Norm of a Gaussian vector

Consider $X \sim N(0, I_d)$. We want to understand the distribution of $\|X\|_2$.

Observation 1

$$\|X\|_2^2 = \sum_{i=1}^{d} X_i^2 \text{, where}$$

$$X_i \sim N(0,1) \text{ are iid.}$$

However, X_i^2 are not bounded, and not even sub-Gaussian (why should one not expect them to be sub-Gaussian?) so the usual Chernoff-Hoeffding bounds do not apply. (and are not even true).

However, it is true that $\|X\|_2$ concentrates around its expectation.

This is in fact a consequence of a general result called the Gaussian concentration inequality: If $f: \mathbb{R}^d \to \mathbb{R}$ is L-Lipschitz and $X \sim N(0, I_d)$ then $\forall t \geq 0$

$$\Pr\left[f(X) - Ef(X) \geq t \right] \leq \exp\left(-\frac{t^2}{2L^2}\right).$$

Ex: Show that this immediately implies

$$\Pr\left[\|X\| > \sqrt{d} + t \right] \leq \exp\left(-\frac{t^2}{2}\right)$$

[Note that $\|X\|$ indeed is sub-Gaussian, although $\|X\|^2$ is not]

We will use a less general direct approach. Note that we have

$$\Pr\left[\|X\|^2 \geq d + t \right] \leq \frac{E\left[\exp(aX_1^2)\right]^d}{\exp(a(d+t))} \quad \forall a \geq 0.$$

Now $E\left[\exp(aX_1^2)\right] = \frac{1}{\sqrt{1-2a}} \quad \forall a < 1/2$

[This is a direct calculation]

An aside : In 1-dimension, if $Z \sim N(0,1)$, then $\forall\, t > 0$

$$\Pr[Z > t] \leq \frac{1}{t\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right).$$

(Why ?).

This gives $\forall\, a \geq 0, t \geq 0$

$$\Pr\left[\|X\|^2 \geq d+t\right] \leq \exp\left(-a(d+t) - \frac{d}{2}\log(1-2a)\right)$$

The exponent is minimized at $a = \frac{t}{2(d+t)}$, so that we have

$$\Pr\left[\|X\|^2 \geq d+t\right] \leq \exp\left(-\frac{t}{2} + \frac{d}{2}\log\left(1+\frac{t}{d}\right)\right)$$

Using $\log(1+x) \leq x - \frac{x^2}{3}$ for $x \in [0, \frac{1}{2}]$, we get, for $t \in [0, \frac{d}{2}]$,

$$\Pr\left[\|X\|^2 \geq d+t\right] \leq \exp\left(-\frac{t^2}{6d}\right)$$

Similar calculations show that for $t \in [0, d]$,

$$\Pr\left[\|X\|^2 \leq d-t\right] \leq \exp\left(-\frac{t^2}{4d}\right)$$

In particular for $t = 6\sqrt{d\alpha \log K}$; s.t. $6\sqrt{\alpha \log K} \leq \frac{\sqrt{d}}{2}$

$$\Pr\left[\|X\|^2 \geq d + 6\sqrt{\alpha d \log K}\right] \leq K^{-6\alpha}.$$

and

$$\Pr\left[\|X\|^2 \leq d - 6\sqrt{\alpha d \log K}\right] \leq K^{-6\alpha}.$$

This also implies that (for $t \leq \frac{\sqrt{d}}{4}$)

$$\Pr\left[\|X\| \geq \sqrt{d} + t\right] \leq \exp\left(-\frac{t^2}{2}\right)$$

$$\Pr\left[\|X\| \leq \sqrt{d} - t\right] \leq \exp\left(-\frac{t^2}{2}\right).$$

Clustering Gaussians

We now consider the following problem: We have $2n$ points in $\mathbb{R}^d$, each sampled independently w.p. $1/2$ from $N(\mu_1, I_d)$ and with probability $1/2$ from $N(\mu_2, I_d)$ where $\mu_2, \mu_1 \in \mathbb{R}^d$ are "far". The goal is to label each pt. with the Gaussian it came from.

A Heuristic: One might expect that there is a threshold d_* s.t. if $X, Y \sim N(\mu_1, I_d)$, then w.h.p.

$$\|X-Y\|_2^2 \leq d_*^2$$ while if

$$X \sim N(\mu_1, I_d) \quad Y \sim N(\mu_2, I_d)$$

$$\|X-Y\|_2^2 >> d_*^2$$

We now show that when $\|\mu_1-\mu_2\| \geq \Omega(d^{1/4})$ such a threshold exists.

Case 1: $X, Y \sim N(\mu_i, I_d)$ $i=1$ or 2.

Then $Z = \frac{X-Y}{\sqrt{2}} \sim N(0, I_d)$.

Hence, from above concentration results, when $\log(2n) \leq \frac{\sqrt{d}}{12}$,

$$\left| \|X-Y\|_2^2 - 2d \right| \leq 12\sqrt{d \log(2n)} \text{ w.p. } \geq 1 - \frac{1}{4n^3}$$

Thus, taking a union bound, we see that for all pairs (X, Y) which were drawn from the same cluster,

$$\left| \|X-Y\|_2^2 - 2d \right| \leq 12\sqrt{d \log(2n)} \quad \text{w.p.} \geq 1 - \frac{1}{2n}. \qquad -(1)$$

Case 2 :- $X \sim N(\mu_1, I_d)$

$Y \sim N(\mu_2, I_d)$ independently.

$\mu_1 - \mu_2 = \Delta$. Then,

$$X - Y \overset{d}{=} \sqrt{2}Z + \Delta \quad \text{where } Z \sim N(0, I_d)$$

(i.e. $X-Y$ has the same distribution as $\sqrt{2}Z + \Delta$)

Then,

$$\|X-Y\|_2^2 \overset{d}{=} 2\|Z\|^2 + \Delta^2 + 2\sqrt{2}\,\Delta \cdot Z$$

$$\geq 2\|Z\|^2 - 2\sqrt{2}\|\Delta\| \cdot |\hat{\Delta} \cdot Z| + \|\Delta\|^2, \text{ where}$$

$\hat{\Delta}$ is the unit vector in the direction of Δ.

Now $Z \sim N(0, I_d)$ so,

$Z \cdot \hat{\Delta} \sim N(0, 1)$, Therefore,

$$\Pr\left(|Z \cdot \hat{\Delta}| > 2 d^{1/4} \sqrt{\log(2n)} \right) \leq \frac{2}{(2n)^{2\sqrt{d}}}$$

(Assuming $n \geq 2$, $d \geq 4$). $\leq \frac{1}{n^{2\sqrt{d}}}$

Let $\Delta = c\, d^{1/4} \sqrt{\log 2n}$ where $c > 6$ is a constant. Then we get that

$$\|X - Y\|_2 \geq 2d - 12\sqrt{d \log 2n} + 2c\sqrt{d \log 2n}$$

with probability $\geq 1 - \frac{5}{4n^3}$.

Put $c = 15$. Then.

$$\|X - Y\|_2 \geq 2d + 18\sqrt{d \log 2n} \quad \text{w.p.} \geq 1 - \frac{5}{4n^3}$$

provided $\|\mu_1 - \mu_2\| \geq 15\, d^{1/4} \sqrt{\log 2n}$

From ① and ② via an union bound, we get: w.p. $\geq 1 - \frac{1}{n}$, we have

$$\|X-Y\|^2 - 2d \begin{cases} \geq 18\sqrt{d\log(2n)} & \text{if } X, Y \text{ sampled from } N(\mu_1, I_d), N(\mu_2, I_d) \text{ respectively} \\ \leq 12\sqrt{d\log(2n)} & \text{if } X, Y \text{ sampled from the same } N(\mu_i, I_d); \end{cases}$$

provided $\boxed{\|\mu_1 - \mu_2\| \geq 15\, d^{1/4}\sqrt{\log n}}$.

Thus we can choose a threshold d^* as above when

$$\|\mu_1 - \mu_2\| \geq d^{1/4}\sqrt{\log 2n}.$$