Computer Science M146, Homework 0

Michael Wu
UID: 404751542

January 18th, 2018

Problem 1

$$y = x\sin(z)e^{-x}$$
$$\frac{\partial y}{\partial x} = \sin(z)e^{-x} - x\sin(z)e^{-x}$$

Problem 2

$$\mathbf{X} = \begin{pmatrix} 2 & 4 \\ 1 & 3 \end{pmatrix} \qquad \mathbf{y} = \begin{pmatrix} 1 \\ 3 \end{pmatrix} \qquad \mathbf{z} = \begin{pmatrix} 2 \\ 3 \end{pmatrix}$$

a)

$$\mathbf{y}^T\mathbf{z} = 2 \times 1 + 3 \times 3 = 11$$

b)

$$\mathbf{Xy} = \begin{pmatrix} 2 & 12 \\ 1 & 9 \end{pmatrix}$$

c) Yes it is invertible.

$$\mathbf{X}^{-1} = \frac{1}{2}\begin{pmatrix} 3 & -4 \\ -1 & 2 \end{pmatrix}$$

d) The rank is 2.

Problem 3

a) The sample mean is $\frac{3}{5}$.

b)

$$\begin{aligned} s^2 &= \frac{1}{n-1}\sum_{i=1} n(x_i - \bar{x})^2 \\ &= \frac{1}{4}\left[2\left(\frac{3}{5}\right)^2 + 3\left(\frac{2}{5}\right)^2\right] \\ &= \frac{3}{10} \end{aligned}$$

c) The probability is $0.5^5 = \frac{1}{32}$.

d) Let $p = P(X_1 = 1)$. Then the probability of sample S occurring is

$$P(S) = p^3(1-p)^2$$

We can find maxima and minima using calculus.

$$\begin{aligned} P'(S) &= 3p^2(1-p)^2 - 2p^3(1-p) \\ P'(S) &= 3(p^2 - 2p^3 + p^4) - 2(p^3 - p^4) \\ P'(S) &= 5p^4 - 8p^3 + 3p^2 \\ P'(S) &= p^2(5p^2 - 8p + 3) \\ P'(S) &= p^2(5p-3)(p-1) \end{aligned}$$

The only critical point exists at $p = \frac{3}{5}$, which must be the maxima. This is the value that maximizes the probability of sample S occurring.

e)

$$P(X = T|Y = b) = \frac{2}{5}$$

Problem 4

a) False.

b) True.

c) False.

d) False.

e) True.

Problem 5

a) v

b) iv

c) ii

d) i

e) iii

Problem 6

a) The mean is p. The variance is $p(1-p)$.

b) We begin with the definition of variance and note that the mean of X is 0, allowing us to write

$$\begin{aligned} \mathrm{Var}(X) &= \mathrm{E}\left(X^2\right) - \mathrm{E}(X)^2 \\ &= \mathrm{E}\left(X^2\right) \\ &= \sigma^2 \end{aligned}$$

Because $\mathrm{E}(X) = 0$, we can find $\mathrm{Var}(2X)$ similarly.

$$\begin{aligned}\mathrm{Var}(2X) &= \mathrm{E}\left((2X)^2\right) - \mathrm{E}(2X)^2 \\ &= \mathrm{E}\left((2X)^2\right) \\ &= 4\,\mathrm{E}\left(X^2\right) \\ &= 4\sigma^2\end{aligned}$$

and similarly for $\mathrm{Var}(X+2)$,

$$\begin{aligned}\mathrm{Var}(X+2) &= \mathrm{E}\left((X+2)^2\right) - \mathrm{E}(X+2)^2 \\ &= \mathrm{E}\left(X^2 + 4X + 4\right) - 4 \\ &= \mathrm{E}\left(X^2\right) + \mathrm{E}(4X) + \mathrm{E}(4) - 4 \\ &= \mathrm{E}\left(X^2\right) \\ &= \sigma^2\end{aligned}$$

Problem 7

a)

i) Both $f(n) = O(g(n))$ and $g(n) = O(f(n))$ are true. This is because

$$f(n) = \ln(n) = \frac{\log_2(n)}{\log_2(e)}$$

which is less than $g(n) = \log_2(n)$ for all $n > 1$. In reverse, we have

$$g(n) = \log_2(n) = \frac{\ln(n)}{\ln(2)}$$

which is less than $\frac{2}{\ln(2)} f(n) = 2\frac{\ln(n)}{\ln(2)}$ for all $n > 1$.

ii) $g(n) = O(f(n))$. $f(n)$ is exponential while $g(n)$ is polynomial.

iii) $g(n) = O(f(n))$. $f(n) > g(n)$ for all $n > 1$. $f(n) \neq O(g(n))$ because for any positive constant c,

$$\frac{f(n)}{cg(n)} = c\left(\frac{3}{2}\right)^n$$

which is an increasing function that will always surpass 1 as n becomes large. Thus $f(n)$ will always outgrow any constant multiple of $g(n)$.

b) Use a binary search algorithm. Begin by checking the location halfway through the array, if it contains a zero with a one following it we have the transition index. If it is a zero with a zero following it, move to the location halfway through the latter half of the array. Otherwise if it is a one with a one following it, check the location halfway through the beginning half of the array. In this new location check if it contains a zero with a one following it in order to determine if it is the transition index. Otherwise recursively continue checking and moving halfway through the unchecked parts of the array, moving up or down depending on whether ones or zeros were found, until the transition index is found. This algorithm is correct because each iteration removes half the array from being searched, and it only terminates upon finding the transition index. It never removes the transition location, so eventually it will terminate and produce the correct output. The runtime is $O(\log(n))$ because each iteration doubles the size of the array that can be searched. After x iterations, we can search an array of size 2^x. Thus for an array of size $n = 2^x$, we only require $\log_2(n) = x$ iterations. Thus our runtime is logarithmic.

Problem 8

a)

$$\begin{aligned}
\mathrm{E}(XY) &= \sum_{x,y\in\mathbb{R}} xy\,\mathrm{P}(X=x)\,\mathrm{P}(Y=y) \\
&= \sum_{x\in\mathbb{R}} x\,\mathrm{P}(X=x) \sum_{y\in\mathbb{R}} \mathrm{P}(Y=y) \\
&= \mathrm{E}(X)\,\mathrm{E}(Y)
\end{aligned}$$

b)

i) E(Number of 3's) $= 6000 * \frac{1}{6} = 1000$. The law of large numbers states that the results obtained from a large number of trials should be close to the expected value.

ii) The coin toss X has a Bernoulli distribution with $p = \frac{1}{2}$, so its mean is $\mu = p = \frac{1}{2}$ and its variance is $\sigma^2 = p(1-p) = \frac{1}{4}$. The central limit

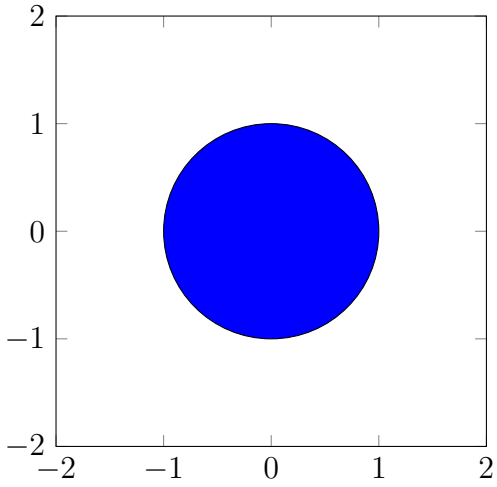
theorem states that as n increases, $\sqrt{n}\left(\bar{X}-\mu\right) \to \mathcal{N}\left(0, \sigma^2\right)$, thus

$$\sqrt{n}\left(\bar{X}-\frac{1}{2}\right) \overset{n\to\infty}{\longrightarrow} \mathcal{N}\left(0, \frac{1}{4}\right)$$

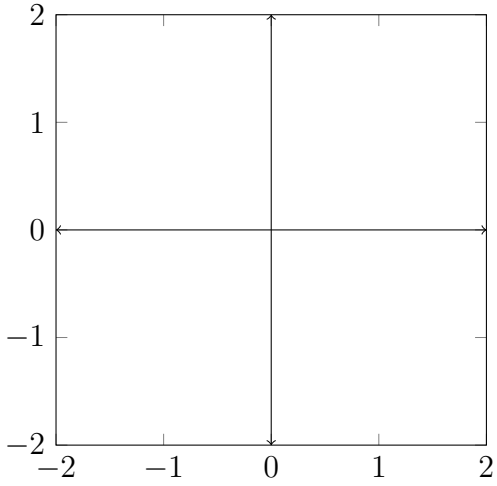
Problem 9

a)

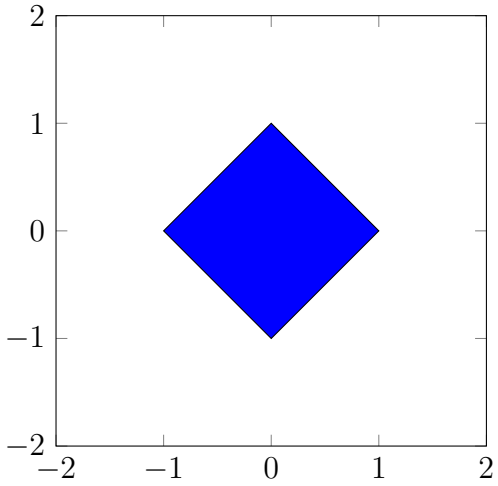
i) $\|x\|_2 \leq 1$



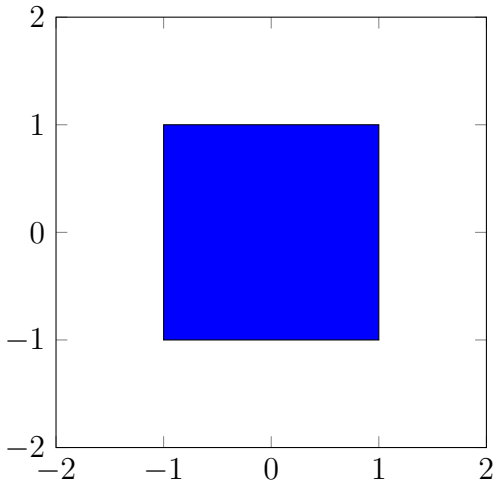
ii) $\|x\|_0 \leq 1$



iii) $\|x\|_1 \leq 1$



iv) $\|x\|_\infty \leq 1$



b)

i) Given a square matrix $\mathbf{A}$, an eigenvector $\vec{\mathbf{x}}$ is a vector such that

$$\mathbf{A}\vec{\mathbf{x}} = \lambda\vec{\mathbf{x}}$$

where λ is a scalar constant. λ is an eigenvalue of the matrix $\mathbf{A}$.

ii) The characteristic equation is

$$(2-\lambda)^2 - 1 = \lambda^2 - 4\lambda + 3 = (\lambda - 3)(\lambda - 1)$$

giving us eigenvalues $\lambda_1 = 1$ and $\lambda_2 = 3$. To find the eigenvectors we need to find the null space of

$$\mathbf{A}_{\lambda_1} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \qquad \mathbf{A}_{\lambda_2} = \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}$$

This gives us eigenvectors $\vec{\mathbf{x}}_1 = \langle 1, -1 \rangle$ and $\vec{\mathbf{x}}_2 = \langle 1, 1 \rangle$

iii) Take any eigenvalue λ and eigenvector $\vec{\mathbf{x}}$ pair of the matrix $\mathbf{A}$. Then

$$\begin{aligned} \mathbf{A}\vec{\mathbf{x}} &= \lambda\vec{\mathbf{x}} \\ \mathbf{A}^2\vec{\mathbf{x}} &= \lambda\mathbf{A}\vec{\mathbf{x}} = \lambda^2\vec{\mathbf{x}} \\ &\vdots \\ \mathbf{A}^k\vec{\mathbf{x}} &= \lambda^{k-1}\mathbf{A}\vec{\mathbf{x}} = \lambda^k\vec{\mathbf{x}} \end{aligned}$$

Thus λ^k and $\vec{\mathbf{x}}$ form and eigenvalue and eigenvector pair for the matrix $\mathbf{A}^k$. This holds for every eigenvalue and eigenvector pair, showing that the eigenvalues of $\mathbf{A}^k$ are $\lambda_1^k, \lambda_2^k, \ldots, \lambda_n^k$ and that each eigenvector of $\mathbf{A}$ is still an eigenvector of $\mathbf{A}^k$.

c)

i)

$$\frac{\partial \mathbf{a}^T\mathbf{x}}{\partial \mathbf{x}} = \mathbf{a}^T$$

ii) We find the first derivative by doing

$$\begin{aligned}
\frac{\partial \mathbf{x}^T\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} &= \frac{\partial}{\partial \mathbf{x}} \begin{pmatrix} x_1 & x_2 & \dots & x_n \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & & \\ \vdots & & \ddots & \\ a_{n1} & & & a_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \\
&= \frac{\partial}{\partial \mathbf{x}} \begin{pmatrix} x_1 & x_2 & \dots & x_n \end{pmatrix} \begin{pmatrix} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n \\ \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n \end{pmatrix} \\
&= \frac{\partial}{\partial \mathbf{x}} \left[\sum_{i=1}^{n} a_{ii}{x_i}^2 + \sum_{i,j \in (1,n), i \neq j} a_{ij}x_i x_j \right] \\
&= \begin{pmatrix} 2a_{11}x_1 + 2a_{12}x_2 + \ldots + 2a_{1n}x_n \\ \vdots \\ 2a_{n1}x_1 + 2a_{n2}x_2 + \ldots + 2a_{nn}x_n \end{pmatrix}^T \\
&= 2 \begin{pmatrix} x_1 & x_2 & \dots & x_n \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & & \\ \vdots & & \ddots & \\ a_{n1} & & & a_{nn} \end{pmatrix} \\
&= 2\mathbf{x}^T\mathbf{A}
\end{aligned}$$

We find the second derivative by doing

$$\begin{aligned}
\frac{\partial \mathbf{x}^T\mathbf{A}\mathbf{x}}{\partial \mathbf{x}^2} &= \frac{\partial}{\partial \mathbf{x}} 2\mathbf{x}^T\mathbf{A} \\
&= 2\mathbf{A}
\end{aligned}$$

d)

i) Because $\mathbf{w}^T\mathbf{x} + b = 0$ is a line, both $\mathbf{w}$ and $\mathbf{x}$ must be vectors in $\mathbb{R}^2$. Let $\mathbf{x}_1$ and $\mathbf{x}_2$ be points on the line. Then

$$\begin{aligned}
\mathbf{w}^T\mathbf{x}_1 + b &= 0 \\
\mathbf{w}^T\mathbf{x}_2 + b &= 0 \\
\mathbf{w}^T(\mathbf{x}_1 - \mathbf{x}_2) &= 0
\end{aligned}$$

Thus $\mathbf{w}$ is orthogonal to the line defined by $\mathbf{w}^T\mathbf{x} + b = 0$, because it is orthogonal to the vector $\mathbf{x}_1 - \mathbf{x}_2$ that lies along the line.

ii) The shortest distance to the line from the origin must be along a vector perpendicular to the line. Thus let $\mathbf{x} = c\frac{\mathbf{w}}{\|\mathbf{w}\|_2}$. We need to solve

$$c\frac{\mathbf{w}^T\mathbf{w}}{\|w\|_2} + b = 0$$

for c to find the distance to the origin. Because $\mathbf{w}^T\mathbf{w} = \|\mathbf{w}\|_2^2$, we get $c = -\frac{b}{\|\mathbf{w}\|_2}$. Taking the absolute value of this, we get the distance $\frac{b}{\|\mathbf{w}\|_2}$ to the origin.

Problem 10

a) See figure 1.

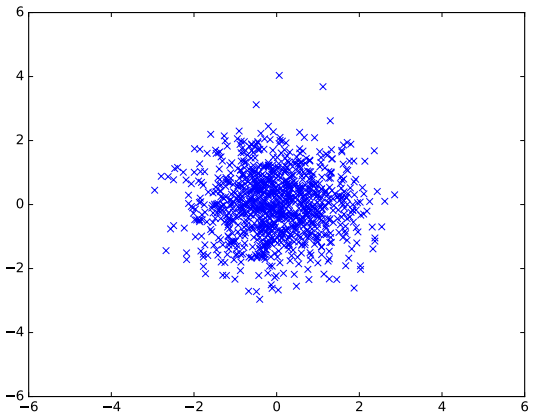


Figure 1: 1000 samples with mean $(0, 0)$ and identity covariance matrix.

b) See figure 2. The center shifts to $(1, 1)$.

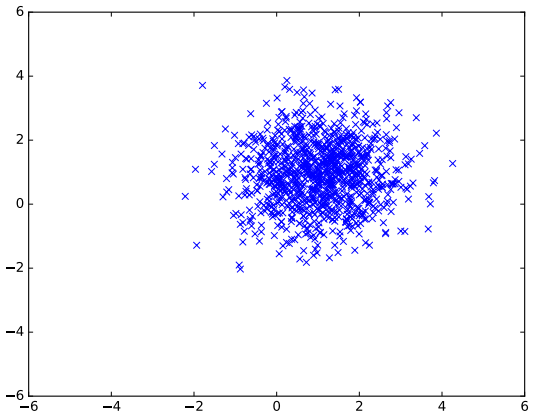


Figure 2: 1000 samples with mean $(1, 1)$ and identity covariance matrix.

c) See figure 3. It becomes more spread out.

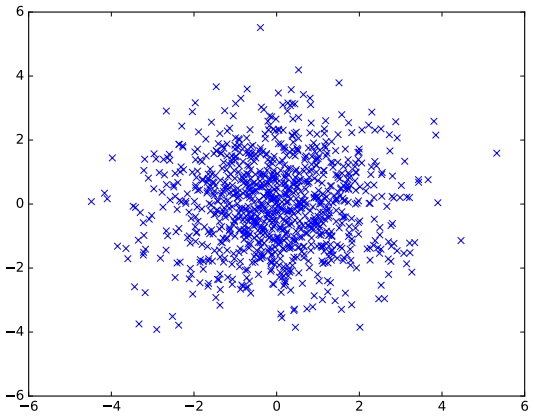


Figure 3: 1000 samples with mean $(0, 0)$ and covariance matrix $2I_2$.

d) See figure 4. x and y become positively correlated.

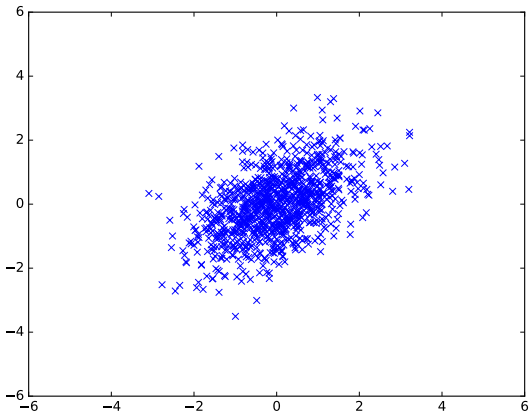


Figure 4: 1000 samples with mean $(0, 0)$ and positively correlated covariance matrix.

e) See figure 5. x and y become negatively correlated.

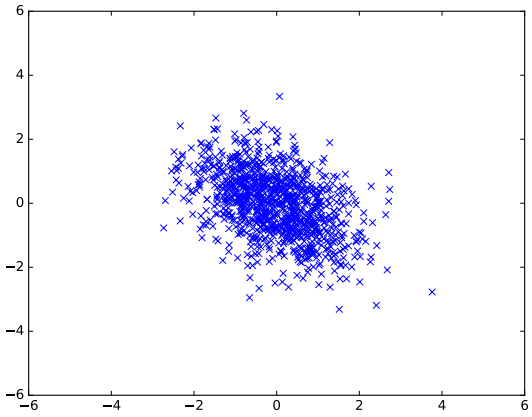


Figure 5: 1000 samples with mean $(0, 0)$ and negatively correlated covariance matrix.

Problem 11

The eigenvector is $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$ which corresponds to the eigenvalue 3.

Problem 12

a) SIDO: A pharmacology dataset.

b) http://www.causality.inf.ethz.ch/data/SIDO.html

c) The features are descriptions of molecules. The labels are binary and can either be $+1$ to represent molecular activity indicating the AIDS/HIV virus, or -1 to represent activity not indicating the AIDS/HIV virus.

d) There are 12678 training examples and 10000 test examples.

e) There are 4932 features in the dataset, each one is a molecular descriptor such as number of carbon atoms. These are binary, so they are 0 for the feature not being present and 1 for the feature being present.