Assignment 9B

The volume of a ball of radius R in d dimensions is given by:

$$V_d(R) = \frac{R^d}{d}\frac{2\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}\right)} = \frac{\pi^{\frac{d}{2}}R^d}{\Gamma\left(\frac{d}{2}+1\right)}\,. \tag{1}$$

In this subsection, we discuss this result in relation to the so-called "curse of dimensionality" in machine learning, particularly in the context of regression.

Consider a hypercube of side length $2R$ centered at the origin. The ratio of the volume of the inscribed sphere to the hypercube is:

$$\text{Ratio}(d) = \frac{V_d(R)}{(2R)^d} = \frac{2^{-d}\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}+1\right)}\,. \tag{2}$$

For example, we have:

- For $d=1$: $\text{Ratio}(1) = 1$,
- For $d=2$: $\text{Ratio}(2) = \frac{\pi}{4} \approx 0.785$,
- For $d=3$: $\text{Ratio}(3) = \frac{\pi}{6} \approx 0.524$.

Figure 1 shows how this ratio decreases rapidly as the dimension d increases, as well as the estimated ratio based on the Stirling formula. This rapid decrease illustrates a fundamental feature of high-dimensional spaces: *as dimensionality increases, an increasingly small fraction of the hypercube's volume is occupied by the inscribed hypersphere.*

In the context of regression, this implies that most points in the space are actually far from any given set of training data. Imagine we have a dataset of function values at certain points in a d-dimensional space, and we aim to predict the value of the function outside this training dataset. Our ability to make predictions is presumably higher closer to the training points. However, the volume ratio shown above reveals that as dimensionality increases, an increasingly large portion of the space is far from any given set of data points. This geometric insight gives us a clue as to why regression is harder in high-dimensional problems:

1. A regular mesh in d dimensions contains $[L/(2R)]^d$ points, where L is the side length of the domain. This represents an exponential growth in the number of points needed to maintain a given mesh density as d increases.

2. Even with this large amount of data, most of the space is at a distance greater than R from the training points, where our predictive ability is limited.

These factors combine to make function approximation in high-dimensional spaces a challenging task. The only way to overcome the curse of dimensionality, introduced above, is to imbue "prior knowledge" about the function that we aim to learn within our regression model, using appropriate machine-learning techniques.

In this assignment, you will estimate the volume of a hypersphere using the Monte Carlo mean-value method and analyze how the effectiveness of the method changes with increasing dimension.

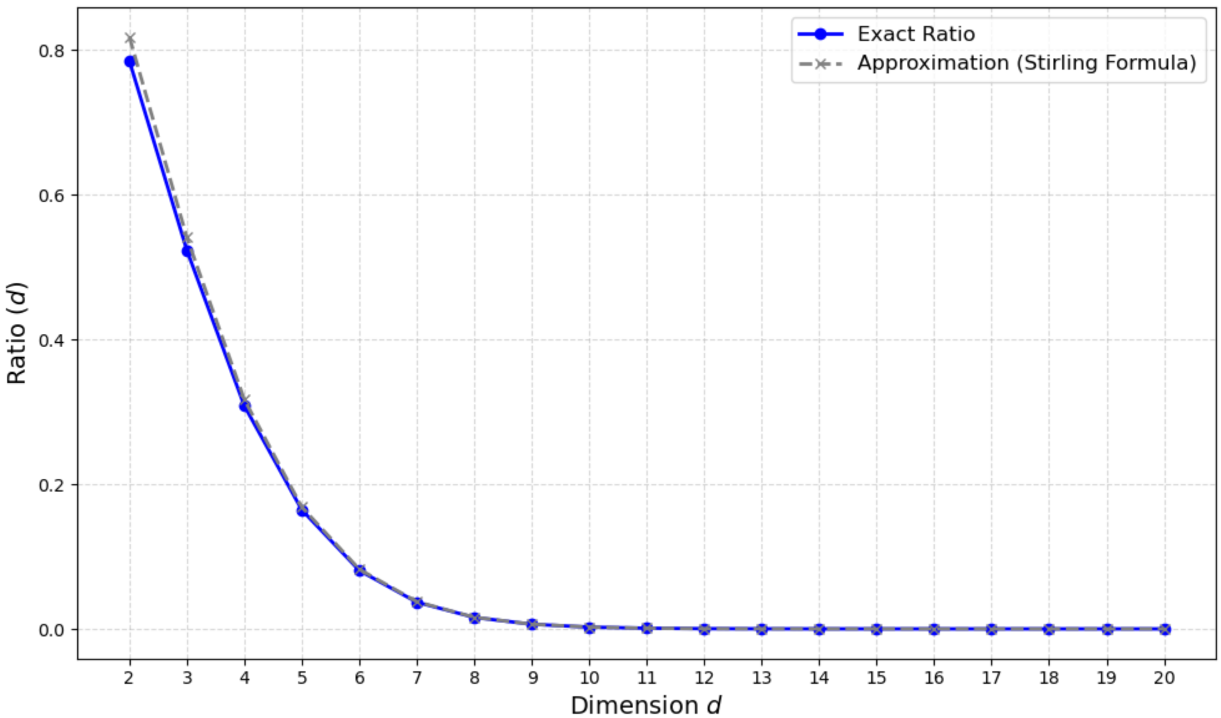

Figure 1: Ratio of sphere volume to hypercube volume as a function of dimension d.

- **Problem 1 (30 points):** Monte Carlo Estimation in Low Dimensions

  Use the mean-value method, as explained in Lecture 8, to estimate the volume of a unit hypersphere for dimensions $d = 2$ and $d = 10$. The volume can be expressed as the integral:

  $$V_d = \int_{[-1,1]^d} f(\mathbf{x})\, d^d x, \tag{3}$$

  where $f(\mathbf{x}) = 1$ inside the hypersphere ($\sum_{i=1}^{d} x_i^2 \leq 1$) and 0 otherwise.

  - Compute the volume using the Monte Carlo mean-value estimate:

    $$V_d \approx V_{\text{cube}} \langle f \rangle, \quad \text{where} \quad V_{\text{cube}} = 2^d\,, \tag{4}$$

    using $N = 100$ samples, $N = 10^6$ samples, and a few intermediate values.
  - Compute the statistical uncertainty and compare your results with the exact formula. Discuss the accuracy of your estimate.

- **Problem 2 (20 points):** High-Dimensional Effects and the Curse of Dimensionality

  Repeat the Monte Carlo calculation for $d = 15$ and $d = 20$, using the same number of samples. Compare your estimates with the exact formula and analyze how the relative error changes as d increases.

  Explain why Monte Carlo integration becomes increasingly unreliable as d increases and relate your observations to the curse of dimensionality.