# Asmt 3: Distances and LSH

Vai Suliafu, u0742607
Monday, February 3

## Overview

In this assignment you will explore LSH and Euclidean distances.

You will use a data set for this assignment:

- `http://www.cs.utah.edu/~jeffp/teaching/cs5140/A3/R.csv`

*It is recommended that you use LaTeX for this assignment (or other option that can properly digitally render math). If you do not, you may lose points if your assignment is difficult to read or hard to follow. Find a sample form in this directory:* ***http://www.cs.utah.edu/~jeffp/teaching/latex/***

## 1 Choosing $r, b$ (35 points)

Consider computing an LSH using $t = 160$ hash functions. We want to find all object pairs which have Jaccard similarity above $\tau = .85$.

**A: (15 points)** Use the trick mentioned in class and the notes to estimate the best values of hash functions $b$ within each of $r$ bands to provide the S-curve

$$f(s) = 1 - (1 - s^b)^r$$

with good separation at $\tau$. Report these values:

$b = 16$
$r = 160/b = 10$

**B: (15 points)** Consider the 4 objects $A, B, C, D$, with the following pairwise similarities:

|   | A | B | C | D |
|---|---|---|---|---|
| A | 1 | 0.77 | 0.25 | 0.33 |
| B | 0.77 | 1 | 0.20 | 0.55 |
| C | 0.25 | 0.20 | 1 | 0.91 |
| D | 0.33 | 0.55 | 0.91 | 1 |

Using your choice of $r$ and $b$ and $f(\cdot)$, what is the probability of each pair of the four objects for being estimated to having similarity greater that $\tau = 0.85$? Report 6 numbers. *(Show your work.)*

$P(h(A) = h(B)) = 1 - (1 - 0.77^{16})^{10} = 0.14262712129925814$
$P(h(A) = h(C)) = 1 - (1 - 0.25^{16})^{10} = 2.3283064365386963e - 09$
$P(h(A) = h(D)) = 1 - (1 - 0.33^{16})^{10} = 1.977985021328621e - 07$
$P(h(B) = h(C)) = 1 - (1 - 0.20^{16})^{10} = 6.553646514362299e - 11$
$P(h(B) = h(D)) = 1 - (1 - 0.55^{16})^{10} = 0.0007009160597826192$
$P(h(C) = h(D)) = 1 - (1 - 0.91^{16})^{10} = 0.9178498654389164$

## 2 Generating Random Directions (30 points)

**A: (10 points)** Describe how to generate a single random unit vector in $d = 10$ dimensions using only the operation $u \leftarrow \mathsf{unif}(0, 1)$ which generates a uniform random variable between 0 and 1: *(This can be called multiple times.)*

We can begin by using the operation $u$ to generate two random variables $u_1$ and $u_2$. We can then use $u_1$ and $u_2$ as inputs in the *Box Muller Transform*, which generates a Gaussian RV $g$. Then by repeating the previous step $d$ times, we get a vector $\overrightarrow{g}$ in $R^d$, where each element $g_i$ is a Gaussian RV. Finally, normalizing the vector $\overrightarrow{g}$ results in a random uniform unit vector.

**B: (20 points)** Generate $t = 160$ unit vectors in $R^d$ for $d = 100$. Plot of cdf of their pairwise dot products (yes, you need to calculate $\binom{t}{2}$ dot products).

The CDF plot for pairwise dot products is shown below in Figure 1.

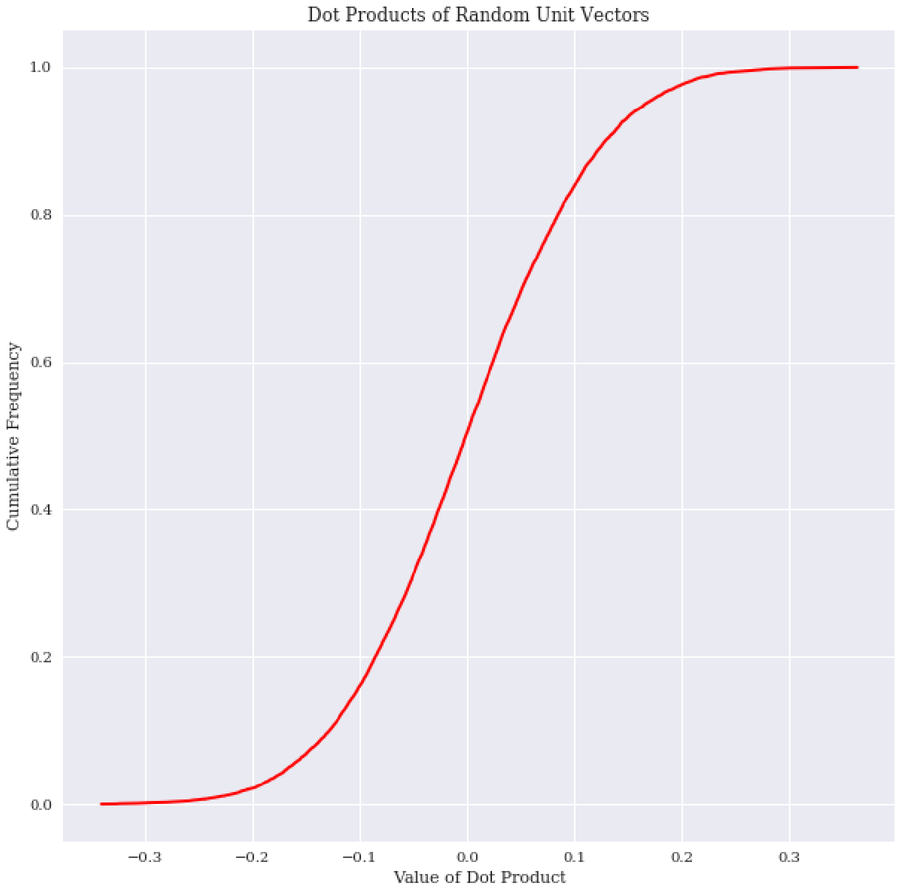

Figure 1: Cumulative Density of Dot Products from Random Unit Vectors

## 3 Angular Hashed Approximation (35 points)

Consider the $n = 500$ data points in $R^d$ for $d = 100$ in data set $R$, given at the top. We will use the angular similarity, between two vectors $a, b \in R^d$:

$$s_{\text{ang}}(a, b) = 1 - \frac{1}{\pi} \arccos(\langle \bar{a}, \bar{b} \rangle)$$

If $a, b$ are not unit vectors (e.g., in $S^{d-1}$), then we convert them to $\bar{a} = a/\|a\|_2$ and $\bar{b} = b/\|b\|_2$. The definition of $s_{\text{ang}}(a, b)$ assumes that the input are unit vectors, and it takes a value between 0 and 1, with as usual 1 meaning most similar.

**A: (15 points)** Compute all pairs of dot products *(Yes, compute $\binom{n}{2}$ values)*, and plot a cdf of their angular similarities. Report the number with angular similarity more than $\tau = 0.85$.

The CDF plot for angular similarities within R.csv is shown below in Figure 2.

The count of angle similarities greater than $\tau$ = 39283.

**B: (20 points)** Now compute the dot products and angular similarities among $\binom{t}{2}$ pairs of the $t$ random unit vectors from Q2.B. Again plot the cdf, and report the number with angular similarity above $\tau = 0.85$.

The CDF plot for angular similarities between all random unit vectors generated is shown below in Figure 3.

The count of angle similarities greater than $\tau$ = 0

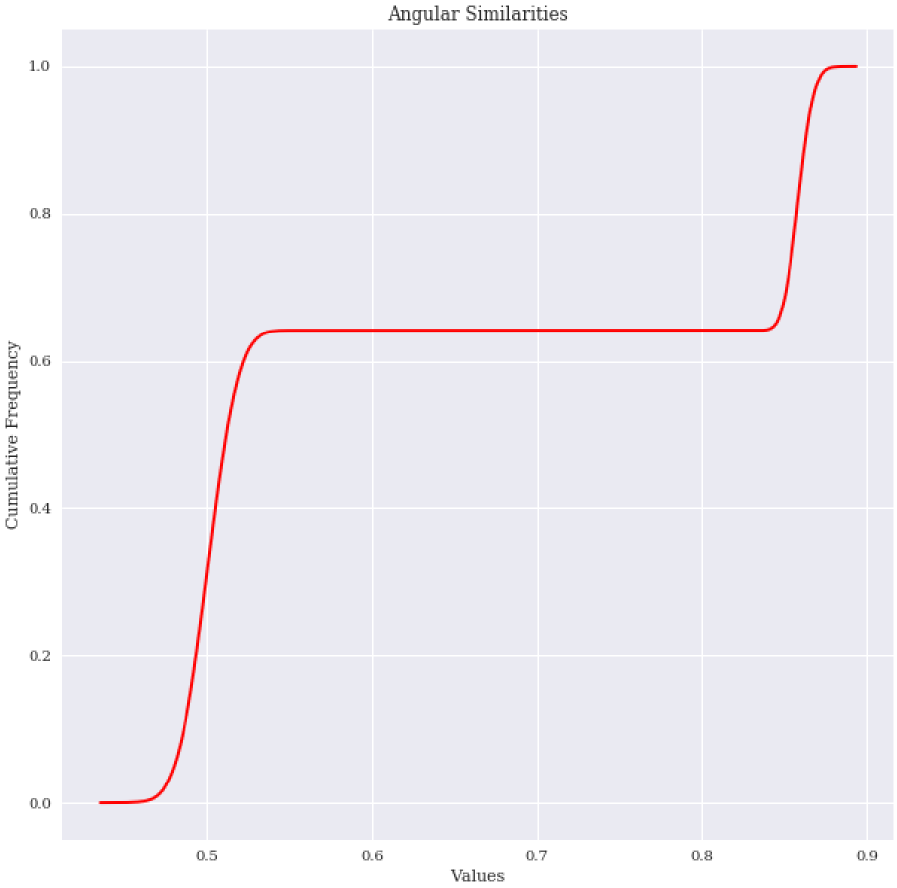

Figure 2: Cumulative Density of Angular Similarities in R

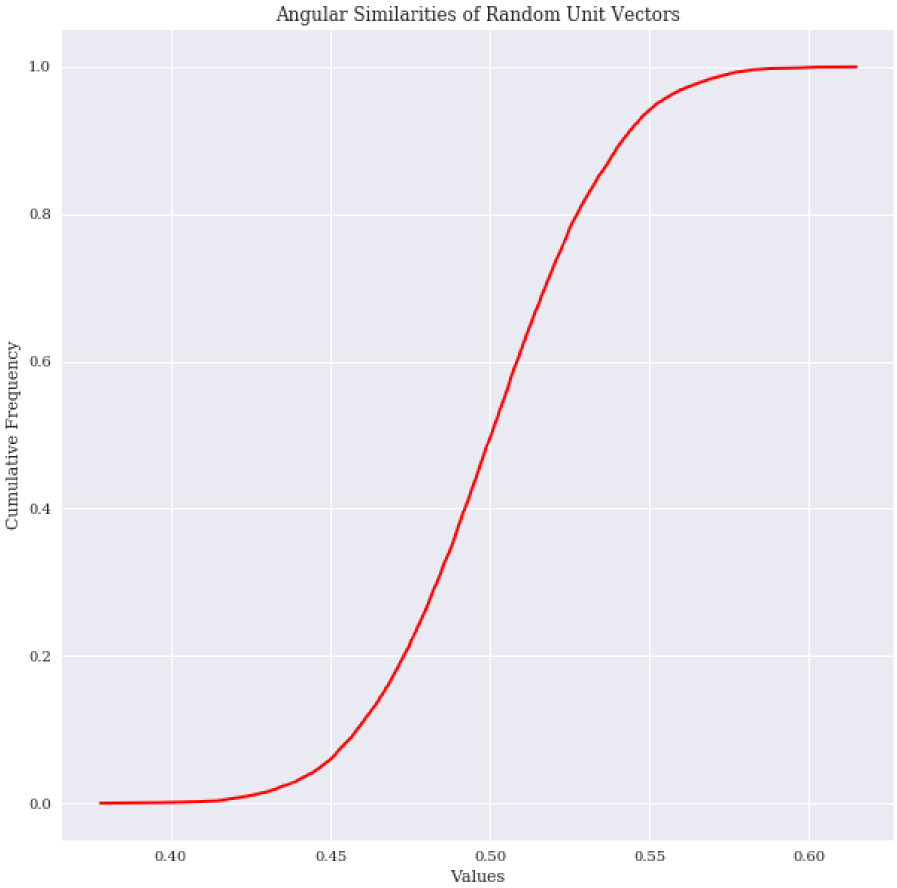

Figure 3: Cumulative Density of Angular Similarities from Random Unit Vectors