1. Linear algebra refresher.

(a) i.

Because Q is orthogonal, $Q^T = Q^{-1}$. Also, $(Q^T)^T Q^T = QQ^T = I$ (since Q is orthogonal), $Q^T (Q^T)^T = Q^T Q = I$ (since Q is orthogonal)

So Q^{-1} and Q^T are also orthogonal

ii.

assume eigenvector is V

~~eigenvalue is λ~~

Then we have $Av = \lambda v$

Then $\|Av\|^2 = \|\lambda v\|^2 = |\lambda|^2 \|v\|^2$

$$\|Av\|^2 = \overline{(Av)}^T (Av) \text{ by definition of the length}$$

$= \bar{v}^T A^T A v$ because A is real

$= \bar{v}^T v$ because $A^T A = I$ as A is orthogonal

$= \|v\|^2$ by definition of the length

$\|v\|^2 = |\lambda|^2 \|v\|^2$

Since V is an eigenvector, it is non-zero, and hence $\|v\| \neq 0$

Canceling $\|v\|$, we have $|\lambda|^2 = 1$. Since the length is non-negative, we get $|\lambda| = 1$

iii.

Since Q is orthogonal, $QQ^T = I = Q^T Q$ by definition

Using the fact that $\det(AB) = \det(A)\det(B)$, we have

$\det(I) = 1 = \det(QQ^T) = \det(Q)\det(Q^T) = \det(Q)\det(Q) = [\det Q]^2$

Since we have $[\det(Q)]^2 = 1$, then $\det(Q) = \pm\sqrt{1} = \pm 1$

iv.

Say that Q is orthogonal. Take arbitrary $\vec{x} \in R^n$. We want to show $\|\vec{x}\| = \|Q(\vec{x})\|$. By definition $\|\vec{x}\| = (\vec{x} \cdot \vec{x})^{\frac{1}{2}}$ and $\|T(\vec{x})\| = (T(\vec{x}) \cdot T(\vec{x}))^{\frac{1}{2}}$

Since Q is orthogonal, we know that $\vec{x} \cdot \vec{x} = Q(\vec{x}) \cdot Q(\vec{x})$, So the results show that Q defines a length preserving transformation

(b) Assume we have n eigenvalues $[\lambda_1, \lambda_2, \ldots, \lambda_n]$, the corresponding eigenvectors $[x_1, x_2, \ldots, x_n]$. Then we have

$$\Sigma = \begin{bmatrix} \lambda_1 & & \\ & \lambda_2 & \\ & & \ddots \\ & & & \lambda_n \end{bmatrix} \quad W = [x_1 \ x_2 \ \ldots \ x_n]$$

Since $Ax = \lambda x$, we have $AW = W\Sigma \Rightarrow A = W\Sigma W^{-1}$

since $\|x_i\|_2^2 = 1$, then $W^T = W^{-1}$, $A = W\Sigma W^T$

Since A's SVD decomposition is $A = UDV^T = U\begin{bmatrix}\Sigma \\ 0\end{bmatrix}V^T$

Thus, $A^T = VD^TU^T = V[\Sigma \ \ 0]U^T$

Then $A^TA = VD^TU^TUDV^T$

Since U is orthogonal: $A^TA = VD^TDV^T = V\Sigma^2V^T$

Thus, A's right singular vector V is the W built by A^TA's eigenvector.

and A^TA's eigenvalue is the square of the singular value of A

Same: $AA^T = UDV^TVD^TU^T = UDD^TU^T = U\Sigma^2U^T$

Thus, A's left singular vector is the W built by AA^T's eigenvector

and AA^T's eigenvalue is the square of the singular value of A

(c) i. False. At most n distinct eigenvalues

ii. False. If v_1 and v_2 are eigenvectors of A corresponding to eigenvalues λ_1 and λ_2 respectively, then the sum v_1+v_2 is not guaranteed to be an eigenvector.
In fact, unless v_1 and v_2 are scalar multiples of each other (i.e. $v_1=\lambda v_2$).
v_1+v_2 will not satisfy the eigenvector equation $Av=\lambda v$ for any single eigenvalue λ

iii. Correct

iv. Correct

v. Corret

2 (a) since $n = 1, 2, 3, 4 \dots$ where n denotes the round in which the duel ends, partitions the sample space so by law of total probability.

i. $P(A \text{ not hit}) = \sum_{n=1}^{\infty} P(A \text{ not hit}, n)$

Now, if duel ends in n rounds and A isn't hit,

then $\underset{1}{\underline{BM}}\ \underset{2}{\underline{BM}}\ \underset{3}{\underline{BM}}\ \underset{4}{\underline{BM}} \cdots \underset{n-1}{\underline{BM}}\ \underset{n}{\underline{A}}$

so, $P(A \text{ not hit}, n) = (1-P_A)^{n-1}(1-P_B)^{n-1} P_A(1-P_B)$

$= P_A(1-P_A)^{n-1}(1-P_B)^n$

so $P(A \text{ not hit}) = P_1 \sum_{n=1}^{\infty} (1-P_A)^{n-1}(1-P_B)^n$

$= P_1 \sum_{n=1}^{\infty} (1-P_A)^{n-1}(1-P_B)^n$

Since $\sum_{n=1}^{\infty} (1-P_A)^{n-1}(1-P_B)^n$ is the sum of a geometric series with $a = 1-P_B$, $r = (1-P_A)(1-P_B)$

so, $P(A \text{ not hit}) = \dfrac{P_A(1-P_B)}{1-(1-P_A)(1-P_B)}$

ii. $P(\text{both duelists are hit}) = \sum_{n=1}^{\infty} (1-P_A)^{n-1}(1-P_B)^{n-1} P_A P_B$

$= P_A P_B \sum_{n=1}^{\infty} (1-P_A)^{n-1}(1-P_B)^{n-1}$

$= \dfrac{P_A P_B}{1-(1-P_A)(1-P_B)}$

iii. Since duel can end after n th round of ~~Shots in 3 possible ways: ①A hit ② B hit ③ Both hit~~

Then, by laws of total probability

$P(\text{duel ends after } n^{th} \text{ round})$

$= (1-P_A)^{n-1}(1-P_B)^{n-1}(1-P_A)P_B + (1-P_A)^{n-1}(1-P_B)^{n-1}P_A(1-P_B)$

$+ (1-P_A)^{n-1}(1-P_B)^{n-1}P_AP_B$

$= [(1-P_A)(1-P_B)]^{n-1}[1-(1-P_A)(1-P_B)]$

iv. $P(\text{Duel ends after } n^{th} \text{ round} \mid A \text{ is not hit})$

$$= \frac{P(\text{Duel ends after } n^{th} \text{ round} + A \text{ is not hit})}{P(A \text{ is not hit})}$$

$$= \frac{[(1-P_A)(1-P_B)]^{n-1} [(1-P_B)P_A]}{[(1-P_B)\cdot P_A / 1-(1-P_B)(1-P_A)]}$$

$$= \frac{[(1-P_A)(1-P_B)]^{n-1}}{\frac{1}{1-(1-P_B)(1-P_A)}} = [(1-P_A)(1-P_B)]^{n-1} [1-(1-P_A)(1-P_B)]$$

v. $P(\text{end in the } n\text{-th} \mid \text{both are shot}) = \frac{P(\text{end in the } n\text{-th}, \text{both are shot})}{P(\text{both are shot})}$

From ii, we get $P(\text{both are shot}) = \frac{P_A P_B}{P_A + P_B - P_A P_B}$

$$P(\text{end in } n\text{th} \mid \text{both hit}) = \frac{[(1-P_A)(1-P_B)]^{n-1} P_A P_B}{\frac{P_A P_B}{P_A + P_B - P_A P_B}}$$

$$= [(1-P_A)(1-P_B)]^{n-1} \cdot (P_A + P_B - P_A P_B)$$

2(b)

i. suppose we define X_i as the bernoulli random variable

$$X_i = \begin{cases} 1, & \text{the ith faculty is isolated} \\ 0, & \text{the ith faculty is not isolated} \end{cases}$$

Then

$$X = \sum_{i=1}^{18} X_i$$

Now, $E[X_i] = P(X_i = 1)$

$P(X_i = 1) = P(\text{ith faculty is isolated})$

Define E: ith facluty is ECE

C: ith faculty is CSE

M: ith faculty is Math

Since E, C, M partitions the sample space so by law of total probability,

$P(\text{ith faculty is isolated})$

$= P(i|E)P(E) + P(i|C)P(C) + P(i|M)P(M)$

$= \left(\frac{12}{17}\right) \cdot \left(\frac{11}{16}\right) \frac{1}{3} \cdot 3 = \frac{33}{68}$

Hence by Linearity of expectations

$E[X] = 18E[X_i] = 18 \cdot \frac{33}{68} = 8.735$

ii. Define Y_i as the Bernoulli random variable

$$Y_i = \begin{cases} 1, & \text{ith faculty is semi-happy} \\ 0, & \text{ith faculty is not semi-happy} \end{cases}$$

$Y = \sum_{i=1}^{15} Y_i$ Now, $E[Y_i] = P(Y_i = 1)$

$P(Y_i=1) = P(\text{ith faculty is semi-happy})$

Use the same event defined: E, C, M

$P(\text{ith faculty is semi-happy})$

$= P(i|E)P(E) + P(i|C)P(C) + P(i|M)P(M)$

$= \left[\frac{12}{17} \cdot \frac{5}{16} + \frac{5}{17} \cdot \frac{12}{16}\right] \cdot \frac{1}{3} \cdot 3 = \frac{30}{68}$

By linearity of expectations

$E[Y] = 15E[Y_i] = 18 \cdot \frac{30}{68} = 7.941$

ii. Define Z_i as the Bernoulli random variable

$$Z_i = \begin{cases} 1, & \text{ith faculty is joyous} \\ 0, & \text{ith faculty is not joyous} \end{cases}$$

Then $Z = \sum_{i=1}^{15} Z_i$ $E[Z_i] = P(Z_i = 1)$

$P(Z_i=1) = P(\text{ith faculty is joyous})$

Use the same event defined before: E, C, M

$P(\text{ith faculty is joyous}) = P(i|E)P(E) + P(i|C)P(C) + P(i|M)P(M)$

$= \left[\frac{5}{17} \cdot \frac{4}{16}\right] \cdot \frac{1}{3} \cdot 3 = \frac{5}{68}$

$E[Z] = 15E[Z_i] = 18 \times \frac{5}{68} = 1.324$

2. (C) Let's define the following events:

D: man has a dangerous type of the disease

T: man has a positive LSA test

From the problem statement, we are given the following quantities

$$P(T|D) = 0.9 \quad P(T|D^c) = 0.01 \quad P(D) = 0.0005$$

i. By Bayes law,

$$P(D|T) = \frac{P(T|D)P(D)}{P(T|D)P(D) + P(T|D^c)P(D^c)}$$

$$= \frac{0.9 \times 0.0005}{0.9 \times 0.0005 + 0.01 \times 0.9995} = 0.043$$

ii. By Bayes Law,

$$P(D|T^c) = \frac{P(T^c|D)P(D)}{P(T^c|D)P(D) + P(T^c|D^c)P(D^c)} = \frac{0.1 \times 0.0005}{0.1 \times 0.0005 + 0.99 \times 0.9995}$$

$$= 0.000050528$$

(d) $E(X)$ is the expected value of vector X. $E(Ax+b)$ is the expectation of the vector $Ax+b$. Where we have known the dimension of x is n. Then A and b are deterministic and the dimension of A is $m \times n$, and the dimension of b is m.

Using the linearity of expectation to break down the calculation:

$$E(Ax+b) = \mathbb{E}([Ax_1+b, Ax_2+b \dots Ax_n+b]^T)$$
$$= [\mathbb{E}(Ax_1+b), \mathbb{E}(Ax_2+b), \dots, \mathbb{E}(Ax_n+b)]^T$$

if A and b are deterministic, we can pull them out of the expectation:

$$E(Ax+b) = A[\mathbb{E}(x_1), \mathbb{E}(x_2), \dots, \mathbb{E}(x_n)]^T + b$$

Since $X_1, X_2 \dots X_n$ are identically distributed random variables, their expectations are the same. Let's denote this common expectation as μ.

$$\mathbb{E}(Ax+b) = A[\mathbb{E}(X_1), E(X_2) \dots \mathbb{E}(X_n)]^T + b = A[\mu, \mu, \dots \mu]^T + b$$
$$= A\mu + b$$

(e) $$\operatorname{cov}(Ax+b) = \mathbb{E}((Ax+b-\mathbb{E}(Ax+b))(Ax+b-\mathbb{E}(Ax+b))^T)$$
$$= \mathbb{E}((Ax+b-A\mathbb{E}x-b)(Ax+b-A\mathbb{E}x-b)^T)$$
$$= \mathbb{E}((Ax-A\mathbb{E}x)(Ax-A\mathbb{E}x)^T)$$

If A is a deterministic matrix and $E(x)$ is a constant vector, we can further simplify.

$$\operatorname{cov}(Ax+b) = \mathbb{E}(A(x-\mathbb{E}x)(x-\mathbb{E}x)^T A^T)$$

Using the properties of covariance, we can rewrite the expression as:

$$\operatorname{cov}(Ax+b) = A \operatorname{cov}(x) A^T$$

3. (a) $\nabla_x x^T A y = Ay$

(b) $\nabla_y x^T A y = \nabla_y (Ay)^T x = \nabla_y y^T A^T x = A^T x$

(c) $\nabla_A x^T A y = x y^T$

(d) $f = x^T A x + b^T x$, $\nabla_x f = \nabla_x x^T A x + \nabla_x b^T x = Ax + A^T x + b$

(e) $f = tr(AB)$ $\nabla_A f = B^T$

(f) $f = tr(BA + A^T B + A^2 B)$

$\nabla_A f = \nabla_A [tr(BA) + tr(A^T B) + tr(A^2 B)]$

$= \nabla_A [tr(BA) + tr(A^T B) + tr(AAB)]$

$= \nabla_A [tr(BA) + tr(A^T B) + tr(BAIA)]$ I is the identity matrix $d\,tr(AXBX) = A^T X^T B^T + B^T X^T A^T$

$= B^T + B + B^T A^T + A^T B^T$

(g) $f = \|A + \lambda B\|^2$

$\nabla_A f = \nabla_A tr[(A+\lambda B)(A+\lambda B)^T]$

$= \nabla_A tr[(A+\lambda B)(A+\lambda B)^T]$

$= \nabla_A tr[(A+\lambda B)(A^T + \lambda B^T)]$

$= \nabla_A tr[AA^T + \lambda AB^T + \lambda BA^T + \lambda^2 BB^T]$

$= 2A + \lambda B + \lambda B^T$

4.

To find the optimal W, we take the derivate of the loss function

$$L(W) = \frac{1}{2}\sum_{i=1}^{n} \|y^{(i)} - Wx^{(i)}\|^2$$

with respect to W and set the derivate to zero

Expand the loss function $L(W)$:

$$L(W) = \frac{1}{2}(y - Wx)^T (y - Wx) = \frac{1}{2}(y^T - x^T W^T)(y - Wx)$$

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

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

Take the derivative of the above formula with respect to W and make the derivative zero.

$$\frac{\partial L(W)}{\partial W} = \frac{1}{2}[0 - yx^T - yx^T + W(xx^T + xx^T)] = \frac{1}{2}(-2yx^T$$

$$+ 2Wxx^T) = -yx^T + Wxx^T = 0$$

Then $Wxx^T = yx^T$

Now we can find the optimal parameter W by solving the above equation

A common method for solving equations is to use Normal Equation:

$$W = YX^T (XX^T)^{-1}$$

5. $$J(\theta) = \frac{1}{2}\sum_{i=1}^{N}\left[(y^{(i)})^2 - 2y^{(i)}\theta^T \hat{x}^{(i)} + (\theta^T x^{(i)})^2 \right) + \frac{\lambda}{2}\sum_{j=1}^{M}\theta_j^2$$

$$\nabla_\theta J(\theta) = -\sum_{i=1}^{N}(y^{(i)} - \theta^T x^{(i)})x^{(i)} + \lambda\theta$$

$$0 = -\sum_{i=1}^{N}(y^{(i)} - \theta^T x^{(i)})x^{(i)} + \lambda\theta$$

Matrix form:

$$-(X^T Y) - X^T X\theta) + \lambda\theta = 0$$

$$-X^T Y + X^T X\theta + \lambda\theta = 0$$

$$X^T Y = X^T X\theta + \lambda\theta$$

$$\theta = (X^T X + \lambda I)^{-1} X^T Y$$