COMPSCI 289A Homework 2 **Mitch Negus**
2/13/2017 **3032146443**

Problem 1

a.)

Say $P(T)$ is the probability of hitting the target, and $P(W)$ is the probability of it being windy. We are given that $P(T|W) = 0.4$ and $P(T|W^C) = 0.7$, as well as that $P(W) = 0.3$. We note that conditional probability theory states that in general, $P(A|B) = \frac{P(A \cap B)}{P(B)}$.

(i) on a given shot there is a gust of wind and she hits her target; $P(T \cap W)$

$$P(T|W) = \frac{P(T \cap W)}{P(W)}$$
$$P(T \cap W) = P(T|W) \cdot P(W)$$
$$P(T \cap W) = (0.4)(0.3)$$
$$\boxed{P(T \cap W) = 0.12}$$

(ii) she hits the target with her first shot; $P(T)$

$$P(T) = P(T \cap W) + P(T \cap W^C)$$
$$P(T) = P(T|W) \cdot P(W) + P(T|W^C) \cdot P(W^C), \quad P(W^C) = 1 - P(W) = 0.7$$
$$P(T) = (0.4)(0.3) + (0.7)(0.7)$$
$$\boxed{P(T) = 0.61}$$

(iii) she hits the target exactly once in two shots; $P(T) \cdot P(T^C) + P(T^C) \cdot P(T)$

$$P(T) \cdot P(T^C) + P(T^C) \cdot P(T) = 2 \cdot P(T) \cdot P(T^C), \quad P(T^C) = 1 - P(T) = 0.39$$
$$P(T) \cdot P(T^C) + P(T^C) \cdot P(T) = 2(0.61)(0.39)$$
$$\boxed{P(T) \cdot P(T^C) + P(T^C) \cdot P(T) = 0.4758}$$

(iv) there was no gust of wind on an occasion when she missed; $\frac{P(T^C \cap W^C)}{P(T^C)}$

$$\frac{P(T^C \cap W^C}{P(T^C} = \frac{P(T^C|W^C)P(W^C)}{P(T^C)}, \quad P(T^C|W^C) = 1 - P(T|W^C) = 0.3$$
$$\frac{P(T^C \cap W^C)}{P(T^C)} = \frac{(0.3)(0.7)}{(0.39)}$$
$$\boxed{\frac{P(T^C \cap W^C)}{P(T^C)} = 0.538}$$

b.)

We are given

$$P(A|B,C) = P(A|B \cap C) > P(A|B)$$

and from the properties of conditional probability we find

$$P(A|B \cap C) = \frac{P(A \cap B \cap C)}{P(B \cap C)} > \frac{P(A \cap B)}{P(B)}. \tag{1}$$

We are looking to show

$$P(A|B,C^C) = P(A|B \cap C^C) < P(A|B).$$

or equivalently

$$P(A|B \cap C^C) = \frac{P(A \cap B \cap C^C)}{P(B \cap C^C)} < \frac{P(A \cap B)}{P(B)}. \tag{2}$$

Returning to equation 1, we find through algebraic manipulation that

$$P(A \cap B \cap C) > \frac{P(A \cap B)P(B \cap C)}{P(B)}$$

$$-P(A \cap B \cap C) < -\frac{P(A \cap B)P(B \cap C)}{P(B)}$$

$$P(A \cap B) - P(A \cap B \cap C) < P(A \cap B) - \frac{P(A \cap B)P(B \cap C)}{P(B)}$$

$$P(A \cap B) - P(A \cap B \cap C) < \frac{P(A \cap B)}{P(B)} \left(P(B) - P(B \cap C)\right)$$

$$\frac{P(A \cap B) - P(A \cap B \cap C)}{P(B) - P(B \cap C)} < \frac{P(A \cap B)}{P(B)}. \tag{3}$$

Since in general,

$$P(X \cap Y) + P(X \cap Y^C) = P(X),$$

$$\text{(also expressed as } P(X \cap Y^C) = P(X) - P(X \cap Y))$$

we can express equation 3 as

$$\frac{P(A \cap B \cap C^C)}{P(B \cap C^C)} < \frac{P(A \cap B)}{P(B)},$$

equivalent to equation 2. ■

Problem 2

a.)

Assume a positive semidefinite matrix $A \in \mathbb{R}^{n \times n}$ so that $x^T Ax \geq 0$. By definition, we say $A \succeq 0$ (i).

(i)⇒(ii):

Since, by definition, $A \succeq 0 \Rightarrow x^T Ax \geq 0$, we could let $x = By$ where $B \in \mathbb{R}^{n \times n}$ is invertible. Then

$$x^T Ax = (By)^T A(By) \geq 0$$

$$y^T B^T ABy \geq 0$$

This is of the form defining a semidefinite matrix, implying that $B^T AB \succeq 0$. $\blacksquare$.

(ii)⇒(i):

We assume now that $B^T AB \succeq 0$ and $B \in \mathbb{R}^{n \times n}$ is invertible. By the definition of semidefinite matrices,

$$x^T B^T ABx \geq 0.$$

Utilizing the properties of transpose matrices, we find

$$(Bx)^T A(Bx) \geq 0.$$

Let $Bx = y$, then this becomes

$$y^T Ay \geq 0$$

and equivalent statement to $A \succeq 0$. $\blacksquare$

(i)⇒(iii):

The eigenvalues of A are defined as λ when $Ax = \lambda x$. Using this equality in the definition of semidefinite matrix A ($A \succeq 0$), we find

$$x^T Ax = x^T \lambda x \geq 0.$$

As a constant, we can reexpress this as

$$\lambda x^T x \geq 0$$

$$\lambda |x|^2 \geq 0$$

We note $|x|^2 = \sum_i x_i^2$ and $x_i^2 \geq 0 \;\therefore\; |x|^2 \geq 0$. With $|x|^2 \geq 0$, $\lambda \geq 0$ in order for $\lambda |x|^2 \geq 0$ to be true. $\blacksquare$

(iii)⇒(iv):

Since matrix A is symmetric (given), we can apply the spectral theorem for symmetric matrices (as stated by Mark Gockenbach on his website. This theorem states that for symmetric $A \in \mathbb{R}^{n \times n}$ there exists a diagonal matrix $D \in \mathbb{R}^{n \times n}$ and an orthogonal matrix $P \in \mathbb{R}^{n \times n}$ such that $A = PDP^T$. Furthermore, the diagonal entries of D are the eigenvalues of A.

Any nonnegative diagonal matrix $E \in R^{n \times n}$ can be equivalently represented as F^2 where $F \in R^{n \times n}$ is another diagonal matrix with elements $F_i = \sqrt{E_i}$. We showed in (iii) that all eigenvalues of A are nonnegative. Therefore, if all entries of D are the eigenvalues of A, then D is a nonnegative diagonal matrix and $D = F^2$. Since all diagonal matrices are symmetrical, $D = F^2 = FF^T$.

Together with the spectral theorem, we find

$$A = PDP^T = PFF^T P^T$$

$$A = PF(PF)^T$$

If we let $U = PF$ ($U \in \mathbb{R}^{n \times n}$), then

$$A = UU^T. \;\blacksquare$$

(iv)⇒(i):

Assume that $\exists$ matrix $U \in \mathbb{R}^{n\times n}$ such that $A = UU^T$. We want to show that $A = UU^T \succeq 0$.

$$A = UU^T$$

$$x^T Ax = x^T UU^T x$$

$$x^T Ax = (U^T x)^T U^T x$$

We can say $U^T x = v$, where $v \in \mathbb{R}^n$, so

$$x^T Ax = v^T v$$

$$x^T Ax = |v|^2$$

$|v|^2 = \sum_i v_i^2$ and for all real numbers, $v_i^2 \geq 0 \ \therefore \ |v|^2 \geq 0$.

$$x^T Ax \geq 0$$

$$A \succeq 0. \ \blacksquare$$

b.)

$A \in \mathbb{R}^{n\times n}$ is a positive definite matrix, such that $A \succ 0$.

(i)

We are given that every $\lambda > 0$ and want to prove that $A + \lambda I \succ 0$. By definition, this is true iff

$$x^T (A + \lambda I)x > 0$$

$$x^T Ax + x^T \lambda Ix > 0$$

By definition, A is positive definite, so $x^T Ax > 0$. Furthermore, $x^T \lambda Ix = x^T \lambda x = \lambda x^T x = \lambda|x|^2$. $|x|^2 > 0$ and it is given that $\lambda > 0$, so the product $\lambda|x|^2 > 0$. Additionally, if $\lambda|x|^2 > 0$ and $x^T Ax > 0$, then their sum must also be greater than 0.

$$x^T Ax + x^T \lambda Ix > 0 \quad \therefore \quad A + \lambda I \succ 0. \ \blacksquare$$

(ii)

We are attempting to prove $\exists \gamma > 0$ such that $A - \gamma I \succ 0$.

$$A - \gamma I \succ 0 \implies x^T (A - \gamma I)x > 0.$$

$$x^T Ax - x^T \gamma Ix > 0$$

Like in 2.b.iii, we note that the eigenvalues of A are given by λ, when $Ax = \lambda x$.

$$x^T \lambda x - x^T \gamma Ix > 0$$

$$\lambda x^T x - \gamma x^T x > 0$$

$$(\lambda - \gamma)x^T x > 0$$

$$(\lambda - \gamma) \cdot |x|^2 > 0$$

Since $|x|^2 > 0$, dividing it out yields

$$\lambda - \gamma > 0.$$

Now, using the definition of positive definite matrices,

$$x^T Ax = x^T \lambda x > 0$$

$$\lambda x^T x > 0$$

$$\lambda |x|^2 > 0.$$

Since $|x|^2 > 0,\ \ \lambda > 0$ for the product $\lambda |x|^2 > 0$.

Now, since we have shown $\lambda > 0$, then when $\lambda - \gamma > 0,\ \ \{\gamma \in \mathbb{R} \mid 0 < \gamma < \lambda\}$. ■

(iii)

First, we note that the procedure for calculating $a = x^T Ax$ is $a = \sum_{j=1}^{n} \sum_{k=1}^{n} x_j x_k A_{jk}$.

By definition, $A \succ 0 \implies x^T Ax > 0 \ \ \forall\, x \in \mathbb{R}^n - \{0\}$. Since all $x \in \mathbb{R}^n - \{0\}$ must satisfy this equation, showing that the diagonal entries of A must be greater than zero for any set of x vectors is sufficient to prove this case.

With this in mind, consider the basis of unit vectors $\boldsymbol{e_1}, ...\ \boldsymbol{e_n}$, where $\boldsymbol{e_1^T} = (1\ \ 0\ \ ...\ \ 0)$, $\boldsymbol{e_2^T} = (0\ \ 1\ \ ...\ \ 0)$, $\boldsymbol{e_n^T} = (0\ \ 0\ \ ...\ \ 1)$, etc.

Letting each of these unit vectors $\boldsymbol{e_i} = x$,

$$x^T Ax > 0$$

$$\boldsymbol{e_i}^T A \boldsymbol{e_i} > 0$$

$$\sum_{j=1}^{n} \sum_{k=1}^{n} x_j x_k A_{jk} > 0$$

Since in $\boldsymbol{e_i},\ \ x_i = 1$ if $i = k$, otherwise $x_i = 0$, we can rewrite this summation as

$$x_i^2 A_{ii} > 0$$

$x_i = 1 \therefore x_i^2 = 1$, so

$$A_{ii} > 0.\ ■$$

(iv)

By definition of a positive definite matrix, $A \succ 0$, we have $x^T Ax > 0$. As noted in the proof of (iii), the procedure for calculating $a = x^T Ax$ is $a = \sum_i^n \sum_j^n x_i x_j A_{ij}$.

Again, since $A \succ 0 \implies x^T Ax > 0 \ \ \forall\, x \in \mathbb{R}^n - \{0\}$, it is sufficient to show that $\sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} > 0$ for any vector.

Consider the vector $v^T = (1\ \ 1\ \ ...\ \ 1)$, a vector in $\mathbb{R}^n$ consisting of all ones.

$$v^T Av > 0$$

$$\sum_i^n \sum_j^n x_i x_j A_{ij} > 0, \text{ where } x_i = 1, x_j = 1 \ \ \forall\, i, j.$$

$$\sum_i^n \sum_j^n A_{ij} > 0.\ ■$$

Problem 3

a.)

Let $x, a \in \mathbb{R}^n$. We can express the gradient of a function $f(x)$ as the vector $\nabla_x f(x)$ where the i^{th} element is given by $\frac{df}{dx_i}$. This gives

$$\nabla_x(a^T x) = \left(\frac{d}{dx_1}(a^T x) \quad \frac{d}{dx_2}(a^T x) \quad ... \quad \frac{d}{dx_n}(a^T x) \right)^T$$

Noting that $a^T x = \sum_{j=1}^n a_j x_j$,

$$\nabla_x(a^T x) = \left(\frac{d}{dx_1} \sum_{i=1}^n a_i x_i \quad \frac{d}{dx_2} \sum_{i=1}^n a_i x_i \quad ... \quad \frac{d}{dx_n} \sum_{i=1}^n a_i x_i \right)^T$$

. For any summation,

$$\frac{d}{dx_i} \sum_{j=1}^n C_j x_j = \sum_{j=1}^n \frac{d}{dx_i} C_j x_j = \sum_{j=1}^n (C_j \delta_{ij}) = C_i$$

where C_j is a constant and δ_{ij} is the Kronecker delta. Then,

$$\nabla_x(a^T x) = (a_1 \quad a_2 \quad ... \quad a_n)^T$$

$$\boxed{\nabla_x(a^T x) = a}$$

b.)

Let $A \in \mathbb{R}^{n \times n}, x \in \mathbb{R}^n$. We note that $x^T A x = \sum_{i=1}^n \sum_{j=1}^n x_i x_j A_{ij}$. Then,

$$\nabla_x(x^T A x) = \left(\frac{d}{dx_1}(x^T A x) \quad \frac{d}{dx_2}(x^T A x) \quad ... \quad \frac{d}{dx_n}(x^T A x) \right)^T$$

The i^{th} row of $x^T A x$ is given by

$$\frac{d}{dx_i}(x^T A x) = \frac{d}{dx_i} \sum_{j=1}^n \sum_{k=1}^n A_{jk} x_j x_k$$

For any summation,

$$\frac{d}{dx_i} \sum_{j=1}^n \sum_{k=1}^n C_{jk} x_j x_k = \sum_{j=1}^n \sum_{k=1}^n \frac{d}{dx_i} C_{jk} x_j x_k$$

$$\frac{d}{dx_i} \sum_{j=1}^n \sum_{k=1}^n C_{jk} x_j x_k = \sum_{j=1}^n \sum_{k=1}^n C_{jk} (x_j \delta_{ij} + x_k \delta_{ik})$$

$$\frac{d}{dx_i} \sum_{j=1}^n \sum_{k=1}^n C_{jk} x_j x_k = \sum_{j=1}^n \sum_{k=1}^n C_{jk} x_j \delta_{ij} + \sum_{j=1}^n \sum_{k=1}^n C_{jk} x_k \delta_{ik}$$

$$\frac{d}{dx_i} \sum_{j=1}^n \sum_{k=1}^n C_{jk} x_j x_k = \sum_{k=1}^n C_{ik} x_i + \sum_{j=1}^n C_{ji} x_i$$

Furthermore, for any matrix $C \in \mathbb{R}^{n \times n}$,

$$Cx = \left(\sum_{l=1}^n C_{l1} x_1 \quad \sum_{l=1}^n C_{l2} x_2 \quad ... \quad \sum_{l=1}^n C_{ln} x_n \right)^T$$

$$(Cx)_i = \sum_{l=1}^n C_{l1} x_1$$

Then,

$$\frac{d}{dx_i}\sum_{j=1}^{n}\sum_{k=1}^{n} C_{jk}x_jx_k = (Cx)_i + (C^Tx)_i$$

We can use this conclusion in our above expression for $\nabla_x(x^TAx)$

$$\nabla_x(x^TAx) = \begin{bmatrix}(Ax)_1 + (A^Tx)_1 & (Ax)_2 + (A^Tx)_2 & \dots & (Ax)_n + (A^Tx)_n\end{bmatrix}^T$$

$$\nabla_x(x^TAx) = (Ax + A^Tx)$$

$$\nabla_x(x^TAx) = (A + A^T)x$$

In the case that A is symmetric, $A = A^T$ so

$$\nabla_x(x^TAx) = (A + A^T)x = (A + A)x$$

$$\boxed{\nabla_x(x^TAx) = 2Ax}$$

c.)

Let $A, X \in \mathbb{R}^{n\times n}$. Explicitly, $(A^TX)_{ij} = \sum_{k=1}^{n} A_{ki}X_{kj}$. The diagonals of (A^TX) are

$$(A^TX)_{ii} = \sum_{k=1}^{n} A_{ki}X_{ki}.$$

As the sum of the diagonals,

$$\operatorname{tr}\left(A^TX\right) = \sum_{i=1}^{n}(A^TX)_{ii}$$

$$\operatorname{tr}\left(A^TX\right) = \sum_{i=1}^{n}\sum_{k=1}^{n} A_{ki}X_{ki}$$

We can further show that

$$\frac{d}{dX_{lm}}\operatorname{tr}\left(A^TX\right) = \frac{d}{dX_{lm}}\sum_{i=1}^{n}\sum_{k=1}^{n} A_{ki}X_{ki}$$

$$\frac{d}{dX_{lm}}\operatorname{tr}\left(A^TX\right) = \sum_{i=1}^{n}\sum_{k=1}^{n}\frac{d}{dX_{lm}} A_{ki}X_{ki}$$

$$\frac{d}{dX_{lm}}\operatorname{tr}\left(A^TX\right) = \sum_{i=1}^{n}\sum_{k=1}^{n} A_{ki}\delta_{(ki)(lm)}$$

$$\frac{d}{dX_{lm}}\operatorname{tr}\left(A^TX\right) = A_{lm}$$

Then, since

$$\nabla_X(\operatorname{tr}\left(A^TX\right)) = \begin{pmatrix} \frac{d}{dX_{11}}(\operatorname{tr}\left(A^TX\right)) & \frac{d}{dX_{12}}(\operatorname{tr}\left(A^TX\right)) & \dots & \frac{d}{dX_{1n}}(\operatorname{tr}\left(A^TX\right)) \\ \frac{d}{dX_{21}}(\operatorname{tr}\left(A^TX\right)) & \frac{d}{dX_{22}}(\operatorname{tr}\left(A^TX\right)) & \dots & \frac{d}{dX_{2n}}(\operatorname{tr}\left(A^TX\right)) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{d}{dX_{n1}}(\operatorname{tr}\left(A^TX\right)) & \frac{d}{dX_{12}}(\operatorname{tr}\left(A^TX\right)) & \dots & \frac{d}{dX_{nn}}(\operatorname{tr}\left(A^TX\right)) \end{pmatrix}$$

we can can say

$$\nabla_X(\operatorname{tr}\left(A^TX\right)) = \begin{pmatrix} A_{11} & A_{12} & \dots & A_{1n} \\ A_{21} & A_{22} & \dots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \dots & A_{nn} \end{pmatrix}$$

$$\boxed{\nabla_X(\operatorname{tr}\left(A^TX\right)) = A}$$

d.)

To be a norm, the distance metric $\delta(x, y) = f(x - y)$ must satisfy the triangle inequality, $\delta(x, z) \leq \delta(x, y) + \delta(y, z)$. In this case, it follows that

$$f(x - z) \leq f(x - y) + f(y - z). \tag{4}$$

Here, $x + y = z$. For vectors $x \in \mathbb{R}^2$, we can test if $f(x) = (\sqrt{|x_1|} + \sqrt{|x_2|})^2$ is a norm using equation 4.

$$(\sqrt{|x_1 - z_1|} + \sqrt{|x_2 - z_2|})^2 \leq (\sqrt{|x_1 - y_1|} + \sqrt{|x_2 - y_2|})^2 + (\sqrt{|y_1 - z_1|} + \sqrt{|y_2 - z_2|})^2$$

$$|x_1-z_1|+|x_2-z_2|+2\sqrt{|x_1 - z_1| \cdot |x_2 - z_2|} \leq |x_1-y_1|+|x_2-y_2|+2\sqrt{|x_1 - y_1| \cdot |x_2 - y_2|}+|y_1-z_1|+|y_2-z_2|+2\sqrt{|y_1 - z_1| \cdot |y_2 - z_2|}$$

Since $x + y = z$, $x_i + y_i = z_i$. From this,

$$x_i - z_i = -y_i \text{ and } y_i - z_i = -x_i$$

$$|x_i - z_i| = |y_i| \text{ and } |y_i - z_i| = |x_i|$$

Then, we have

$$|y_1| + |y_2| + 2\sqrt{|y_1| \cdot |y_2|} \leq |x_1 - y_1| + |x_2 - y_2| + 2\sqrt{|x_1 - y_1| \cdot |x_2 - y_2|} + |x_1| + |x_2| + 2\sqrt{|x_1| \cdot |x_2|}$$

$$|y_1| - |x_1| + |y_2| - |x_2| + 2(\sqrt{|y_1| \cdot |y_2|} - \sqrt{|x_1| \cdot |x_2|}) \leq |x_1 - y_1| + |x_2 - y_2| + 2\sqrt{|x_1 - y_1| \cdot |x_2 - y_2|}$$

Since it is always true that for any $a, b \in \mathbb{R}$, $|a-b| = |b-a| \geq |a|-|b|$, we know it is true that both $|x_1-y_1| \geq |y_1|-|x_1|$ and $|x_2 - y_2| \geq |y_2| - |x_2|$. This leaves it only necessary to further prove

$$2(\sqrt{|y_1| \cdot |y_2|} - \sqrt{|x_1| \cdot |x_2|}) \leq 2\sqrt{|x_1 - y_1| \cdot |x_2 - y_2|}$$

$$\sqrt{|y_1 y_2|} - \sqrt{|x_1 x_2|} \leq \sqrt{|x_1 - y_1| \cdot |x_2 - y_2|}$$

$$\sqrt{|y_1 y_2|} - \sqrt{|x_1 x_2|} \leq \sqrt{|x_1 x_2 + y_1 y_2 - x_1 y_2 - x_2 y_1|}$$

$$\sqrt{|y_1 y_2|} \leq \sqrt{|x_1 x_2 + y_1 y_2 - x_1 y_2 - x_2 y_1|} + \sqrt{|x_1 x_2|}$$

If we wish to provide a counter example, it must satisfy the condition that

$$\sqrt{|y_1 y_2|} > \sqrt{|x_1 x_2 + y_1 y_2 - x_1 y_2 - x_2 y_1|} + \sqrt{|x_1 x_2|}$$

To satisfy this condition, we could try x, y such that $|y_1 y_2|$ is large, $|x_1 x_2|$ is small, and $(-x_1 y_2 - x_2 y_1)$ is large, but smaller than $y_1 y_2$. Specifically, we try

$$x = \begin{pmatrix} 1 \\ 64 \end{pmatrix} \qquad y = \begin{pmatrix} 10 \\ 1000 \end{pmatrix}$$

Then,

$$\sqrt{|y_1 y_2|} > \sqrt{|x_1 x_2 + y_1 y_2 - x_1 y_2 - x_2 y_1|} + \sqrt{|x_1 x_2|}$$

$$\sqrt{10^4} > \sqrt{|64 + 10^4 - 1000 - 640|} + \sqrt{|64|}$$

$$10^2 > \sqrt{10064 - 1640} + 8$$

$$92 > 91.782...$$

Since this satisfies this condition, we can try it in our original equation,

$$(\sqrt{|x_1 - z_1|} + \sqrt{|x_2 - z_2|})^2 \leq (\sqrt{|x_1 - y_1|} + \sqrt{|x_2 - y_2|})^2 + (\sqrt{|y_1 - z_1|} + \sqrt{|y_2 - z_2|})^2$$

Now $z = \begin{pmatrix} 11 \\ 1064 \end{pmatrix}$ and so

$$(\sqrt{|1 - 11|} + \sqrt{|64 - 1064|})^2 \leq (\sqrt{|1 - 10|} + \sqrt{|64 - 1000|})^2 + (\sqrt{|10 - 11|} + \sqrt{|1000 - 1064|})^2$$

$$(\sqrt{10} + \sqrt{1000})^2 \leq (\sqrt{9} + \sqrt{936})^2 + (\sqrt{1} + \sqrt{64})^2$$

But, instead we find $1210 \not\leq 1209.565$, and so we have found a counterexample. The given function is **not** a norm.

e.)

Let $x \in \mathbb{R}^n$. We know that $||x||_\infty = \max_i |x_i|$, and $||x||_2 = \sum_i^n |x_i|^2$.

Minimum $||x||_2$:
The minimum (nontrivial) value of $||x||_2$ would be given for any vector with only one single nonzero element (say this element is the j^{th} element). This can be shown as:

$$\sqrt{x_1^2} \leq \sqrt{x_1^2 + ... + x_n^2}.$$

In this case,

$$||x||_2 = \sqrt{\sum_i^n |x_i|^2} = \sqrt{\sum_i^n |x_i \delta_{ij}|^2}$$

$$||x||_2 = \sqrt{|x_j|^2} = |x_j|$$

With all other elements of x being zero, $|x_j| = \max_i |x_i|$.

Maximum $||x||_2$
Similarly, the maximum value of $||x||_2$ would be given when every $|x_i| = \max_i |x_i| = |x_j|$. This can be shown as:

$$||x||_2 = \sqrt{\sum_i^n |x_i|^2}$$

$$||x||_2 = \sqrt{n|x_j|^2} = \sqrt{n}|x_j|$$

With all elements of x being nonzero, $|x_j| = \max_i |x_i|$.

In both cases, we have defined $|x_j| = \max_i |x_i|$. Using proper substitutions, we find that

$$|x_j| \leq \sqrt{x_1^2 + ... + x_n^2} \leq \sqrt{n}|x_j|$$

$$\max_i |x_i| \leq \sqrt{x_1^2 + ... + x_n^2} \leq \sqrt{n} \max_i |x_i|$$

$$||x||_\infty \leq ||x||_2 \leq \sqrt{n}||x||_\infty \quad \blacksquare$$

f.)

Let $x \in \mathbb{R}^n$. We know that $||x||_1 = \sum_i^n |x_i|$, and $||x||_2 = \sum_i^n |x_i|^2$.

Minimum $||x||_1$:
The minimum (nontrivial) value of $||x||_1$ would be given for any vector with only one single nonzero element (say this element is the j^{th} element). This can be shown as:

$$|x_1| \leq |x_1| + ... + |x_n|.$$

In this case,

$$||x||_1 = \sum_i^n |x_i \delta_{ij}| = |x_j|$$

$$||x||_2 = \sqrt{\sum_i^n |x_i \delta_{ij}|^2} = |x_j|$$

Maximum $||x||_1$
Similarly, the maximum value of $||x||_1$ would be given when every $|x_i| = \max_i |x_i| = |x_j|$. This can be shown as:

$$||x||_1 = \sum_i^n |x_i|$$

$$||x||_1 = n|x_j|.$$

From the previous problem, we also showed that in this case

$$||x||_2 = \sqrt{n}|x_j|$$

$$||x||_1 = \sqrt{n}||x||_2$$

Combining these minimum an maximum values:

$$||x||_2 \leq ||x||_1 \leq \sqrt{n}||x||_2 \quad \blacksquare$$

Problem 4

Let $A \in \mathbb{R}^{n \times n}$ be a symmetric matrix with $A \succeq 0$.

a.)

As discussed in problem 2.a, the spectral theorem for symmetric matrices states that for symmetric $A \in \mathbb{R}^{n\times n}$ there exists a diagonal matrix $D \in \mathbb{R}^{n\times n}$ and an orthogonal matrix $P \in \mathbb{R}^{n\times n}$ such that $A = PDP^T$. Furthermore, the diagonal entries of D are the eigenvalues of A.

We can use this to show

$$\lambda_{\max}(A) = \max_{||x||_2=1} x^T Ax.$$

or equivalently (using the spectral theorem)

$$\lambda_{\max}(A) = \max_{||x||_2=1} x^T PDP^T x.$$

Since $x \in \mathbb{R}^n$ and $P \in \mathbb{R}^{n\times n}$,

$$P^T x = y, \text{ and } (x^T P)^T = y^T$$

where $y \in \mathbb{R}^n$. We then find

$$\lambda_{\max}(A) = \max_{||x||_2=1} y^T Dy.$$

We note that since P is orthogonal, the columns of P are orthonormal, and it follows that

$$||(P^T x)||_2 = 1 \text{ and } ||(x^T P)||_2 = 1$$

$$||y||_2 = 1 \text{ and } ||y^T||_2 = 1$$

$$\lambda_{\max}(A) = \max_{||y||_2=1} y^T Dy.$$

Now,

$$\max_{||y||_2=1} y^T Dy = \max_{||y||_2=1} \sum_i^n y_i^2 D_{ii}$$

Since $\sum_i y_i^2 = 1$ (it is constant), we will maximize $\max_{||y||_2=1} \sum_i^n y_i^2 D_{ii}$ by choosing the configuration of y that favors $\max_i D_{ii}$. If we let $\max_i D_{ii} = D_{jj}$, then this would be the vector where $y_{i=j} = 1$ and $y_{i\neq j} = 0$. Since $y_j = 1$,

$$\max_{||y||_2=1} y^T Dy = \max_{||y||_2=1} \sum_i^n D_{ii}\delta_{ij}$$

$$\max_{||y||_2=1} y^T Dy = \max_{||y||_2=1} D_{ii}$$

Then,

$$\lambda_{\max}(A) = \max_{||y||_2=1} D_{ii}.$$

Returning to the assertion of the spectral theorem that the elements of D are the eigenvalues of A, this statement is true. ■

b.)

Using the procedure established in part (a) we can repeat to show

$$\lambda_{\min}(A) = \min_{||x||_2=1} x^T Ax.$$

or equivalently (using the spectral theorem)

$$\lambda_{\min}(A) = \min_{||x||_2=1} x^T PDP^T x.$$

Since $x \in \mathbb{R}^n$ and $P \in \mathbb{R}^{n\times n}$,

$$P^T x = y, \text{ and } (x^T P)^T = y^T$$

where $y \in \mathbb{R}^n$. We then find

$$\lambda_{\min}(A) = \min_{||x||_2=1} y^T Dy.$$

We note that since P is orthogonal, the columns of P are orthonormal, and it follows that

$$||(P^T x)||_2 = 1 \text{ and } ||(x^T P)||_2 = 1$$

$$||y||_2 = 1 \text{ and } ||y^T||_2 = 1$$

$$\lambda_{\min}(A) = \min_{||y||_2=1} y^T Dy.$$

Now,

$$\min_{||y||_2=1} y^T Dy = \min_{||y||_2=1} \sum_i^n y_i^2 D_{ii}$$

Since $\sum_i y_i^2 = 1$ (it is constant), we will minimize $\min_{||y||_2=1} \sum_i^n y_i^2 D_{ii}$ by choosing the configuration of y that favors $\min_i D_{ii}$. If we let $\min_i D_{ii} = D_{jj}$, then this would be the vector where $y_{i=j} = 1$ and $y_{i\neq j} = 0$. Since $y_j = 1$,

$$\min_{||y||_2=1} y^T Dy = \min_{||y||_2=1} \sum_i^n D_{ii}\delta_{ij}$$

$$\min_{||y||_2=1} y^T Dy = \min_{||y||_2=1} D_{ii}$$

Then,

$$\lambda_{\min}(A) = \min_{||y||_2=1} D_{ii}.$$

Returning to the assertion of the spectral theorem that the elements of D are the eigenvalues of A, this statement is true. ■

c.)

The conditions which must be satisfied for a minimization (maximization) problem to be satisfied are:
(1) the objective function to be a convex (concave) function
(2) the feasible region to be a convex set

By definition, a set $\mathcal{C} \subseteq \mathbb{R}^n$ is convex iff $\forall x, y \in \mathcal{C},\ \forall t \in [0,1],\ tx + (1-t)y \in \mathcal{C}$.

Say we choose 2 vectors $x = \begin{pmatrix}1\\0\end{pmatrix}$ and $y = \begin{pmatrix}0\\1\end{pmatrix}$. Furthermore, choose $t = 0.75$. Note that $||x||_2 = 1$ and $||y||_2 = 1$, where $||\cdot||_2 = 1$ is the condition defining our set.

$$0.75\begin{pmatrix}1\\0\end{pmatrix} + (1-0.75)\begin{pmatrix}0\\1\end{pmatrix} = \begin{pmatrix}0.75\\0.25\end{pmatrix} \notin \mathcal{C}.$$

The set is not convex, and so neither program is convex.

d.)

Again using the spectral theorem for symmetric matrices, we know $A = PDP^T$ where A is symmetric, P is orthogonal, and D is diagonal and its diagonals are the eigenvalues of A. Then, we can see

$$AA = (PDP^T)(PDP^T)$$

$$A^2 = PD\mathbb{1}DP^T, \text{ since } P^T = P^{-1} \text{ for orthogonal matrices}$$

$$A^2 = PD^2P^T$$

And so, since A^2 is also symmetric, D^2 has diagonals λ^2 which are the eigenvalues of A^2.

Using this fact, we know from (a) and (b) that

$$\lambda_{\max}(A^2) = \max(D_{ii}^2) \text{ and } \lambda_{\min}(A^2) = \min(D_{ii}^2)$$

$$\lambda_{\max}(A^2) = (\max D_{ii})^2 \text{ and } \lambda_{\min}(A^2) = (\min D_{ii})^2$$

$$\lambda_{\max}(A^2) = \lambda_{\max}(A)^2 \text{ and } \lambda_{\min}(A^2) = \lambda_{\min}(A)^2$$

e.)

Noting that $||Ax||_2 = \sqrt{(Ax)^T(Ax)}$, we find

$$||Ax||_2 = \sqrt{x^TA^TAx}$$

Since A is symmetric, $AA^T = A^2$, and

$$||Ax||_2 = \sqrt{x^TA^2x}$$

It is by definition that

$$\min_{||x||_2=1} ||Ax||_2 \leq ||Ax||_2 \leq \max_{||x||_2=1} ||Ax||_2. \tag{5}$$

From this, we can deduce

$$\min_{||x||_2=1} ||Ax||_2 = \min_{||x||_2=1} \sqrt{x^TA^2x} \text{ and } \max_{||x||_2=1} ||Ax||_2 = \max_{||x||_2=1} \sqrt{x^TA^2x}$$

$$\min_{||x||_2=1} ||Ax||_2 = \sqrt{\min_{||x||_2=1} x^TA^2x} \text{ and } \max_{||x||_2=1} ||Ax||_2 = \sqrt{\max_{||x||_2=1} x^TA^2x}.$$

By substituting our answer from parts (a) and (b) we have

$$\min_{||x||_2=1} ||Ax||_2 = \sqrt{\lambda_{\min}(A^2)} \text{ and } \max_{||x||_2=1} ||Ax||_2 = \sqrt{\lambda_{\max}(A^2)}$$

and by then substituting our answer from part (d) we have

$$\min_{||x||_2=1} ||Ax||_2 = \sqrt{\lambda_{\min}(A)^2} \text{ and } \max_{||x||_2=1} ||Ax||_2 = \sqrt{\lambda_{\max}(A)^2}$$

$$\min_{||x||_2=1} ||Ax||_2 = \lambda_{\min}(A) \text{ and } \max_{||x||_2=1} ||Ax||_2 = \lambda_{\max}(A).$$

When consiered with equation 5, we have

$$\lambda_{\min}(A) \leq ||Ax||_2 \leq \lambda_{\max}(A)$$

f.)

As in part (e) we know $||Ax||_2 = \sqrt{x^T Ax}$, $\forall x \in \mathbb{R}^n$. The result found in (d) is a special case for unit vectors, where

$$\lambda_{\min}(A) \leq ||Ax||_2 \leq \lambda_{\max}(A).$$

Instead, in the general case, we can force any vector x to be a unit vector by dividing it by its magnitude: $\frac{x}{||x||_2}$. Then for non-unit vectors, we have

$$||Ax||_2 = \sqrt{y^T Ay||x||_2^2} = ||x||_2\sqrt{y^T Ay}$$

where $y = \frac{x}{||x||_2}$. Using this in conjunction with the part (d), we find that in general, $\forall x \in \mathbb{R}^n$,

$$\lambda_{\min}(A)||x||_2 \leq ||Ax||_2 \leq \lambda_{\max}(A)||x||_2.$$

Problem 5

a.)

First order optimality conditions require $\nabla_x f(x) = 0$ (where $f(x)$ is the objective function). When computing this gradient, the i^{th} element is given by $\frac{df}{dx_i}$.

The objective function in this case is

$$\frac{1}{2}x^T Ax - b^T x,$$

so

$$\nabla_x \left(\frac{1}{2}x^T Ax - b^T x \right) = 0.$$

From problem 3.a we have $\nabla_x(a^T x) = a$, and from problem 3.b we have $\nabla_x(x^T Ax) = 2Ax$. We then find for our current objective function:

$$\nabla_x \left(\frac{1}{2}x^T Ax - b^T x \right) = \frac{1}{2}(2Ax^*) - b = 0$$

$$Ax^* - b = 0$$

$$Ax^* = b$$

$$\boxed{x^* = A^{-1}b}$$

b.)

The update rule for gradient descent is given by

$$w \leftarrow w - \epsilon \nabla_w R(w)$$

where $R(w)$ is the risk function and ϵ is the step size.

For this problem, the update rule becomes

$$x \leftarrow x - (1)\nabla_x \left(\frac{1}{2}x^T Ax - b^T x \right)$$

where $R(w) = \frac{1}{2}x^T Ax - b^T x$ and $\epsilon = 1$. Then,

$$x^{(k)} = x^{(k-1)} - (Ax^{(k-1)} - b).$$

$$\boxed{x^{(k)} = x^{(k-1)} - Ax^{(k-1)} + b}.$$

c.)

$$x^{(k)} - x^* = (x^{(k-1)} - Ax^{(k-1)} + b) - x^*$$

$$x^{(k)} - x^* = x^{(k-1)} - Ax^{(k-1)} + Ax^* - x^*$$

$$x^{(k)} - x^* = x^{(k-1)} - Ax^{(k-1)} - x^* + Ax^*$$

$$x^{(k)} - x^* = (I - A)(x^{(k-1)} - x^*)$$

$$\boxed{x^{(k)} - x^* = (I - A)(x^{(k-1)} - x^*)}$$

d.)

We know from (c) that $x^{(k)} - x^* = (I - A)(x^{(k-1)} - x^*)$. It follows that $||x^{(k)} - x^*||_2 = ||(I - A)(x^{(k-1)} - x^*)||_2$. If we let $y = x^{(k-1)} - x^*$ then this equation becomes

$$||x^{(k)} - x^*||_2 = ||(I - A)y||_2$$

If we can show $(I - A) \succeq 0$ then we can use the inequality found in problem 4.f. For $(I - A) \succeq 0$, $x^T(I - A)x \geq 0$. By the definition of an eigenvalue, we know that

$$Ax = \lambda x$$

$$Ax - \lambda x = 0$$

$$(A - \lambda I)x = 0$$

$$(\lambda I - A)x = 0$$

We are told that all eigenvalues of A are on the interval (0,1), so we know that $(\lambda I)_{ii} < I_{ii} \quad \forall i \in 1, ..., n$. Then $(I - A) > (\lambda I - A)$ and if $x^T(\lambda I - A)x = 0$ then $x^T(I - A)x > 0$ (assuming $x \neq 0$). It becomes evident that indeed $(I - A) \succeq 0$.

Now that we have shown that $(I - A)$ is semi-positive definite, we can use the result of problem 4.f to show

$$||(I - A)y||_2 \leq \lambda_{\max}(I - A)||y||_2.$$

Since we know that $(I - A)x = \lambda x$ where λ represents the eigenvalues of $(I - A)$, we can show

$$(I - A)x = \lambda x$$

$$Ax - x = -\lambda x$$

$$Ax = (1 - \lambda)x$$

(*i.e.* $(1 - \lambda)$ are the eigenvalues of A). Since we are told $0 \leq \lambda_{\min}(A)$ and $\lambda_{\max}(A) \leq 1$ we can deduce that $0 \leq \lambda_{\min}(I - A)$ and $\lambda_{\max}(I - A)$ as well. Let ρ represent the maximum eigenvalue (still, $0 < \rho < 1$).

Now,

$$||(I - A)y||_2 \leq \rho||y||_2.$$

Substituting $||(I - A)y||_2 = ||x^{(k)} - x^*||_2$ and $y = x^{(k-1)} - x^*$,

$$\boxed{||x^{(k)} - x^*||_2 \leq \rho||x^{(k-1)} - x^*||_2}.$$

e.)

Assuming the "worst" case scenario, where

$$||x^{(k')} - x^*||_2 = \rho||x^{(k'-1)} - x^*||_2,$$

we find that

$$||x^{(1)} - x^*||_2 = \rho||x^{(0)} - x^*||_2$$

$$||x^{(2)} - x^*||_2 = \rho||x^{(1)} - x^*||_2 = \rho^2||x^{(0)} - x^*||_2$$

$$...$$

$$||x^{(k)} - x^*||_2 = \rho||x^{(k-1)} - x^*||_2 = \rho^k||x^{(0)} - x^*||_2$$

We want our solution $||x^{(k')} - x^*||_2 \leq \epsilon$, so

$$||x^{(n)} - x^*||_2 = \rho^k||x^{(0)} - x^*||_2 \leq \epsilon$$

Solving for k-iterations,

$$\rho^k \leq \frac{\epsilon}{||x^{(0)} - x^*||_2}$$

$$k \log \rho \leq \frac{\epsilon}{||x^{(0)} - x^*||_2}.$$

Since $\log \rho < 0$, convergence to tolerance ϵ will occur for

$$k \geq \frac{\epsilon}{\log \rho ||x^{(0)} - x^*||_2}$$

Since we are dealing with the worst case scenario, we can be sure that our algorithm will converge in this many iterations.

f.)

The iteration of gradient descent is dominated by a matrix-vector product, where a $n \times n$ matrix is multiplied by a vector of length n. Computing this product requires n multiplications and $n-1$ additions per each of the n-rows in the matrix. This is a total of $(n+n-1)n = 2n^2 - n$ operations per iteration. For $k = \frac{\epsilon}{\log \rho ||x^{(0)} - x^*||_2}$ iterations, the overall running time is

$$t \propto \frac{\epsilon(2n^2 - 2)}{\log \rho ||x^{(0)} - x^*||_2}$$

Problem 6

The risk function is defined as:

$$R(f(x) = i|x) = \sum_{j=1}^{c} L(f(x) = i, y = j)P(Y = j|x).$$

We can try to minimize R by selecting a policy that chooses class i if $P(Y = i|x) \geq P(Y = j|x)\forall j$. Then

$$R(f(x) = i|x) = L(f(x) = i, y = i)P(Y = i|x) + \sum_{j=1, j\neq i}^{c} L(f(x) = i, y = j \neq i)P(Y = j \neq i|x)$$

Without doubt, this becomes

$$R(f(x) = i|x) = 0 + \lambda_s(1 - P(Y = i|x))$$
$$R(f(x) = i|x) = \lambda_s(1 - P(Y = i|x))$$

Since $P(Y = i|x) \geq P(Y = j|x)\ \forall j$, we can state $(1 - P(Y = i|x)) \leq (1 - P(Y = j|x))\ \forall j$.

Introducing doubt may allow us to minimize this risk function further.

Imposing the condition that we only choose doubt when $P(Y = i|x) \leq 1 - \lambda_r/\lambda_s$, we find that

$$\lambda_s(1 - P(Y = i|x)) \geq \lambda_r$$

$$R(f(x) = i|x) \geq \lambda_r.$$

Our old risk function is no longer a minimum, and so choosing doubt will minimize the risk function. Otherwise, our old risk function is, in fact, minimized.

b.)

If $\lambda_r = 0$ then the second condition of part (1) of the policy is never satisfied (except when $P(Y = i|x) = 1$) since a probability can not be greater than 1. In this case, doubt will always be chosen This makes sense intuitively because there is no longer any penalty to choosing doubt.

If $\lambda_r > \lambda_s$ then the second condition of part (1) of the policy is always satisfied, since a probability cannot be less than 0. This also makes sense intuitively, as it makes no sense to choose doubt if you will be more harshly penalized for it than a misclassification.

Problem 7

a.)

Using Gaussian Discriminant Analysis, we aim to maximize $P(X = x|Y = i)\pi_i$ using Bayes decision rule. Given our Gaussiant probability distributions, $P(x|\omega_i) \sim \mathcal{N}(\mu_i, \sigma^2)$, it is equivalent to maximize $Q_i(x) = \ln\left(\left(\sqrt{2\pi}\right)^d P(x)\pi_i\right)$ instead.

We can also express $Q_i(x)$ as

$$Q_i(x) = \ln\left(\left(\sqrt{2\pi}\right)^d P(x)\pi_i\right) = -\frac{|x-\mu_i|^2}{2\sigma_i^2} - d\ln\sigma_i + \ln\pi_i.$$

The Bayes optimal decision boundary is given by

$$Q_1(x) - Q_2(x) = -\frac{|x-\mu_1|^2}{2\sigma_1^2} - d\ln\sigma_1 + \ln\pi_1 + \frac{|x-\mu_2|^2}{2\sigma_2^2} + d\ln\sigma_2 - \ln\pi_2 = 0.$$

Since the problem is one-dimensional, $d = 1$. Also, $\sigma_1 = \sigma_2 = \sigma$. The decision boundary equation simplifies to:

$$Q_1(x) - Q_2(x) = \frac{|x-\mu_2|^2 - |x-\mu_1|^2}{2\sigma^2} + \ln\pi_1 - \ln\pi_2 = 0.$$

We are given that $\pi_1 = P(x|\omega_1) = \pi_2 = P(x|\omega_2) = \frac{1}{2}$, so we find

$$\begin{aligned} Q_1(x) - Q_2(x) = \frac{|x-\mu_2|^2 - |x-\mu_1|^2}{2\sigma^2} &= 0 \\ |x-\mu_2|^2 - |x-\mu_1|^2 &= 0 \\ |x-\mu_2|^2 &= |x-\mu_1|^2 \\ |x-\mu_2| &= |x-\mu_1| \end{aligned}$$

For this to be true, either $\mu_1 = \mu_2$, or

$$\begin{aligned} x - \mu_2 &= -x + \mu_1 \\ 2x &= \mu_1 + \mu_2 \\ x &= \frac{\mu_1 + \mu_2}{2} \end{aligned}$$

The Bayes decison rule which corresponds to this boundary is

$$r^*(x) = \begin{cases} 1 & Q_1(x) - Q_2(x) > 0 \\ 2 & \text{otherwise} \end{cases}$$

b.)

Using the given definition

$$P_\ell = P((\text{misclassified as } \omega_1)|\omega_2)P(\omega_2) + P((\text{misclassified as } \omega_2)|\omega_1)P(\omega_1)$$

We can say

$$P((\text{misclassified as } \omega_1)|\omega_2) = \int_{-\infty}^{\frac{\mu_1+\mu_2}{2}} \frac{1}{(\sqrt{2\pi}\sigma)^d} e^{\frac{-|x-\mu_2|^2}{2\sigma^2}}\, dx$$

and

$$P((\text{misclassified as } \omega_2)|\omega_1) = \int_{\frac{\mu_1+\mu_2}{2}}^{\infty} \frac{1}{(\sqrt{2\pi}\sigma)^d} e^{\frac{-|x-\mu_1|^2}{2\sigma^2}}\, dx$$

We can then use the fact that $d = 1$ to remove it from the equation.

If we use a change of variables such that $z(x) = \frac{-x+\mu_2}{\sigma}$ on $P((\text{misclassified as } \omega_1)|\omega_2)$, we find

$$P((\text{misclassified as } \omega_1)|\omega_2) = \int_{-\infty}^{z\left(\frac{\mu_1+\mu_2}{2}\right)} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(-z)^2}{2}} (-\sigma) dz$$

$$P((\text{misclassified as } \omega_1)|\omega_2) = -\int_{\infty}^{\frac{\mu_2-\mu_1}{2\sigma}} \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} dz$$

$$P((\text{misclassified as } \omega_1)|\omega_2) = \int_{\frac{\mu_2-\mu_1}{2\sigma}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} dz$$

We can use a similar change of variables on $P((\text{misclassified as } \omega_2)|\omega_1)$, where $z(x) = \frac{x-\mu_1}{\sigma}$. We find

$$P((\text{misclassified as } \omega_2)|\omega_1) = \int_{z\left(\frac{\mu_1+\mu_2}{2}\right)}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-z^2}{2}} \sigma dz$$

$$P((\text{misclassified as } \omega_2)|\omega_1) = \int_{\frac{\mu_2-\mu_1}{2\sigma}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} dz$$

Then,

$$P_\ell = \frac{1}{2}\int_{\frac{\mu_2-\mu_1}{2\sigma}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} dz + \frac{1}{2}\int_{\frac{\mu_2-\mu_1}{2\sigma}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} dz$$

$$P_\ell = \int_{\frac{\mu_2-\mu_1}{2\sigma}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} dz$$

$$P_\ell = \frac{1}{\sqrt{2\pi}} \int_{a}^{\infty} e^{\frac{-z^2}{2}} dz, \text{ where } a = \frac{\mu_2 - \mu_1}{2\sigma}$$

Problem 8

We can analyze each probability ($P(X=1),\ P(X=2),\ P(X=3)$), by treating each likelihood function as if it were only of two outcomes, $P(X=i)$ and $P(X\neq i)$.

Then, we can use the binomial distribution to say

$$\mathscr{L}(p_i) = \binom{n}{k_i} p_i^{k_i}(1-p_i)^{n-k_i}$$

$$\frac{d\mathscr{L}(p_i)}{dp_i} = \frac{n!}{k!(n-k)!}\left(k_i p_i^{k_i-1}(1-p_i)^{n-k_i} - p_i^{k_i}(n-k_i)(1-p_i)^{n-k_i-1}\right)$$

To maximize this likelihood function, we find $\frac{d\mathscr{L}}{dp_i}=0$.

$$0 = \frac{n!}{k!(n-k)!}\left(k_i p_i^{k_i-1}(1-p_i)^{n-k_i} - p_i^{k_i}(n-k_i)(1-p_i)^{n-k_i-1}\right)$$

$$\begin{aligned}
p_i^{k_i}(n-k_i)(1-p_i)^{n-k_i-1} &= k_i p_i^{k_i-1}(1-p_i)^{n-k_i} \\
p_i(n-k_i) &= k_i(1-p_i) \\
np_i - k_i p_i &= k_i - k_i p_i \\
np_i &= k_i \\
p_i &= \frac{k_i}{n}
\end{aligned}$$

We find then, that $p_1 = k_1/n,\ \ p_2 = k_2/n,\ \ \text{and}\ p_3 = k_3/n$.