1 Surprises in high dimensions

Our intuition about space is based on two and three dimensions and can often be misleading in high dimensions. It is instructive to analyze the shape and properties of some basic geometric forms, which we understand very well in dimensions two and three, in high dimensions. To that end, we will look at the sphere and the cube as their dimension increases.

1.1 Geometry of the d-dimensional Sphere

Consider the unit sphere is d dimensions. Its volume is given by

$$V(d) = \frac{\pi^{\frac{d}{2}}}{\frac{d}{2}\,\Gamma\left(\frac{d}{2}\right)}$$

where Γ is the Gamma function. Recall that for positive integers n, $\Gamma(n) = (n-1)!$. Using Stirling's Formula,

$$\Gamma(n) \sim \sqrt{\frac{2\pi}{n}}\left(\frac{n}{e}\right)^n$$

we can see $\Gamma\left(\frac{d}{2}\right)$ grows much faster than $\pi^{\frac{d}{2}}$, and hence

$$V(d) \to 0 \qquad \text{as} \qquad d \to \infty.$$

In words, the volume of the d-dimensional sphere with radius 1 goes (very quickly) to 0 as the dimension d increases to infinity, see Figure 1.1. That means a unit sphere in high dimensions has almost no volume (compare this to the volume of the unit cube, which is always 1).

One can also show that "most" of the volume of the d-dimensional sphere is contained near the boundary of the sphere. That is, for a d-dimensional sphere of radius r, most of the volume is contained in an annulus of width proportional to $\frac{r}{d}$. This means that if you peel a high-dimensional orange, then there is almost nothing left, since almost all of the orange's mass is in the peel.

1.2 Geometry of the d-dimensional Cube

Claim: Most of the volume of the high-dimensional cube is located in its corners.

Proof (probabilistic argument). You do not need to know or follow this proof. But for completeness, I include it anyway.

We assume that the cube is given by $[-1,1]^d$, our argument easily extends to cubes of any size. Pick a point at random in the box $[-1,1]^d$. We want to calculate the probability that the point is also in the sphere.

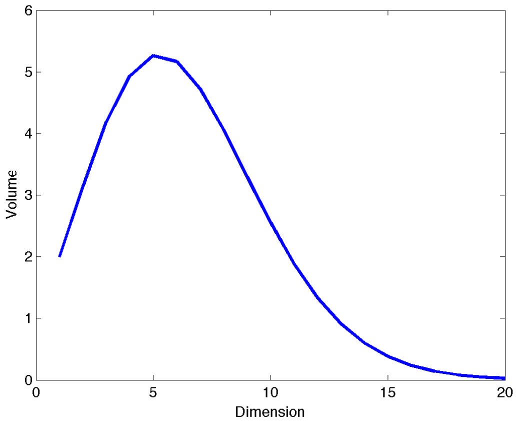

Figure 1: Volume of a unit sphere when the dimension of the sphere increases

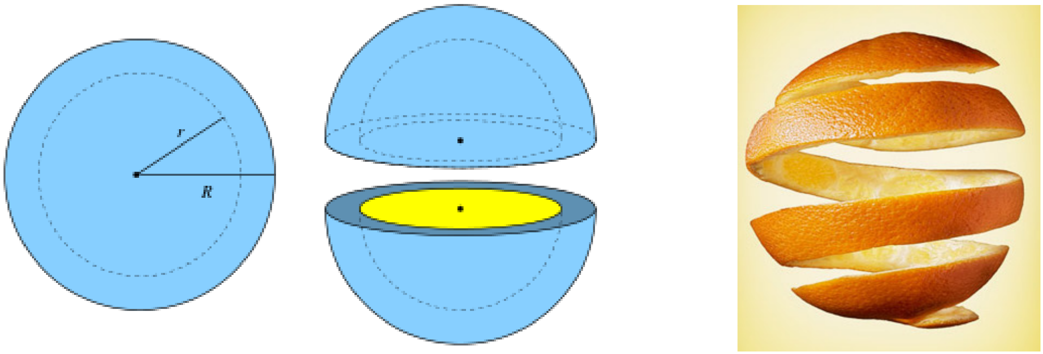

Figure 2: Most of the volume of the d-dimensional sphere is contained near its boundary. Hence, if you peel a high-dimensional orange, then there is almost nothing left.

Let $\boldsymbol{x} = [x_1, \ldots, x_d] \in \mathbb{R}^d$ and each $x_i \in [-1, 1]$ is chosen uniformly at random. The event that $\boldsymbol{x}$ also lies in the sphere means

$$\|\boldsymbol{x}\|_2 = \sqrt{\sum_{i=1}^{d} x_i^2} \leq 1.$$

Let $z_i = x_i^2$ and note that the expection of z_i is given by

$$\mathbb{E}(z_i) = \frac{1}{2}\int_{-1}^{1} t^2 \mathrm{d}t = \frac{1}{3} \implies \mathbb{E}\left(\|\boldsymbol{x}\|_2^2\right) = \frac{d}{3}$$

and the variance of z_i is

$$\mathrm{Var}(z_i) = \frac{1}{2}\int_{-1}^{1} t^4 \mathrm{d}t - \left(\frac{1}{3}\right)^2 = \frac{1}{5} - \frac{1}{9} = \frac{4}{45} \leq \frac{1}{10}$$

Using Chernoff's Inequality,

$$\begin{aligned}
\mathbb{P}\left(\|\boldsymbol{x}\|_2^2 \leq 1\right) &= \mathbb{P}\left(\sum_{i=1}^{d} x_i^2 \leq 1\right) \\
&= \mathbb{P}\left(\sum_{i=1}^{d} (z_i - \mathbb{E}(z_i)) \leq 1 - \frac{d}{3}\right) \\
&\leq \exp\left[-\frac{(\frac{d}{3} - 1)^2}{\frac{4d}{10}}\right] \\
&\leq \exp\left[-\frac{d}{10}\right].
\end{aligned}$$

Since this value converges to 0 as the dimension d goes to infinity, this shows random points in high cubes are most likely outside the sphere. In other words, almost all the volume of hypercubes lie in their corners. □

For completeness, here is Chernoff's Inequality.

Theorem 1.1 (Chernoff's Inequality)**.** *Let $X_1, X_2, \ldots, X_n$ be independent random variables with zero mean such that for all i there holds $|X_i| \leq 1$ almost surely. Furthermore, define $\sigma_i := \mathbb{E}\,|X_i|^2$ and $\sigma^2 := \sum_{i=1}^{n} \sigma_i^2$. Then,*

$$\mathbb{P}\Big\{\sum_{i=1}^{n} X_i \geq t\Big\} \leq \max\big\{e^{\frac{-t^2}{4\sigma^2}}, e^{-\frac{t}{2}}\big\},$$

and

$$\mathbb{P}\Big\{\sum_{i=1}^{n} X_i \leq -t\Big\} \leq \max\big\{e^{\frac{-t^2}{4\sigma^2}}, e^{-\frac{t}{2}}\big\},$$

1.3 Comparisons between the d-dimensional Sphere and Cube

We compare a unit d-dimensional cube (a cube with sidelength 1) with a unit d-dimensional sphere (a sphere with radius 1) as the dimension d increases.

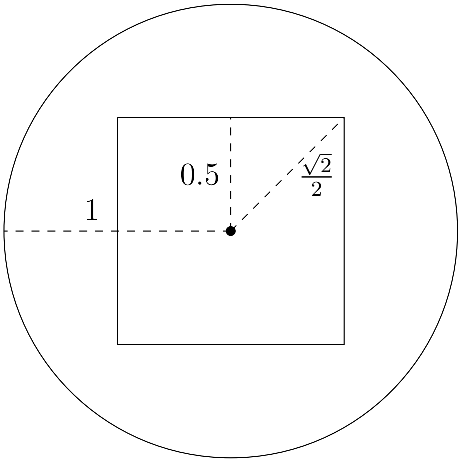

Figure 3: 2-dimensional unit sphere and unit cube, centered at the origin.

In two dimensions (Fig. 3), the unit square is completely contained in the unit sphere. The distance from the center to a vertex (radius of the circumscribed sphere or length of the diagonal of the cube) is $\frac{\sqrt{2}}{2}$ and the apothem (radius of the inscribed sphere) is $\frac{1}{2}$. In four dimensions (Fig. 4), the distance from the center to a vertex (i.e., the lenght of the cross-diagonal) is 1, so the vertices of the cube touch the surface of the sphere. However, the length of the apothem is still $\frac{1}{2}$. The result, when projected in two dimensions no longer appears convex, however all hypercubes are convex. This is part of the strangeness of higher dimensions - hypercubes are both convex and "pointy." In dimensions greater than 4 the distance from the center to a vertex is $\frac{\sqrt{d}}{2} > 1$, and thus the vertices of the hypercube extend far outside the sphere.

Hence, we can picture a cube in high dimensions to look like a regular cube, but also something like the spiky solid in the right panel of Figure 6. Maybe the Renaissance artist Pablo Uccello had hypercubes in mind when he drew the picture shown in in the right panel of Figure 7?

2 Curses and Blessings of Dimensionality

2.1 Curses

Bellman, in 1957, coined the term *the curse of dimensionality.* It describes the problem caused by the exponential increase in volume associated with adding extra dimensions to Euclidean space.

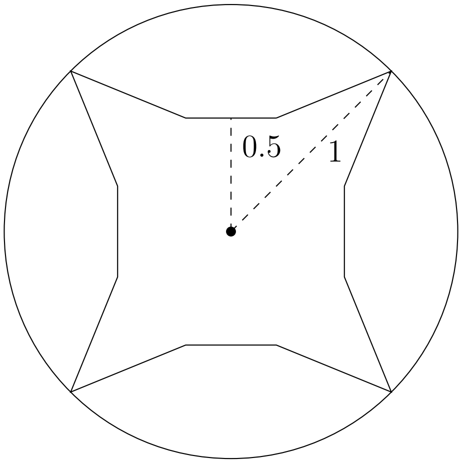

Figure 4: Projections of the 4-dimensional unit sphere and unit cube, centered at the origin (4 of the 16 vertices of the hypercube are shown).

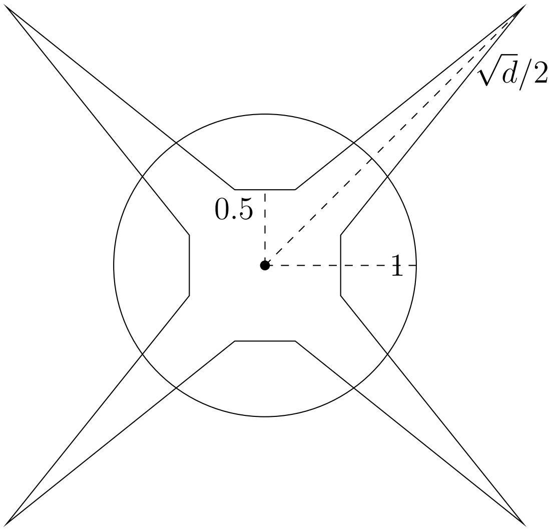

Figure 5: Projections of the d-dimensional unit sphere and unit cube, centered at the origin (4 of the 2^d vertices of the hypercube are shown).

For example, 100 evenly-spaced grid points suffice to sample the interval $[0, 1]$ with a distance of 0.01 between the grid points. For the unit square $[0, 1] \times [0, 1]$ we would need 100^2 grid points to ensure a distance of 0.01 between adjacent points. If we want sample the 10-dimensional unit cube with a grid with a distance of 0.01 between adjacent points, we would need 100^{10} points. Thus we need a factor of 10^{18} more points, even though the dimension increased only by a factor of 10. Any algorithm would now need to process a factor of 10^{18} more points than in dimension 1. This exponential increase of complexity vs a linear increase in dimension is a manifestation of the curse of dimensionality.

But we can feel the curse already even if the complexity does not increase exponentially. Often the computation time of algorithms scales badly with dimension. For example, the

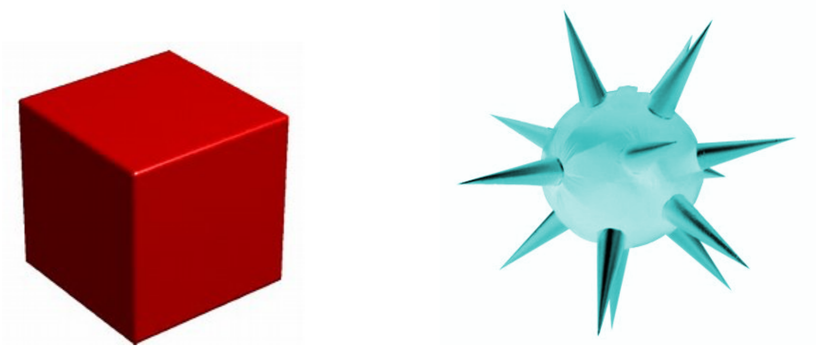

Figure 6: Representations of a 16-dimensional hypercube. Both representations are valid, each captures different features of the hypercube.

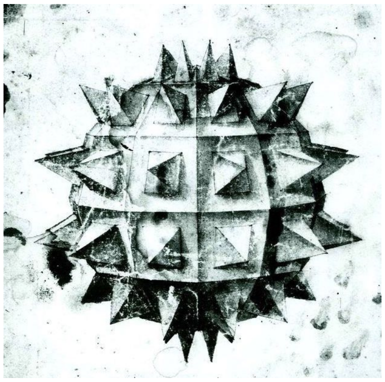

Figure 7: Artistic representations of a 72-dimensional hypercube? (Drawing by Paolo Uccello from the 15th Century.)

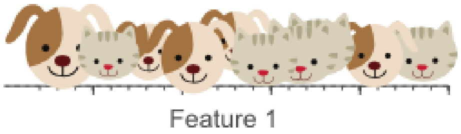

Figure 8: Classification using 1 feature. (Image by Vincent Spruyt)

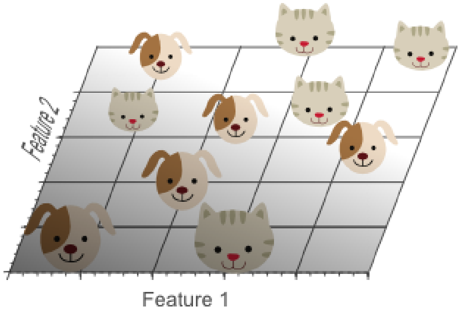

Figure 9: Classification using 2 features. (Image by Vincent Spruyt)

naive approach for finding nearest neighbors among n points in d dimensions requires $\mathcal{O}(n^2 d)$ operations, but for $d = 1$ we can simply sort in $\mathcal{O}(n \log n)$ operations. For large data sets ($n \gg 1$) in high dimensions ($d \gg 1$), the complexity of $n^2 d$ can be too high to be feasible.

The curse of dimensionality manifests itself also in the number of training data versus the number features needed for successful classification. This example and the images below are from the blog of Vincent Spruyt (VP Chief Scientist at Sentiance).

Assume we have images of cats and dogs and we want to construct an algorithm that can correctly classify those images into, well, cats and dogs. To be concrete, assume we have a training set of 10 labeled cat/dog-images, and a larger number of unlabeled images. We start out using one feature to attempt to distinguish between cats and dogs. For example we could use the number of red pixels. With one feature alone we will not be able to separate the training images into two separate classes, see Figure 8.

Therefore, we might decide to add another feature, e.g. the average green color in the image. But adding a second feature still does not result in a linearly separable classification problem: No single line can separate all cats from all dogs in this example, see Figure 9.

Finally we decide to add a third feature, e.g. the average blue color in the image, yielding a three-dimensional feature space, see Figure 10. Adding a third feature results in a linearly separable classification problem in our example. A plane exists that perfectly separates dogs from cats, see Figure 10.

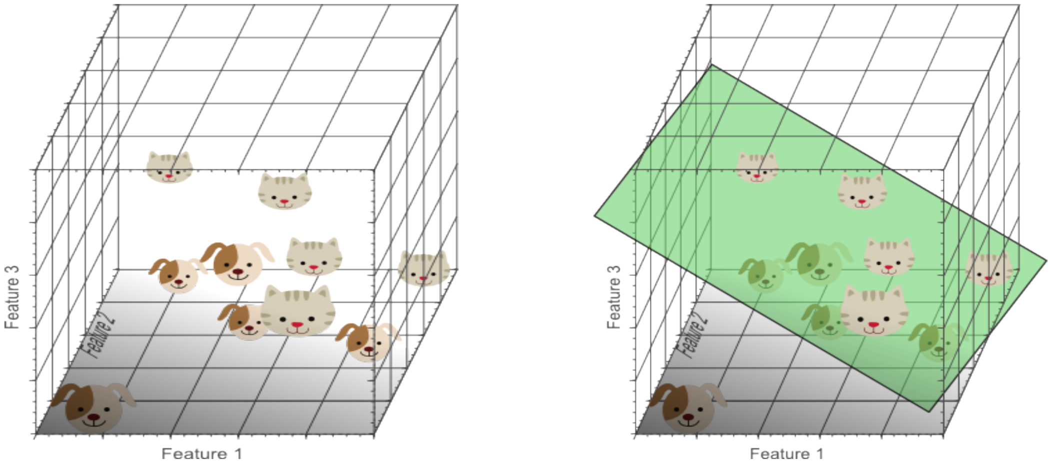

Figure 10: Classification using 3 features. (Images by Vincent Spruyt)

In the three-dimensional feature space, we can now find a plane that perfectly separates dogs from cats. This means that a linear combination of the three features can be used to obtain perfect classification results on our training data of 10 images:

The above illustrations might seem to suggest that increasing the number of features until perfect classification results are obtained is the best way to train a classifier. But this is not the case. If we would keep adding features, the dimensionality of the feature space grows, and becomes sparser and sparser. Due to this sparsity, it becomes much more easy to find a separable hyperplane because the likelihood that a training sample lies on the wrong side of the best hyperplane becomes infinitely small when the number of features becomes infinitely large. However, if the amount of available training data is fixed, then overfitting occurs if we keep adding dimensions. That means, we perfectly separate the training data, but when we try it on the rest of the data, the classification will fail terribly. On the other hand, if we keep adding dimensions, the amount of training data needs to grow exponentially fast to maintain the same coverage and to avoid overfitting.

2.2 Blessings

The blessings of dimensionality include the concentration of measure phenomenon (so-called in the geometry of Banach spaces), which means that certain random fluctuations are very well controlled in high dimensions and the success of asymptotic methods, used widely in mathematical statistics and statistical physics, which suggest that statements about very high-dimensional settings may be made where moderate dimensions would be too complicated.

The most well known example is the Law of Large Numbers, which says in a nutshell that the sum of independent, identically distributed scalar-valued random variables is approximately normal distributed. We observe such blessings, or such concentration of measure phenomena,

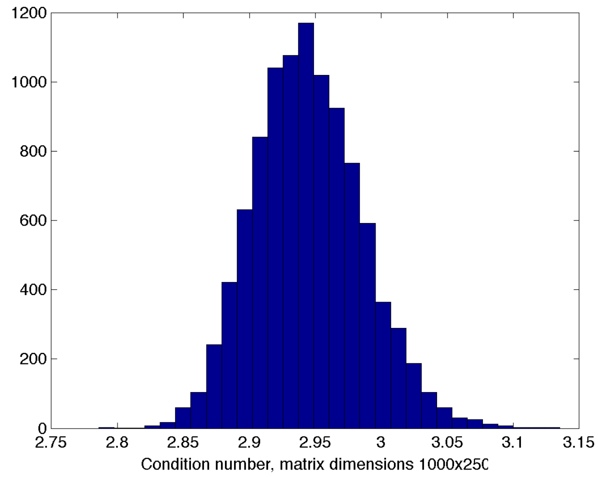

Figure 11: Concentration of the condition number of a Gaussian random matrix.

for example in physics. It is very difficult to predict the behavior of individual atoms in a gas. But we nevertheless predict very precisely how a gas will behave, because we can make quite accurate statements about the behavior averaged over millions of atoms.

There is a vast number of such concentration phenomena that extend from scalars to vectors and matrices. These generalizations often make our life in machine learning much easier. Here is one such example of the blessings of dimensionality from random matrix theory. The condition number of a matrix (the ratio of its largest singular value and its smallest singular value) is an important measure of its stability with respect to noise when solving a linear system of equations. In general it is very difficult or even impossible to tell in advance what the condition number of a matrix will be without having explicit access to all the entries of that matrix. However, for matrices whose entries are chosen randomly, we can often give a fairly precise prediction of the condition number of that matrix based on very few parameters, such as the dimensions of the matrix.

Theorem 2.1. *Let A be a $d \times n$ matrix (with $d \geq n$) whose entries are independent standard normal random variables . Let $\sigma_{\min}$ and σ_{max} be the smallest and largest singular value of A (i.e., the smallest and largest eigenvalue of A^*A), respectively. Let $\gamma = \frac{d}{n}$ and let $\kappa(A) = \frac{\sigma_{\max}}{\sigma_{\min}}$ denote the condition number of A. Then, as $d \to \infty, n \to \infty$ (but with $\frac{d}{n} = \gamma$ fixed),*

$$\kappa(A) \to \frac{1+\sqrt{\gamma}}{1-\sqrt{\gamma}}.$$

In Matlab we can generate A via `A = randn(d,n)`. Figure 11 shows the distribution of $\kappa(A)$ of a 1000×250 Gaussian random matrix for 10000 different realizations of A. One can clearly see how well $\kappa(A)$ is concentrated near the value $\frac{1+\sqrt{\gamma}}{1-\sqrt{\gamma}}$ (which is equal to 3 in this case).