1 Sampling in High-Dimensional Space

Consider generating points in $\mathbb{R}^d$ using a spherical Gaussian distribution. We claim that their pairwise distances are essentially the same for large d. This is because for two randomly sampled points $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, their ℓ_2 distance

$$\|\mathbf{x} - \mathbf{y}\|_2 = \Bigg(\underbrace{\sum_{i=1}^{d} |x_i - y_i|^2}_{\text{sum of independent r.v.}} \Bigg)^{-1/2}$$

can be expressed as a summation of independent random variables. So, we can apply the Law of Large Numbers.

Theorem (Law of Large Numbers). *Let $x_1, x_2, \ldots, x_n \in \mathbb{R}$ be independent samples of a random variable x. Then,*

$$\Pr\left[\left| \frac{x_1 + x_2 + \ldots + x_n}{n} - \mathrm{E}(x) \right| > \varepsilon \right] \le \frac{\mathrm{Var}(x)}{n\varepsilon^2}.$$

This can be proved using Chebyshev inequality, which can be derived from Markov inequality.

In addition to the Law of Large Number, the Central Limit Theorem says that in the limit as the sample size $n \to \infty$, the distribution of the sample average is Gaussian, provided that the random variable x has finite variance.

The d-variate spherical Gaussian distribution with variance σ^2 is given by the probability density function

$$G(\mathbf{x}) = \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\Bigg[-(\underbrace{x_1^2 + x_2^2 + \ldots + x_d^2}_{\text{symmetric quadratic form}})/(2\sigma^2) \Bigg].$$

Notice that the exponent is a symmetric quadratic form. Therefore, its level sets $\{G^{-1}(c)\}$ are spheres (Figure 1(a)).

1.1 Examples of High-Dimensional Data

It is useful to think of n random variables $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$ as a data matrix

$$D = \begin{pmatrix} x_1^1 & x_1^2 & \cdots & x_1^d \\ \vdots & \vdots & \ddots & \vdots \\ x_n^1 & x_n^2 & \cdots & x_n^d \end{pmatrix},$$

where each row represents a data sample. Some examples include protein structures, pixels in spectral imaging, and genes.

In protein molecules, there is a backbone of atoms connected to many side-chains. It is specified by $\{P, Q, R\}$, where P represents the positions of the atoms, Q represents the partial charges (they are like dipoles), and R represents the types of atoms. A high degree of freedom comes from the dihedral angles $\{\phi_i, \psi_i, \chi_i\}$. See [B1, Appendix A] for more details.

The protein folding problem is to find a configuration that minimizes free energy. The ℓ_2-distance r_{ij} is important in the calculation of the Coulombic energy term given by $E_\mathrm{q} = \sum_{ij} q_i q_j / r_{ij}$.

1.2 Quadratic Form

In general, a d-variate polynomial of degree 2 has the form

$$P(\mathbf{x}) = \sum_{i_1 + i_2 + \ldots + i_d \le 2} a_{\mathbf{i}} \mathbf{x}^{\mathbf{i}}.$$

(If all $a_{\mathbf{i}}$ are in $\mathbb{R}$, then P is called a *real polynomial.* Similarly, if all $a_{\mathbf{i}}$ are in $\mathbb{C}$, then P is called a *complex polynomal.* Polynomial equations are also known as *algebraic equations.*) When $d = 2$, the degree 2 monic monomials are $\{x_1^2, x_2^2, x_3^2, x_1x_2, x_1x_3, x_2x_3\}$, the degree 1 monic monomials are $\{x_1, x_2, x_3\}$ and the degree 0 monic monomial is $\{1\}$. More generally, total number of monic monomials of degree at most e is given by the binomial coefficient $\binom{d+e}{d} = \binom{d+e}{e}$. The monic monomials forms a basis for a vector space.

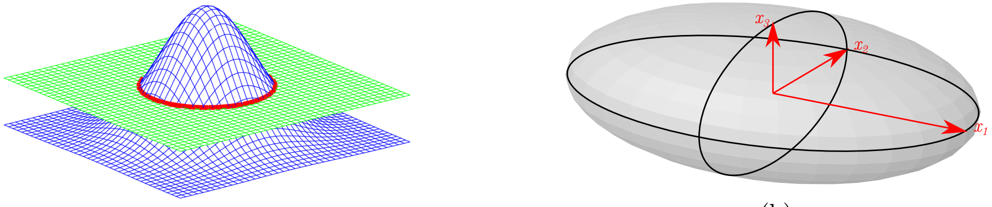
The equation of the unit sphere centered at origin

$$x_1^2 + x_2^2 + x_3^2 = 1$$

is an example of quadratic equation. If it is centered at (c_1, c_2, c_3) instead, then the equation becomes

$$(x_1 - c_1)^2 + (x_2 - c_2)^2 + (x_3 - c_3)^2 = 1.$$

The equations of spheres have the property that all coefficients of x_i^2 are equal. A spherical Gaussian is also called an *isotropic* Gaussian.



(b)

(a)

Figure 1: (a) A circular level set of a bi-variate Gaussian. (b) The three semi-axes of an ellipsoid.

Using spectral theory, an ellipsoid corresponds to a quadratic form where the semi-axes correspond to the eigenvectors (Figure 1(b)). The degree of the polynomial is also related to the *geometric degree* of the manifold. In the case of sphere, its geometric degree equals 2.

2 Properties of the Euclidean Ball

We claim that the volume of a unit ball in $\mathbb{R}^d$ tends to 0 as $d \to \infty$. Meanwhile, it is trivial to see that a unit hypercube $\{\mathbf{x} \in \mathbb{R}^d : |x_i| \leq 1/2\}$ $\mathbb{R}^d$ has volume exactly 1.

When $d = 2$, the unit square having a semi-diagonal of length $1/\sqrt{2}$ is totally contained inside the unit disk which has area π (Figure 2(a)). When $d = 3$, the unit cube having a semi-diagonal of length $\sqrt{3}/2$ is also totally contained inside the unit ball which has volume $4\pi/3$ (Figure 2(b)). When $d = 4$, the unit hypercube having a semi-diagonal of length 1 is contained inside the unit ball and has its corners lying on the boundary of the ball (Figure 2(c)). For large d, the unit hypercube having semi-diagonal of length $\sqrt{0.5^2 + 0.5^2 + \ldots 0.5^2} = \sqrt{d}/2$ extends outside the unit ball. In fact, most volume of the unit hypercube is outside the unit ball (Figure 2(d)).

The following lemma shows that most volume of a unit d-ball is within a narrow band of width $O(1/\sqrt{d})$ (Figure 3(a)). Notice that the volume decays like a Gaussian as the band width increases.

Lemma ([BHK, Theorem 2.7])**.** *Let $c \geq 1$ and $d \geq 3$. Then, at least a $1 - \frac{2}{c}e^{-c^2/2}$ fraction of the volume of the unit d-ball centered at origin has $|x_1| \leq c/\sqrt{d-1}$.*

Let $V(d)$ and $A(d)$ be the volume and the surface area of the unit d-ball respectively. The following lemma shows that most volume of a unit d-ball is contained in an annulus of width $O(1/d)$ (Figure 3(b)).

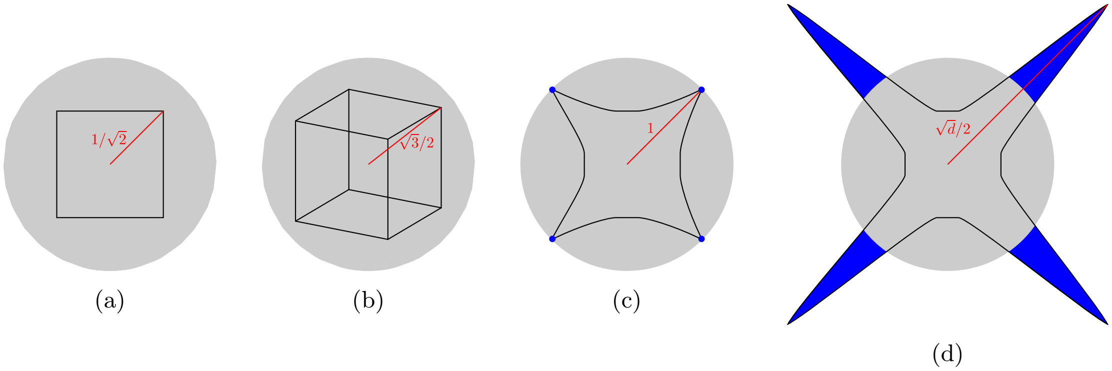


Figure 2: (a) A unit square contained inside a unit disk in $\mathbb{R}^2$. (b) A unit cube contained inside a unit ball in $\mathbb{R}^3$. (c) A unit hypercube circumscribed inside a unit ball in $\mathbb{R}^4$. (d) A unit hypercube extending outside a unit ball in $\mathbb{R}^d$.

Lemma. *For any $c > 0$, all but e^{-c} of the volume of a unit d-ball is contained in an annulus of width $\varepsilon = c/d$.*

Proof.
$$\frac{\text{volume of } d\text{-ball with radius } (1-\varepsilon)}{\text{volume of } d\text{-ball with radius } 1} = \frac{(1-\varepsilon)^d V(d)}{V(d)} = (1-\varepsilon)^d \leq e^{-c}. \qquad \square$$

In Cartesian coordinates, the volume $V(d)$ can be expressed as the following integral.

$$V(d) = \int\limits_{x_1=-1}^{1} \int\limits_{x_2=-\sqrt{1-x_1^2}}^{\sqrt{1-x_1^2}} \cdots \int\limits_{x_d=-\sqrt{1-x_1^2-x_2^2-\ldots-x_{d-1}^2}}^{\sqrt{1-x_1^2-x_2^2-\ldots-x_{d-1}^2}} \mathrm{d}\, x_d \ldots \mathrm{d}\, x_1 \,\mathrm{d}\, x_2$$

In polar coordinates, the volume $V(d)$ can be expressed as the following integral, where $\mathbf{\Omega}$ is the solid angle.

$$V(d) = \int_{S^d} \int_{r=0}^{1} r^{d-1} \,\mathrm{d}\, r \,\mathrm{d}\,\mathbf{\Omega} = \frac{1}{d} A(d)$$

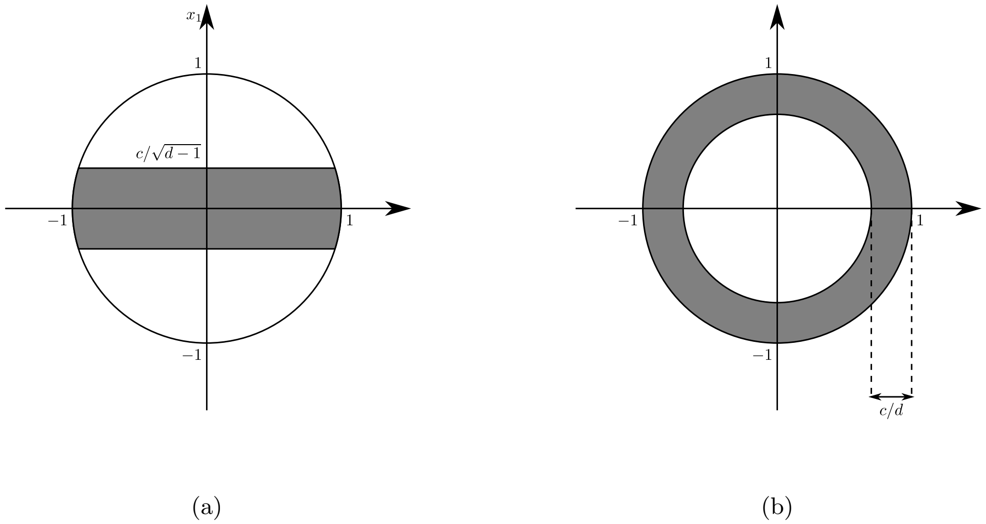
The formula of the surface area $A(d)$ is given by

$$A(d) = \frac{2\pi^{d/2}}{\Gamma(d/2)},$$

where the gamma function satisfies $\Gamma(x) = (x-1)\Gamma(x-1)$, $\Gamma(1) = \Gamma(2) = 1$ and $\Gamma(1/2) = \sqrt{\pi}$. Notice that the denominator is growing like factorial, which is faster than the exponential growth of the numerator. So, its value tends to 0 as $d \to \infty$.

3 Random Samples from the unit ball (or Gaussians) in R^d

From Theorem 2.7 and the lemma above we learn that for the unit ball in high dimensions, most of its volume is **concentrated** near its equator. This also implies that if we draw two points (vectors) x and y at random from the unit ball in d-dimensions, with high probability they will be nearly orthogonal to each other. Specifically, with high probability, since both will have length 1-$O(\frac{1}{d})$, $\mathrm{E}(x^2) = \mathrm{E}(y^2) \approx 1$. Further since most of the volume of the unit ball lies in the thin slab $O(\frac{1}{\sqrt{d}})$ of points (i.e. near the equator), $\mathrm{E}(x \cdot y) \approx O(\frac{1}{\sqrt{d}})$ (i.e. projection or dot product or cosine of the angle between the two is very small . This



(a) (b)

Figure 3: (a) A band of width $2c/\sqrt{d-1}$ inside a unit ball. (b) An annulus of width c/d include a unit ball.

implies the angle between the two vectors is nearly $\frac{\pi}{2}$. Another way to say this that $\mathrm{E}(x-y)^2 \approx 2$ and using Pythagorean rule the angle between the two random vectors x and y is almost $\frac{\pi}{2}$.

The following theorem formalizes this and states that if we draw n points at random in the unit ball, with high probability all points will be close to unit length, and each pair of points will be almost orthogonal.

In the next lecture we shall formally extend this and study additional geometric properties of random samples from probability distributions, including Gaussians.

Lemma ([BHK, Theorem 2.8]). *Consider drawing n points $x_1, x_2, ..., x_n$ at random from a unit ball with probability $1 - O(\frac{1}{n})$*

- $|x_i| \geq$ *(1 - $\frac{2ln\ n}{d}$) for all i and d*
- $|x_i \cdot x_j| \leq \frac{\sqrt{6ln\ n}}{\sqrt{d-1}}$ *for all $i \neq j$*

References

[B1] Chandrajit Bajaj. *Molecular Structural Bioinformatics: A Computational Science Perspective.*

[BHK] Avrim Blum, John Hopcroft and Ravindran Kannan. *Foundations of Data Science.*