Foundations of Data Science and Machine Learning

Isaac Martin

Last compiled January 24, 2023

EXERCISE 1. Let X be a random variable with probability density 1/4 for $0 \leq X \leq 4$ and zero elsewhere.

(a) Use Markov's inequality to bound the probability that $X \geq 3$.

(b) Make use of $\mathrm{Prob}*(|X| \geq a) = \mathrm{Prob}(X^2 \geq a^2)$ to get a tighter bound.

(c) What is the bound obtained using $\mathrm{Prob}(|X| \geq a) = \mathrm{Prob}(X^r \geq a^r)$ where r is a positive even integer?

Proof:

(a) Since X has a probability density function $f(x) = \frac{1}{4}\,\mathrm{id}_{[0,4]}$, X is a uniformly distributed random variable and has expectation

$$\mathbb{E}[X] = \int_0^4 xf(x)dx = \frac{1}{4}\left(\frac{x^2}{x}\right)_{x=0}^{x=4} = 2.$$

Hence, by Markov's inequality, we have

$$\mathbb{P}[X \geq 3] \leq \frac{\mathbb{E}[X]}{3} = \frac{2}{3}.$$

(b) We may use the fact that $\mathbb{P}[X \geq a] = \mathbb{P}[X^r \geq a^r]$ to get a tighter bound on $\mathbb{P}[X \geq 3]$. The moment-generating function for X is $m(x) = \frac{e^{4t}-1}{4t}$, so

$$\mathbb{E}[X^2] = \lim_{t\to 0}\frac{d}{dt}m(x) = \lim_{t\to 0}\frac{e^{4t}(8t^2 - 4t + 1) - 1}{2t^3} = 16/3,$$

so Markov gives us

$$\mathbb{P}[X \geq 3] = \mathbb{P}[X^2 \geq 9] = \frac{16/3}{9} = \frac{16}{27} < \frac{2}{3},$$

a better bound.

(c) I'm not sure why I felt the need to use the moment generating function in part (b); we know the distribution function of X so we can just calculate the rth moment:

$$\mathbb{E}[X^r] = \int_0^4 \frac{1}{4}x^r dx = \frac{1}{4(r+1)}x^{r+1}\Big|_0^4 = \frac{4^r}{r+1}.$$

Hence, Markov's inequality gives us the bound

$$\mathbb{P}[|X| \geq a] = \mathbb{P}[|X|^r \geq a^r] \leq \frac{\mathbb{E}[|X|^r]}{a^r} = \frac{4^r}{(r+1)a^r}$$

if we utilize the higher moments.

□

EXERCISE 2. A *Rademacher* random variable X has the probability mass function $\mathrm{Prob}(X = -1) = 1/2$. Suppose that X and Y are independent Rademacher random variables, and form the random variable $Z = XY$. Show that the random variables X, Y, Z are pairwise independent but no mutually independent.

Proof: We first note that X and Y are independent by definition and that, because X and Y are both Rademacher random variables, it suffices to show that X and Z are independent by symmetry. Recall that X and Z are independent if the σ-algebras they generate are independent. Equivalently, they are independent if $\mathbb{P}(X \leq x \quad \text{and} \quad Z \leq z) = \mathbb{P}(X \leq x) \cdot \mathbb{P}(Z \leq z)$. We have that

$$\mathbb{P}(X \leq x) = \begin{cases} 0 & x < -1 \\ 1/2 & -1 \leq x < 1 \\ 1 & 1 \leq x \end{cases}$$

and similarly

$$\mathbb{P}(Z \leq z) = \mathbb{P}(XY \leq z) = \begin{cases} 0 & z < -1 \\ 1/2 & -1 \leq z < 1 \\ 1 & 1 \leq z \end{cases}.$$

Now we calculate $\mathbb{P}(X \leq x \text{ and } XY \leq z)$ by examining cases. We assume without loss of generality that X and Y only contain -1 and 1 in their domain; if not, then we can replace X and Y with such functions which are almost surely equivalent. If either x or z is less than -1 then this evaluates to zero. If $-1 \leq x < 1$ then X must be -1. If $-1 \leq z < 1$ then XY also must be -1, meaning $Y = 1$ and $(X, Y) = (-1, 1)$. This occurs with probability $1/4$. If $z \geq 1$, then $XY = -1$ or 1 so $Y = -1$ or 1 and $\mathbb{P}(X \leq x \quad \text{and} \quad XY \leq z) = 1/2$. In either case, $\mathbb{P}(X \leq x \quad \text{and} \quad XY \leq z) = \mathbb{P}(X \leq x) \cdot \mathbb{P}(XY \leq z)$.

For the final two cases we get probabilities of $1/2$ and 1, which also agree with the necessary product. This means X and Z are indeed independent, and hence so are Y and Z.

However, the product XYZ is 1 with probability 1. This means that

$$\mathbb{E}[XYZ] = 1 \neq 0 = \mathbb{E}[X] \cdot \mathbb{E}[Y] \cdot \mathbb{E}[Z].$$

Hence XYZ are not mutually independent. □

EXERCISE 3. Consider drawing a random point $\mathbf{x}$ on the unit sphere in $\mathbb{R}^d$. What is the variance of x_1 (i.e. The first coordinate of $\mathbf{x}$)? Bonus [1pt]: See if you can give an argument without doing integrals.

Proof: I first thought to uniformly sample spherical coordinates $\theta_i \in [0, \pi]$ for $1 \leq i \leq d-2$ and $\theta_{d-1} \in [0, 2\pi)$. This is incorrect; points on S^{d-1} chosen in this way will tend to bunch points up at the poles. Instead, we utilize the spherical symmetry of the Gaussian distribution as shown in class. Let $X_1, ..., X_d \sim N(0, 1)$ be i.i.d. and define

$$Y_i = \frac{X_i}{\|(X_1, ..., X_d)\|}.$$

Then $Y = (Y_1, ..., Y_d)$ lies on S^{d-1} by definition and is uniformly distributed on S^{d-1} as shown in class.

Notice that this means

$$Y_1^2 + ... + Y_d^2 = 1 \implies \mathbb{E}[Y_1^2] + ... + \mathbb{E}[Y_d^2] = \mathbb{E}[Y_1^2 + ... + Y_d^2] = 1.$$

Since the coordinates $Y_1, ..., Y_d$ are identically distributed, the expectations of Y_i^2 are all equal and hence

$$d \cdot \mathbb{E}[Y_1^2] = 1 \implies \mathbb{E}[Y_1^2] = \frac{1}{d}.$$

Due to spherical symmetry, $\mathbb{E}[Y_1] = 0$, and so we have actually calculated the variance already: $\mathrm{Var}(Y_1) = \mathbb{E}[(Y_1 - \mathbb{E}[Y_1])^2] = \mathbb{E}[Y_1^2] = 1/d$. □

EXERCISE 4. Consider the probability distribution $\mathrm{Prob}(X = 0) = 1 - 1/a$ and $\mathrm{Prob}(X = a) = 1/a$. Plot the probability that X is greater than or equal to a as a function of a for the bound given by Markov's inequality, as well as the bounds given by Markov's inequality applied to X^2 and X^4.

Proof: We assume that $a > 1$ – otherwise, $\mathrm{Prob}(X = a) = 1/a$ wouldn't make sense. Since the probability that $X \in \{0, a\}$ is 1, the expectation of X is given by

$$\mathbb{E}[X] = \mathrm{Prob}(X = 0) \cdot 0 + \mathrm{Prob}(X = a) \cdot a = 1.$$

The higher moments of X are given by

$$\mathbb{E}[X^r] = \mathrm{Prob}(X = 0) \cdot 0 + \mathrm{Prob}(X = a) \cdot a^r = a^{r-1}.$$

This means that for $r \geq 1$ Markov yields

$$\mathbb{P}[X \geq a] = \mathbb{P}[X^r \geq a^r] \leq \frac{\mathbb{E}[X^r]}{a^r} = \frac{a^{r-1}}{a^r} = \frac{1}{a}.$$

Since $\mathbb{P}[X \geq a] = \mathbb{P}[X > a] + \mathbb{P}[X = a] = 0 + \frac{1}{a}$, Markov's inequality is actually an *equality* here.

I'm worried that I'm missing something here, because Markov seems to tell us something weaker than what we already know: $\mathbb{P}[X \geq a] = \mathbb{P}[X = a] = 1/a$. Here's the figure:

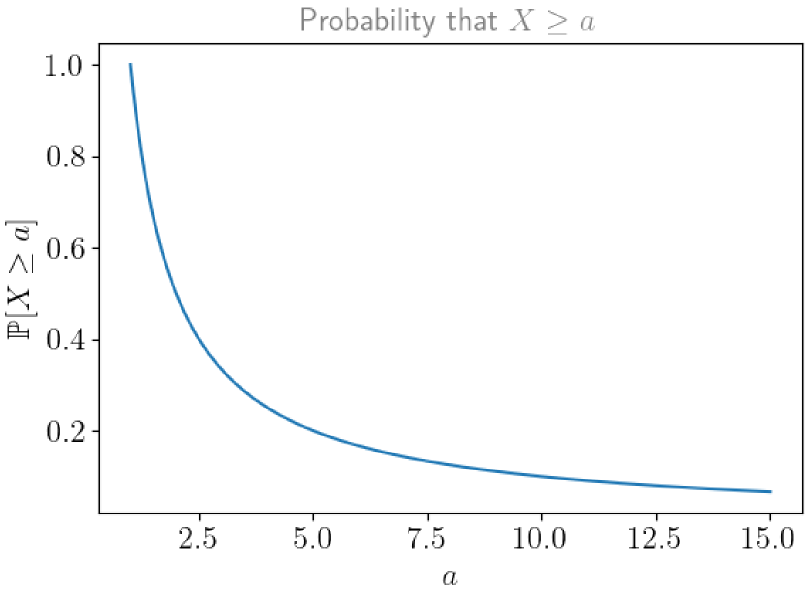


Figure 1: Probability that $X \geq a$ plotted as a function of a

□

EXERCISE 5. If one randomly generates points in $\mathbb{R}^d$ with each coordinate a unit variance Gaussian, the points will approximately lie on the surface of a sphere of radius $\sqrt{d}$. What is the distribution when the points are projected onto the line through the origin in the direction of $(1, 0, 0, ..., 0)$? How about when they are projected onto an arbitrary line through the origin?

Proof: If one chooses $X_1, ..., X_d \sim N(0, 1)$ all i.i.d. normal random variables as in problem 3, then $X = (X_1, ..., X_d)$ satisfies $X \sim N(0, I_d)$ where I_d is the identity matrix. From this it follows that OX is identically distributed to X for any orthogonal matrix $O \in O(d)$. Thus, if ℓ is an arbitrary line through the origin, we may apply a rotation matrix A so that $A\ell$ is the line passing through the origin and $(1, 0, ..., 0)$ without affecting the distribution of X. It therefore suffices to find the distribution of X projected onto the line passing through $(1, 0, ..., 0)$ and the origin to answer both parts of this question. However, this is easy: X projected onto this line is simply the projection of $X \mapsto X_1$, so the distribution is $N(0, 1)$. □

EXERCISE 6. In your preferred programming language, uniformly generate 100 random points on the surface of a sphere in 3-dimensions and in 100-dimensions. Create a histogram of all distances between the pairs of points in both cases and discuss.

Solution: To generate points on the unit sphere we follow the same procedure as in problem 3 and 5: we sample a data point X from $N(0, I_d)$ (which is equivalent to sampling the ith coordinate X_i of X from $N(0, 1)$) and then normalize its components. This yields a point uniformly sampled from S^{d-1}, the unit sphere in $\mathbb{R}^d$.

Before we show the desired plots, we first list some observations.

(1) For any two points $x, y \in S^n$, the Euclidean distance $\|x - y\|$ falls within the interval $[0, 2]$

(2) If we fix a point $x \in S^n$ and a real number $r \in [0, 2]$, then the set $L_x(r)$ of all points $y \in S^n$ distance r from x is a copy of S^{n-1} obtained by intersecting some hyperplane $H \subseteq \mathbb{R}^n$ orthogonal to the line between x and its antipode $-x$.

(3) There is exactly one $r \in [0, 2]$ such that set $L_x(r)$ is a great circle, and it occurs for $r = \sqrt{2}$.

With these claims in mind, let's see the requested plots.

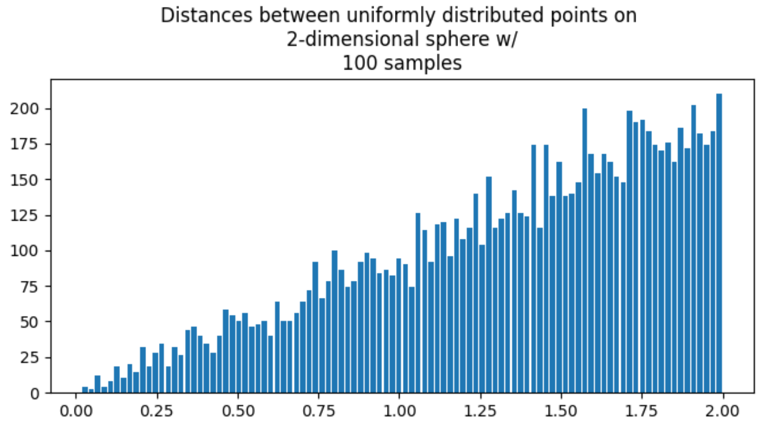


Figure 2: Distribution of distances between points on the unit sphere in $\mathbb{R}^3$

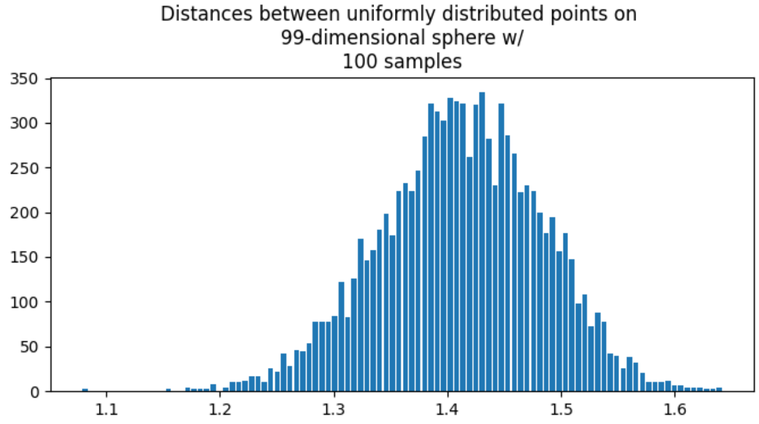
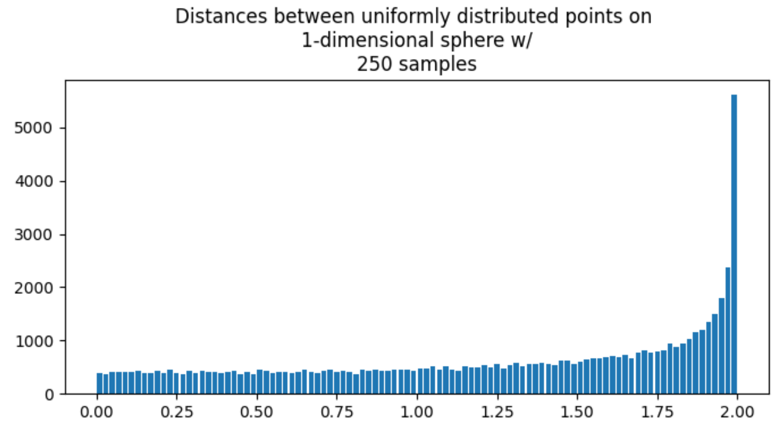
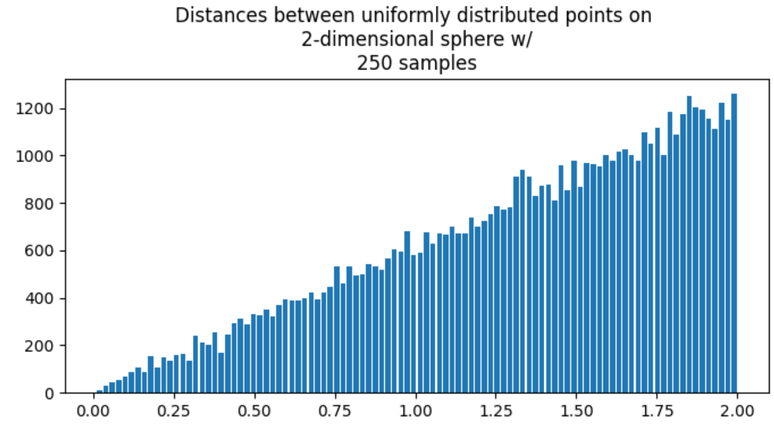
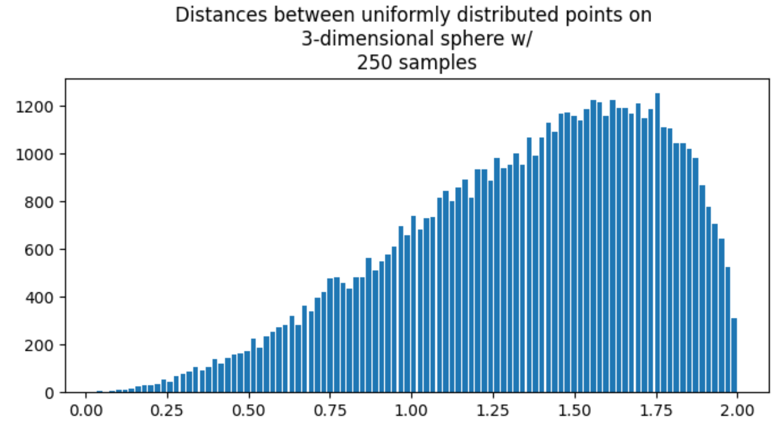


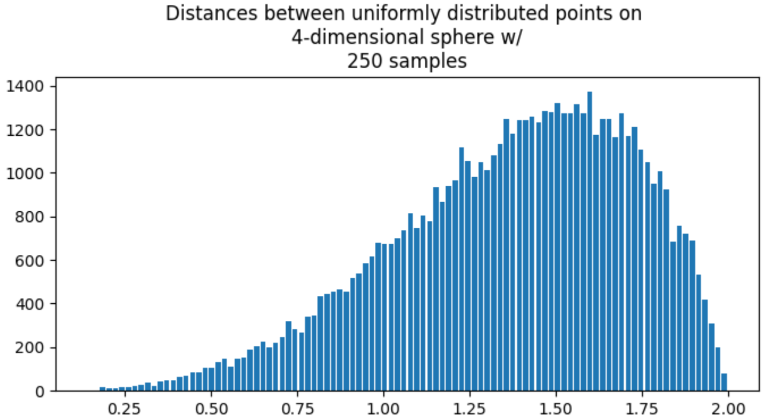
Figure 3: Distribution of distances between points

For $d = 3$ the distribution seems to be linear, while for $d = 100$ it appears to be normal. To see what's going on here, let's plot in several more dimensions with more samples.

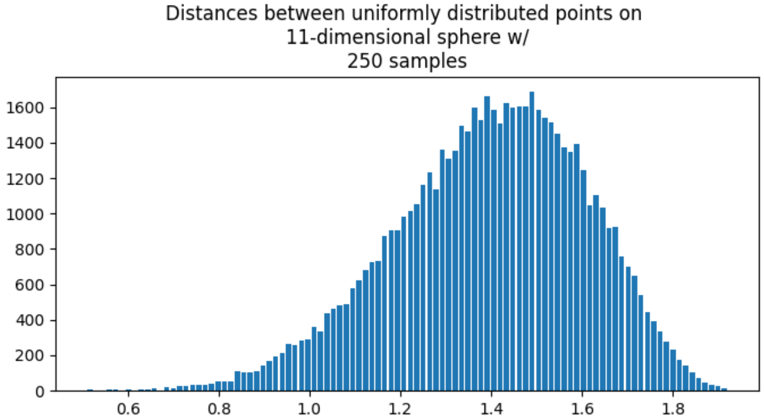

Distances between uniformly distributed points on
1-dimensional sphere w/
250 samples


Distances between uniformly distributed points on
2-dimensional sphere w/
250 samples

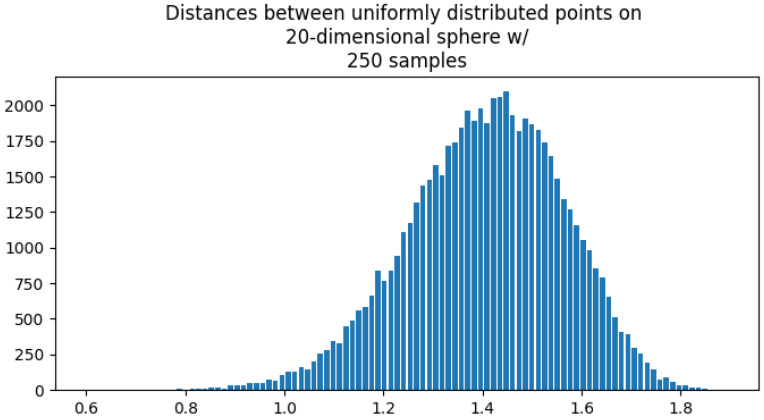

Distances between uniformly distributed points on
3-dimensional sphere w/
250 samples



Distances between uniformly distributed points on
4-dimensional sphere w/
250 samples



Distances between uniformly distributed points on
11-dimensional sphere w/
250 samples



Distances between uniformly distributed points on
20-dimensional sphere w/
250 samples

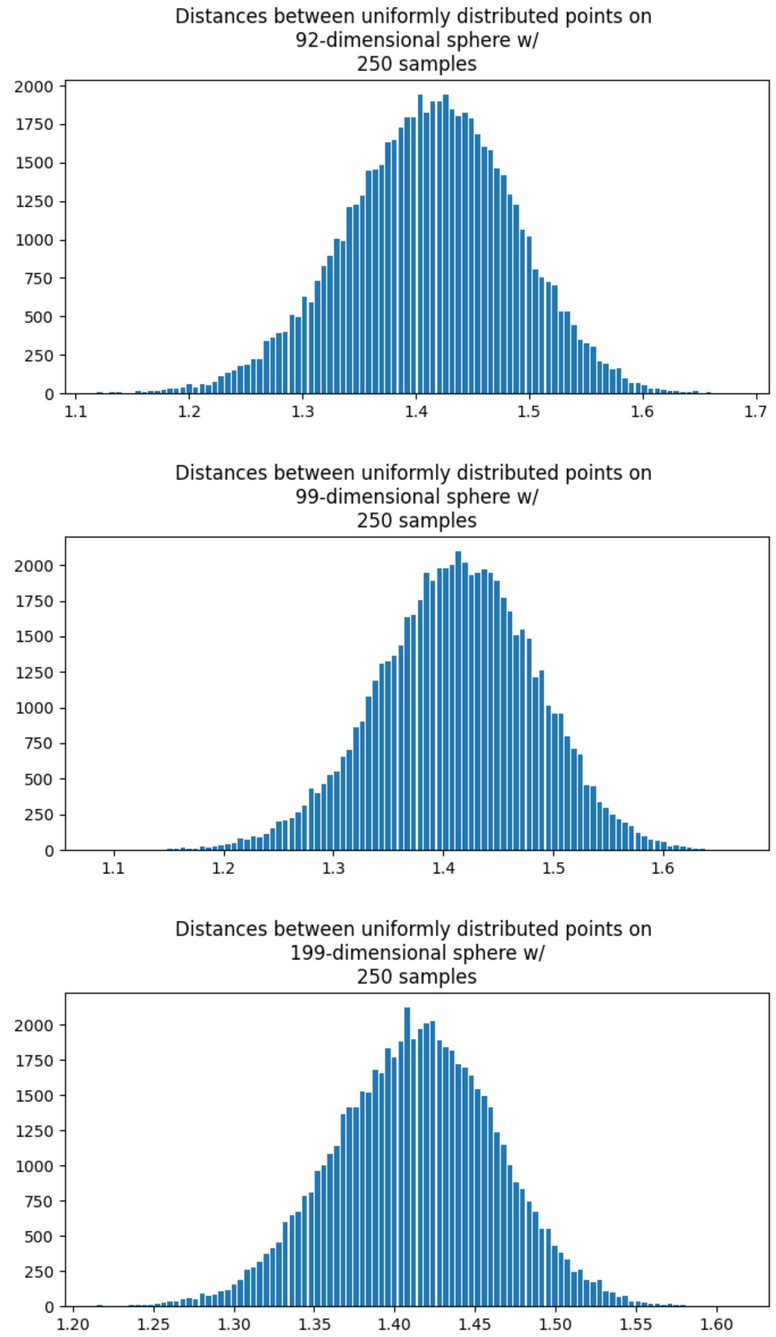


Figure 4: Distribution of distances between points for many values of d

It seems that as the dimension d increases, the distribution of distances between uniformly sampled points on the unit sphere becomes a better approximation of a bell curve centered around $\sqrt{2}$. Of course, this isn't actually a normal distribution – the support of our distribution is compact, $[0, 2]$ – but perhaps it's something like a truncated normal distribution in the limit. What is surprising, at least to me, is that in low dimensions we have a significant bias towards the larger distances. What's going on here?

Consider two points P and Q randomly chosen on the surface of S^{d-1}. Since the special orthogonal group SO_d acts isometrically on S^{d-1}, we can assume that $P = (1, 0, ..., 0)$ is the north pole of S^{d-1} after perhaps applying a rotation. Now consider the map $\varphi_P : S^{d-1} \longrightarrow \mathbb{R}$ which sends a point on the sphere to its distance from P: $\varphi_P(x) = \|P - x\|$. We would like to understand the level sets $L_P(a) = \varphi_P^{-1}(a)$ of this function. Actually, that's not quite right; we want to understand the measure of $L_P([a - \varepsilon, a + \varepsilon])$ for small ε as a function of $a \in [0, 2]$. This should tell us the probability that two points randomly chosen on S^{d-1} have a distance within ε of a.

For small values of d, there are many points on S^{d-1} which are far from P and fewer which are close to P. Another way to state this is $\varphi^{-1}([a - \varepsilon, a + \varepsilon])$ has large measure when a is large and small measure when a is small. This is illustrated in Figure 5 and is demonstrated empirically in Figure 6.

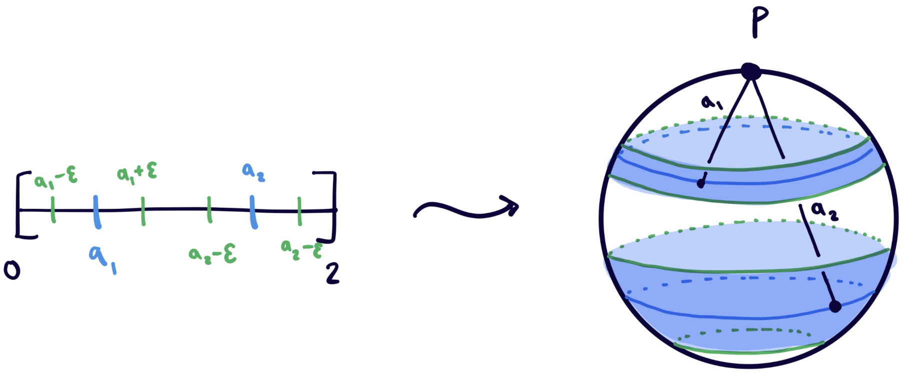
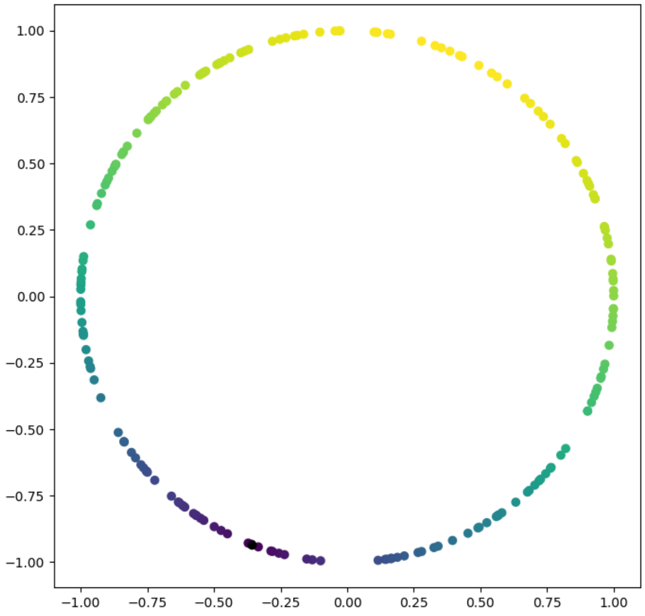


Figure 5: Epsilon neighborhoods around small a values correspond to thin strips on the surface of S^{d-1}



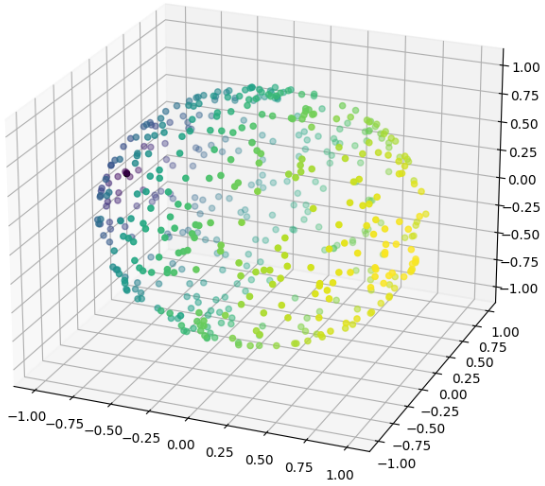


Figure 6: Points uniformly distributed on S^{d-1} for $d = 2, 3$ colored by their distance from the black point. Violet means a point is close, yellow means a point is far.

As the dimension d increases however, the width of the ε-neighborhood around a and the thickness of the strip around the level set $L_P(a)$ ceases to depend as heavily on a. Though we have demonstrated this empirically, we have not shown it rigorously. What follows is an attempt to do that, although it got quite messy.

As previously noted, the level set $L_P(a) \subseteq S^{d-1}$ is a $(d-2)$-dimensional sphere of some radius R. Hopefully the geometric intuition is clear enough (see Figure 7) to claim this without proof.

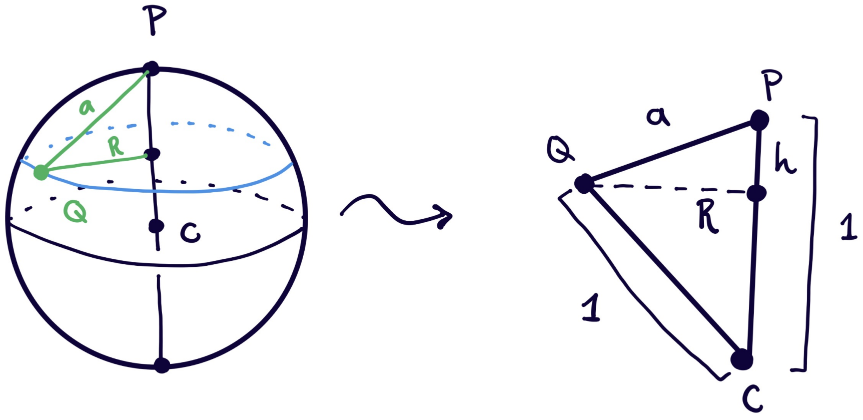


Figure 7: Geometric relationship between distance a between P and Q and the spherical level set $L_P(a)$ of radius R, seen here in blue.

The relationship between a and R can be derived from two applications of Pythagorean's theorem. The triangle on the right of Figure 7 gives us that

$$R^2 + h^2 = a^2 \quad \text{and} \quad R^2 + (1-h)^2 = 1^2.$$

Solving for R in terms of a gives us

$$R^2 = a^2 - \frac{a^4}{4} \implies R = a \cdot \frac{\sqrt{4-a^2}}{2}. \tag{1}$$

Now, we'd like to understand the measure of $L_P([a-\varepsilon, a+\varepsilon]) = \varphi_P^{-1}([a-\varepsilon, a+\varepsilon])$, which is a strip around the sphere enclosing the "circle" $L_P(a)$ of radius $a \cdot \sqrt{(4-a^2)/2}$. A first guess might be that

$$\mu(\varphi_P^{-1}([a-\varepsilon, a+\varepsilon])) \approx \mu(\varphi_P^{-1}(a)) \times 2\varepsilon,$$

i.e. that the surface area of the strip on the sphere is approximately the surface area of a cylinder with height 2ε and base $\varphi_P^{-1}(a)$ (see Figure)

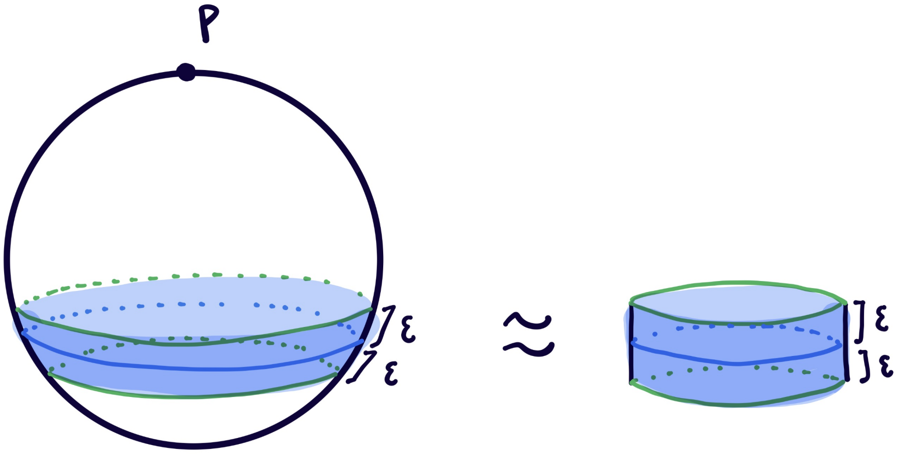


Figure 8: This strip of the sphere looks a lot like a cylinder

The problem is that the change in height of this cylinder changes depending on the distance of the central circle $L_P(a)$ from P. The distance "changes much faster" near P than it does far from P.

Formalizing this a little, we see that the path $\gamma : [0, \pi] \longrightarrow S^{d-1}$ defined $\gamma(\theta) = (\cos\theta, \sin\theta, 0, ..., 0)$ has exactly one point in each level set $L_P(a)$ for $a \in [0, 2]$. What's more, the path it traces is orthogonal to all of these level sets. If we denote by a_t the distance between $\gamma(t)$ and P, then the length between $\gamma(t_2)$ and $\gamma(t_1)$ is equal to the "height of the cylinder" which approximates the area of the strip bounded by $L_P(a_{t_1})$ and $L_P(a_{t_2})$.

By noticing we're simply tracing a portion of a great circle around S^{d-1}, we see that

$$\frac{|t_2 - t_1|}{2\pi} = \text{ arc length between } \gamma(t_2) \text{ and } \gamma(t_1), \tag{2}$$

and after some simplifications we get that

$$a_t^2 = 2 - 2\cos(t) \implies t = \arccos\left(1 - \frac{a_t^2}{2}\right),$$

where the range of $\arccos$ is taken to be $[0, \pi]$. This is gross, so we approximate by the Taylor series of $\arccos$ to get

$$t = \frac{\pi}{2} - 1 + \frac{a_t^2}{2} - \frac{1}{6}\left(1 - \frac{a_t^2}{2}\right)^3 .$$

This is a good approximation when $a_t \geq 1/\sqrt{2}$, otherwise $t = |a_t|$ is a better approximation.

We can now find the interval for t which corresponds to the interval $a_t \in [a - \varepsilon, a + \varepsilon]$. If $a \leq 1/\sqrt{2}$, then $a_t \in [a - \varepsilon, a + \varepsilon] \implies t \in [1/\sqrt{2} - \varepsilon, 1/\sqrt{2} + \varepsilon]$. Otherwise, we have that

$$\begin{aligned} t_{\varepsilon^+} &= \frac{1}{48}\Big((a+\varepsilon)^6 - 6(a+\varepsilon)^4 + 36(a+\varepsilon)^2 + 8(3\pi - 7)\Big) \\ t_{\varepsilon^-} &= \frac{1}{48}\Big((a-\varepsilon)^6 - 6(a-\varepsilon)^4 + 36(a-\varepsilon)^2 + 8(3\pi - 7)\Big) \end{aligned}$$

are the right and left endpoints of the interval. This is also disgusting, but since we're really just after the length of this interval, we'll denote it by $\delta_a = t_{\varepsilon^+} - t^{\varepsilon^-}$, retaining the a to remind ourselves that this is really a function of a.

Putting everything together, the estimate we get for the measure of the strip on S^{d-1} containing all points whose distance from P is between $a - \varepsilon$ and $a + \varepsilon$ is

$$A = \frac{2\pi^{d/2}}{\Gamma(d/2)} R^{d-1} \cdot \delta(a)$$

where $\frac{2\pi^{d/2}}{\Gamma(d/2)} R^{d-1}$ is the surface area of a $(d-2)$-dimensional sphere of radius R. This term is dominated by $\delta(a)$ only for small values of d, otherwise, its contribution is small.

It still remains to show that in the limit $d \longrightarrow \infty$ the distribution of distances between uniformly sampled points on the unit sphere in $\mathbb{R}^d$ is a bell curve. If there is ample time before the submission deadline, I will try to prove this final claim and improve the previous argument. Perhaps there are some estimates that can clean up the math in places. □

EXERCISE 7. Suppose you wish to estimate the unknown center (i.e. mean) of a Gaussian in d-dimensional space with independent coordinates each with variance one. Show that $\mathcal{O}(\log(d)/\varepsilon^2)$ random samples from the Gaussian are sufficient to get an estimate $\hat{\mu}$ of the true center $\mu \in \mathbb{R}^d$, such that with probability at least 99% it holds

$$\|\mu - \hat{\mu}\|_\infty \leq \varepsilon?$$

How many samples are sufficient to ensure that with probability at least 99% it holds

$$\|\mu - \hat{\mu}\|_2 \leq \varepsilon?$$

Proof: This problem was by far the most fun in the set. To aid organization of the problem, we first prove some lemmas.

Lemma 1.1. Let $X \sim N(0,1)$ be a normal random variable. Then the higher moments of X are

$$\mathbb{E}[X^n] = \begin{cases} 0 & n \text{ is odd} \\ \frac{1}{\sqrt{2\pi}}\Gamma\left(\frac{n+1}{2}\right) & n \text{ is even} \end{cases}$$

Proof. The distribution of X is given by $f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$, an even function. If n is odd, then $x^n \cdot f(x)$ is an odd function and hence

$$\begin{aligned} \mathbb{E}[X^n] &= \int_{\mathbb{R}} x^n f \, dx \\ &= \int_0^{\infty} x^n f \, dx \; - \; \int_0^{\infty} x^n f \, dx = 0. \end{aligned}$$

If instead n is even, then

$$\begin{aligned} \mathbb{E}[X^n] &= \int_{\mathbb{R}} x^n f \, dx = \frac{2}{\sqrt{2\pi}} \int_0^{\infty} x^n e^{-x^2/2} \, dx \\ &= \frac{2}{\sqrt{2\pi}} \int_0^{\infty} u^{(n-1)/2} e^{-u} \, du \; = \; \frac{1}{\sqrt{2\pi}} \cdot \Gamma\left(\frac{n+1}{2}\right). \end{aligned}$$

□

Corollary 1.2. If $X_1, ..., X_n$ are i.i.d. normal random variables with mean 0 and variance 1 then $X = \frac{1}{n} \cdot (X_1 + ... + X_n)$ satisfies the master tail bound.

Proof. If $s \geq 3$ is odd then we have

$$|\mathbb{E}[X_i^s]| = 0 < s!$$

for all $1 \leq i \leq n$ by Lemma 1.1. If instead $s = 2k \geq 3$ is even then

$$|\mathbb{E}[X_i^s]| = \frac{1}{\sqrt{2\pi}}\Gamma\left(k + \frac{1}{2}\right) = \frac{1}{\sqrt{2\pi}} \frac{(2k)!\sqrt{\pi}}{4^k \cdot k!} = \frac{s!}{\sqrt{2} \cdot 4^k \cdot (s/2)!} \leq s!,$$

and hence X satisfies the conditions for the master tail bound. □

Suppose now that $X_{ij} \sim N(\mu_i, 1)$ is a normally distributed random variable with mean μ_i and variance 1, that $X_i = (X_{i1}, ..., X_{id})$ and that $\{X_{ij}\}$ is a mutually independent collection for $1 \leq j \leq d$ and $1 \leq i \leq n$. That is, each X_i is a point sampled from a d-dimensional Gaussian centered at $\mu = (\mu_1, ..., \mu_d)$. Define also a random variable $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$, the mean of the n points. Note that

$$\|\mu - \hat{\mu}\|_\infty = \max_{1 \leq j \leq d} |\mu_j - \hat{\mu}_j|,$$

where $\hat{\mu}_j$ is the jth component of $\hat{\mu}$.

Choose $\varepsilon > 0$. We first make the observation

$$\mathbb{P}\big[\|\mu - \hat{\mu}\|_\infty < \varepsilon\big] \geq \frac{99}{100} \iff \mathbb{P}\big[\|\mu - \hat{\mu}\|_\infty \geq \varepsilon\big] \leq \frac{1}{100}$$

and

$$\begin{aligned}\mathbb{P}\big[\|\mu - \hat{\mu}\|_\infty \geq \varepsilon\big] &= \mathbb{P}\big[\max_{1\leq j\leq d} |\mu_j - \hat{\mu}_j| \geq \varepsilon\big] \\ &= \mathbb{P}\Big[(|\mu_1 - \hat{\mu}_1| \geq \varepsilon) \text{ OR } (|\mu_2 - \hat{\mu}_2| \geq \varepsilon) \text{ OR ... OR } (|\mu_d - \hat{\mu}_d| \geq \varepsilon)\Big] \\ &\leq \sum_{j=1}^{d} \mathbb{P}\big[|\mu_j - \hat{\mu}_j| \geq \varepsilon\big].\end{aligned}$$

We're going to apply the master tail bound to the expression above. First, we argue that we might as well assume $\mu_j = 0$ for all $1 \leq j \leq d$. Define $Y_{ij} = X_{ij} - \mu_j$; then $Y_{ij} \sim N(0,1)$. If we further set

$$Z_j = \frac{1}{n}(Y_{1j} + Y_{2j} + ... + Y_{nj})$$

then

$$|\mu_j - \hat{\mu}_j| = \left|\sum_{i=1}^{n}(\mu_j - X_{ij})\right| = \left|-\sum_{i=1}^{n} Y_{ij}\right| = |Z_j|$$

for all $j \in \{1..d\}$. This means that

$$\mathbb{P}\big[|\mu_j - \hat{\mu}_j| \geq \varepsilon\big] = \mathbb{P}\big[|Z_j| \geq \varepsilon\big]$$

for all j. Even better, the Z_j are i.i.d. since each is the mean of the Y_{ij} for varying i, and hence

$$\mathbb{P}\big[|\mu_j - \hat{\mu}_j| \geq \varepsilon\big] = \mathbb{P}\big[|Z_j| \geq \varepsilon\big] = \mathbb{P}\big[|Z_k| \geq \varepsilon\big] = \mathbb{P}\big[|\mu_k - \hat{\mu}_k| \geq \varepsilon\big]$$

for all pairs $j, k \in \{1..d\}$. This implies

$$\sum_{j=1}^{d} \mathbb{P}\big[|\mu_j - \hat{\mu}_j| \geq \varepsilon\big] = d \cdot \mathbb{P}\big[|Z| \geq \varepsilon\big],$$

where we've chosen $Z = Z_1$ arbitrarily – we could have defined $Z = Z_j$ for any $1 \leq j \leq d$. By the Corollary, $Z = \frac{1}{n}(Y_{11} + ... + Y_{n1})$ satisfies the hypotheses of the master tail bound, so as long as we choose ε small enough such that $0 \leq \varepsilon \leq \sqrt{2}n$ we get

$$d \cdot \mathbb{P}\big[|Z| \geq \varepsilon\big] = d \cdot \mathbb{P}\big[|n \cdot Z| \geq n \cdot \varepsilon\big] \leq d \cdot 3e^{-\frac{n^2\varepsilon^2}{12n}} = 3d \cdot e^{-\frac{n\varepsilon^2}{12}}.$$

Notice that we've multiplied Z by n since $n \cdot Z = Y_{11} + ... + Y_{Y_{n1}}$, which better fits the form of the master tail bound. After completing a bit of algebra, we see that

$$3d \cdot e^{-\frac{n\varepsilon^2}{12}} \leq \frac{1}{100} \quad \text{if and only if} \quad n \geq \frac{12(\log(d) + \log(300))}{\varepsilon^2}.$$

Putting all this together, we conclude that for a number of samples n on the order of $\mathcal{O}(\log(d)/\varepsilon^2)$ we have that

$$\mathbb{P}\big[\|\mu - \hat{\mu}\|_\infty \geq \varepsilon\big] \leq \sum_{j=1}^{d} \mathbb{P}\big[|\mu_j - \hat{\mu}_j| \geq \varepsilon\big] \leq 3d \cdot e^{-n\varepsilon^2/12} \leq \frac{1}{100}.$$

Now let's consider the case of the l^2 norm. As we did in the last portion of the problem, we can translate our mean μ to the origin without affecting the relevant underlying probabilities; hence it suffices to assume $\mu = 0$. Suppose first that $d = 1$, in which case $\hat{\mu} = \sum_{i=1}^{n} \frac{X_i}{n}$. Squaring this yields

$$\hat{\mu}^2 = \frac{1}{n^2} \sum_{\substack{1 \le i \le n \\ 1 \le j \le n}} X_i X_j = \frac{1}{n^2} \sum_{i=1}^{n} X_i^2 \; + \; \frac{1}{n^2} \sum_{i \neq j} 2 X_i X_j,$$

but because the X_i are independent,

$$\begin{aligned} \mathbb{E}[\hat{\mu}^2] &= \frac{1}{n^2} \mathbb{E}\left[\sum_{i=1}^{n} X_i^2\right] \; + \; \frac{1}{n^2} \mathbb{E}\left[\sum_{i \neq j} 2 X_i X_j\right] \\ &= \frac{1}{n^2} \sum_{i=1}^{n} \mathbb{E}[X_i^2] \; + \; \frac{1}{n^2} \sum_{i \neq j} 2\mathbb{E}[X_i]\mathbb{E}[X_j] = \frac{1}{n^2} \sum_{i=1}^{n} 1 = \frac{1}{n} \end{aligned}$$

since $\mathbb{E}[X_i] = 0$. For arbitrary d, we have that

$$\mathbb{E}[\hat{\mu}^2] = \mathbb{E}[\hat{\mu} \cdot \hat{\mu}] = \mathbb{E}\left[\hat{\mu}_1^2 + ... + \hat{\mu}_d^2\right] = d \cdot \mathbb{E}\left[\hat{\mu}_1^2\right] = \frac{d}{n}$$

because we can perform an identical calculation for all $\hat{\mu}_i$. Applying Markov's inequality yields

$$\mathbb{P}[\|\hat{\mu}\|_2 \ge \varepsilon] = \mathbb{P}[\hat{\mu}^2 \ge \varepsilon^2] \le \frac{\mathbb{E}[\hat{\mu}^2]}{\varepsilon^2} = \frac{d}{n\varepsilon^2},$$

so if $\mathbb{P}[\|\hat{\mu}\|_2 \ge \varepsilon]$ is to be less than $1/100$ we must have that $n \ge \frac{100d}{\varepsilon^2}$. This means we must take $\mathbb{O}(100d/\varepsilon^2)$ random samples to ensure with 99% probability that $\hat{\mu}$ is within ε of the true center, measured in the l^2 norm. □