Lecture 17

1 Measure concentration and Johnson Lindenstrauss lemma

The phenomenon of measure concentration is one reason that high dimensional data can often be compressed to low dimensional data while preserving the essential features of the problem. The geometric information about a set of high dimensional point set $x_1, x_2, \cdots, x_n \in \mathbb{R}^d$ is captured by the pairwise distances between them. For example, the points could be feature vectors and distances between them could be measures of similarity. The Johnson Lindenstrauss lemma shows that projecting the data points on to a random low dimensional subspace approximately preserves the distance between them.

Lemma 1
If y_i are the projections of the $x_i \in \mathbb{R}^d$ onto a random $k = \frac{c \log n}{\epsilon^2}$ dimensional subspace then with probability $1 - \frac{1}{n^{c-2}}$,

$$(1-\epsilon)\sqrt{\frac{k}{d}}|x_i - x_j|^2 \leq |y_i - y_j|^2 \leq (1+\epsilon)\sqrt{\frac{k}{d}}|x_i - x_j|^2 \tag{1}$$

i.e. projecting and scaling by $\sqrt{\frac{d}{k}}$ preserves all pairwise distances within a factor of $1+\epsilon$.

The notion of a random subspace: A k dimensional subspace can be specified as the span of k orthonormal unit vectors $v_1, v_2, \cdots, v_k$. A random k dimensional subspace is obtained by choosing v_1 to be a uniformly random unit vector and recursively choosing v_i to be a uniformly random unit vector from the $d - i + 1$ dimensional subspace orthogonal to $v_1, v_2, \cdots, v_{i-1}$.

Alternatively we can select k uniformly random unit vectors $v_1, v_2, \cdots, v_k \in \mathbb{R}^d$ and apply the Gram Schmidt process to the $k \times d$ matrix V that has vectors v_i as the rows, to obtain a set of k random orthonormal vectors.

One of the goals of this lecture is to understand the notion of a uniformly random unit vector, we will see that a random unit vector can be obtained by sampling from the n dimensional normal distribution and normalizing to unit length.

A picture of random projections: Multipliying by the matrix V projects $x \in \mathbb{R}^d$ onto the k dimensional subspace spanned by $v_1, v_2, \cdots, v_k$,

$$Vx = \begin{bmatrix} v_{11} & v_{12} & \dots & v_{1d} \\ v_{21} & v_{22} & \dots & v_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ v_{k1} & v_{k2} & \dots & v_{kd} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ \vdots \\ x_d \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_k \end{bmatrix} \tag{2}$$

A random projection of $x \in \mathbb{R}^d$ is equivalent to taking the inner product of x with k uniformly random mutually orthogonal unit vectors.

Inverting the perspective: Inner products are invariant under rotations, so the distribution of the length of the projection of a fixed vector onto a random k dimensional subspace is the same as the distribution of the projection of a random vector z onto the subspace spanned by the standard basis vectors $e_1, e_2, \cdots, e_k$.

More formally, we can chose an orthogonal matrix U that maps $v_i \to e_i$ and is extended arbitrarily on the remaining space and reason as follows,

$$y_i = \langle v_i | x \rangle = \langle Uv_i | Ux \rangle = \langle e_i | Ux \rangle = \langle e_i | z \rangle \tag{3}$$

The inverse of U maps the standard basis e_i to the random orthonormal basis v_i and therefore is a random orthogonal transformation. Hence, the vector $z = Ux$ is uniformly distributed on the d dimensional unit sphere for all $x \in \mathbb{R}^d$.

Using the alternative view of a random projection we can compute the expected values of the coordinates y_i. As z is a unit vector $E[\sum_{i \in [d]} z_i^2] = 1$, by symmetry the z_i are identically distributed we have,

$$E[\sum_{i \in [k]} z_i^2] = \frac{k}{d} \tag{4}$$

The expected length of the projection of a unit vector $x \in \mathbb{R}^d$ onto a random k dimensional subspace is $\sqrt{\frac{k}{d}}$. The Johnson Lindenstrauss lemma states that the length of the projection of x is sharply concentrated around the expectated value, we will prove a weaker concentration bound below.

Almost all the volume of the d dimensional sphere lies close to the equator, the measure (volume) of the sphere is thus concentrated close to the origin.

Claim 2
If z is a uniformly random unit vector from the d dimensional unit sphere,

$$\Pr[|z_1| > \frac{t}{\sqrt{d}}] \leq e^{-t^2/2} \tag{5}$$

Proof: The probability that $|z_1|$ is greater than $\frac{t}{\sqrt{d}}$ is equal to the ratio of the surface area of two spherical caps of radius $r = \sqrt{1 - \frac{t^2}{d}}$ to the surface area of the d dimensional sphere. We will provide an argument that illustrates the idea while avoiding explicit calculations.

The area of two spherical caps of radius r is less that the area of a sphere of the same radius. In d dimensions the surface area of a sphere radius R scales as $c.R^{d-1}$ where c is some constant that we do not state explicitly.

The claim follows by comparing the area of the sphere $c.1^d$ to the area of the spherical caps which is at most $c.\left(1 - \frac{t^2}{d}\right)^{d/2} \leq e^{-t^2/2}$ using the approximation $1 - x \leq e^{-x}$. $\square$

The claim tells us that for a random unit vector in $\mathbb{R}^d$, with overwhelming probability one coordinate is at most $\log d$ times bigger than another. The Johnson Lindenstrauss lemma follows from the following stronger concentration bound that we will not prove,

$$\Pr[\left|\sqrt{z_1^2 + z_2^2 + \cdots + z_k^2} - \sqrt{\frac{k}{d}}\right| > t] \leq e^{-t^2 d} \tag{6}$$

Substituting $t = \epsilon\sqrt{\frac{k}{d}}$ where $k = \frac{c \log n}{\epsilon^2}$ we have,

$$\Pr[\left|\sqrt{z_1^2 + z_2^2 + \cdots + z_k^2} - \sqrt{\frac{k}{d}}\right| > \epsilon\sqrt{\frac{k}{d}}] \leq e^{-\epsilon^2 k} = e^{-c \log n} = \frac{1}{n^c} \tag{7}$$

Proof of Lemma 1: The lemma asserts that all the $\binom{n}{2}$ pairwise distances $|x_i - x_j|$ are preserved by random projection. Equation (3) shows that it is suffices to bound the probability that the length of the projection of a random unit vector lies in $(1 \pm \epsilon)\sqrt{\frac{k}{d}}$ to show that the distance $|x_i - x_j|$ is preserved. Equation (7) shows that the projection of a random unit vector fails to lie in $(1 \pm \epsilon)\sqrt{\frac{k}{d}}$ with probability $\frac{1}{n^c}$. Using the union bound we conclude that the probability of some pairwise distance not being preserved is bounded by $\frac{\binom{n}{2}}{n^c} = \frac{1}{n^{c-2}}$.

1.1 Locality preserving hashing

One application of the Johnson Lindenstrauss lemma is locality preserving hashing. Given geometric data we want to hash the data such that points that are close in Euclidean distance have the same hash value. The setting makes sense for applications where data points are noisy and a hash function should map x and $x + \delta$ to the same bucket. The idea is to project onto a random low dimensional space and divide the low dimensional space into cells defined by a grid with $h(x)$ being the label of the cell into which the projection of x falls.

The distances between points are preserved approximately so far off points are not hashed to the same bucket. Data can be hashed with multiple hash functions to boost accuracy.

1.2 Implementing Johnson Lindenstrauss

The method for choosing a random k dimensional subspace required a large number of random bits, it is natural to ask if a random projection can be implemented with a smaller amount of randomness. Instead of choosing a random k dimensional subspace we choose k random ± 1 vectors from the hypercube $\{-1, 1\}^d$, we do not care about orthogonality as random vectors are close to orthogonal. The argument will be sketched below, refer to [**?**] for details.

Consider the random variable $C_1 = \frac{1}{\sqrt{d}} \sum_{i \in d} b_i z_i$ representing the inner product of a unit vector z with uniformly random vector b from the hypercube $\{-1, 1\}^d$. The expectation $E[C_1^2] = \frac{1}{d}$ as the coordinates of b are independent random variables. The expected length of the projection of z onto the span of k vectors from the hypercube is $E[C] := E[\sum_{i \in [k]} C_i^2] = k/d$, and a concentration result similar to (6) would suffice to prove

the the Johnson Lindenstrauss lemma,

$$\Pr\left[|C - \frac{k}{d}| \geq \epsilon \frac{k}{d}\right] \leq e^{-\epsilon^2 k} \tag{8}$$

The Chernoff bounds are a popular method for establishing concentration results for sums of independent random variables,

THEOREM 3
Chernoff bound: If $X_1, X_2, \cdots, X_n$ are independent random variables such that $0 \leq X_i \leq 1$, $X = \sum_i X_i$ and $E[X] = \mu$,

$$\Pr[|X - \mu| \geq \epsilon\mu] \leq e^{-\epsilon^2 \mu/3} \tag{9}$$

Applying the Chernoff bound for the sum of k independent random variables $\sum_{i\in[k]} C_i^2$ we have,

$$\Pr[C \geq \frac{k}{d}(1 \pm \epsilon)] \leq e^{-\epsilon^2 k/3d} \tag{10}$$

The concentration result obtained using the Chernoff bound depends on d and does not establish the Johnson Lindenstrauss lemma. Concentration bounds (6) and (8) follow from analogs of tail bounds on the χ^2 distribution, while the Chernoff bounds are an analog of tail bounds for the normal distribution.