ECE 247 HW # 1 - Ashwin Ranade

① a) i) Let $A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, then the characteristic polynomial is

$\det(A - \lambda I) = \det\begin{pmatrix} -\lambda & 1 \\ 1 & -\lambda \end{pmatrix} = \lambda^2 - 1$. Note $AA^T = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.

Setting the characteristic polynomial to 0, we get $\lambda^2 - 1 = 0 \rightarrow \lambda = \pm 1$.

$\boxed{\text{Hence, the 2 eigenvalues are } \pm 1}$

Let $\vec{v}$ and $\vec{w}$ be the eigenvectors corresponding to $1, -1$ respectively.
Then,

$$(A - (1)I_2)\vec{v} = 0$$

$$\left(\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} - \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)\vec{v} = 0$$

$$\begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}\vec{v} = 0 \rightarrow \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \rightarrow -v_1 + v_2 = 0 \rightarrow v_2 = v_1$$

Then, $\vec{v} = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} v_1 \\ v_1 \end{pmatrix} = v_1\begin{pmatrix} 1 \\ 1 \end{pmatrix} = k\begin{pmatrix} 1 \\ 1 \end{pmatrix}, k \in \mathbb{R}$.

Similarly, for $\lambda = -1$,

$$(A - (-1)I_2)\vec{w} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$w_1 + w_2 = 0 \rightarrow w_1 = -w_2$$

$$\vec{w} = \begin{pmatrix} w_1 \\ -w_1 \end{pmatrix} = w_1\begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

$\boxed{\text{Hence, } \vec{v} = k_1\begin{pmatrix} 1 \\ 1 \end{pmatrix}, k_1 \in \mathbb{R}, \text{ and } \vec{w} = k_2\begin{pmatrix} 1 \\ -1 \end{pmatrix}, k_2 \in \mathbb{R}}$

$\boxed{\text{We notice that each eigenvalue has norm 1, and the eigenvectors are orthogonal}}$

ii) Since $AB = I \rightarrow BA = I$, $AA^T = I \rightarrow A^TA = I$
Let $\vec{v}$ be an eigenvector with eigenvalue λ such that $\lambda\vec{v} = A\vec{v}$. Then,

$$\vec{v}^T A^T A\vec{v} = (A\vec{v})^T(A\vec{v}) = (\lambda\vec{v})^T(\lambda\vec{v}) = (\lambda\vec{v})\cdot(\lambda\vec{v})$$
$$= \lambda^2\, \vec{v}\cdot\vec{v} = \lambda^2|\vec{v}|^2.$$

Also, $\vec{v}^T A^T A\vec{v} = \vec{v}^T I\vec{v} = \vec{v}\cdot\vec{v} = |\vec{v}|^2$ since $A^TA = I$.

Hence, $\lambda^2|\vec{v}|^2 = |\vec{v}|^2$, or $\lambda^2 = 1 \rightarrow \lambda = \pm 1$.

$\boxed{\text{Hence all eigenvalues must be } \pm 1\text{, so all eigenvalues must have norm 1.}}$

iii) From Piazza, we only need to focus on real eigenvalues, so we can ignore the adjoint case.

From ii), we know we can only have eigenvalues ± 1.

Let $\vec{x}$ be an eigenvector with eigenvalue λ_1, and $\vec{y}$ be an eigenvector with λ_2. Then $A\vec{x} = \lambda_1 \vec{x}$, $A\vec{y} = \lambda_2 \vec{y}$, with $\lambda_1 \neq \lambda_2$.

Then, $\vec{x} \cdot \vec{y} = \vec{x}^T \vec{y} = \vec{x}^T I \vec{y} = \vec{x}^T A^T A \vec{y} = (A\vec{x})^T (A\vec{y})$
$= (\lambda_1 \vec{x})^T (\lambda_2 \vec{y}) = \lambda_1 \lambda_2 (\vec{x} \cdot \vec{y})$.

Hence either $\lambda_1 \lambda_2 = 1$ or $\vec{x} \cdot \vec{y} = 0$. Since $\lambda_1 \cdot \lambda_2 = 1 \cdot -1 = -1 \neq 1$, $\vec{x} \cdot \vec{y} = 0$, and hence $\vec{x}$ and $\vec{y}$ are orthogonal, as needed.

iv) Since the eigenvalues are ± 1, the transformation does not scale the input in any way. Since length and angles must be preserved,
$A\vec{x}$ will simply be $\vec{x}$ after a rotation (around the origin), or a flip / reflection across some subspace.

If $\det(A) = 1$, we will not have a reflection, while if $\det(A) = -1$, we will (note we can still rotate before / after a rotation).

b) We utilize the Spectral Thereom, which tells us that if A is $n \times n$ and symmetric, $A = PDP^{-1} = PDP^T$, where P's columns are an orthonormal eigenbasis of $\mathbb{R}^n$, and D is a diagonal matrix with entries $\lambda_1, \dots, \lambda_n$.

Let $A = U \Sigma V^T$, where $A \in \mathbb{R}^{m \times n}$. Then,

$$AA^T = (U \Sigma V^T)(U \Sigma V^T)^T = U \Sigma V^T V \Sigma^T U^T.$$

Since V is orthogonal, $V^T V = I$, so

$AA^T = U \Sigma \Sigma^T U^T$. Since $(AA^T)^T = A^{T^T} A^T = AA^T$, AA^T is symmetric, so we can use spectral decomposition to conclude $\boxed{\text{the left singular vectors of } A \text{ are the eigenvectors of } AA^T.}$

Similarly, since $A^T A = (U \Sigma V^T)^T (U \Sigma V^T) = V \Sigma^T U^T U \Sigma V^T$
$= V \Sigma^T \Sigma V^T$,
$\boxed{\text{the right singular vectors of } A \text{ are the eigenvectors of } A^T A.}$

ii) From i), we know $AA^T = U \Sigma \Sigma^T U^T$, so by the spectral thereom $\Sigma \Sigma^T$ is the eigenvalue matrix for AA^T.

If A is $m \times n$, then Σ is also $m \times n$, so $\Sigma \Sigma^T$ is $m \times m$. However, A only has $\min(m,n)$ singular values.

Hence, the first $\min(m,n)$ eigenvalues of AA^T correspond to the singular values of A, and any remaining eigenvalues (any remaining diagonal entries in $\Sigma \Sigma^T$) will be 0.

The eigenvalues of $A^T A$ are the same, except we are filling in $\Sigma^T \Sigma$ instead (so we might not need to pad with 0s depending on the dimension of $\Sigma^T \Sigma$).

c) i) False
ii) False
iii) True
iv) True
v) True

②

a) i) Using the probability chain rule, we get

$$P(H50 \mid Tails) = \frac{P(H50 \wedge Tails)}{P(Tails)} = \frac{0.5 \cdot 0.5}{P(Tails \mid H50)\, P(H50) + P(Tails \mid H60)\, P(H60)}$$

$$= \frac{0.25}{0.4 \cdot 0.5 + 0.5 \cdot 0.5} \boxed{= \frac{5}{9}}$$

ii) We want to find $P(H50 \mid THHH)$. Using Bayes' Thm, we can rewrite as

$$P(H50 \mid THHH) = \frac{P(THHH \mid H50)\, P(H50)}{P(THHH)}$$

Then using the law of probability we can rewrite $P(THHH)$ as

$$P(THHH) = P(THHH \mid H50)\, P(H50) + P(THHH \mid H60)\, P(H60)$$
$$= 0.5^4 \; 0.5 + 0.4 \cdot 0.6^3 \cdot 0.5$$

Hence, $P(H50 \mid THHH) = \dfrac{0.5^5}{0.5^5 + 0.5 \cdot 0.4 \cdot 0.6^3} \approx \boxed{42\%}$

iii) We first need to find $P(9 \text{ heads}) = P(H50)\, P(9 \text{ heads} \mid H50) + P(H55)\, P(9 \text{ heads} \mid H55) + P(60)\, P(9 \text{ heads} \mid H60)$.

We can do this with the binomial distribution. We use

$$\binom{n}{k} p^k (1-p)^{n-k} \quad \text{with } n = 10,\ k = 9, \text{ and } p = 0.5, 0.55, \text{ and } 0.6.$$

Then, $P(9 \text{ heads} \mid H50) = \binom{10}{9} 0.5^9 \; 0.5^1 = 10 \cdot 0.5^{10} = \frac{5}{512}$,

$P(9 \text{ heads} \mid H55) = \binom{10}{9} 0.55^9 \; 0.45^1 \approx 0.020724$, and

$P(9 \text{ heads} \mid H60) = \binom{10}{9} 0.6^9 \; 0.4^1 \approx 0.04031$

Hence, $P(9 \text{ heads}) \approx 0.0236$.

Then, $P(H50 \mid 9 \text{ heads}) = \dfrac{P(9 \text{ heads} \mid H50)\, P(H50)}{P(9 \text{ heads})} = \dfrac{5/512 \cdot \frac{1}{3}}{0.0236} \boxed{\approx 14\%}$

$P(H55 \mid 9 \text{ heads}) = \dfrac{0.020724 \cdot 1/3}{0.0236} \boxed{\approx 29\%}$

$P(H60 \mid 9 \text{ heads}) = \dfrac{0.04031 \cdot 1/3}{0.0236} \boxed{\approx 57\%}$

b) We once again use Bayes' Thm.

$$P(\text{pregnant} \mid \text{positive}) = \frac{P(\text{positive} \mid \text{pregnant})\, P(\text{pregnant})}{P(\text{positive})}$$, and then

$$P(\text{positive}) = P(\text{positive} \mid \text{pregnant})\, P(\text{pregnant}) + P(\text{positive} \mid \text{not pregnant})\, P(\text{not pregnant})$$

$$P(\text{positive}) = 0.99\,(0.01) + 0.1\,(0.99) = 0.1089$$

Hence $$P(\text{pregnant} \mid \text{positive}) = \frac{0.99\,(0.01)}{0.1089} \approx \boxed{9.1\%}$$

The number is low due to the high rate of false positives; since only 1% of the population is pregnant, most positive tests will come from false positives instead of true positives.

c) Since expected value is a linear operator, we have
$\mathbb{E}(A\vec{x} + \vec{b}) = \mathbb{E}(A\vec{x}) + \mathbb{E}(\vec{b})$.
Then since $\vec{b}$ is deterministic (not random), $\mathbb{E}(\vec{b}) = \vec{b}$.

Finally, since A is deterministic, we have
$$\mathbb{E}(A\vec{x}) = \mathbb{E}\left[\begin{pmatrix} a_{11} & & a_{1n} \\ & \ddots & \\ a_{n1} & & a_{nn} \end{pmatrix}\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}\right] = \begin{pmatrix} a_{11} & & a_{1n} \\ & \ddots & \\ a_{n1} & & a_{nn} \end{pmatrix} \mathbb{E}\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$
$$= A\,\mathbb{E}(\vec{x})$$

Hence $$\boxed{\mathbb{E}(A\vec{x} + \vec{b}) = A\,\mathbb{E}(\vec{x}) + \vec{b}}$$

d) We get $$\operatorname{cov}(A\vec{x} + \vec{b}) = \mathbb{E}\left((A\vec{x} + \vec{b} - \mathbb{E}(A\vec{x} + \vec{b})) * (A\vec{x} + \vec{b} - \mathbb{E}(A\vec{x} + \vec{b}))^T \right)$$
$$= \mathbb{E}\left((A\vec{x} + \vec{b} - A\mathbb{E}(\vec{x}) - \vec{b}) * (A\vec{x} + \vec{b} - A\mathbb{E}(\vec{x}) - \vec{b})^T \right)$$
$$= \mathbb{E}\left((A\vec{x} - A\mathbb{E}(\vec{x}))(A\vec{x} - A\mathbb{E}(\vec{x}))^T \right)$$
$$= \mathbb{E}\left(A(\vec{x} - \mathbb{E}(\vec{x}))(\vec{x} - \mathbb{E}(\vec{x}))^T A^T \right)$$
$$= A\,\mathbb{E}\left((\vec{x} - \mathbb{E}(\vec{x}))(\vec{x} - \mathbb{E}(\vec{x}))^T\right) A^T$$
$$\boxed{= A \operatorname{cov}(\vec{x}) A^T}$$

we are able to cancel $\vec{b}$ and factor A since both are deterministic.

(3) a) $\nabla_{\vec{x}} \vec{x}^T A\vec{y} = \nabla_{\vec{x}} \vec{x} \cdot A\vec{y}$. Let $\vec{d} = A\vec{y}$, then

$$\nabla_{\vec{x}} (\vec{x} \cdot \vec{d}) = \begin{pmatrix} \partial/\partial x_1 (\vec{x} \cdot \vec{d}) \\ \partial/\partial x_2 (\vec{x} \cdot \vec{d}) \\ \partial/\partial x_n (\vec{x} \cdot \vec{d}) \end{pmatrix} = \begin{pmatrix} \vec{d}_1 \\ \vdots \\ \vec{d}_n \end{pmatrix} = \vec{d} \boxed{= A\vec{y}}$$

b) $\nabla_y x^T Ay = \nabla_y (A^T x)^T y$. Let $A^T x = d$, then

$$\nabla_y x^T A y = \nabla_y d^T y = \nabla_y (d_1 y_1 + \dots + d_m y_m) = d = \boxed{A^T x}$$

c) $\vec{x}^T A y = \sum_i^n \sum_j^m a_{ij} x_i y_j$ Then,

$$\nabla_A (\vec{x}^T A y) = \begin{bmatrix} \frac{\partial f}{\partial a_{11}} x_1 y_1 & & \\ & \ddots & \\ & & \frac{\partial f}{\partial a_{nm}} x_n y_m \end{bmatrix} = \begin{bmatrix} x_1 y_1 & & x_1 y_m \\ & \ddots & \\ x_n y_1 & & x_n y_m \end{bmatrix}$$

$$\boxed{= \vec{x}\,\vec{y}^T}$$

d) $\nabla_x f = \nabla_x (\vec{x}^T A \vec{x}) + \nabla_x(\vec{b}^T \vec{x})$.

From class, we know the gradient of the first term is $A\vec{x} + A^T\vec{x}$.

From (a), we know $\nabla_x (\vec{b}^T \vec{x}) = \vec{b}$. Since gradient is a linear operator,

$$\boxed{\nabla_x f = A\vec{x} + A^T \vec{x} + \vec{b}}$$

e) $tr(AB) = \sum_i \sum_j (ab)_{ij} = \sum_i \sum_j b_{ji}\, a_{ij}$

Then $\frac{\partial f}{\partial a_{11}} = \frac{\partial}{\partial a_{11}} \left(a_{11} b_{11} + a_{12} b_{21} + \dots + a_{nm} b_{mn} \right) = b_{11}$,

$\frac{\partial f}{\partial a_{12}} = \frac{\partial}{a_{12}} \left(a_{11} b_{11} + a_{12} b_{21} + \dots + a_{nm} b_{mn} \right) = b_{21}$, and

$$\nabla_A f = \begin{bmatrix} b_{11} & b_{21} & b_{31} & \dots & b_{m1} \\ b_{12} & & & & \\ \vdots & & & & \\ b_{1m} & & & & b_{mn} \end{bmatrix}$$

these are swapped in B

As we can see, the Matrix is flipped from B.

Hence, $\boxed{\nabla_A f = B^T}$

(4) Since $\|\vec{x}\|^2 = tr(\vec{x}^T x)$,

$$f = \frac{1}{2} \sum_{i=1}^{n} \| y^i - W x^i \|^2 = \frac{1}{2} \sum_{i=1}^{n} tr\left((y^i - Wx^i)^T (y^i - Wx^i) \right).$$

Then, using FOIL / distributing,

$$f = \frac{1}{2} \sum_{i=1}^{n} tr\left(y^{iT} y^i - y^{iT} W x^i - (Wx^i)^T y^i + (Wx^i)^T (Wx^i) \right).$$

Then, since $tr(A+B) = tr(A) + tr(B)$, and $tr(AB) = tr(BA)$,

$$f = \frac{1}{2} \sum_{i=1}^{n} \left(tr(y^{iT} y^i) - tr(y^{iT} W x^i) - tr((Wx^i)^T y^i) + tr((Wx^i)^T (Wx^i)) \right).$$

Now we apply $\partial / \partial W$, noting $\frac{\partial}{\partial X} tr(AXB) = A^T B^T$ and $\frac{\partial}{\partial X} tr(AX^T B) = BA$.

Then $\frac{\partial f}{\partial W} = \frac{1}{2} \sum_{i=1}^{n} \left(0 - y^{iTT} x^{iT} - y^i x^{iT} + \frac{\partial}{\partial W} tr\left((Wx^i)^T (Wx^i) \right) \right).$

Using the hint, we get $\frac{\partial}{\partial W} tr\left((Wx^i)^T (Wx^i) \right) = \frac{\partial}{\partial W} tr\left((Wx^i)(Wx^i)^T \right)$

$= \frac{\partial}{\partial W} tr\left(W x^i x^{iT} W \right) = W (x^i x^{iT})^T + W (x^i x^{iT})$

Putting it all together, we get

$$D = -2 y^i x^{iT} + W (x^i x^{iT})^T + W (x^i x^{iT})$$

$$D = -2 y^i x^{iT} + 2W x^i x^{iT}, \text{ where}$$

$$\frac{\partial f}{\partial W} = \frac{1}{2} \sum_{i=1}^{n} D = 0$$

Now we solve for W. We have

$$\frac{1}{2} (-2 y^i x^{iT} + 2W x^i x^{iT}) = 0$$

$$W x^i x^{iT} = y^i x^{iT}$$

$$W = y^i x^{iT} (x^i x^{iT})^{-1}$$ for $i \in [1, n]$, since $(x^i x^{iT})$ is symmetric so is invertible.

Since this holds for all $i \in [1, n]$, we can collapse the x^i and y^i to get:

$$\boxed{W = y x^T (x x^T)^{-1}}$$