Name: Kevin Zhang

Problem 1.

(a) We are given two sets P, N with n unit vectors on opposite sides of a hyperplane through the origin. $< a, x >= 0$. Moreover, the distance of each point from the hyperplane is at least ε. Let $S = \{0, a\} \cup P \cup N \cup P' \cup N'$ with P', N' being their respective points reflected across the origin. We want to show it is possible to represent the points in S in lower dimension $O(\log n/\varepsilon^2)$ with distances being preserved up to a $1 \pm (\varepsilon/10)$.

This is possible with the JL Lemma. The JL Lemma states that for n points in d dimensions, it is possible to represent these points in $m = O(\log n/\varepsilon^2)$ dimensions. If we plug into our lemma ($n = 4n$ and $\varepsilon = \varepsilon/10$), we end up with a bound of $m = O(100 \log 4n/\varepsilon^2)$ with distances being preserved up to $1 \pm (\varepsilon/10)$. With Big-O notation, we can simplify $m = O(100 \log 4n/\varepsilon^2) = O(\log n/\varepsilon^2)$.

(b) We want to show that the margin for the above transformation is still preserved up to $\varepsilon/2$. To do so, we will use the identity $< a, x >= \frac{||a+x||^2 - ||a-x||^2}{4}$. From this, we have the following:

$$
\begin{aligned}
< \frac{T(a)}{||T(a)||}, T(x) > &= \frac{1}{||T(a)||} < T(a), T(x) > \\
< T(a), T(x) > &= \frac{||T(a) + T(x)||^2 - ||T(a) - T(x)||^2}{4} && \overset{?}{\geq} \varepsilon/2 \\
&= ||T(a) + T(x)||^2 - ||T(a) - T(x)||^2 && \overset{?}{\geq} 2\varepsilon
\end{aligned}
$$

From the JL-Lemma, we have that $T(a) = (1 \pm \frac{\varepsilon}{10}) \times a$ and that $T(x) = (1 \pm \frac{\varepsilon}{10}) \times x$. Since we want to find the lower bound, we can push the bounds as far as we can go. This nets us:

$$
\begin{aligned}
(1 - \frac{\varepsilon}{10})^2 ||a + x||^2 - (1 + \frac{\varepsilon}{10})^2 ||a - x||^2 &\overset{?}{\geq} 2\varepsilon \\
(1 - \frac{\varepsilon}{10})^2 ||a + x||^2 - (1 + \frac{\varepsilon}{10}) ||a + x||^2 - (1 + \frac{\varepsilon}{10})^2 ||a - x||^2 + (1 + \frac{\varepsilon}{10}) ||a + x||^2 &\overset{?}{\geq} 2\varepsilon \\
\left[(1 - \frac{\varepsilon}{10})^2 - (1 + \frac{\varepsilon}{10})^2\right] ||a + x||^2 + (1 + \frac{\varepsilon}{10})^2 \left[||a + x||^2 - ||a - x||^2\right] &\overset{?}{\geq} 2\varepsilon
\end{aligned}
$$

We can examine this expression piecewise. First, we know that $a^2 - b^2 = (a + b)(a - b)$.

$$
\left[(1 - \frac{\varepsilon}{10})^2 - (1 + \frac{\varepsilon}{10})^2\right] ||a + x||^2 = (\frac{-4\varepsilon}{10}) ||a + x||^2
$$

Next we know that $||a + x||^2 - ||a - x||^2 = 4\varepsilon$. We know this because of the identity $< a, x >= \frac{||a+x||^2 - ||a-x||^2}{4}$.

$$
(1 + \frac{\varepsilon}{10})^2 \left[||a + x||^2 - ||a - x||^2\right] = 4\varepsilon(1 + \frac{\varepsilon}{10})^2
$$

If combine these expressions, shuffle terms around, and add back in the $\frac{1}{||T(a)||}$ from earlier, we get the following:

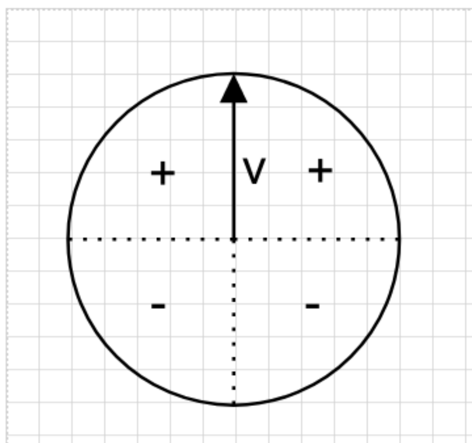
$$(1 + \frac{\varepsilon}{10})(4\varepsilon) \overset{?}{\geq} 2\varepsilon + \frac{(\frac{4\varepsilon}{10})||a + x||^2}{1 + \frac{\varepsilon}{10}}$$

Even if we stretch $||a+x||^2$ as far as it can go, which is 4, the RHS is bounded by 4ε. Meanwhile, the LHS is $4\varepsilon + (\frac{\varepsilon}{10})(4\varepsilon)$. As such, we are done, and we have proved that the margin is still preserved up to $\varepsilon/2$.

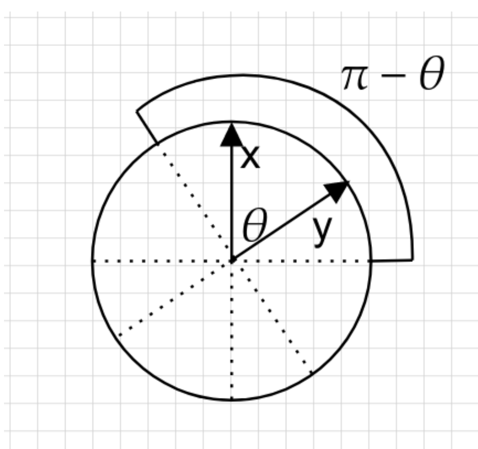
Problem 2.

(a) We are designing an LSH family as follows: A hash function is to pick a random unit vector v from the unit sphere, and then $h(x)$ is the sign of $< v, x >$. If $< v, x >= 0$, then the hash value is 1. We want to express the probability that two unit vectors x and y shared the same hash value (they collide) as a function of r, the distance between x and y.

We can do this by focusing on the 2-dimensional space spanned by x and y. The reasoning for this is because we can project v onto this plane and that will the primary decider on the hash value. The first thing we want to define is collision. In this case, two vectors x and y collide when $< v, x >$ and $< v, y >$ share the same sign. And in terms of v, the regions of $< v, x >$ are defined below:



Thus, in order for $< v, x >$ and $< v, y >$ to share the same sign, x and y must fall into same sign regions. The region that v is allowed to be in can be expressed as a function of θ, the angle between vectors x and y. This angle of this region is $\pi - \theta$, and can be visualized below:



Thus, the probability of collision can be expressed in terms of θ. We double the region size, since we the signs can either be both + or both $-$, and then divide over the entire circle.

$$\Pr[\text{collision}] = \frac{2(\pi - \theta)}{2\pi} = 1 - \frac{\theta}{\pi}$$

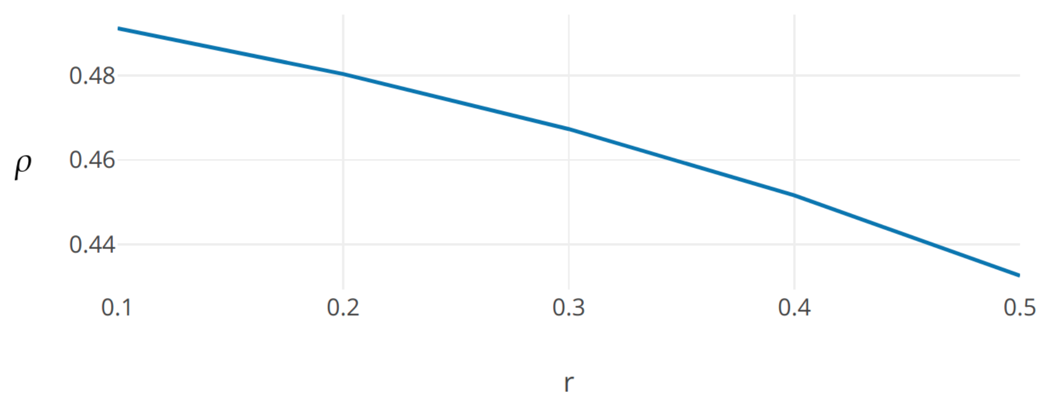
To get theta from r, we can use the law of cosines:

$$\begin{aligned} r^2 &= ||x||^2 + ||y||^2 + 2||x||||y|| \cos\theta \\ r^2 &= 1 + 1 + 2\cos\theta \\ 2 - r^2 &= 2\cos\theta \\ \cos\theta &= 1 - \frac{r^2}{2} \\ \theta &= \arccos\left(1 - \frac{r^2}{2}\right) \end{aligned}$$

And so, the final probability is:

$$\Pr[\text{collision}] = 1 - \frac{\arccos\left(1 - \frac{r^2}{2}\right)}{\pi}$$

(b) We would like to evaluate the parameter ρ for the approximate near neighbor problem with $c = 2$. If we let $p_1 = \Pr[\text{collision for r}]$ and $p_2 = \Pr[\text{collision for 2r}]$ and $\rho = \frac{\log p_1}{\log p_2}$, we get the following graph:



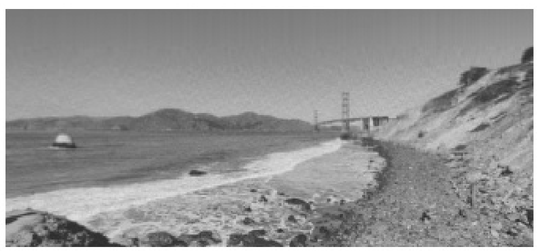
Problem 3.

Followed the instructions at: https://www.youtube.com/watch?v=H7qMMudo3e8

```
from PIL import Image
from matplotlib.image import imread
import matplotlib.pyplot as plt
import numpy as np

# Import image
A = imread('images/sf-gray.jpg')

# SVD computations
U, S, VT = np.linalg.svd(A,full_matrices=False)
S = np.diag(S)

j = 1
for k in (50, 60, 70, 80, 90, 100):
    # appr image
    X = U[:,:k] @ S[0:k,:k] @ VT[:k,:]
    plt.figure(j)
    img = plt.imshow(X)
    img.set_cmap('gray')
    plt.title('k = ' + str(k))
    plt.show()
    j += 1

# Best Approximation
k = 90
B = U[:,:k] @ S[0:k,:k] @ VT[:k,:]
img  = plt.imshow(B)
img.set_cmap('gray')
plt.axis('off')

plt.savefig('images/sf-compressed.jpg')
```

Original Image:

Compressed Image (at rank 90):



Problem 4.

(a) We are given $A = U\Sigma V^T$ with singular values $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_n$. We would like to find a matrix B of at most rank k such that $\|A - B\|_2 \leq \frac{\|A\|_F}{\sqrt{k}}$.

We can define B similarly to A, but at rank k:

$$B = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \sigma_k & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}^T$$

Thus, the result of $\|A - B\|_2$ can be expressed as follows:

$$\begin{aligned} \|A - B\|_2 &= \sigma_{k+1} \\ \sigma_{k+1} &\overset{?}{\leq} \frac{\|A\|_F}{\sqrt{k}} \\ \sigma_{k+1} &\overset{?}{\leq} \frac{\sqrt{\sum_i \sigma_i^2}}{\sqrt{k}} \\ \sigma_{k+1}^2 &\overset{?}{\leq} \frac{\sum_i \sigma_i^2}{k} \\ k\sigma_{k+1}^2 &\overset{?}{\leq} \sum_i \sigma_i^2 \end{aligned}$$

The last statement is true because $\sigma_1 \geq \sigma_2 \geq ...\sigma_k \geq \sigma_{k+1} \geq ....$ Since there are k singular values from σ_1 to σ_{k+1}, the following is true:

$$k\sigma_{k+1}^2 \leq \sum_i \sigma_i^2 \checkmark$$

(b) We want to find a matrix C that is a good approximate for A, such that the margin of error is $||(A-C)x||_2 \leq \varepsilon ||A||_F ||x||_2$. If we define C like we defined B as in Part A, but with rank $k = \frac{1}{\varepsilon^2}$, we have the following:

$$C = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \sigma_k & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}^T$$

We can also prove the bounds:

$$\begin{aligned} ||(A-C)x||_2 &= ||A-C||_2 ||x||_2 \\ &\leq \frac{||A||_F}{\sqrt{k}} ||x||_2 \\ &\leq \frac{||A||_F}{1/\varepsilon} ||x||_2 \\ &\leq \varepsilon ||A||_F ||x||_2 \end{aligned}$$

The reason we want to approximate A with C is for performance reasons. The runtime for using A is $O(n^2)$. We can do better with C, which has a runtime of $O(nk)$ or $O(\frac{n}{\varepsilon^2})$.