IE 531 Algorithms for Data Analytics H.W.2.

KARTHIK VENKATA
kvn3

1.(a.) For a given discrete Random Variable,

$X \in \{-1, 0, 1\}$

$Prob(X=-1) = Prob(X=1) = \frac{1}{2k^2}$

$E(X) = -1 \times \frac{1}{2k^2} + 0 + 1 \times \frac{1}{2k^2} = 0 = \sum p(x_i) x_i$

~~Var(X) = E(X)(X-E(X))~~

$Var(X) = E[(X-\mu)^2] = \frac{1}{2k^2} \times 1 + \frac{1}{2k^2} \times 1 = \frac{1}{k^2}$

The chebyshev bound by observation 1.2 (Notes)

is $Prob\{|X - E(X)| \geq \delta\} \leq \frac{Var\{X\}}{\delta^2}$

$\Rightarrow Prob(|X| \geq 1) \leq \frac{1}{k^2}$ $\delta = 1$

Hence $Prob(|X| \geq 1) = \frac{1}{k^2}$

(b.) We can start with an assumption of a continuous distribution with Mean $\mu = 0$, Standard deviation $\sigma = 1$. Here, we need $P(|X| > x) = \frac{1}{x^2}$. For our case, we can consider $x > 0$, Negative Values being symmetric. Cumulative distribution Function of distribution is $1 - \frac{1}{x^2}$.

~~Pdf~~ Pdf $= d(1 - \frac{1}{x^2}) = \frac{2}{x^3}$ $x > 0$

For this distribution to follow through with the hypothesis that chebyshev bound is tight, we need to have a finite variance distribution which is not possible. This distribution is not valid. To get a continuous distribution over the real axis with a finite variance, E(X) must be finite and $E[X^2]$ is then undefined as it includes a logarithm behavior of an integral which is a part of it.

REF - Stack Exchange.

In this case, Chebyshev's Bound can't be satisfied.
For satisfaction of Chebyshev's bound,

$P(|X| > x) = x^{-(2+\varepsilon)}$ for small ε.

2. $B(1,d,4) = \{(x_1, x_2, \ldots, x_d) \in \mathbb{R}^d \mid x_1^4 + x_2^4 + \cdots + x_d^4 \leq 1\}$

(a.) $S = \{(x_1, x_2, \ldots, x_d) \in \mathbb{R}^d \mid x_1^4 + x_2^4 + \cdots + x_d^4 \leq \frac{1}{2}\}$

From the notes, $Vol(B(r,d)) = r^d \times Vol(B(1,d))$

In our case, $r = \left(\frac{1}{2}\right)^{1/4}$

$\Rightarrow Vol(B(r,d)) = Vol(S) = \left(\frac{1}{2}\right)^{d/4} Vol(B(1,d))$

This is a direct extension of the formula that has been derived in the lecture notes.

(b.) We can consider V_{d-1} to denote the volume of the l_4 norm in $(d-1)$ dimension.

We need to consider the portion of the ball with $|x_1| \geq \frac{c}{d^{1/4}}$ [R- Region]

For volume of R, consider a cylinder of width dx_1 and radius $\sqrt[4]{1-x_1^4}$ and hence it is a ball of dimension $d-1$ and radius $\sqrt[4]{1-x_1^4}$. Volume of R $= 2\int_{\frac{c}{d^{1/4}}} (1-x_1^4)^{\frac{d-1}{4}} V(d-1)\, dx_1$

$$\leq 2\int_{c/d^{1/4}}^{\infty} (1-x_1^4)^{\frac{d-1}{4}} V_{d-1} dx$$

$$\leq 2V_{d-1} \int_{c/d^{1/4}}^{\infty} \exp(-x^4(d-1)/4)\, dx$$

$$\leq 2V_{d-1} \int_{c/d^{1/4}}^{\infty} \frac{x^3}{(c/d^{1/4})^3} \exp(-x^4(d-1)/4)\, dx$$

$$= 2V_{d-1} \frac{d^{3/4}}{c^3} \frac{1}{d-1} \Big(-\exp(-x^4(d-1)/4)\Big]_{c/d^{1/4}}^{\infty}$$

$$\leq \frac{3V_{d-1}}{c^3(d-1)^{1/4}} \exp(-c^4/4) \quad \text{(For d being large)}$$

Now, we can construct a lower bound for the volume of K.

$$\mathrm{Vol}(K) = 2\int_0^1 (1-x^4)^{\frac{d-1}{4}} V_{d-1}\,dx$$

$$\geq 2V_{d-1}\int_0^{1/(d-1)^{1/4}} (1-x^4)^{d-1/4}\,dx$$

$$\geq \frac{2V_{d-1}}{(d-1)^{1/4}}\left(1-\frac{1}{d-1}\right)^{\frac{d-1}{4}} dx$$

$$\geq \frac{2V_{d-1}}{(d-1)^{1/4}}\left(1-\frac{1}{d-1}\frac{d-1}{4}\right)$$ For d being large, Approximation

$$\geq \frac{2V_{d-1}}{(d-1)^{1/4}}\times\frac{3}{4} \geq \frac{3V_{d-1}}{2(d-1)^{1/4}}$$

Hence, Fraction of volume of K outside the slab $|x_1| \leq c/d^{1/4}$ is at most

$$\frac{\mathrm{Vol}(R)}{\mathrm{Vol}(K)} = \frac{\dfrac{3V_{d-1}\exp(-c^4/4)\times 2}{c^3(d-1)^{1/4}}}{\dfrac{3V_{d-1}}{2(d-1)^{1/4}}} = \frac{\exp(-c^4/4)}{c^3}$$

Hence, Max. Fraction is $\frac{1}{c^3}e^{-c^4/4}$

3. (a.) The usual ~~[illegible]~~ scheme of random sample generation on unit sphere surface involves generating $x_1, \dots, x_d$ with a zero mean, and a unit gaussian variance i.e. $\frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ Normalizing this, we need to convert into a unit vector [$x = (x_1, \dots, x_d)$ to unit vector]

(Referred to Solutions Manual of Book)

Therefore, we have a distribution that is uniform over surface.

$$E[x] = 0$$

(b.) $$\mathrm{Var}[x_i] = E[(x_i - E(x_i))^2] = E[x_i^2] - E[x_i]^2$$

$$E[x_i^2] = \frac{1}{d} E\left[\sum_{i=1}^{d} x_i^2\right] = \frac{1}{d} \quad \text{[From symmetric considerations]}$$

$$\mathrm{Var}[x_i] = E[x_i^2] - E[x_i]^2 = \frac{1}{d} - 0 = \frac{1}{d}$$

(c.) Variance of real valued random variable $u^T x$ is $\sum_{i=1}^{d} u_i^2 E\{x_i^2\}$, compute Var & Std.dev.

$$\mathrm{Var}[u^T x] = E[(u^T x)^2] - E[u^T x]^2 = E[(u^T x)^2] \quad \because E[x_i] = 0 \ \forall i$$

$$= E[(u^T x)^2]$$

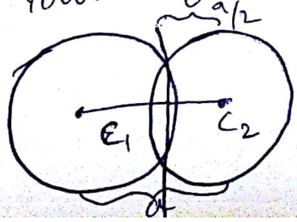
$$= \sum_{i,j} E[u_i u_j x_i x_j] = \frac{1}{d} \sum_i u_i^2 = \frac{1}{d}$$

$E[x_i x_j] = 0$ when $i \neq j$ and $E[x_i^2] = \frac{1}{d}$

Std-Deviation of $u^T x$ = $\sqrt{\mathrm{Var}[u^T x]} = 1/\sqrt{d}$

(d.) From Lemma 1.2 of the book,

$$\frac{\text{Volume of Intersection}}{\text{Volume of unit ball}} = 2 \times \text{Fraction of hemisphere above plane } x_1 = \frac{a}{2} \text{ (of a unit ball centered at origin.)}$$



which is maximum,

$$\frac{2\times 2}{\sqrt{d-1}\,a/2}\, e^{\left(-\frac{(a/2)^2(d-1)}{2}\right)} = \frac{8}{a\sqrt{d-1}}\, e^{\left(-\frac{a^2(d-1)}{8}\right)}$$

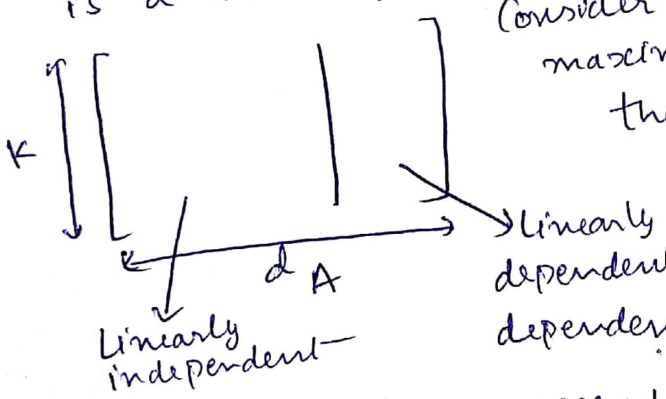
(e.) We can start by considering the case of

$a = \frac{c}{\sqrt{d-1}}$ ($c >> 1$, Assume radius $r = 1$),

$\frac{8}{c} e^{-\frac{c^2}{8}}$ would be the maximum value of the fraction of the intersection which is very small exponentially in c. For r-radius, we have,

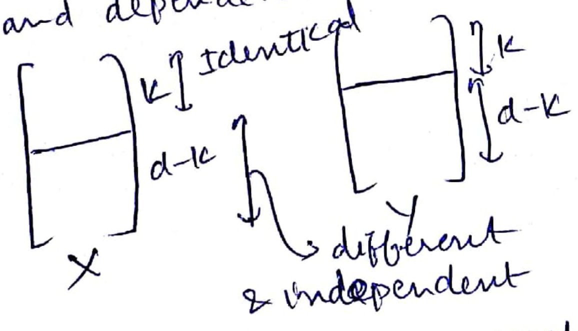
~~$\frac{Y}{1} = \frac{2r}{\epsilon\sqrt{d-1}} e^{-\frac{(d-1)}{2}\frac{\epsilon^2}{r^2}}$ $\epsilon = \frac{r}{2\sqrt{d}}$ [Half the distance of separation]~~

4. We know that every fixed dimension matrix $A \in R^{k \times d}$ is a matrix of maximum rank k.



Consider that the A matrix is of maximum rank K (full rank). If the rank is less than k,
We can divide A as shown, into a set of linearly independent and linearly dependent column vectors.

X can also be expressed as a set of independent and dependent vectors.



Upon multiplying both A & x and A & y; As some of the elements (k) are identical in X and Y (Assumption), Even AX and AY would point to the same image if the linearly dependent vectors were used for matrix multiplication. Hence, Although X and Y are different & (with non-identical X & Y) used for multiplication, we get he same mapping Ax and AY which are similar and point to the same image.

$k < d$, A has a non-trivial null space,

There are two vectors $\bar{x} \neq \bar{y}$ such that $Ax = Ay$.

Hence, $\|\bar{x} - \bar{y}\| \neq 0$ and $\|Ax - Ay\| = 0$ Hence,

Unbounded distortion.