Project 3

Random Projection: Johnson-Lindenstrauss Lemma

- The goal of this project is to gain hands-on understandings of the JL lemma.
- Consider that there are $n = 100$ points $x_1, x_2, ..., x_n$ in high dimensional space R^d, with $d >> n$. That is, each $x_i \in R^d$ is a d-dimensional vector. Each entry of x_i is distributed i.i.d. according to $N(0, 1)$.
- Fix the tolerance level ϵ to be $1/3$.
- The main task is to identify the smallest k needed to preserve the pair-wise distance within $[1 - \epsilon, 1 + \epsilon]$ with $\epsilon = 1/3$ as in JL lemma.
- The projection matrix is $\frac{1}{\sqrt{k}}R$, where R is a $k \times d$ random matrix with each entry $R_{i,j}$ distributed i.i.d. according to $N(0, 1)$.
- Two Tasks:
  - For $d = 10^3, 10^4, 2 * 10^4$, plot how does the tolerance level changes with k. Apparently, smaller tolerant value is likely to require larger k.
  - Based on the plots, claim what would be the minimum k needed to preserve the pair-wise distance within $[1 - 1/3, 1 + 1/3]$. Explain your findings.
- WARNING: this project is NOT to plot the lower bound of k from JL lemma. You are suppose to recover the value of k based on computing all the pair-wise distances of the original data vs. the projected data and see how much they differ (quantified by the tolerance ϵ) for different k.