Today

Johnson-Lindenstrass

Points: $x_1,\dots,x_n \in \mathbb{R}^d$.

Random $k = \frac{c\log n}{\varepsilon^2}$ dimensional subspace.

Claim: with probability $1-\frac{1}{n^{c-2}}$,

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

"Projecting and scaling by $\sqrt{\frac{d}{k}}$ preserves all pairwise distances w/in factor of $1\pm\varepsilon$."

Random subspace.

Method 1:
Pick unit v_1 ,
v_2 orthogonal to v_1,
...
v_k orthogonal to previous vectors...

Method 2:
Choose k vectors $v_1,\dots,v_k$
Gram Schmidt orthonormalization of $k\times d$ matrix where rows are v_i.
remove projection onto previous subspace.

Projections.

Project x into subspace spanned by $v_1, v_2, \cdots, v_k$.

$y_1 = x\cdot v_1, y_2 = x\cdot, v_2, \cdots, y_k = x\cdot v_k$

Projection: $(y_1,\dots,y_k)$.

Have: Arbitrary vector, random k-dimensional subspace.

View As: Random vector, standard basis for k dimensions.

Orthogonal U - rotates $v_1,\dots,v_k$ onto $e_1,\dots,e_k$

$y_i = \langle v_i|x\rangle = \langle Uv_i|Ux\rangle = \langle e_i|Ux\rangle = \langle e_i|z\rangle$

Inverse of U maps e_i to random vector v_i and $U^{-1}=U$.

$z = Ux$ is uniformly distributed on d sphere for unit $x\in\mathbb{R}^d$.

y_i is ith coordinate of random vector z.

Expected value of y_i.

Random projection: first k coordinates of random unit vector, z_i.

$E[\sum_{i\in[d]} z_i^2] = 1$. Linearity of Expectation.

By symmetry, each z_i is identically distributed.

$E[\sum_{i\in[k]} z_i^2] = \frac{k}{d}$. Linearity of Expectation.

Expected length is $\sqrt{\frac{k}{d}}$.

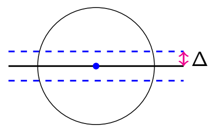
Johnson-Lindenstrass: close to expectation.
k is large enough $\rightarrow$
$\approx (1\pm\varepsilon)\sqrt{\frac{k}{d}}$ with decent probability.

Concentration Bounds.

z is uniformly random unit vector.
Random point on the unit sphere. $E[\sum_{i\in[k]} z_i^2] = \frac{k}{d}$.

Claim: $\Pr[|z_1| > \frac{t}{\sqrt{d}}] \le e^{-t^2/2}$

Sphere view: surface "far" from equator defined by e_1.



$|z_1| \ge \Delta$ if
$z \ge \Delta$ from equator of sphere.
Point on "Δ-spherical cap".

Area of caps
$\le$ S.A. of sphere of radius $\sqrt{1-\Delta^2}$
$\propto r^d = (1-\Delta^2)^{d/2}$
$\propto \left(1-\frac{t^2}{d}\right)^{d/2} \approx e^{-t^2}2d$
Constant of $\propto$ is unit sphere area. □

$\Pr[\text{any } z_i^2 > \sqrt{2\log d}E[z_i^2]]$ is small.

Many coordinates.

Proved Pr[any $z_i^2 > \sqrt{2\log d}E[z_i^2]$] is small.

Length? $z = z_1^2 + z_2^2 + \cdots z_k^2$.

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

Substituting $t = \varepsilon\sqrt{\frac{k}{d}}$, $k = \frac{c\log n}{\varepsilon^2}$.

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

Johnson-Lindenstraus: For n points, $x_1, \ldots, x_n$, all distances preserved to within $1 \pm \varepsilon$ under $\sqrt{\frac{k}{d}}$-scaled projection above.

View one pair $x_i - x_j$ as vector.
Scale to unit.
Projection fails to preserve $|x_i - x_j|$
with probability $\le \frac{1}{n^c}$
Scaled vector length also preserved.

$\le n^2$ pairs plus union bound
$\rightarrow$ prob any pair fails to be preserved with $\le \frac{1}{n^{c-2}}$.

Locality Preserving Hashing

Find nearby points in high dimensional space.
Points could be images!

Hash function $h(\cdot)$ s.t. $h(x_i) = h(x_j)$ if $d(x_i, x_j) \le \delta$.

Low dimensions: grid cells give $\sqrt{d}$-approximation.
Not quite a solution. Why?
Close to grid boundary.
Find close points to x:
Check grid cell and neighboring grid cells.

Project high dimensional points into low dimensions.

Use grid hash function.

Implementing Johnson-Lindenstraus

Random vectors have many bits

Use random bit vectors: $\{-1,+1\}^d$ instead.

Almost orthogonal.

Project z.

Coordinate for bit vector b.

$C_i = \frac{1}{\sqrt{d}}\sum_i b_i z_i$

$E[C_i^2] = E[\frac{1}{d}\sum_{i,j} b_i b_j z_i z_j] = \frac{1}{d}\sum_{i,j} E[b_i b_j] z_i z_j = \frac{1}{d}\sum_i z_i^2 = \frac{1}{d}$

$E[\sum_i C_i^2] = \frac{k}{d}$

Binary Johnson-Lindenstrass

Project onto $[-1,+1]$ vectors.

$E[C] = E[\sum_i C_i^2] = \frac{k}{d}$

Concentration?

$$\Pr\left[|C - \frac{k}{d}| \ge \varepsilon\frac{k}{d}\right] \le e^{-\varepsilon^2 k}$$

Choose $k = \frac{c\log n}{\varepsilon^2}$.
$\rightarrow$ failure probability $\le 1/n^c$.

Analysis Idea.

$$\Pr\left[|C - \frac{k}{d}| \ge \varepsilon\frac{k}{d}\right] \le e^{-\varepsilon^2 k}$$

Variance of C_i^2? $\left(\frac{k}{d^2}\right)(\sum_i z_i{}^4 + 4\sum_{i,j} z_i^2 z_j^2) \le \left(\frac{k}{d^2}\right)2(\sum_i z_i^2)^2 \le \frac{2k}{d^2}$.

Roughly normal (gaussian):
Density $\propto e^{-t^2}/2$ for t std deviations away.

So, assuming normality
$\sigma = \frac{\sqrt{k}}{d}$, $t = \frac{\varepsilon\frac{k}{d}}{\frac{\sqrt{2k}}{d}} = \varepsilon\sqrt{k}/\sqrt{2}$.

Probability of failure roughly $\le e^{-t^2/2}$
$\rightarrow e^{\varepsilon^2 k/4}$

"Roughly normal." Chernoff, Berry-Esseen, Central Limit Theorems.

Sum up

Have a good break!