Working in High-Dimensional Spaces

Prof. Dan A. Simovici

UMB
1. The Euler Functions and the Volume of a Sphere
2. The Dimensionality Curse
3. Other Areas of Difficulty
4. Mathematical Background
The integrals

\[ B(a, b) = \int_0^1 x^{a-1}(1 - x)^{b-1} \, dx \quad \text{and} \quad \Gamma(a) = \int_0^\infty x^{a-1}e^{-x} \, dx, \]

are known as Euler’s integral of the first type and Euler’s integral of the second type, respectively. We assume here that \( a \) and \( b \) are positive numbers to ensure that the integrals are convergent. Mathematical properties of \( B(a, b) \) are discussed on Slide 24.
$B$ satisfies the equality:

$$B(a, b) = \frac{a - 1}{a + b - 1} \cdot B(a - 1, b).$$

Note that $B(a, 1) = \frac{1}{a}$ and

$$B(m, n) = \frac{(n - 1)!(m - 1)!}{(m + n - 1)!}$$

for $m, n \in \mathbb{N}$. See Silde 25.
The connection between Euler’s integral functions: is

\[ B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a + b)}. \]  

(1)

(see 26 and 27)
The Γ function is a generalization of the factorial. Starting from the definition of Γ and integrating by parts, we obtain

\[
\Gamma(x) = \int_0^\infty x^{a-1} e^{-x} \, dx = \frac{x^a}{a} e^{-x} \bigg|_0^\infty + \frac{1}{a} \int_0^\infty x^a e^{-x} \, dx = \frac{1}{a} \Gamma(a + 1).
\]

Thus, \(\Gamma(a + 1) = a \Gamma(a)\). Since \(\Gamma(1) = \int_0^\infty e^{-x} \, dx = 1\), it is easy to see that \(\Gamma(n + 1) = n!\) for \(n \in \mathbb{N}\).
\( \Gamma \) has derivatives of arbitrary order and that we can compute these derivatives by deriving the function under the integral sign. Namely, we can write:

\[
\Gamma'(a) = \int_0^\infty x^{a-1}(\ln x)e^{-x} \, dx,
\]

and, in general, \( \Gamma^{(n)}(a) = \int_0^\infty x^{a-1}(\ln x)^n e^{-x} \, dx \). Thus, \( \Gamma^{(2)}(a) > 0 \), which shows that the first derivative is increasing.
Since \( \Gamma(1) = \Gamma(2) = 1 \), there exists \( a \in [1, 2] \) such that \( \Gamma'(a) = 0 \). For \( 0 < x < a \), we have \( \Gamma'(x) \leq 0 \), so \( \Gamma \) is decreasing. For \( x > a \), \( \Gamma'(x) \geq 0 \), so \( \Gamma \) is increasing. It is easy to see that
\[
\lim_{x \to 0^+} \Gamma(x) = \frac{\Gamma(x + 1)}{x} = \infty,
\]
and \( \lim_{x \to \infty} \Gamma(x) = \infty \).

An integral that is useful for a variety of applications is
\[
I = \int_{\mathbb{R}} e^{-\frac{1}{2}t^2} dt.
\]

We prove that \( I = \sqrt{2\pi} \) on Slide 28.
Thus, $I = \sqrt{2\pi}$. Since $e^{-\frac{1}{2}t^2}$ is an even function, it follows that

$$\int_0^\infty e^{-\frac{1}{2}t^2} dt = \sqrt{\frac{\pi}{2}}.$$ 

Using this integral, we can compute the value of $\Gamma\left(\frac{1}{2}\right)$. Note that Since $\Gamma\left(\frac{1}{2}\right) = \int_0^\infty \frac{e^{-x}}{\sqrt{x}} dx$, by applying the change of variable $x = \frac{t^2}{2}$, we have

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{2} \cdot \int_0^\infty e^{-\frac{1}{2}t^2} dt = \sqrt{\pi}. \quad (2)$$

The last equality allows us to compute the values of the form $\Gamma\left(\frac{2p+1}{2}\right)$. It is easy to see that

$$\Gamma\left(\frac{2p+1}{2}\right) = \frac{(2p-1) \cdot (2p-3) \cdots 3 \cdot 1}{2^p} \sqrt{\pi} = \frac{(2p)!}{p!2^{2p}} \sqrt{\pi}. \quad (3)$$
A *closed sphere* centered in \((0, \ldots, 0)\) and having the radius \(R\) in \(\mathbb{R}^n\) is defined as the set of points:

\[
S_n(R) = \left\{ (x_1, \ldots, x_n) \in \mathbb{R}^n \mid \sum_{i=1}^{n} x_i^2 = 1 \right\}.
\]

The volume of this sphere is denoted by \(V_n(R)\) and is

\[
\frac{\pi^{\frac{n}{2}} R^n}{\Gamma \left( \frac{n}{2} + 1 \right)}.
\]

For \(n = 1, 2, 3\), by applying Formula (3), we obtain the well-known values \(2R, \pi R^2\), and \(\frac{4\pi R^3}{3}\), respectively. For \(n = 4\), the volume of the sphere is \(\frac{\pi^2 R^4}{2}\).

See Slides 30–32 for mathematical details.
The term “dimensionality curse,” invented by Richard Bellman, is used to describe the difficulties of exhaustively searching a space of high dimensionality for an optimum value of a function defined on such a space. These difficulties stem from the fact that the size of the sets that must be searched increases exponentially with the number of dimensions. Phenomena that are at variance with the common human intuition acquired in two- or three-dimensional spaces become more significant.
The dimensionality curse impacts many data mining tasks, including classification and clustering. Thus, it is important to realize the limitations that working with high-dimensional data impose on designing data mining algorithms.

Let $Q_n(\ell)$ be an $n$-dimensional cube in $\mathbb{R}^n$. The volume of this cube is $\ell^n$. Consider the $n$-dimensional closed sphere of radius $R$ that is centered in the center of the cube $Q_n(2R)$ and is tangent to the opposite faces of this cube. We have:

$$\lim_{n \to \infty} \frac{V_n(R)}{2^n R^n} = \frac{\pi^{n/2}}{2^n \Gamma \left( \frac{n}{2} + 1 \right)} = 0.$$ 

In other words, as the dimensionality of the space grows, the fraction of the cube volume that is located inside the sphere decreases and tends to become negligible for very large values of $n$. 
It is interesting to compare the volumes of two concentric spheres of radii $R$ and $R(1 - \epsilon)$, where $\epsilon \in (0, 1)$. The volume located between these spheres relative to the volume of the larger sphere is

$$\frac{V_n(R) - V_n(R(1 - \epsilon))}{V_n(R)} = 1 - (1 - \epsilon)^n,$$

and we have

$$\lim_{n \to \infty} \frac{V_n(R) - V_n(R(1 - \epsilon))}{V_n(R)} = 1.$$

Thus, for large values of $n$, the volume of the sphere of radius $R$ is concentrated mainly near the surface of this sphere.
Let $Q_n(1)$ be a unit side-length $n$-dimensional cube, $Q_n(1) = [0, 1]^n$, centered in $c_n = (0.5, \ldots, 0.5) \in \mathbb{R}^n$. The $d_2$-distance between the center of the cube $c_n$ and any of its vertices is $\sqrt{0.5^2 + \cdots 0.5^2} = 0.5 \sqrt{n}$, and this value tends to infinity with the number of dimensions $n$ despite the fact that the volume of the cube remains equal to 1. On the other hand, the distance from the center of the cube to any of its faces remains equal to 0.5. Thus, the $n$-dimensional cube is exhibits very different properties in different directions; in other words the $n$-dimensional cube is an anisotropic object.
Let \( P = (p, \ldots, p) \in \mathbb{R}^n \) be a point located on the main diagonal of \( Q_n(1) \) and let \( K \) be the subcube of \( Q_n(1) \) that includes \((0, \ldots, 0)\) and \( P \) and has a side of length \( p \); similarly, let \( K' \) be the subcube of \( Q_n(1) \) that includes \( P \) and \((1, \ldots, 1)\) and has side of length \( 1 - p \).

The ratio of the volumes \( V \) and \( V' \) of the cubes \( K \) and \( K' \) is

\[
r(p) = \left( \frac{p}{1 - p} \right)^n .
\]
To determine the increase $\delta$ of $p$ needed to double the volume of this ratio, we must find $\delta$ such that $\frac{r(p+\delta)}{r(p)} = 2$, that is

$$\frac{p(1 - p) + \delta(1 - p)}{p(1 - p) - \delta p} = \sqrt{2}.$$ 

Equivalently, we have

$$\delta = \frac{p(1 - p)(\sqrt{2} - 1)}{1 - p + p\sqrt{2}}.$$ 

The first factor $\frac{p(1-p)}{1-p+p\sqrt{2}}$ remains almost constant for large values of $n$. However, the second factor $\sqrt{2} - 1$ tends toward 0, which shows that within large dimensionality smaller and smaller moves of the point $p$ are needed to double the ratio of the volumes of the cubes $K$ and $K'$. 
This suggests that the division of $Q_n(1)$ into subcubes is very unstable. If data classifications are attempted based on the location of data vectors in subcubes, this shows in turn the instability of such classification schemes.
Another interesting example of the counterintuitive behavior of spaces of high dimensionality: Now let $Q_n(1)$ be the unit cube centered in the point $c_n \in \mathbb{R}^n$, where $c_n = (0.5, \ldots, 0.5)$. For $n = 2$ or $n = 3$, it is easy to see that every sphere that intersects the sides of $Q_2(1)$ or all faces of $Q_3(1)$ must contain the center of the cube $c_n$. 

We shall see that, for sufficiently high values of $n$ a sphere that intersects all $(n - 1)$-dimensional faces of $Q_n(1)$ does not necessarily contain the center of $Q_n(1)$.

Consider the closed sphere $B(q_n, r)$, whose center is the point $q_n = (q, \ldots, q)$, where $q \in [0, 1]$. Clearly, we have $q_n \in Q_n(1)$ and $d_2(c_n, q_n) = \sqrt{n(q^2 - q + 0.25)}$. 
If the radius \( r \) of the sphere \( B(q_n, r) \) is sufficiently large, then \( B(q_n, r) \) intersects all faces of \( Q_n \). Indeed, the distance from \( q_n \) to an \((n - 1)\)-dimensional face is no more than \( \max\{q, 1 - q\} \), which shows that \( r \geq \max\{q, 1 - q\} \) ensures the nonemptiness of all these intersections. Thus, the inequalities

\[
    n(q - 0.5)^2 > r^2 > \max\{q^2, (1 - q)^2\}
\]

(4)

ensure that \( B(q_n, r) \) intersects every \((n - 1)\)-dimensional face of \( Q_n \), while leaving \( c_n \) outside \( B(q_n, r) \). This is equivalent to requiring

\[
    n > \frac{\max\{q^2, (1 - q)^2\}}{(q - 0.5)^2}.
\]
For example, if we choose $q = 0.3$, then $n > \frac{0.7^2}{0.2^2} = 12.25$. Thus, in the case of $R^{13}$, Inequality (4) amounts to $0.52 > r^2 > 0.49$. Choosing $r = \frac{\sqrt{2}}{2}$ gives the sphere with the desired “paradoxical” property.

The examples discussed suggest that precautions and sound arguments are needed when trying to extrapolate familiar properties of two- or three-dimensional spaces to spaces of higher dimensionality.
Combinatorial Explosion

In some problems, each variable can take one of several discrete values. Taking the variables as an aggregate, a huge number of combinations of values must be considered.

This effect is also known as the **combinatorial explosion**. Even in the simplest case of binary variables, the number of possible combinations already is $2^d$, exponential in the dimensionality. Each additional dimension doubles the effort needed to try all combinations.
There is an exponential increase in volume associated with adding extra dimensions to a mathematical space.

**Example**

100 evenly spaced sample points suffice to sample a unit interval $[0, 1]$ with no more than 0.01 distance between points. An equivalent sampling of a 10-dimensional unit hypercube with a lattice that has a spacing of 0.01 between adjacent points would require $10^{20} = (10^2)^{20}$ sample points. In the above example $n=2$: when using a sampling distance of 0.01 the 10-dimensional hypercube appears to be $10^{18}$ "larger" than the unit interval.
Replacing $x$ by $1 - x$ yields the equality

$$B(a, b) = -\int_{1}^{0} (1 - x)^{a-1}(x)^{b-1} dx = B(b, a),$$

which shows that $B$ is symmetric.

Integrating $B(a, b)$ by parts, we obtain

$$B(a, b) = \int_{0}^{1} x^{a-1}(1 - x)^{b-1} dx = \int_{0}^{1} (1 - x)^{b-1} d\frac{x^a}{a}$$

$$= \frac{x^a(1 - x)^{1-b}}{a} \bigg|_{0}^{1} + \frac{b-1}{a} \int_{0}^{1} x^a(1 - x)^{b-2} dx$$

$$= \frac{b-1}{a} \int_{0}^{1} x^{a-1}(1 - x)^{b-2} dx - \frac{b-1}{a} \int_{0}^{1} x^{a-1}(1 - x)^{b-1} dx$$

$$= \frac{b-1}{a} B(a, b - 1) - \frac{b-1}{a} B(a, b),$$

which yields

$$B(a, b) = \frac{b-1}{a + b - 1} B(a, b - 1). \quad (5)$$
The symmetry of the function $B$ allows us to infer the formula

$$B(a, b) = \frac{a - 1}{a + b - 1} \cdot B(a - 1, b).$$

If $b$ is a natural number $n$, a repeated application of Equality (5) allows us to write

$$B(a, n) = \frac{n - 1}{a + n - 1} \cdot \frac{n - 2}{a + n - 2} \cdots \frac{1}{a + 1} \cdot B(a, 1).$$

The last factor of this equality, $B(a, 1)$, is easily seen to equal $\frac{1}{a}$. Thus,

$$B(a, n) = B(n, a) = \frac{1 \cdot 2 \cdots (n - 1)}{a \cdot (a + 1) \cdots (a + n - 1)}.$$

If $a$ is also a natural number, $a = m \in \mathbb{N}$, then

$$B(m, n) = \frac{(n - 1)!(m - 1)!}{(m + n - 1)!}.$$
Next, we show the connection between Euler’s integral functions:

\[ B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a + b)}. \]  

Replacing \( x \) in the integral

\[ \Gamma(a) = \int_0^\infty x^{a-1} e^{-x} \, dx \]

by \( x = ry \) with \( r > 0 \) gives \( \Gamma(a) = r^a \int_0^\infty y^{a-1} e^{-ry} \, dy \).

Replacing \( a \) by \( a + b \) and \( r \) by \( r + 1 \) yields the equality

\[ \Gamma(a + b)(r + 1)^{-(a+b)} = \int_0^\infty y^{a+b-1} e^{-(r+1)y} \, dy. \]

By multiplying both sides by \( r^{a-1} \) and integrating, we have

\[ \Gamma(a+b) \int_0^\infty r^{a-1}(r+1)^{-(a+b)} \, dr = \int_0^\infty r^{a-1} \left( \int_0^\infty y^{a+b-1} e^{-(r+1)y} \, dy \right) \, dr. \]

By the definition of \( B \), the last equality can be written

\[ \Gamma(a + b)B(a, b) = \int_0^\infty r^{a-1} \left( \int_0^\infty y^{a+b-1} e^{-(r+1)y} \, dy \right) \, dr. \]
By permuting the integrals from the right member (we omit the justification of this manipulation), the last equality can be written as

$$\Gamma(a + b)B(a, b) = \int_0^\infty y^{a+b-1}e^{-y} \left( \int_0^\infty r^{a-1}e^{-ry} dr \right) dy.$$ 

Note that $\int_0^\infty r^{a-1}e^{-ry} dr = \frac{\Gamma(a)}{y^a}$. Therefore,

$$\Gamma(a+b)B(a, b) = \int_0^\infty y^{a+b-1}e^{-y} \frac{\Gamma(a)}{y^a} dy = \int_0^\infty y^{b-1}e^{-y}\Gamma(a)dy = \Gamma(a)\Gamma(b),$$
We can write

\[ I^2 = \int_{\mathbb{R}} e^{-\frac{1}{2}x^2} \, dx \cdot \int_{\mathbb{R}} e^{-\frac{1}{2}y^2} \, dy = \int_{\mathbb{R}^2} e^{-\frac{x^2+y^2}{2}} \, dxdy. \]

Changing to polar coordinates by using the transformation \( x = \rho \cos \theta \) and \( y = \rho \sin \theta \) whose Jacobian is

\[
\begin{vmatrix}
\frac{\partial x}{\partial \rho} & \frac{\partial x}{\partial \theta} \\
\frac{\partial y}{\partial \rho} & \frac{\partial y}{\partial \theta}
\end{vmatrix} = \begin{vmatrix}
\cos \theta & -\rho \sin \theta \\
\sin \theta & \rho \cos \theta
\end{vmatrix} = \rho,
\]

we have

\[ I^2 = \int_{\mathbb{R}^2} e^{-\frac{\rho^2}{2}} \rho \, d\rho \, d\theta = \int_0^{2\pi} d\theta \int_0^\infty e^{-\frac{\rho^2}{2}} \rho \, d\rho = 2\pi. \]
Computation of the Volume of a Sphere in $\mathbb{R}^n$

We approximate the volume of an $n$-dimensional sphere of radius $R$ as a sequence of $n-1$-dimensional spheres of radius $r(u) = \sqrt{R^2 - u^2}$, where $u$ varies between $-R$ and $R$. This allows us to write

$$V_{n+1}(R) = \int_{-R}^{R} V_n(r(u)) du.$$
We seek $V_n(R)$ as a number of the form $V_n(R) = k_n R^n$. Thus, we have

$$V_{n+1}(R) = k_n \int_{-R}^{R} (r(u))^n du = k_n \int_{-R}^{R} (R^2 - u^2)^{\frac{n}{2}} du$$

$$= k_n R^n \int_{-R}^{R} \left(1 - \left(\frac{u}{R}\right)^2\right)^{\frac{n}{2}} du$$

$$= V_n(R) \int_{-R}^{R} \left(1 - \left(\frac{u}{R}\right)^2\right)^{\frac{n}{2}} du = RV_n(R) \int_{-1}^{1} (1 - x^2)^{\frac{n}{2}} dx.$$ 

In turn, this yields the recurrence

$$k_{n+1} = k_n \int_{-1}^{1} (1 - x^2)^{\frac{n}{2}} dx.$$
Note that
\[ \int_{-1}^{1} (1 - x^2)^{\frac{n}{2}} \, dx = 2 \cdot \int_{0}^{1} (1 - x^2)^{\frac{n}{2}} \, dx \]
because the function \( (1 - x^2)^{\frac{n}{2}} \) is even. To compute the latest integral, substitute \( u = x^2 \). We obtain
\[ \int_{0}^{1} (1 - x^2)^{\frac{n}{2}} \, dx = \frac{1}{2} \int_{0}^{1} u^{-\frac{1}{2}} (1 - u)^{\frac{n}{2}} \, du, \]
which equals \( \frac{1}{2} \cdot B\left(\frac{1}{2}, \frac{n}{2} + 1\right) \). Using the \( \Gamma \) function, the integral can be written as
\[ \int_{0}^{1} (1 - x^2)^{\frac{n}{2}} \, dx = \frac{1}{2} \cdot \frac{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n}{2} + 1\right)}{\Gamma\left(\frac{n}{2} + \frac{3}{2}\right)}. \]
Thus,

\[ k_{n+1} = k_n \frac{\Gamma \left( \frac{1}{2} \right) \Gamma \left( \frac{n}{2} + 1 \right)}{\Gamma \left( \frac{n+1}{2} + 1 \right)} . \]

Since \( k_1 = 2 \), this implies

\[
k_n = 2 \left( \frac{\Gamma \left( \frac{1}{2} \right)}{\Gamma \left( \frac{n}{2} + 1 \right)} \right)^{n-1} \frac{\Gamma \left( \frac{1}{2} + 1 \right)}{\Gamma \left( \frac{n}{2} + 1 \right)} = \left( \frac{\Gamma \left( \frac{1}{2} \right)}{\Gamma \left( \frac{n}{2} + 1 \right)} \right)^n \frac{1}{\Gamma \left( \frac{n}{2} + 1 \right)} = \pi^{\frac{n}{2}} \frac{1}{\Gamma \left( \frac{n}{2} + 1 \right)} .
\]

Thus, the volume of the \( n \)-dimensional sphere of radius \( R \) equals

\[
\frac{\pi^{\frac{n}{2}} R^n}{\Gamma \left( \frac{n}{2} + 1 \right)} .
\]

For \( n = 1, 2, 3 \), by applying Formula (3), we obtain the well-known values 2\( R \), \( \pi R^2 \), and \( \frac{4\pi R^3}{3} \), respectively. For \( n = 4 \), the volume of the sphere is \( \frac{\pi^2 R^4}{2} \).