Assignment - 1

Name : Gavali Deshabhakt Nagnath

Roll No. : 202CD005

Branch : CD (MACS-Dept)

Subject : Advanced Data Science

Q.1 State and Prove Markov's Inequality

→ Statement :-

Let x be a non-negative random variable. Then for $a > 0$,

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

Proof :-

For continuous non-negative r.v. x with probability density p,

$$E(x) = \int_0^{\infty} x p(x)\, dx$$

$$= \int_0^{a} x p(x)\, dx + \int_a^{\infty} x p(x)\, dx$$

$$\geq \int_a^{\infty} x p(x)\, dx \geq a \int_a^{\infty} p(x)\, dx$$

$$= a \,\text{Prob}(x \geq a)$$

$$\therefore \quad \boxed{\text{Prob}(x \geq a) \leq \frac{E(x)}{a}}$$

Hence proved.

Q.2 State and prove chebyshev's inequality

→ **Statement :-**

Let x be a random variable. Then for $c>0$,

$$P(|x-E(x)| \geq c) \leq \frac{Var(x)}{c}$$

Proof :-

Here, $P(|x-E(x)| \geq c)$ can also be written as $P(|x-E(x)|^2 \geq c^2)$.

Let, $y = |x-E(x)|^2$.

Note that y is a non-negative random variable & $E(y) = Var(x)$

So, markov's inequality can be

$$P(|x-E(x)| \geq c) = P(|x-E(x)|^2 \geq c^2)$$

$$\leq \frac{E(|x-E(x)|^2)}{c^2} = \frac{Var(x)}{c^2}$$

$$\therefore \boxed{P(|x-E(x)| \geq c) \leq \frac{Var(x)}{c}}$$

Q.3 State and prove Law of large Numbers

→ Statement :-

Let $x_1, x_2, \dots, x_n$ be n independent samples of a random var. x, then

$$P\left(\left|\frac{x_1+x_2+\dots+x_n}{n} - E(x)\right| \geq \varepsilon\right) \leq \frac{Var(x)}{n\varepsilon^2}$$

Proof :-

By chebyshev's inequality,

$$P\left(\left|\frac{x_1+x_2+\dots+x_n}{n} - E(x)\right| \geq \varepsilon\right) \leq \frac{Var\left(\frac{x_1+x_2+\dots+x_n}{n}\right)}{\varepsilon^2}$$

$$= \frac{1}{n^2\varepsilon^2} Var(x_1+x_2+\dots+x_n)$$

$$= \frac{1}{n^2\varepsilon^2}(V(x_1)+V(x_2)+\dots+V(x_n))$$

$$= \frac{Var(x)}{n^2\varepsilon^2}$$

$$\boxed{\therefore P\left(\left|\frac{x_1+x_2+\dots+x_n}{n} - E(x)\right| \geq \varepsilon\right) \leq \frac{Var(x)}{n^2\varepsilon^2}}$$

Hence proved.

Q.4 Show that most of the volume of unit ball in R^d is contained in an annulus of width $O(1/d)$ near the boundry if 'd' is large.

→ Consider any object A in R^d. Now shrink A by small amount ε to produce new object

$$(1-\varepsilon)A = \{(1-\varepsilon)x \mid x \in A\}.$$

Then the following equality holds:

$$\text{Volume}((1-\varepsilon)A) = (1-\varepsilon)^d \text{ volume}(A)$$

To see that this is true, partition A into infinitesimal cubes. Then $(1-\varepsilon)A$ is the union of a set of cubes obtained by shrinking the cubes in A by a factor $(1-\varepsilon)$.

When we shrink each of 2d sides of a d-dimensional cube by a factor f, its volume shrinks by a factor f^d.

Using the fact that $1-x \le e^{-x}$, for any object in R^d we have:

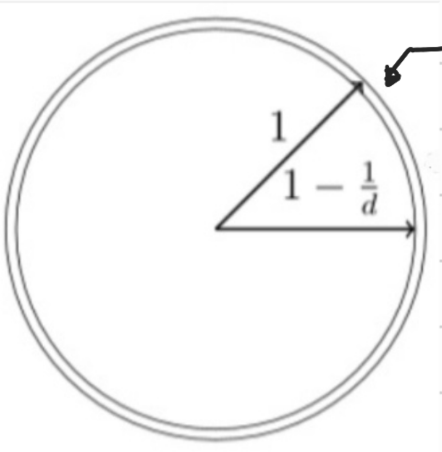
$$\boxed{\frac{\text{Volume}((1-\varepsilon)A)}{\text{Volume}(A)} = (1-\varepsilon)^d \le e^{-\varepsilon d}}$$

Fixing ε and letting $d \to \infty$, the above qty. rapidly approaches 0. This means that nearly all of the volume of A must be in portion of A that does not belong to the region $(1-\varepsilon)A$.

Let S denote the unit ball in d-dimension. that is, the set of points within distance one of the origin.

An immediate implication of above observation is that atleast a $1-e^{-\varepsilon d}$ fraction of volume of unit ball is concentrated in $S \setminus (1-\varepsilon)S$, namely in a small annulus of width ε at the boundary.

In particular, most of the volume of the d-dimensional unit ball is contained in an annulus of width $O(1/d)$ near boundry.



Annulus of width $O(1/d)$

Note:- If radius of ball is r, then width of annulus will be $O(r/d)$

Q.5 Derive formulas for the volume and surface area of unit ball in $\mathbb{R}^d$. Also show that both $\to 0$ as $d \to \infty$

→ Volume of the unit ball :-

The volume, $V(d)$, of unit ball in cartesian co-ordinate system is given by

$$V(d) = \int_{x_1=-1}^{x_1=1} \int_{x_2=-\sqrt{1-x_1^2}}^{x_2=\sqrt{1-x_1^2}} \cdots \int_{x_d=-\sqrt{1-x_1^2-\cdots-x_{d-1}^2}}^{x_d=\sqrt{1-x_1^2-\cdots-x_{d-1}^2}} dx_d \ldots dx_2 dx_1$$

As, the limits in cartesian co-ordinate system are complicated, we will follow polar co-ordinate system.

$\therefore$ $V(d)$ in polar co-ordinate system is,

$$V(d) = \int_{S^d} \int_{r=0}^{1} r^{d-1} \, dr \, d\Omega$$

Since variables Ω & r donot interact we can separate integrals.

i.e. $V(d) = \int_{S^d} V(d) \int_{r=0}^{1} r^{d-1} dr = \frac{1}{d} \int_{S^d} d\Omega$

$$\therefore \quad \boxed{V(d) = \frac{A(d)}{d}}$$

where $A(d)$ is surface area of d-dimensional unit ball.

For instance, for 3-dimensional unit ball $(d=3)$,

$$\boxed{A(d) = 4\pi \quad \& \quad V(d) = \frac{4\pi}{3}}$$

Surface area of unit ball :-

To calculate $A(d)$, consider diff^n integral

$$I(d) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{-(x_1^2 + x_2 + \cdots + x_d)^2} \, dx_d \cdots dx_2 \, dx_1$$

Including exponential allows integration to infinity rather than stopping at the surface of the sphere.
Thus, $I(d)$ can be computed in both cartesian and polar co-ordinates.
Integrating in polar co-ordinates will relate $I(d)$ to $A(d)$. & Equating 2-results for $I(d)$ allows one to solve for $A(d)$

In cartesian co-ordinate system,

$$I(d) = \left[\int_{-\infty}^{\infty} e^{-x^2} dx\right]^d = (\sqrt{\pi})^d = \pi^{d/2}$$

$$\therefore \; I(d) = \pi^{d/2} \quad - (1)$$

& In polar co-ordinate system,

$$I(d) = \int_{S^d} d\Omega \int_0^{\infty} e^{-r^2} r^{d-1} dr.$$

Here, integral $\int_{S^d} d\Omega$ is integral over entire solid angle and gives surface area $A(d)$ of a unit sphere.

Thus,

$$I(d) = A(d) \int_0^{\infty} e^{-r^2} r^{d-1} dr. \quad - (2)$$

Now,

let, $I = \int_0^{\infty} e^{-r^2} r^{d-1} dr.$

put $r^2 = t \Rightarrow r = \sqrt{t}$

limits.

r	0	∞
t	0	∞

$$\therefore \; dr = \frac{1}{2\sqrt{t}} dt$$

$$\therefore I = \int_0^{\infty} e^{-t}\, t^{\frac{d-1}{2}} \left(\frac{1}{2\sqrt{t}}\right) dt$$

$$\therefore I = \frac{1}{2} \int_0^{\infty} e^{-t}\, t^{\frac{d}{2}-1}\, dt$$

$$\therefore I = \frac{1}{2} \Gamma \frac{d}{2}$$

$\therefore$ eq^n (2) becomes,

$$I(d) = A(d) \times \frac{1}{2} \Gamma d/2 \qquad -(3)$$

Solving eq^n (1) & (3) we get,

$$\boxed{A(d) = \frac{2\pi^{d/2}}{\Gamma d/2}}$$

&

$$\therefore V(d) = \frac{A(d)}{d}$$

$$\therefore \boxed{V(d) = \frac{2\pi^{d/2}}{d\, \Gamma d/2}}$$

from formulas, it can be seen that as $d \to \infty$, both $A(d)$ & $r(d) \to 0$ because,

as d increases; the denominator term gets bigger & bigger than the numerator term & Hence the ratio approaches to zero.

Q.6 Show that most of the volume of unit ball in R^d is near equator, if d is large.

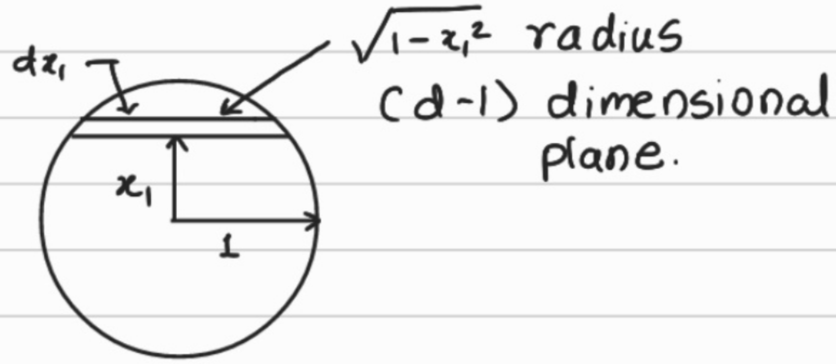
→ Consider a high-dimensional unit ball & fix north pole on the x_1-axis at $x_1=1$ Divide the ball in half by intersecting it with plane $x_1=0$. The intersection of the plane with ball forms a region of one lower dimension, namely $\{x \mid |x| \leq 1, x_1=0\}$ which we call equator. The intersection is a sphere of dimension d-1 and has volume $V(d-1)$. In 3-dimensions this region is circle, in four dimensions the region is a 3-dimensional sphere/ball, etc.
In general, the intersection is a ball of 'd-1'-dimension.
It turns out that essentially all of the mass of upper hemisphere lies between the plane $x_1=0$ & a parallel plane $x_1=\varepsilon$, that is slightly higher. For what value of ε does essentially all the mass lie between $x_1=0$ & $x_1=\varepsilon$?
The answer depends on the dimension.

For dimension d it is $O(1/\sqrt{d-1})$. To see this, calculate the volume of portion of the ball above the slice lying between $x_1=0$ & $x_1=\varepsilon$.

Let $T=\{x \mid |x| \leq 1, x_1 \geqslant \varepsilon\}$ be the portion of the sphere above the slice. To calculate volume of T, integrate over x_1 from ε to 1.

The incremental volume is a disk of width dx_1 whose face is a sphere of dimension $d-1$ of radius $\sqrt{1-x_1^2}$ & therefore, the surface area of disk is

$$(1-x_1^2)^{\frac{d-1}{2}} V(d-1)$$



volume of cross sectional slab of a d-dimensional ball

Thus,

$$\text{Volume}(T) = \int_{\varepsilon}^{1} (1-x_1^2)^{\frac{d-1}{2}} V(d-1)\, dx_1$$

$$= V(d-1) \int_{\varepsilon}^{1} (1-x_1^2)^{\frac{d-1}{2}}\, dx_1$$

Note that $V(d)$ denotes the volume of the d-dimensional unit ball. For the volume of other sets such as the set T, we use the notation Volume (T) for the volume. The above integral is difficult to evaluate so we use some approximations.

First, we use the inequality $1+x \leq e^x$ for all real x & change the upper bound on integral to be infinity. Since x_1 always greater than ε over the region of integration we can insert x_1/ε in the integral. This gives,

$$\text{Volume}(T) \leq V(d-1) \int_{\varepsilon}^{\infty} e^{-\frac{d-1}{2}x_1^2} \cdot dx_1$$

$$\leq V(d-1) \int_{\varepsilon}^{\infty} \frac{x_1}{\varepsilon} e^{-\left(\frac{d-1}{2}\right)x_1^2} dx_1$$

Now,

$$\int x_1 e^{-\frac{d-1}{2}x_1^2} dx_1 = -\frac{1}{d-1} e^{-\frac{d-1}{2}x_1}$$ & hence,

$$\text{Volume}(T) \leq \frac{1}{\varepsilon(d-1)} e^{-\frac{d-1}{2}\varepsilon^2} V(d-1)$$

Next, we lower bound the volume of the entire upper hemisphere. Clearly the volume of upper hemisphere is atleast the volume between the slabs $x_1=0$ &

$x_1 = \frac{1}{\sqrt{d-1}}$, which is at least the volume

of cylinder of radius $\sqrt{1-\frac{1}{d-1}}$ & height

$\frac{1}{\sqrt{d-1}}$.

The volume of cylinder is $1/\sqrt{d-1}$ times the $d-1$ dimensional volume of the disk

$R = \{x \mid |x| \leq 1;\ x_1 = 1/\sqrt{d-1}\}$.

Now, R is a $d-1$-dimensional ball of radius $\sqrt{1-\frac{1}{d-1}}$ and so its volume is,

$$\text{Volume}(R) = V(d-1)\left(1-\frac{1}{d-1}\right)^{(d-1)/2}$$

Using $(1-x)^{\alpha} \geq 1-\alpha x$,

$$\therefore \text{Volume}(R) \geq V(d-1)\left(1-\frac{1}{d-1}\cdot\frac{d-1}{2}\right)$$

$$\therefore \boxed{\text{Volume}(R) \geq \frac{1}{2} V(d-1)}$$

Thus, the volume of upper hemisphere is at least $\frac{1}{2\sqrt{d-1}} V(d-1)$. The fraction of the volume above the plane $x_1 = \varepsilon$ is upper bounded by the ratio of upper bound on the volume of the hemisphere above the plane $x_1 = \varepsilon$ to the lower bound on the total volume. This ratio is

$$\frac{2}{\varepsilon\sqrt{d-1}} e^{-\left(\frac{d-1}{2}\right)\varepsilon^2}$$

Now, substitute, $\varepsilon = \frac{c}{\sqrt{d-1}}$

$\therefore$ The ratio becomes $\frac{2}{c} e^{-c^2/2}$.

Now, For $c \geq 2$, the fraction of the volume of the hemisphere above

$x_1 = \frac{c}{\sqrt{d-1}}$ is less than $e^{-2} \approx 0.14$

& for $c \geq 4$, the fraction is less than $0.5 \times e^{-8} \approx 0.0003$.

This essentially means, all the mass as well as volume of ball in $\mathbb{R}^d$ is in a narrow slice at the equator.

Q.7 State and prove Gaussian annulus thm.

Statement :-

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 probability mass lies within the annulus

$$\sqrt{d} - \beta \leq |x| \leq \sqrt{d} + \beta, \text{ where}$$

c is a fixed positive constant.

Proof :-

Let $x = (x_1, x_2, \dots, x_d)$ be a point selected from a unit variance Gaussian centered at the origin, and let $r = |x|$.

$\sqrt{d} - \beta \leq |y| \leq \sqrt{d} + \beta$ is equivalent to

$|r - \sqrt{d}| \geq \beta$. If $|r - \sqrt{d}| \geq \beta$, then multiplying both sides by $r + \sqrt{d}$ gives,

$$|r^2 - d| \geq \beta(r + \sqrt{d}) \geq \beta\sqrt{d}$$

So, it suffices to bound the probability that $|r^2 - d| \geq \beta\sqrt{d}$.

Rewrite $r^2 - d = (x_1^2 + \dots + x_d^2)$
$= (x_1^2 - 1) + \dots + (x_d^2 - 1)$

and perform a change of variables

as $y_i = x_i^2 - 1$.

We want to bound the probability that $|y_1 + \dots + y_d| \geq \beta\sqrt{d}$.

Notice that $E(y_i) = E(x_i^2) - 1 = 0$.

To apply tails theorem, we need to bound the s^{th} moment of y_i

For $|x_i| \leq 1$, $|y_i|^s \leq 1$ and for $|x_i| \geq 1$, $|y_i|^s \leq |x_i|^{2s}$.

Thus,

$$|E(y_i^s)| = E(|y_i|^s)$$

$$\leq E(1 + x_i^{2s}) = 1 + E(x_i^{2s})$$

$$= 1 + \sqrt{\frac{2}{\pi}} \int_0^\infty x^{2s} e^{-x^2/2} \cdot dx.$$

Using the substitution $2z = x^2$,

$$|E(y_i^s)| = 1 + \frac{1}{\sqrt{\pi}} \int_0^\infty 2^s z^{s-(1/2)} e^{-z} dz$$

$$\leq 2^s \cdot s!$$

/ using Gamma integral inequality

Since $E(y_i) = 0$, $Var(y_i) = E(y_i^2) \leq 2^2 \cdot 2 = 8$.

Unfortunately, we do not have

$|E(y_i^s)| \leq 8s!$ as required by tail's theorem. To fix this problem, perform one more change of variables, using

$\omega_i = y_i/2$. Then $var(\omega_i) \leq 2$ and

$|E(\omega_i^s)| \leq 2s!$ and our goal is now to bound the probabity that $|\omega_1 + \dots + \omega_d| \geqslant \beta\sqrt{d}/2$.

Applying tails theorem, where $\sigma^2 = 2$ and $n=d$ this occurs with probability less than or equal to $3e^{-\beta^2/96}$.

Q.8 Explain in detail random projection theorem.

Statement :-

Let v be a fixed vector in R^d and let $f: R^d \to R^k$. There exists constant $c>0$ such that for $\varepsilon \in (0,1)$,

$$\text{Prob}\left(\Big| |f(v)| - \sqrt{k}\,|v| \Big| \geq \varepsilon \sqrt{k}\,|v|\right) \leq 3e^{-ck\varepsilon^2}$$

where the probability is taken over the random draws of vectors u_i used to construct f.

Proof :-

By scaling both sides of the inner inequality by $|v|$, we may assume that $|v| = 1$.

The sum of independent normally distributed real variables is also normally distributed where the mean and variance are the sum of individual means and variances.

since $u_i \cdot v = \sum_{j=1}^{d} u_{ij} \cdot v_j$, the random variable

$u_i \cdot v$ has Gaussian density with zero mean and unit variance, in particular,

$$Var(u_i \cdot v) = Var\left(\sum_{j=1}^{d} u_{ij} \cdot v_j\right)$$

$$= \sum_{j=1}^{d} v_j^2 \cdot Var(u_{ij})$$

$$= \boxed{\sum_{j=1}^{d} v_j^2 = 1}$$

Since $u_1 \cdot v, u_2 \cdot v, \ldots, u_k \cdot v$ are independen- Gaussian random variables, $f(v)$ is a random vector from a k-dimensional spherical Gaussian with unit variance in each co-ordinate, and so the thm follows from the Gaussian Annulus theorem with d replaced by k.

The random projection theorem establishes that the probability of the length of projection of a single vector differeing significantly from its expected value is exponentially small in k, the dimension of the target subspace. By a union bound, the probability that any of $O(n^2)$ pairwise differences $|v_i - v_j|$ among n vectors $v_1, \ldots, v_n$ differs significantly from their expected values is small, provided $k \geq (3/c\varepsilon^2) \ln(n)$. Thus, this random projection preserves all relative pairwise distances between points in a set of n-points with high probability. This is the content of Johnson-Lindenstrauss Lemma.

Q.9 Explain the statement and proof of the Johnson-Lindenstrauss lemma.

Statement:-

For any $0 < \varepsilon < 1$ and any integer n, let $k \geq \frac{3}{c\varepsilon^2} \ln n$. with c as fixed positive constant. For any set of 'n' points in R^d, the random projection $f: R^d - R^k$ defined above has the property that for all pairs of points v_i and v_j, with probability at least $1 - (3/2n)$,

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

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

Proof:-

Applying the random projection th^m for any fixed v_i and v_j, the probability that $|f(v_i - v_j)|$ is outside the range

$$[(1-\varepsilon)\sqrt{k}\,|v_i - v_j|, (1+\varepsilon)\sqrt{k}\,|v_i - v_j|]$$

is at most $3e^{-ck\varepsilon^2} \leq 3/n^3$ for

$k \geq (3\ln n)/c\varepsilon^2$. Since there are

$\binom{n}{2} < n^2/2$ pairs of points, by the

union bound, the probability that any pair has a large distortion is less than $(3/2n)$.

It is important to note that the conclusion of above theorem asserts for all v_i and v_j, not just for most of them. The weaker assertion for most v_i & v_j is typically less useful since our algorithm for a problem such as nearest-neighbour search might return one of the bad points. A remarkable aspect of the theorem is that the number of dimensions in projection is only dependent logarithmically on n. Since k is often much less than d, this is called dimension reduction technique. In applications, the dominant term is typically the $1/\varepsilon^2$ term.

For nearest neighbour problem, if the database has n_1 points and n_2 queries are expected during the lifetime of algorithm take $n = n_1 + n_2$ & project database to random k-dimensional space, for k as in above theorem. On receiving a query, project the query to same subspace & compute nearby database points. The Johnson-Lindenstrauss Lemma says that with high probability this will yield the right answer whatever the query.

Note that the exponentially small in k-probability was useful here in making k only dependent on $\ln n$, rather than n.

Q.10 Derive in detail the algorithm for separating points from 2-gaussians in R^d if d-is large by calculating pairwise distances between them.

Algorithm :-

Calculate all pairwise distances between points. The cluster of smallest pairwise distances must come from single Gaussian. Remove these points. The remaining points come from second Gaussian.

Derivation :-

First consider just one spherical unit-variance Gaussian centered at origin. From Gaussian annulus theorem, most of its probability mass lies on an annulus of width $O(1)$ at radius $\sqrt{d}$. Also

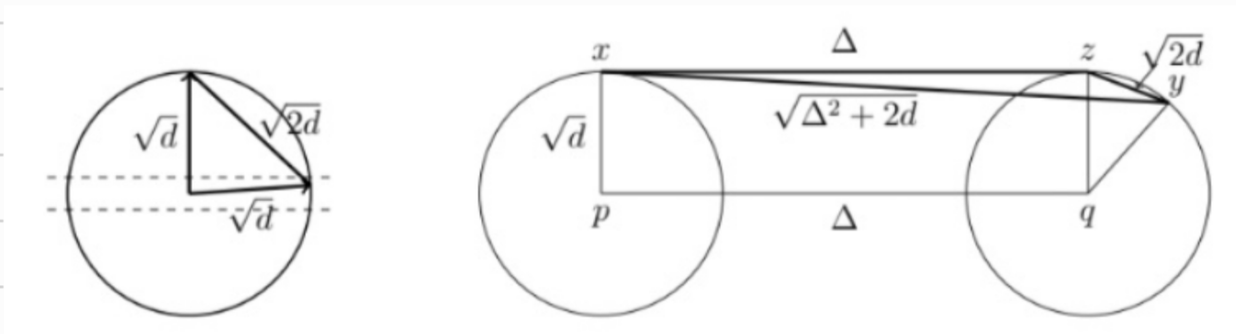
$$e^{-|x|^2/2} = \prod_i e^{-x_i^2/2}$$ and almost all of the

mass is within the slab $\{x \mid -c \leq x_1 \leq c\}$ for $c \in O(1)$. Pick a point x from this Gaussian. After picking x, rotate the concentrate system to make the first axis align with x. Independently pick a second point y from this Gaussian.

The fact that almost all of probability mass of the Gaussian is within the slab $\{x \mid -c \le x_1 \le c,\ c \in O(1)\}$ at equator implies that y's component along x's dir^n is $O(1)$ with high probability. Thus, y is nearly perpendicular to x.

So,

$$|x-y| \approx \sqrt{|x|^2 + |y|^2}$$



(a) (b)

(a) indicates that two randomly chosen points in high dimension are surely almost nearly orthogonal.

(b) Indicates the distance between a pair of random points from two different unit balls approximating the annuli of 2-Gaussians.

More precisely, since co-ordinate system has been rotated so that x is at north pole, $x = (\sqrt{d} \pm O(1), 0, \dots, 0)$. Since y is almost on the equator, further rotate the co-ordinate system so that the component of y that is perpendicular to the axis of north pole is in the 2^{nd} co-ordinate.

Then $y = (O(1), \sqrt{d} \pm O(1), 0, \dots, 0)$.

Thus,

$$(x-y)^2 = d \pm O(\sqrt{d}) + d \pm O(\sqrt{d})$$

$$\therefore \boxed{(x-y)^2 = 2d \pm O(\sqrt{d})}$$

and $|x-y| = \sqrt{2d} \pm O(1)$ with high probability.

Consider 2-spherical unit variance Gaussians with centers p & q separated by a distance Δ. The distance between a randomly chosen point x from the first Gaussian and randomly chosen point y from second is close to $\sqrt{\Delta^2 + 2d}$, since

$x-p$, $p-q$ & $q-y$ are nearly mutually $\perp$er.

Pick x & rotate the co-ordinate system so that x is at north pole. Let z be the North Pole of ball approximating the second Gaussian.

Now pick y. Most of the mass of 2nd Gaussian is within $O(1)$ of the equator $\perp$er to $z-q$. Also most of the mass of each Gaussian is within distance $O(1)$ of the respective equators $\perp$er to line $q-p$. Thus,

$$|x-y|^2 \approx \Delta^2 + |z-q|^2 + |q-y|^2$$

$$\therefore \boxed{|x-y|^2 = \Delta^2 + 2d \pm O(\sqrt{d})}$$

To ensure that the distance 2-poits picked from same Gaussian are closer to each other than two points picked from different Gaussians requires that the upper limit of distance between a pair of points from the same Gaussian is at most the lower limit of distance between points from different Gaussians. This requires that

$$\sqrt{2d} + O(1) \leq \sqrt{2d+\Delta^2} - O(1)$$

or

$$2d + O(\sqrt{d}) \leq 2d + \Delta^2$$

which holds when $\Delta \in \omega(d^{1/4})$. Thus, mixtures of spherical Gaussians can be separated in this way, provided their centers are separated $\omega(d^{1/4})$.

If we have n points and want to correctly separate all of them with high probability we need our individual high-probability statements to hold with probability

$1 - 1/\text{poly}(n^3)$, which means our $O(1)$ terms from Gaussian annulus theorem become $O(\sqrt{\log n})$. So we need to include an extra $O(\sqrt{\log n})$ term in the separation distance.