1a) i. $(.4)(.3) = .12$

ii. $P(T) = (.4)(.3) + (.7)(.7) = .61$

iii. $P(A_1)P(A_2^c) + P(A_1^c)P(A_2) = (.61)(.39) + (.61)(.39) = .4758$

iv. $P(W^c \mid T^c) = \frac{(.3)(.7)}{.39} = .538.$

1b) $P(A \mid B, C) > P(A \mid B)$

$P(A, B, C) / P(B, C) > P(A \mid B)$

$P(C \mid A, B) P(A, B) / P(B, C) > P(A, B) / P(B)$

$P(C \mid A, B) > P(B, C) / P(B)$

$1 - P(C^c \mid A, B) > P(B, C) / P(B)$

$-P(C^c \mid A, B) > P(B, C) / P(B) - 1$

$P(C^c \mid A, B) < 1 - \frac{P(B, C)}{P(B)}$

$\frac{P(C^c, A, B)}{P(A, B)} < 1 - \frac{P(B, C)}{P(B)}$

$\frac{P(A \mid B, C^c) P(B, C^c)}{P(A, B)} < 1 - \frac{P(B, C)}{P(B)}$

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

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

$P(A \mid B, C^c) < P(A, B) / P(B)$

$P(A \mid B, C^c) < P(A \mid B)$

2a. ii ⇒ i We have $B^TAB \geq 0$, and can transform it s.t:

$x^TB^TABx \geq 0$ } By definition of positive semidefinite.

let $y = Bx$, and $y^T = (Bx)^T = x^TB^T$

From this we can reformulate our equation as:

$$yAy^T \geq 0$$

which follows the form of (i). Therefore we have shown that ii ⇒ i

2a) **i. ⟺ ii** We can take the transpose of (ii) such that:

$(B^TAB)^T = (AB)^T(B)^T = B^TAB \longrightarrow (B^TAB)^T = B^TAB.$

Because (i) tells us that A is symmetric and positive semi-definite, we can say that because B^TAB is symmetric for some invertible $B \in \mathbb{R}^{n \times n}$, it must also be positive semi-definite.

iii ⟺ iv. $A = PDP^T = P\sqrt{D}\sqrt{D}P^T$ (valid b/c of iii's implications)

Let $U = P\sqrt{D}$ and $U^T = (P\sqrt{D})^T = \sqrt{D}^T P^T = \sqrt{D}P^T$ (valid bc D is diagonal)

$\therefore A = UU^T$

i ⟺ iii The definition of an eigenvalue and eigenvector is as follows:

$$A\vec{v} = \lambda\vec{v}$$

The definition of SPD is: $\vec{x}^TA\vec{x} \geq 0 \quad \forall x \in \mathbb{R}^n$. If $\vec{x}$ is then an eigenvector of A, it follows that:

$$\vec{x}^T\vec{x}\lambda \geq 0$$

Because $\vec{x}^Tx$ is positive, in order for $\vec{x}^TA\vec{x}$ to be ≥ 0, we know that λ must be ≥ 0. (We know x^Tx is positive because it is the sum of squares)

iv. ⟺ i Because is symmetric $\in \mathbb{R}^{n \times n}$, it holds that:

$A = UU^T \Rightarrow x^TAx = x^TUU^Tx \Rightarrow (U^Tx)^T(U^Tx) = \|U^Tx\|_2^2 \geq 0$

Therefore UU^T is positive semidefinite.

2b) (i) We know that $x^TAx \geq 0$ by the definition of symm. positive definite, $x^Tx > 0$ because it is composed of the sum of squared, and $\lambda > 0$ as given in the problem. From this we can also ascertain that $A > 0$.

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

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

$$x^T(A + \lambda I)x > 0 \quad \text{holds b/c we know } \lambda \text{ is positive}$$

$$\therefore A + \lambda I > 0$$

(ii) $x^T A x - x^T \gamma x > 0$

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

$A - \gamma > 0$

$A > \gamma$ or $A > \gamma I$ ∴

Therefore for $A - \gamma I > 0$, $A > \gamma$ where $\gamma \neq 0$.

(iii) Say there is a vector e_i that is all zeros except for a 1 in the i^{th} position. By the definition of A being SPD, it holds that $e_i^T A e_i > 0$ for all i. If the i^{th} entry of A wasn't positive, then $e_i^T A e_i = 0 + \cdots 1 \cdot a_{ii} + \cdots 0$, which means that $e_i^T A e_i$ would be ≤ 0, a direct violation of the definition of SPD.

Therefore in order for $A > 0 \in \mathbb{R}^{n \times n}$, $A_{ii} > 0$ for $i = 1, \ldots, n$

(iv) Say our vector x is $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$. By the definition of SPD, we know that $x^T A x > 0$. With our definition of X, we then have:

$$\begin{bmatrix} 1 & 1 \end{bmatrix} A \begin{bmatrix} 1 \\ 1 \end{bmatrix} > 0$$

Because x is composed of 1's, the sum of A_{ij} for all $i, j = 1, \ldots, n$ must be > 0. We see this in the following example:

$$\begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} a_1 & a_3 \\ a_2 & a_4 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} a_1 + a_2 & a_3 + a_4 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} > 0$$

$$\begin{bmatrix} a_1 + a_2 \\ a_3 + a_4 \end{bmatrix} > 0$$, where a_{1-4} are component of $A \in \mathbb{R}^{2 \times 2}$

In the 2x2 matrix where $i = 2$ and $j = 2$ we see that $\sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} > 0$, therefore this expands to the greater case where $i = n$ and $j = n$, such that $\sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} > 0$.

3.

a) $f(x) = a^T x$

$\nabla_x f(x) = a$

b) $f(x) = x^T A x = \sum_{i=1}^{n} \sum_{j=1}^{n} x_i a_{ij} x_j$

$= \sum_{i=1}^{n} a_{ii} x_i^2 + \sum_{i \neq j}^{n} \sum^{n} a_{ij} x_i x_{ji}$

$\frac{\partial f(x)}{x_k} = 2 \sum_{i=1}^{n} a_{ki} x_i$ } where x_k for $k \in \{1, 2 \ldots, n\}$

Therefore $\nabla f(x) = 2Ax$

c) $tr(A^T X) = \sum_{j=1}^{n} \sum_{i=1}^{n} a_{ij} x_{ij}$

$\frac{\partial (\sum_{j} \sum a_{ij} x_{ij})}{\partial x_{ij}} = a_{ij} \Rightarrow \nabla_x tr(A^T X) = A$

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

Counterexample: $x = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$ $y = \begin{bmatrix} 4 \\ 7 \end{bmatrix}$

e) $\|x\|_\infty \leq \|x\|_2 \leq \sqrt{n}\,\|x\|_\infty$

$$\|x\|_\infty = \max_{1\leq i\leq n} |x_i| \leq \sqrt{\max_{1\leq i\leq n}(x_i)^2 + \sum_{j=1, j\neq i}^{n} x_i^2}$$

$$= \sqrt{\sum_{i=1}^{n} x_i^2}$$

$$= \|x\|_2$$

$\therefore \|x\|_\infty \leq \|x\|_2$

$$\|x\|_2 = \sqrt{\sum_{j=1}^{n} x_j^2} \leq \sqrt{n \cdot \max_{1\leq i\leq n}(x_i^2)}$$

valid because x_i is greatest from $x_i \to n$, therefore $n \cdot x_i \geq x_1 + \cdots + x_j$

$$= \sqrt{n}\,\|x\|_\infty$$

$\therefore \|x\|_2 \leq \sqrt{n}\,\|x\|_\infty$

Therefore we have shown that $\|x\|_\infty \leq \|x\|_2 \leq \sqrt{n}\,\|x\|_\infty$

f) $\|x\|_2 = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2} \leq \sqrt{x_1^2 + \cdots + x_n^2 +}$

$$= \sqrt{(|x_1| + |x_2| + \cdots |x_n|)^2}$$

$$= |x_1| + |x_2| + \cdots |x_n|$$

$$= \|x\|_1$$

$\therefore \|x\|_2 \leq \|x\|_1$

$\|x\|_1 = |x_1| + \cdots + |x_n|$ ← let this be vector $\underline{x}$

$= \underline{x}^T * 1$

$\|x\|_1 \leq \|\underline{x}\|_2 \|1\|_2$ ← By Cauchy-Schwarz, where $\|\underline{x}\|_2 = \|x\|_2$ and $\|1\|_2 = \sqrt{n}$.

$\|x\|_1 \leq \sqrt{n}\,\|x\|_2$

$\therefore \|x\|_1 \leq \sqrt{n}\,\|x\|_2$

Therefore we have shown that $\|x\|_2 \leq \|x\|_1 \leq \sqrt{n}\,\|x\|_2$

4

a) Because A is symmetric, we can use the Spectral Thm. such that:

$A = UDU^T \Longrightarrow A = U\lambda U^T$] where U and U^T are orthogonal

We can apply this to our orginal question:

$$\lambda_{max}(A) = \max_{\|x\|_2=1} x^T U \lambda U^T x \quad \text{] let } Y = x^T U,\ Y^T = U^T x$$

$$= \max_{\|x\|_2=1} Y \lambda Y^T$$

$$= \lambda_{max} Y_1^2 + \lambda Y_2^2 + \cdots + \lambda_{min} Y_n^2 \leq \lambda_{max}(Y_1^2 + \cdots + Y_N^2)$$

$$= \max_{\|x\|_2=1} Y \lambda_{max} Y^T$$

From here we know that $YY^T = 1$ because $\|x\|_2 = 1$, and if we reduce A to the diagonal case we get the following:

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

b) Using spectral theorem in our orginal equation, we get:

$$\lambda_{min}(A) = \min_{\|x\|_2=1} x^T U \lambda U^T x \quad \text{] let } Y = x^T U \quad Y^T = U^T x$$

$$= \min_{\|x\|_2=1} Y \lambda Y^T$$

$$= \lambda_{min} Y_1^2 + \cdots + \lambda_{max} Y_n^2 \geq \lambda_{min}(Y_1^2 + \cdots + Y_n^2)$$

$$= \min_{\|x\|_2=1} Y \lambda_{min} Y^T$$

From here we know that $YY^T = 1$, because $\|x\|_2 = 1$, and if we reduce A to the diagonal case we get:

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

c) Neither are convex programs. Since A is positive semi-definite, the objective fn is convex, proving that the maximization problem is not convex. The minimization is also not convex because there exists no linear combination of vectors in the feasible set that satisfies $\|x\|_2 = 1$.

d) $Ax = \lambda x \Rightarrow A^2x = A\lambda x \Rightarrow A^2x = \lambda^2 x$

From this we see that λ is an eigenvalue of A, and λ^2 is an eigenvalue of A^2. From this we can use similar logic in part b to deduce that:

$$\lambda_{max}(A^2) = \max_{\|x\|_2=1} x^T A^2 x = \left(\max_{\|x\|_2=1} x^T A x\right)^2 = \lambda_{max}(A)^2$$] Because all λ are > 0

We can make a similar argument to prove that:

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

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

$$\sqrt{(\lambda_{min}(A)^T Ax} \le \sqrt{(Ax)^T Ax} \le \sqrt{(\lambda_{max}(A))^T Ax}$$

$$\sqrt{\lambda_{min}(A)^T Ax} \le \|Ax\|_2 \le \sqrt{\lambda_{max}(A)^T Ax}$$

$$\sqrt{\lambda^2_{min}(A)} \le \|Ax\|_2 \le \sqrt{\lambda^2_{max}(A)}$$

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

f) $$\lambda_{min}(A)\|x\|_2 \le \|x\|_2 \sqrt{A^2(x/\|x\|_2 \cdot x/\|x\|_2} \le \lambda_{max}(A)\|x\|_2$$

$$\lambda_{min}(A) \le \sqrt{A^2(x/\|x\|_2 \cdot x/\|x\|_2} \le \lambda_{max}(A)$$] $\|x\|_2 = 1$

$$\lambda_{min}(A) \le \sqrt{(Ax)^T Ax} \le \lambda_{max}(A)$$

Because $\|x\|_2$ is the unit vector, we simplify our orginal equation to the one that we proved in 4e.

5

a) Because A is symmetric and positive semi-definite, we can also infer that A is convex. This tells us that A must have a global minimum. We can find this minimum by taking the derivative of the optimization problem and setting it equal to 0.

$f(x) = x^T Ax - b^T x$

$\nabla_x f(x) = 2Ax - b$ If we sub this back into the original problem

$\frac{1}{2}(2Ax) - b = 0 \Rightarrow Ax - b = 0 \Rightarrow Ax = b \Rightarrow x^* = b/A$

Therefore we know that $x^* = b/A$

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

$= x - (\frac{1}{2}(2Ax) - b)$

$= x - Ax + b$.

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

$x^k - x^* = x^{k-1} - Ax^{k-1} + b - b/a$] sub out b w/ Ax^*

$= x^{k-1} - Ax^{k-1} + Ax^* - b/a$

$= (I - A)(x^{k-1} - x^*)$ ∴

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

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

$\|(I-A)\|_2 \|x^{k-1} - x^*\|_2 \leq \rho \|x^{k-1} - x^*\|_2$] holds by matrix-norm rules

$\|I - A\|_2 \leq \rho$

Because the $\lambda_{min}(A) > 0$ and $\lambda_{max}(A) < 1$, we know that $\|I - A\|_2$ will also be between 0 and 1. From this we can deduce that $\exists \rho$ where $0 < \rho < 1$ s.t.

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

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

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

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

f) $n^2 \log_\rho {\epsilon}/{\|x^{(0)} - x^*\|_2}$

This is because the matrix - vector multiplication of Ax takes up n^2 time.

6.

a) Suppose $\exists i, j$ s.t $P(Y=i|x) \geq P(Y=j|x)$. From this we can tell that $P(Y=j|X) \leq 1/c$ given $j \neq i$ and $i \neq 1+c$. We can then apply this to the given risk fn s.t.

$R \leq (c-1)(1/c)\lambda_s \Rightarrow R \leq (1-1/c)\lambda_s$

On the other hand, $\forall j$ where $P(Y=i|x) < P(Y=j|x)$, $P(Y=j|x) > 1/c$. We can then apply this to the given risk fn s.t.:

$R > (c-1)(1/c)\lambda_s \Rightarrow R > (1-1/c)\lambda_s$

From this we can see that the policy outlined, where we select class i if $P(Y=i|x) \geq P(Y=j|x)$, is indeed the minimum, as R increases otherwise. For the second part of the policy we can apply the following transformation:

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

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

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

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

This transformation shows that we face an equal or smaller amount of risk by guessing on a classification rather than choosing doubt.

b) If $\lambda_r = 0$, you are better off choosing doubt rather than misclassifying, as reflected by the loss function. If $\lambda_r > \lambda_s$, you are better off misclassifying a data point over choosing doubt - this is, once again, reflected by the risk and loss functions. We can see this reflected in the policy equations as well. For $\lambda_r = 0$, the policy becomes:

$$P(Y = i \mid x) \geq 1 - 0/\lambda_s \Rightarrow P(Y = i \mid x) \geq 1$$

and $\lambda_r > \lambda_s$ results in:

$P(Y = i \mid x) \geq b$, where b is some negative number, because λ_r/λ_s will always be > 1, so $1 - \lambda_r/\lambda_s$ is necessarily negative.

7.

a) $P(\omega_1 | x) = P(\omega_2 | x)$

$P(x|\omega_1) P(\omega_1) = P(x|\omega_2) P(\omega_2)$

$N(\mu_1, \sigma^2) = N(\mu_2, \sigma^2) \rightarrow (x-\mu_1^2) = (x-\mu_2^2)$

$x = \frac{\mu_1 + \mu_2}{2}$

Therefore, if $x < \frac{\mu_1 + \mu_2}{2}$ select ω_1,

else select ω_2.

b) $P_e = 1/2\, P\left(\frac{x-\mu_2}{\sigma^2} \leq \frac{\mu_1-\mu_2}{\sigma^2} \mid x = \omega_2\right) + 1/2\, P\left(\frac{x-\mu_1}{\sigma^2} \geq \frac{\mu_1-\mu_1}{\sigma^2} \mid x = \omega_1\right)$

$= P\left(\frac{x-\mu_1}{\sigma^2} \geq \frac{\mu_2-\mu_1}{\sigma^2} \mid x = \omega_1\right)$

From here we can use the PDF to show the association such that:

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

8. Using the method of maximum likelihood, we can estimate p_1, p_2, and p_3 as follows:

$$L_p(k) = \Pi p_x^{k_x}$$

We use this equation because k_x is multinomially distributed. Because we only have n observations of X, our $L(p)$ has the constraint $\sum_x p_x = 1$, where $x \in \{1, 2, 3\}$. To maximize L, we take the gradient of L and our constraint, C and show that they are colinear such that:

$$\partial/\partial p_x L(p) = \lambda \partial/\partial p_x C(p) \;] \; \exists \lambda \; \forall x$$

After taking the derivative we get:

$$k_x/p_x L(p) = \lambda$$

From this we see that p_x and k_x are proportional. The original problem states that $p_1 + p_2 + p_3 = 1$ (also stated in $C(p)$). With this, we know that $\hat{p}_x = k_x/(k_1 + k_2 + k_3)$. For each p we then have:

$$\hat{p}_1 = k_1/(k_1 + k_2 + k_3)$$

$$\hat{p}_2 = k_2/(k_1 + k_2 + k_3)$$

$$\hat{p}_3 = k_3/(k_1 + k_2 + k_3)$$

Collaborators: Roshni Patel, Sydney McMuldroch, Lynn Kong

Problem 8 Citation: Math Stack Exchange # 421105

1a) i. $(.4)(.3) = .12$

ii. $P(T) = (.4)(.3) + (.7)(.7) = .61$

iii. $P(A_1)P(A_2^c) + P(A_1^c)P(A_2) = (.61)(.39) + (.61)(.39) = .4758$

iv. $P(W^c|T^c) = \frac{(.3)(.7)}{.39} = .538$

1b) $P(A|B,C) > P(A|B)$

$P(A,B,C)/P(B,C) > P(A|B)$

$P(C|A,B)P(A,B)/P(B,C) > P(A,B)/P(B)$

$P(C|A,B) > P(B,C)/P(B)$

$1 - P(C^c|A,B) > P(B,C)/P(B)$

$-P(C^c|A,B) > P(B,C)/P(B) - 1$

$P(C^c|A,B) < 1 - \frac{P(B,C)}{P(B)}$

$\frac{P(C^c,A,B)}{P(A,B)} < 1 - \frac{P(B,C)}{P(B)}$

$\frac{P(A|B,C^c)P(B,C^c)}{P(A,B)} < 1 - \frac{P(B,C)}{P(B)}$

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

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

$P(A|B,C^c) < P(A,B)/P(B)$

$P(A|B,C^c) < P(A|B)$