Uniform law and normalization

- a warning -

by Bernard Beauzamy

May 2008

Warning: if you take a uniform law on the components of a vector, and then normalize the vector, you do not get a uniform law on the normalized vector.

Quite often, people do the following:

- They draw a sample for a random vector \((x_1, \ldots, x_N)\); each \(x_n\) follows a uniform law, for instance on the interval \([0,1]\) (this is useful in order to test values of parameters in a general experiment, or to define some experiment planning);

- And then they normalize, that is they replace each \(x_n\) by

\[
y_n = \frac{x_n}{\sqrt{\sum_{i=1}^{N} x_i^2}}, \text{ so that indeed } \sum_{n=1}^{N} y_n^2 = 1.\]

This normalization is often useful, for example because the whole population is not the same, and so on.

People often think that the vector \(Y = (y_1, \ldots, y_n)\) has a uniform law on the unit sphere, or, if one prefers, that all portions of the unit sphere have the same probability. This is wrong.

In mathematical terms, the radial projection of the uniform law on the unit hypercube is not the uniform law on the unit sphere.

Let us see this on a very simple example, dimension 2.

One takes two random numbers \(x_1\) and \(x_2\) with uniform law between 0 and 1. This means that we have a uniform law in the unit square \([0,1]^2\).

For each \(x_1\) and \(x_2\), we consider the normalized vector:
\[ y_1 = \frac{x_1}{\sqrt{x_1^2 + x_2^2}} \]
\[ y_2 = \frac{x_2}{\sqrt{x_1^2 + x_2^2}} \]

and we want to find the law of this vector, which is on the unit circle.

Fix any \( \theta \), \( 0 \leq \theta \leq \frac{\pi}{2} \):

The point \( Y \) is on the arc AC if and only if the point \( X \) is inside the triangle \( OAB \). The area of this triangle is:

If \( 0 \leq \theta \leq \frac{\pi}{4} \), \( \text{area}(OAB) = \frac{\tan(\theta)}{2} \)

If \( \frac{\pi}{4} \leq \theta \leq \frac{\pi}{2} \), \( \text{area}(OAB) = 1 - \frac{\tan(\frac{\pi}{2} - \theta)}{2} = 1 - \frac{1}{2 \tan(\theta)} \)

Let \( f(\theta) \) be the density function of the vector \( Y \) and \( F(\theta) \) its repartition function. We get:

If \( 0 \leq \theta \leq \frac{\pi}{4} \), \( F(\theta) = \frac{\tan(\theta)}{2} \)

If \( \frac{\pi}{4} \leq \theta \leq \frac{\pi}{2} \), \( F(\theta) = 1 - \frac{1}{2 \tan(\theta)} \)

Since \( f(\theta) = F'(\theta) \), we obtain:
If $0 \leq \theta \leq \frac{\pi}{4}$, $f(\theta) = \frac{1}{2\cos^2(\theta)}$

If $\frac{\pi}{4} \leq \theta \leq \frac{\pi}{2}$, $f(\theta) = \frac{1}{2\sin^2(\theta)}$

Here is the graph of this density of probability:

\[ \text{Graph of the density function.} \]

We see that it is not at all uniform!

Let us compute the probability of the intervals

$I_1 = \left\{ 0 \leq \theta \leq \frac{\pi}{10} \right\}$

$I_2 = \left\{ \frac{2\pi}{10} \leq \theta \leq \frac{3\pi}{10} \right\}$

Both have the same width, namely $\frac{\pi}{10}$, and the second one is centered around $\frac{\pi}{4}$. We get:

$P\{I_1\} = F\left(\frac{\pi}{10}\right) = \frac{\tan(\pi/10)}{2} \approx 0.1625$

$P\{I_2\} = F\left(\frac{\pi}{4} + \frac{\pi}{20}\right) - F\left(\frac{\pi}{4} - \frac{\pi}{20}\right)$

$= 2\left(F\left(\frac{\pi}{4}\right) - F\left(\frac{\pi}{4} - \frac{\pi}{20}\right)\right)$

$= 1 - \tan\left(\frac{\pi}{4} - \frac{\pi}{10}\right) \approx 0.4905$
So there is a considerable difference between both probabilities, despite the fact that both intervals have same width.

If one wants to have a uniform law on the unit circle, one should use polar coordinates:

\[ x = \cos(\theta), \quad y = \sin(\theta) \]

and if you take \( \theta \) with uniform law on the interval \([0, 2\pi]\), then the point \((x, y)\) will have a uniform law on the unit circle. The same holds in higher dimensions: one should use spherical coordinates directly, and not projections of cartesian coordinates.

The same mistake would occur if we would project upon a hyperplane, instead of the unit circle, and would occur as well with other laws (not uniform): the ratio between the area of the sector (on the circle or on a hyperplane) and the area of the triangle is not constant.