COMS 4771 Machine Learning (2018 Fall)
Homework 0

Jing Qian - jq2282@columbia.edu

September 9, 2018

Problem 1.1

(i)

$$
\begin{aligned}
\Pr[X=1] &= \sum_y \Pr[X=1, Y=y] = 0.2+0.2+0.3 = 0.7 \\
\Pr[X=2] &= \sum_y \Pr[X=2, Y=y] = 0.1+0.1+0.1 = 0.3
\end{aligned} \tag{1}
$$

(ii)

$$\Pr[Y=1|X=2] = \frac{\Pr[Y=1, X=2]}{\Pr[X=2]} = \frac{1}{3} \tag{2}$$

(iii)

$$
\begin{aligned}
\Pr[X=1|Y=3] &= \frac{0.3}{0.3+0.1} = 0.75 \\
\Pr[X=2|Y=3] &= \frac{0.1}{0.3+0.1} = 0.25
\end{aligned} \tag{3}
$$

$$\mathbb{E}[f(X)|Y=3] = 1^2 * \Pr[X=1|Y=3] + 2^2 * \Pr[X=2|Y=3] = 1.75 \tag{4}$$

Problem 1.2

(i) To be a probability distribution, g_θ has to satisfy following two conditions: 1) $g_\theta(x) \geq 0$ for any $x \in [0, \infty)$; 2) $\int_0^\infty g_\theta(x)dx = 1$.

Since $\theta > 0$ and $e^{-\frac{x}{\theta}} > 0$, $g_\theta > 0$. First condition is satisfied.

$$\int_0^\infty g_\theta(x)dx = \int_0^\infty \frac{1}{\theta}\, e^{-\frac{x}{\theta}}dx = \int_0^\infty e^{-y}dy = 1 \tag{5}$$

where $y = \frac{x}{\theta}$. Second condition is satisfied too. So g_θ is a probability distribution.

(ii)

$$\mathbb{E}[g_\theta] = \int_0^\infty x\ g_\theta(x)\ dx = \int_0^\infty \frac{x}{\theta}\ e^{-\frac{x}{\theta}} dx = \theta \int_0^\infty y\ e^{-y} dy = \theta \tag{6}$$

where $y = \frac{x}{\theta}$.

(iii)

$$\begin{aligned}
\mathrm{var(g_\theta)} &= \int_0^\infty (x - \mathbb{E}[g_\theta])^2 g_\theta(x)\ dx \\
&= \int_0^\infty \frac{(x-\theta)^2}{\theta}\ e^{-\frac{x}{\theta}} dx \\
&= \theta^2 \int_0^\infty y^2\ e^{-y} dy - 2\theta^2 \int_0^\infty y\ e^{-y} dy + \theta^2 \int_0^\infty\ e^{-y} dy \\
&= \theta^2
\end{aligned} \tag{7}$$

Problem 1.3

Considering that X and Y are jointly distributed Gaussian random variables, $X + Y$ is still Gaussian distributed.

$$\begin{aligned}
\mathbb{E}[X+Y] &= \sum_x \sum_y (x+y) \mathrm{Pr(X = x, Y = y)} \\
&= \sum_x \sum_y x \mathrm{Pr(X = x, Y = y)} + \sum_\mathrm{x} \sum_\mathrm{y} \mathrm{yPr(X = x, Y = y)} \\
&= \sum_x x \mathrm{Pr(X = x)} + \sum_\mathrm{y} \mathrm{yPr(Y = y)} \\
&= \mathbb{E}[X] + \mathbb{E}[Y] \\
&= 0
\end{aligned} \tag{8}$$

$$\begin{aligned}
\mathrm{var(X+Y)} &= \mathbb{E}[(X+Y)^2] - \mathbb{E}[X+Y]^2 \\
&= \mathbb{E}[X^2] + \mathbb{E}[Y^2] + 2\mathbb{E}[XY] - (\mathbb{E}[X] + \mathbb{E}[Y])^2 \\
&= (\mathbb{E}[X^2] - \mathbb{E}[x]^2) + (\mathbb{E}[Y^2] - \mathbb{E}[Y]^2) + 2(\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]) \\
&= \mathrm{var(X) + var(Y) + 2cov(X, Y)} \\
&= 6
\end{aligned} \tag{9}$$

So $X+Y$ has Gaussian distribution with the mean 0 and variance 6, in other words, $N(0,6)$.

Problem 1.4

For a fair coin, which means the possiblity of tossing a head is $\frac{1}{2}$ every time, the expected absolute difference between the number of heads H and that of tails T is:

$$\mathbb{E}[|H-T|] = \sum_{i=0}^{n} \binom{n}{i} (\frac{1}{2})^i (\frac{1}{2})^{n-i}\ |i - (n-i)| = \frac{1}{2^n} \sum_{i=0}^{n} \binom{n}{i}\ |2i - n| \tag{10}$$

where i is the number of heads. Since $\binom{n}{i}\ |2i-n| = \binom{n}{j}\ |2j-n|$ while $j = n-i$, we have:

$$\sum_{i=0}^{n}\binom{n}{i}\ |2i-n| = \begin{cases} 2\sum_{i=0}^{\frac{n}{2}-1}\binom{n}{i}\ (n-2i) + \binom{n}{n/2}(n-n) = 2\sum_{i=0}^{\frac{n}{2}-1}\binom{n}{i}\ (n-2i), & n \text{ is even.} \\ 2\sum_{i=0}^{\frac{n-1}{2}}\binom{n}{i}\ (n-2i), & n \text{ is odd.} \end{cases} \tag{11}$$

Since the odd case and even case have slight differences, we treat them seperately.

$$\begin{aligned} \mathbb{E}_{\text{even}}[|H-T|] &= \frac{1}{2^{(n-1)}}\sum_{i=0}^{\frac{n}{2}-1}\binom{n}{i}\ (n-2i) = \frac{1}{2^{(n-1)}}\sum_{i=0}^{\frac{n}{2}-1}\frac{n!}{i!\ (n-i)!}\ [(n-i)-i] \\ &= \frac{1}{2^{(n-1)}}[\sum_{i=0}^{\frac{n}{2}-1}\frac{n!}{i!\ (n-i)!}\ (n-i) - \sum_{i=0}^{\frac{n}{2}-1}\frac{n!}{i!\ (n-i)!}\ i] \\ &= \frac{1}{2^{(n-1)}}[\sum_{i=0}^{\frac{n}{2}-1}\frac{n!}{i!\ (n-i-1)!} - \sum_{j=0}^{\frac{n}{2}-2}\frac{n!}{j!\ (n-j-1)!}] \end{aligned} \tag{12}$$

where $j = i-1$. After subtraction of the two summations in Eq. (12), only the term $i = \frac{n}{2}-1$ remains.

$$\mathbb{E}_{\text{even}}[|H-T|] = \frac{1}{2^{(n-1)}}\frac{n!}{(\frac{n}{2}-1)!(\frac{n}{2})!} = \frac{n}{2^n}\ \frac{n!}{((\frac{n}{2})!)^2} \tag{13}$$

Similarly, we could get the result when n is an odd number:

$$\mathbb{E}[|H-T|] = \begin{cases} \frac{n}{2^n}\ \frac{n!}{((\frac{n}{2})!)^2}, & n \text{ is even.} \\ \frac{1}{2^{(n-1)}}\ \frac{n!}{((\frac{n-1}{2})!)^2}, & n \text{ is odd.} \end{cases} \tag{14}$$

When $n \to \infty$, we could get an approximate value for Eq. (14) using Stirling's approximation for factorials.

$$\mathbb{E}_{\text{even}}[|H-T|] = \frac{n}{2^n}\ \frac{n!}{((\frac{n}{2})!)^2} \approx \frac{n}{2^n}\ \frac{\sqrt{2\pi n}\ (\frac{n}{e})^n}{\pi n(\frac{n}{2e})^n} = \sqrt{\frac{2n}{\pi}} \tag{15}$$

And

$$\begin{aligned} \mathbb{E}_{\text{odd}}[|H-T|] &= \frac{1}{2^{(n-1)}}\ \frac{n!}{((\frac{n-1}{2})!)^2} \approx \frac{1}{2^{(n-1)}}\ \frac{\sqrt{2\pi n}\ (\frac{n}{e})^n}{\pi(n-1)(\frac{n-1}{2e})^{(n-1)}} \\ &= \sqrt{\frac{2n}{\pi}}\ \frac{1}{e}\ \frac{1}{(1-\frac{1}{n})^n} \\ &\approx \sqrt{\frac{2n}{\pi}} \end{aligned} \tag{16}$$

So when independently toss a fair coin n times, the expected absolute difference between the number of heads H and the number of tails T is about $\sqrt{\frac{2n}{\pi}}$.

Problem 2.1

(i) Let A be a matrix contains vectors in the question and be transformed into row echelon form:

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 8 & 10 \\ 3 & 3 & 6 \\ 4 & 2 & 6 \end{bmatrix} \Longleftrightarrow \mathbf{V} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{17}$$

Since rank($\mathbf{A}$) = rank($\mathbf{V}$) =2, the dimension of the subspace S is 2.

(ii) To get the orthogonal linear projection of the point $\left(\begin{smallmatrix} 6 \\ 5 \\ 9 \\ 2 \end{smallmatrix}\right)$ onto the subspace S, we need to calculate the orthonormal basis of subspace S first. From (i), we could get the orthonormal basis $(\mathbf{v_1}, \mathbf{v_2})$ of subspace S:

$$(\mathbf{v_1}, \mathbf{v_2}) = \left(\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \right) \tag{18}$$

Gram-Schmidt orthogonalization also works but the given orthonormal basis are much more complex.

So the projection P of given point $\mathbf{u} = \left(\begin{smallmatrix} 6 \\ 5 \\ 9 \\ 2 \end{smallmatrix}\right)$ onto S is:

$$\mathbf{P} = \sum_i \frac{< \mathbf{u}, \mathbf{v_i} >}{< \mathbf{v_i}, \mathbf{v_i} >} \mathbf{v_i} =< \mathbf{u}, \mathbf{v_1} > \mathbf{v_1} + < \mathbf{u}, \mathbf{v_2} > \mathbf{v_2} = \begin{pmatrix} 6 \\ 5 \\ 0 \\ 0 \end{pmatrix}. \tag{19}$$

Problem 2.2

A square matrix is invertible if and only if none of its eigenvalues is zero. So to prove $\mathbf{A^T A} + \rho \mathbf{I}$ is invertible, we could try to prove that all its eigenvalues are non-zero. Let λ be one eigenvalue of matrix $\mathbf{A^T A} + \rho \mathbf{I}$. Then we have:

$$\mathbf{det}(\mathbf{A^T A} + \rho \mathbf{I} - \lambda \mathbf{I}) = \mathbf{0} \tag{20}$$

So,

$$\lambda = \mathbf{det}(\mathbf{A^T A}) + \rho \tag{21}$$

Let $\tilde{\lambda}$ and x be one eigenvalue and corresponding eigenvector of matrix $\mathbf{A^T A}$. In other words, $\mathbf{A^T A x} = \tilde{\lambda} \mathbf{x}$. Then,

$$||\mathbf{Ax}|| = (\mathbf{Ax})^\mathbf{T} \mathbf{Ax} = \mathbf{x^T A^T A x} = \tilde{\lambda} \mathbf{x^T x} = \tilde{\lambda} \tag{22}$$

Since $\mathbf{det}(\mathbf{A^T A}) = \tilde{\lambda} = ||\mathbf{Ax}|| \geq \mathbf{0}$ and $\rho > 0$, $\lambda = \mathbf{det}(\mathbf{A^T A}) + \rho > \mathbf{0}$.

Therefore, the matrix $\mathbf{A^T A} + \rho \mathbf{I}$ is invertible because its eigenvalues are all positive.

Problem 3.1

(i) The function could be expanded as following:

$$\begin{aligned} f(\mathbf{x}) &= ||\mathbf{Ax}-\mathbf{b}||^{\mathbf{2}} + ||\mathbf{x}||^{\mathbf{2}} = (\mathbf{Ax}-\mathbf{b})^{\mathbf{T}}(\mathbf{Ax}-\mathbf{b}) + \mathbf{x}^{\mathbf{T}}\mathbf{x} \\ &= \mathbf{x}^{\mathbf{T}}\mathbf{A}^{\mathbf{T}}\mathbf{Ax} + \mathbf{x}^{\mathbf{T}}\mathbf{x} - \mathbf{x}^{\mathbf{T}}\mathbf{A}^{\mathbf{T}}\mathbf{b} - \mathbf{b}^{\mathbf{T}}\mathbf{Ax} + \mathbf{b}^{\mathbf{T}}\mathbf{b} \end{aligned} \tag{23}$$

Using the derivatives of matrices:

$$\begin{aligned} \frac{\partial \mathbf{x}^{\mathbf{T}}\mathbf{B}}{\partial \mathbf{x}} &= \frac{\partial \mathbf{B}^{\mathbf{T}}\mathbf{x}}{\partial \mathbf{x}} = \mathbf{B} \\ \frac{\partial \mathbf{x}^{\mathbf{T}}\mathbf{Bx}}{\partial \mathbf{x}} &= (\mathbf{B}+\mathbf{B}^{\mathbf{T}})\mathbf{x} \\ \frac{\partial \mathbf{x}^{\mathbf{T}}\mathbf{Bx}}{\partial \mathbf{x}} &= \mathbf{2x} \end{aligned} \tag{24}$$

We could get:

$$\begin{aligned} \nabla f(\mathbf{x}) &= \frac{\partial(\mathbf{x}^{\mathbf{T}}\mathbf{A}^{\mathbf{T}}\mathbf{Ax} + \mathbf{x}^{\mathbf{T}}\mathbf{x} - \mathbf{x}^{\mathbf{T}}\mathbf{A}^{\mathbf{T}}\mathbf{b} - \mathbf{b}^{\mathbf{T}}\mathbf{Ax} + \mathbf{b}^{\mathbf{T}}\mathbf{b})}{\partial \mathbf{x}} \\ &= \mathbf{2A}^{\mathbf{T}}\mathbf{Ax} + \mathbf{2x} - \mathbf{2A}^{\mathbf{T}}\mathbf{b} \end{aligned} \tag{25}$$

(ii) To find $\mathbf{x}$ minimizes f, first, we need to find the $\mathbf{x}$ where $\nabla f(\mathbf{x})$ equals to $\mathbf{0}$. In other words, $\nabla f(\mathbf{x}) = \mathbf{2A}^{\mathbf{T}}\mathbf{Ax} + \mathbf{2x} - \mathbf{2A}^{\mathbf{T}}\mathbf{b} = \mathbf{0}$. So $(\mathbf{A}^{\mathbf{T}}\mathbf{A}+\mathbf{I})\mathbf{x} - \mathbf{A}^{\mathbf{T}}\mathbf{b} = \mathbf{0}$. Then $\mathbf{x} = (\mathbf{A}^{\mathbf{T}}\mathbf{A}+\mathbf{I})^{\mathbf{-1}}\mathbf{A}^{\mathbf{T}}\mathbf{b}$.

Because when $||\mathbf{x}|| \to \infty$, $f(\mathbf{x}) \to \infty$, min(f($\mathbf{x}$)) appears when $\nabla f(\mathbf{x}) = \mathbf{0}$. Hence $\mathbf{x} = (\mathbf{A}^{\mathbf{T}}\mathbf{A}+\mathbf{I})^{\mathbf{-1}}\mathbf{A}^{\mathbf{T}}\mathbf{b}$ minimizes f.

Problem 3.2

Suppose $x = [\begin{smallmatrix} x_1 \\ x_2 \end{smallmatrix}]$, then:

$$\begin{aligned} g(x) &= x^{\mathrm{T}}Ax - b^{\mathrm{T}}x + c \\ &= [x_1\ x_2][\begin{matrix} 1 & 3 \\ 3 & 1 \end{matrix}][\begin{matrix} x_1 \\ x_2 \end{matrix}] - [1\ 2][\begin{matrix} x_1 \\ x_2 \end{matrix}] + 3 \\ &= x_1^2 + 6x_1x_2 + x_2^2 - x_1 - 2x_2 + 3 \end{aligned} \tag{26}$$

The first partical derivatives are

$$\begin{aligned} P_{x_1}(x) &= P_{x_1}(x_1, x_2) = 2x_1 + 6x_2 - 1 \\ P_{x_2}(x) &= P_{x_2}(x_1, x_2) = 6x_1 + 2x_2 - 2 \end{aligned} \tag{27}$$

So $\nabla g((1,1)) = (2+6-1, 6+2-2) = (7,6)$.

Problem 4.1

The output is:
(ii) 2

(iii) [85 87 19 21 3 60 62 69] [15 16 22 3 9 90 91 97 78 84]

(iv) 50.5

(vi) [1022.0677061701972, 263.62801935563306, 102.83034619975379]

* If ”the top three eigenvalues” refers to the eigenvalues with largest absolute value, then the output would be: $[1022.0677061701972, 263.62801935563306, -235.74325551868949]$.

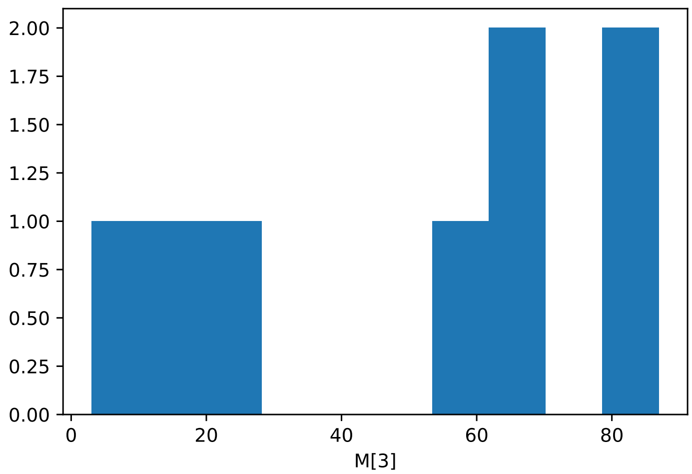

Figure 1: Histogram of the 4th row of **M**.

HW0_4.1.py

```
import numpy as np
import scipy.io as sio
import matplotlib.pyplot as plt
#load data
m = sio.loadmat('hw0data.mat')['M']
#print the dimensions of M.
print(np.ndim(m))
#print the 4th row and 5th column entry of M
print(m[3], m[:, 4])
#print the mean value of the 5th column of M
print(np.mean(m[:, 4]))
#compute the histogram of the 4th row of M and show the figure
plt.hist(m[3])
plt.show()
```

```
#compute and print the top three eigenvalues of the matrix MTM
evals, evecs = np.linalg.eig(np.dot(m.transpose(), m))
print(sorted(evals, reverse = True)[:3])
```

Problem 4.2

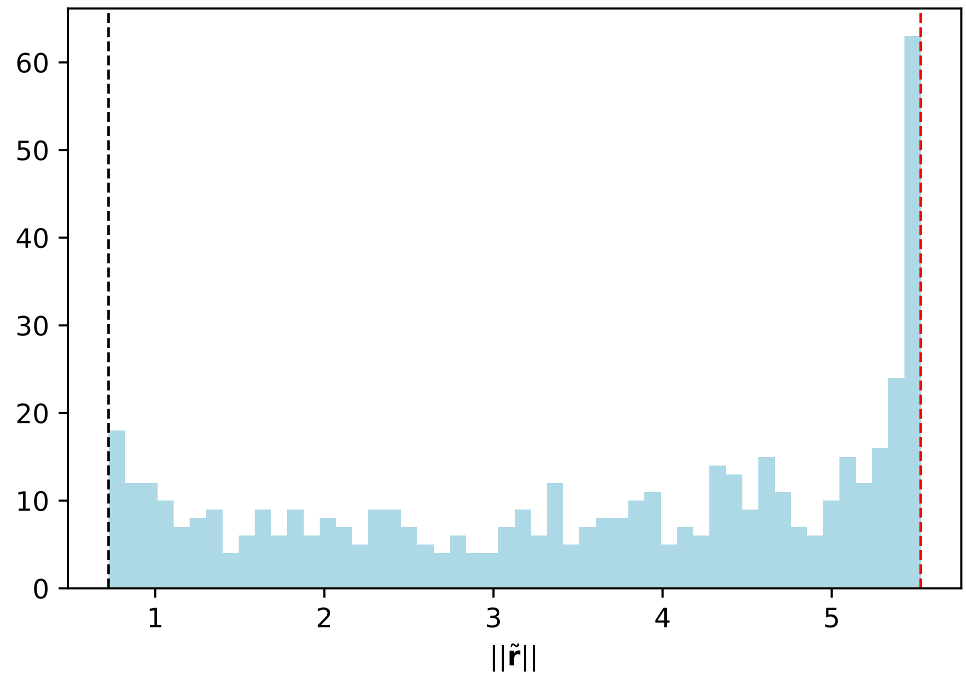

Figure 2: Histogram of values of $||\tilde{\mathbf{r}}||$. Black dashed line is $\lambda_{\min}$ and red dashed line is $\lambda_{\max}$.

(vii) From Fig. (2), we could see that $||\tilde{\mathbf{r}}||$ lie between $\lambda_{\min}$ and $\lambda_{\max}$, which means $\lambda_{\min}$ is the lower bound of $||\tilde{\mathbf{r}}||$ while $\lambda_{\max}$ is the upper bound.

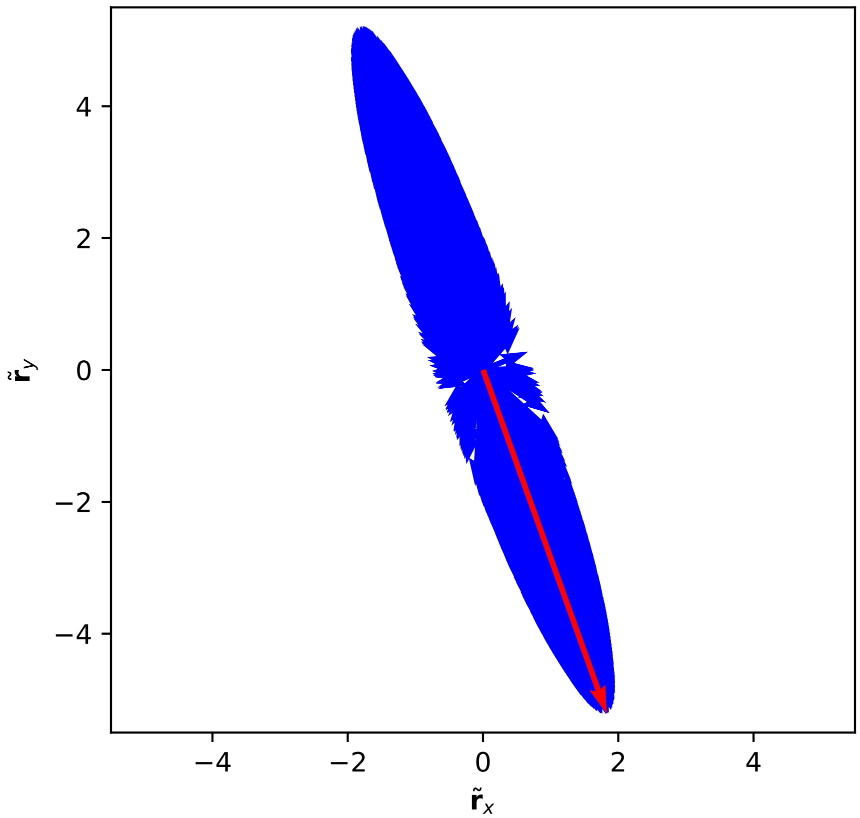

Figure 3: 2-D plot of all the distorted vectors $\tilde{\mathbf{r}}$ (in blue color) and $\mathbf{L}\mathbf{v}_{\mathbf{max}}$ (in red color).

(x) From Fig. (3), we could see that $\mathbf{L}\mathbf{v}_{\mathbf{max}}$ is the semi-major axis of the distribution of all the distorted vectors $\tilde{\mathbf{r}}$. Since eigenvectors point in directions that are streched by the transformation from matrix $\mathbf{L}$, the eigenvector $\mathbf{v}_{\mathbf{max}}$, which corresponds to the largest eigenvalue, points to the direction that is streched most severely. As a result, after the transformation of matrix $\mathbf{L}$, the set of vectors $\mathbf{r}$ which originally have randomly distributed directions became a set of vectors $\tilde{\mathbf{r}}$ which have distorted distribution of directions. Most of them tend to point to the direction (or the reverse direction) of $\mathbf{v}_{\mathbf{max}}$.

HW0_4.2.py

```
import numpy as np
import random
import matplotlib.pyplot as plt
nV = 500
#create a 2*2 matrix L
L = np.matrix([[5/4, -3/2], [-3/2, 5]])
#create a 2*500 array R to represent 500 random, unit length,
   2-d vectors, in which R[:, i] is the i-th vector
```

```
R = np.zeros((2, nV))
for i in range(nV):
    tmp = np.array([random.gauss(0,1), random.gauss(0,1)])
    tmp = tmp/(sum(tmp * tmp) ** 0.5)
    R[:, i] = tmp
#compute R2 = LR, R2[:, i] is the distorted R[:, i]
R2 = np.dot(L,R)
#compute the eigenvalues of L and denote the minimum eigenvalue
    with lmax and lmin.
evals, evecs = np.linalg.eig(L)
[lMax, lMin] = sorted(evals, reverse = True)
#lr[i] is the length of the i-th vector in R2
lr = np.zeros((nV))
for i in range(nV):
    lr[i] = (np.dot(R2[:, i].transpose(), R2[:, i]))[0, 0] **
       0.5
#create a histogram of values of lr (use 50 bins) and compare
   it to lMax and lMin.
plt.hist(lr, bins = 50)
plt.axvline(lMax, color='r', linestyle='dashed', linewidth=1)
plt.axvline(lMin, color='k', linestyle='dashed', linewidth=1)
plt.show()
#compute the eigenvectors of L and Let vmax denote the
   eigenvector corresponding to the maximum eigenvalue lmax.
for i in range(2):
    if evals[i] == lMax:
        vMax = evecs[:, i]
        break
#make a two-dimensional plot of R2 (in blue color) and the
   eigenvector vmax (in red color).
Lv = np.asarray(np.dot(L, vMax))
fig = plt.figure(figsize = (5, 5))
plt.quiver([0], [0], np.asarray(R2[0, :]), np.asarray(R2[1, :])
   , color = 'b', angles='xy', scale_units='xy', scale=1)
plt.quiver([0], [0], Lv[0], Lv[1], color = 'r', angles='xy',
   scale_units='xy', scale=1)
plt.xlim(-5.5, 5.5)
plt.ylim(-5.5, 5.5)
plt.show()
```