**Review: High Dimensional Data**

For high dimensions, almost all the volume of the unit cube is outside the unit sphere.

How far apart can 2 points be on a:

1) sphere: $2$
2) cube: $2\sqrt{d}$

All the probability will be found in the shaded area (the annulus).

Suppose points are placed at random and we want to calculate the distance between points.

$$x = (x_1, x_2, \ldots, x_d)$$
$$y = (y_1, y_2, \ldots, y_d)$$

$$\text{dist}^2(x, y) = \sum_{i=1}^{d} (x_i - y_i)^2$$

Deviation of sum of random variables from expected value of sum

$$\Pr \left( \left| \sum_{i=1}^{n} x_i - E(\sum_{i=1}^{n} x_i) \right| \geq c \right) \leq e^{-\frac{2c^2}{\sigma^2}}$$

**Another Problem:**

How do you generate points at random on the surface of a sphere?

Possible Solution:

In 2-dimensions, you might try to generate points uniformly on a square (i.e. rand function in Matlab).

Solution:

Discard all points outside of circle and project remaining points onto surface.

Why does this method not work in high dimensions?

Most of the area of a hypercube will lie outside the sphere.
Generate points according to the following distribution:

\[(x_1, x_2, \ldots, x_d) \quad e^{-\frac{x_1^2 + x_2^2 + \ldots + x_d^2}{2}} = e^{-\frac{r^2}{2}}\]

Then normalize the points:

\[
x_1 \quad \frac{x_2}{\sqrt{x_1^2 + x_2^2 + \ldots + x_d^2}} \quad \ldots \quad \frac{x_d}{\sqrt{x_1^2 + x_2^2 + \ldots + x_d^2}}
\]

**Now we want to know:**

Generate two points on unit sphere.

After generating first point, rotate coordinate to place it on North Pole.

Generate 2\(^{nd}\) point.

\[
dist^2 = (1 - \frac{x_1}{\sqrt{x_1^2 + x_2^2 + \ldots + x_d^2}})^2 + \left(\frac{x_2^2}{x_1^2 + x_2^2 + \ldots + x_d^2}\right) + \ldots + \left(\frac{x_d^2}{x_1^2 + x_2^2 + \ldots + x_d^2}\right)
\]

\[
= 1 - \frac{2x_1}{\sqrt{x_1^2 + x_2^2 + \ldots + x_d^2}} + \left(\frac{x_1^2}{x_1^2 + x_2^2 + \ldots + x_d^2}\right) + \left(\frac{x_2^2}{x_1^2 + x_2^2 + \ldots + x_d^2}\right) + \ldots + \left(\frac{x_d^2}{x_1^2 + x_2^2 + \ldots + x_d^2}\right)
\]

\[
= 2 - \frac{2x_1}{\sqrt{x_1^2 + x_2^2 + \ldots + x_d^2}} = 2 - 2x_1
\]

**Clarification Question: Why did we rotate in this manner?**

We wanted to make things easier to calculate the distance; we simply rotated the coordinate system.

\[
E(dist^2) = 2
\]

\[
E(dist) = \sqrt{2}
\]

Points on average are perpendicular.

Gaussian distribution used here. Depending on what distribution is used, there may or may not be an annulus.

Two Gaussians, points would be on 2 annuluses.

Pick 2 random points and calculate the distance between them.
Let $\delta = \text{distance between vectors}$
Distance between points $= \sqrt{d^2 - 2}$

**Question: What if spheres of radius $\sqrt{d}$?**

If two points generated by some Gaussian, they will be $\sqrt{2d}$ distance apart.
If two points generated by different Gaussians, they will be $\sqrt{\delta^2 + 2d}$

To determine which Gaussian generated, calculate all pairwise distances and compare distance to $\sqrt{2d}$ or $\sqrt{\delta^2 + 2d}$

If $\sqrt{\delta^2 + 2d} \geq \sqrt{2d} + c$

\[ \sqrt{\delta^2 + 2d} = \sqrt{2d} \left( \sqrt{1 + \frac{\delta^2}{4d}} \right) = \sqrt{2d} \left( 1 + \frac{\delta^2}{4d} + \ldots \right) \geq \sqrt{2d} - c \]

\[ \sqrt{2d} \left( \frac{\delta^2}{4d} \right) \geq c \]

\[ \delta^2 \geq \frac{c \cdot 4d}{\sqrt{2} \sqrt{d}} \geq \frac{4c}{\sqrt{2}} \sqrt{d} \]

\[ \delta \geq c \cdot d^{1/4} \]

But we can do better than this.