Math 126 Problem 9: Lonely Points

Spring 2027

Submission Instructions

Please upload your work to Gradescope by **11:59pm on Monday, February 22, 2027 (Pacific time).**

- You are encouraged to discuss ideas and work with your classmates. However, you *must acknowledge* your collaborators at the top of each solution on which you collaborated with others and you *must write* your solutions independently.
- Your solutions to theory questions must be written legibly, or typeset in LaTeX or markdown.
- I recommend that you write your solutions in a Python notebook using Google Colab.
- You should submit your solutions as a *single PDF* via the assignment on Gradescope.

Grading: The point of the problem set is for *you* to learn. To this end, I hope to disincentivize the use of LLMs by *not* grading your work for correctness. Instead, you will grade your own work by comparing it to my solutions. This self-grade is due the Friday *after* the problem set is due, also on Gradescope.
Original notebook: Google Colab.

Lonely Points

Let $\mathbf{x}$ and $\mathbf{y}$ be two random points picked independently and uniformly from the d-dimensional hypercube on $[0,1]^d$. That is, the ith entry x_i is continuous, uniform random variable on $[0,1]$. We are interested in the Euclidean distance between $\mathbf{x}$ and $\mathbf{y}$, specifically the squared distance $D_d = \|\mathbf{x}-\mathbf{y}\|_2^2$.

Part A: Expected Distance

First, show that for a single dimension i, we have $\mathbb{E}[(x_i - y_i)^2] = 1/6$. Use linearity of expectation to find the expected squared distance $\mu_d = \mathbb{E}[D_d]$ for the full vectors.

Part B: Variance of Distance

We want to know if the distance varies a lot or stays close to the mean.Calculate the variance of one component: $\sigma_i^2 = \mathrm{Var}((x_i - y_i)^2)$. (You don't need the exact number, just show it is a constant k that doesn't depend on d). Then calculate the total variance $\mathrm{Var}(D_d)$. How does the "coefficient of variation" (the ratio of standard deviation to the mean) $\frac{\sqrt{\mathrm{Var}(D_d)}}{\mathbb{E}[D_d]}$ behave?

Part C: Lonely Points

Show that as $d \to \infty$, the coefficient of variation approaches 0. Suppose you draw 1,000 random points. You look for the "nearest neighbor" to $\mathbf{x}$ and the "farthest neighbor" to $\mathbf{x}$. Based on your limit, what can you say about the ratio of these two distances? Why does this make clustering algorithms difficult in high dimensions?