Chapter 2.
High Dimensional space

Gadgets for Proving:

The Law of Large Numbers $$\text{Prob}\left(\left|\frac{x_1 + x_2 + \cdots + x_n}{n} - E(x)\right| \geq \epsilon\right) \leq \frac{Var(x)}{n\epsilon^2}.$$

Theorem 2.1 (Markov's inequality) *Let x be a nonnegative random variable. Then for $a > 0$,*

$$Prob(x \geq a) \leq \frac{E(x)}{a}.$$

Theorem 2.3 (Chebyshev's inequality) *Let x be a random variable. Then for $c > 0$,*

$$Prob\Big(|x - E(x)| \geq c\Big) \leq \frac{Var(x)}{c^2}.$$

Statistical facts about expectation, variance etc.

Important Geometry Conclusions about High Dim Space

- Most volume is near surface

$$\frac{\text{volume}\big((1-\epsilon)A\big)}{\text{volume}(A)} = (1-\epsilon)^d \leq e^{-\epsilon d}.$$

Most of the volume of the d-dim ball of radius r is contained in an annulus of width O(r/d) near surface

- Surface area and volume of a unit sphere in d:

$$A\,(d) = \frac{2\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2})} \quad \textit{and} \quad V\,(d) = \frac{2\pi^{\frac{d}{2}}}{d\,\Gamma(\frac{d}{2})}.$$

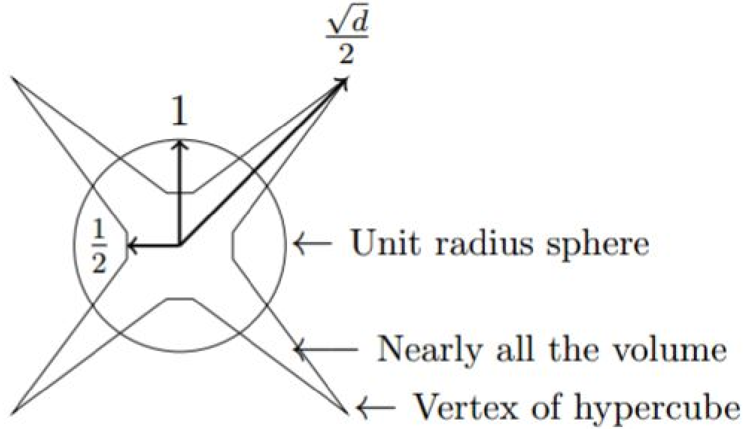
- Volume of sphere is near equator

Theorem 2.7 *For $c \geq 1$ and $d \geq 3$, at least a $1 - \frac{2}{c}e^{-c^2/2}$ fraction of the volume of the d-dimensional unit ball has $|x_1| \leq \frac{c}{\sqrt{d-1}}$.*

Prove by showing the ratio of volume of half equator over volume of hemisphere goes to one in limit or the other way around.

Important Geometry Conclusions about High Dim Space

- Nearly all points in a unit sphere are in a box of side-length $O\left(\frac{\ln d}{d-1}\right)$



Generat points uniformly at random on surface of unit sphere

1. For a point Generate x1, x2, . . . , xd independently using a zero mean, unit variance Gaussian distribution.

   This gives a probabiliy density that is spherically symmetric.

2. Normalize the vector x=<x1, x2, …., xd> so that the point is on surface of unit sphere. (coordinates no longer independent)

To generate points uniformly at random over a unit sphere:

   Introduce scale factor dr^{d-1}

Gaussians in High Dim

- Two random points from a d-dimensional Gaussian with unit variance in each direction are approximately orthogonal. It implies: for the normalized random points, if we pick one as north pole, then all the others will lie on the equator of the unit sphere.
- Gaussian Annulus Theorem

Theorem 2.9 (Gaussian Annulus Theorem) *For a d-dimensional spherical Gaussian with unit variance in each direction, for any* $\beta \leq \sqrt{d}$*, all but at most* $3e^{-c\beta^2}$ *of the probability mass lies within the annulus* $\sqrt{d} - \beta \leq |\mathbf{x}| \leq \sqrt{d} + \beta$*, where c is a fixed positive constant.*

Random Projection

The projection $f : \mathbf{R}^d \to \mathbf{R}^k$ that we will examine (in fact, many related projections are known to work as well) is the following. Pick k Gaussian vectors $\mathbf{u_1}, \mathbf{u_2}, \ldots, \mathbf{u_k}$ in $\mathbf{R}^d$ with unit-variance coordinates. For any vector $\mathbf{v}$, define the projection $f(\mathbf{v})$ by:

$$f(\mathbf{v}) = (\mathbf{u_1} \cdot \mathbf{v}, \mathbf{u_2} \cdot \mathbf{v}, \ldots, \mathbf{u_k} \cdot \mathbf{v}).$$

WIth high probabiliy, $|f(\mathbf{v})| \approx \sqrt{k}|\mathbf{v}|$

Random projection preserves all relative pairwise distances between points in a set of n points with high probability

Theorem 2.11 (Johnson-Lindenstrauss Lemma) *For any $0 < \varepsilon < 1$ and any integer n, let $k \geq \frac{3}{c\varepsilon^2} \ln n$ for c as in Theorem 2.9. For any set of n points in R^d, the random projection $f : R^d \to R^k$ defined above has the property that for all pairs of points $\mathbf{v_i}$ and $\mathbf{v_j}$, with probability at least $1 - 1.5/n$,*

$$(1-\varepsilon)\sqrt{k}\,|\mathbf{v_i} - \mathbf{v_j}| \leq |f(\mathbf{v_i}) - f(\mathbf{v_j})| \leq (1+\varepsilon)\sqrt{k}\,|\mathbf{v_i} - \mathbf{v_j}|\,.$$

Separating Gaussians

Algorithm for separating points from two Gaussians: Calculate all pairwise distances between points. The cluster of smallest pairwise distances must come from a single Gaussian. Remove these points. The remaining points come from the second Gaussian.