $\|HDx\|_\infty$. $(HDx)(1) = \sum_{i=1}^d H(1,i)D(i,i)x_i = \sum_{i=1}^d \frac{x_i}{\sqrt{d}} s_i$ where $s_i \in \{-1,1\}$ uniformly. To bound this we recap Hoeffding's inequality.

Fact 3.1 (Hoeffding's inequality)**.** *Let $X_1, \ldots, X_n$ be independent random variables s.t. $X_i \in [a_i, b_i]$. Let $X = \sum_{i=1}^n X_i$.*

$$\Pr[|X - \mathbb{E}[X]| \geq t] \leq 2e^{-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}} \tag{19}$$

Invoking Hoeffding's inequality and then the union bound we get that if $\|HDx\|_\infty \leq \sqrt{\frac{c \log(n)}{d}}$ for all points x. Remark, for this we assumed $\log(d) = O(\log(n))$ otherwise we should have had $\log(nd)$ in the bound. The situation, however, that the dimension is super polynomial in the number of points is unlikely. Usually it is common to have $n > d$.

4 Fast Random Projecton

Combining fast spreading with sparse projections we get the result in [3]. Randomly project vectors by $x \mapsto \frac{1}{\sqrt{k}} RHDx$. Computing HDx requires $O(d \log(d))$ operations and guaranties that $\|HDx\|_\infty \leq \eta = \sqrt{\frac{c \log(n)}{d}}$. Setting this into the bound on $q \geq \frac{3 \log(n) \eta^2}{\varepsilon^2}$ we get that is is sufficient to have $q \geq \frac{c \log^2(n)}{d\varepsilon^2}$. The expected number of non zeros in R is qkd. Therefore, the expected running time required to compute $x' \mapsto Rx'$ is bounded from above by $O(ck \log^2(n)/\varepsilon^2) = O(\varepsilon^2 k^3)$. Putting this together we get a total running time of $O(d \log(d) + \varepsilon^2 k^3)$ instead of the straight forward $O(kd)$.

References

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

[2] S. DasGupta and A. Gupta. An elementary proof of the Johnson-Lindenstrauss lemma. *Technical Report, UC Berkeley*, 99-006, 1999.

[3] Nir Ailon and Bernard Chazelle. Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform. In *Proceedings of the 38st Annual Symposium on the Theory of Compututing (STOC)*, pages 557–563, Seattle, WA, 2006.