CS189 HW2

David Chen

January 2022

Contents

Write-up

Collaborated with: N/A.

I certify that all solutions are entirely in my own words and that I have not looked at another student's solutions. I have given credit to all external sources I consulted.

Signed: David Chen

1 Identities and Inequalities with Expectation

1.1 Part 1

Proof. Proof by induction on k.
Base Case: For $k = 0$,

$$\begin{aligned} E\left[X^k\right] &= E\left[X^0\right] \\ &= E\left[1\right] \\ &= 1. \end{aligned}$$

Additionally,

$$\begin{aligned} \frac{k!}{\lambda^k} &= \frac{0!}{\lambda^0} \\ &= \frac{1}{1} \\ &= 1, \end{aligned}$$

so the equation holds for k=0.
Inductive Step: Assume

$$E\left[X^k\right] = \frac{k!}{\lambda^k}$$

holds for $k = n - 1$.
Conclusion: For $k = n$,

$$\begin{aligned} E\left[X^k\right] &= E\left[X^n\right] \\ &= \int_{-\infty}^{\infty} x^n f(x)\, dx \\ &= \int_{-\infty}^{\infty} x^n \lambda e^{-\lambda x} \mathbf{1}\left\{x > 0\right\}\, dx \\ &= \int_{0}^{\infty} x^n \lambda e^{-\lambda x}\, dx. \end{aligned} \tag{1}$$

Using integration by parts, we can transform integral (1) into

$$
\begin{aligned}
\int_0^\infty x^n \lambda e^{-\lambda x}\, dx &= -(x^n e^{-\lambda x})|_0^\infty + \int_0^\infty n x^{n-1} e^{-\lambda x}\, dx \\
&= -(x^n e^{-\lambda x})|_0^\infty + \frac{n}{\lambda} \int_0^\infty \lambda x^{n-1} e^{-\lambda x}\, dx \qquad (2) \\
&= -(x^n e^{-\lambda x})|_0^\infty + \frac{n}{\lambda} \frac{(n-1)!}{\lambda^{n-1}} \qquad (3) \\
&= -(x^n e^{-\lambda x})|_0^\infty + \frac{n!}{\lambda^n} \\
&= -\lim_{x \to \infty} x^n e^{-\lambda x} + \lim_{x \to 0} x^n e^{-\lambda x} + \frac{n!}{\lambda^n} \\
&= -0 + 0 + \frac{n!}{\lambda^n} \\
&= \frac{n!}{\lambda^n}.
\end{aligned}
$$

Note that the integral from the right side of equation (2) can be transformed to form equation (3) due to our inductive step. Therefore, for any $k = n$, we have shown that

$$E\left[X^k\right] = \frac{k!}{\lambda^k}$$

holds for any $k \geq 0$.

□

1.2 Part 2

Proof. Consider the inequality

$$\mathbf{1}\{a \geq b\} \leq \frac{a}{b}.$$

Note that this inequality always holds since the inequality holds for both cases of the indicator function. For instance, for the first case where $a \geq b$, $\mathbf{1}\{a \geq b\} = 1$ and $\frac{a}{b} \geq 1$. Furthermore, for the second case where $a < b$, $\mathbf{1}\{a \geq b\} = 0$ and $\frac{a}{b} \geq 0$.

Now, given that this inequality holds, taking the expected value of both sides gives

$$E[\mathbf{1}\{X \geq t\}] \leq E\left[\frac{X}{t}\right]$$
$$\therefore P(a \geq b) \leq \frac{E[X]}{t}.$$

□

1.3 Part 3

Proof. For any non-negative random variable X with corresponding probability distribution function $f_X(x)$

$$\begin{aligned} E[X] &= \int_{-\infty}^{\infty} x f_X(x)\, dx \\ &= \int_{0}^{\infty} x f_X(x)\, dx \\ &= \int_{0}^{\infty} \int_{0}^{x} f_X(t)\, dt\, dx \\ &= \int_{0}^{\infty} \int_{t}^{\infty} f_X(x)\, dx\, dt \\ &= \int_{0}^{\infty} P(X \geq t)\, dt. \end{aligned}$$

□

1.4 Part 4

Proof. For any non-negative random variable X, it can be re-written as

$$\begin{aligned} X &= X \cdot \mathbf{1}\{X > 0\} \\ \therefore E[X] &= E[X] \cdot E[\mathbf{1}\{X > 0\}]. \end{aligned}$$

Applying the Cauchy-Schwarz inequality gives

$$\begin{aligned} (E[X])^2 &\leq E\left[X^2\right] \cdot E\left[(\mathbf{1}\{X > 0\})^2\right] \\ &\leq E\left[X^2\right] \cdot E[\mathbf{1}\{X > 0\}] \\ &\leq E\left[X^2\right] \cdot P(X > 0) \\ \therefore P(X > 0) &\geq \frac{(E[X])^2}{E[X^2]}. \end{aligned}$$

□

1.5 Part 5

Proof. For any $t > 0$, let

$$X' = t - X.$$

We can re-write X' as

$$\begin{aligned} X' &= X' \cdot \mathbf{1}\{X' > 0\} + X' \cdot \mathbf{1}\{X' \leq 0\} \\ &\leq X' \cdot \mathbf{1}\{X' > 0\}. \end{aligned}$$

Applying the Cauchy-Schwarz Inequality gives

$$\begin{aligned} E[X']^2 &\leq E\left[X'^2\right] \cdot E\left[\mathbf{1}\{X' > 0\}^2\right] \\ &\leq E\left[X'^2\right] \cdot E[\mathbf{1}\{X' > 0\}] \\ &\leq E\left[X'^2\right] \cdot P(X' > 0) \\ \therefore P(X' > 0) &\geq \frac{E[X']^2}{E\left[X'^2\right]}. \end{aligned}$$

Converting X' back gives

$$\begin{aligned} P(X < t) &\geq \frac{E[t - X]^2}{E\left[t - X^2\right]} \\ &\geq \frac{E[t]^2 - 2E[tX] + E[X]^2}{E[t^2] - 2E[tX] + E[X^2]} \\ &\geq \frac{t^2}{t^2 + E[X^2]} \\ \therefore P(X \geq t) &\leq 1 - \frac{t^2}{t^2 + E[X^2]} \\ &\leq \frac{E\left[X^2\right]}{t^2 + E[X^2]}. \end{aligned}$$

□

2 Probability Potpourri

2.1 Part 1

Proof. For any vector $a \in \mathbb{R}^n$, where n is the length of Z,

$$\begin{aligned} a^T \Sigma a &= a^T E\left[(Z-\mu)(Z-\mu)^T\right] a \\ &= E\left[a^T (Z-\mu)(Z-\mu)^T a\right] \\ &= E\left[\left((Z-\mu)^T a\right)^2\right] \\ &\geq 0. \end{aligned}$$

Therefore, by definition (a), the covariance matrix is always positive semi-definite.

□

2.2 Part 2

2.2.1 Section (i)

For sake of ease, define W to be the event that it is windy and H to be the event that an archer hits her target.

$$\begin{aligned} P(+W \wedge +H) &= P(+W) \cdot P(+H|+W) \\ &= 0.3 \cdot 0.4 \\ &= 0.12. \end{aligned}$$

2.2.2 Section (ii)

$$\begin{aligned} P(+H) &= P(+W) \cdot P(+H|+W) + P(-W) \cdot P(+H|-W) \\ &= 0.3 \cdot 0.4 + 0.7 \cdot 0.6 \\ &= 0.12 + 0.42 \\ &= 0.54. \end{aligned}$$

2.2.3 Section (iii)

$$\begin{aligned} P(\text{One } +H \text{ and one } -H) &= \binom{2}{1} (P(+H))^1 (P(-H))^{2-1} \\ &= 2 \cdot 0.54 \cdot 0.46 \\ &= 0.4968. \end{aligned}$$

2.2.4 Section (iv)

$$\begin{aligned} P(-W|-H) &= \frac{P(-H|-W) \cdot P(-W)}{P(-H|-W) \cdot P(-W) + P(-H|+W) \cdot P(+W)} \\ &= \frac{0.3 \cdot 0.7}{0.3 \cdot 0.7 + 0.6 \cdot 0.3} \\ &= \frac{0.21}{0.39} \\ &= \frac{7}{13}. \end{aligned}$$

2.3 Part 3

$$
\begin{aligned}
E\left[\text{Score}\right] &= \int_{-\infty}^{\infty} xf\left(x\right)\,dx \\
&= \int_{0}^{\infty} xf\left(x\right)\,dx \\
&= 4\cdot\int_{0}^{\frac{1}{\sqrt{3}}} xf\left(x\right)\,dx + 3\cdot\int_{\frac{1}{\sqrt{3}}}^{1} xf\left(x\right)\,dx + 2\cdot\int_{1}^{\sqrt{3}} xf\left(x\right)\,dx \\
&= 4\cdot\int_{0}^{\frac{1}{\sqrt{3}}} \frac{2x}{\pi x\left(1+x^2\right)}\,dx + 3\cdot\int_{\frac{1}{\sqrt{3}}}^{1} \frac{2x}{\pi x\left(1+x^2\right)}\,dx + 2\cdot\int_{1}^{\sqrt{3}} \frac{2x}{\pi x\left(1+x^2\right)}\,dx \\
&= \frac{4}{\pi}\left(\ln\left(x^2+1\right)\right)|_{0}^{\frac{1}{\sqrt{3}}} + \frac{3}{\pi}\left(\ln\left(x^2+1\right)\right)|_{\frac{1}{\sqrt{3}}}^{1} + \frac{2}{\pi}\left(\ln\left(x^2+1\right)\right)|_{1}^{\sqrt{3}} \\
&= 1.195.
\end{aligned}
$$

2.4 Part 4

For any non-negative integer k that satisfies $k \leq n$,

$$\begin{aligned} P\left(X = k | X + Y = n\right) &= \frac{P\left(X + Y = n | X = k\right)}{P\left(X + Y = n\right)} \\ &= \frac{P\left(X = k\right) \cdot P\left(Y = n - k\right)}{P\left(X + Y = n\right)}. \end{aligned} \tag{1}$$

Note that the probability can be split into two probabilities, as shown in equation (1), because X and Y are independent variables. Now, using Poisson merging, we create a Poisson variable Z with parameter $\gamma = \lambda + \mu$.

$$\begin{aligned} P\left(X = k | X + Y = n\right) &= \frac{\frac{\lambda^k e^{-\lambda}}{k!} \frac{\mu^{n-k} e^{-\mu}}{(n-k)!}}{\frac{(\lambda+\mu)^n e^{-(\lambda+\mu)}}{n!}} \\ &= \frac{n!}{k!\left(n-k\right)!} \frac{\lambda^k \mu^{n-k}}{\left(\lambda+\mu\right)^n} \\ &= \binom{n}{k} \frac{\lambda^k \mu^{n-k}}{\left(\lambda+\mu\right)^k \left(\lambda+\mu\right)^{n-k}} \\ &= \binom{n}{k} \left(\frac{\lambda}{\lambda+\mu}\right)^k \left(\frac{\mu}{\lambda+\mu}\right)^{n-k}. \end{aligned}$$

This is the binomial distribution, with parameter $p = \frac{\lambda}{\lambda+\mu}$.

3 Properties of the Normal Distribution (Gaussians)

3.1 Part 1

Proof.

$$\begin{aligned}
E\left[e^{\lambda x}\right] &= \int_{-\infty}^{\infty} e^{\lambda x} f_X(x)\, dx \\
&= \int_{-\infty}^{\infty} e^{\lambda x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2} dx \\
&= e^{\frac{\sigma^2\lambda^2}{2}} \int_{-\infty}^{\infty} e^{-\frac{\sigma^2\lambda^2}{2}} e^{\lambda x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2} dx \\
&= e^{\frac{\sigma^2\lambda^2}{2}} \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2 + \lambda x + -\frac{\sigma^2\lambda^2}{2}} dx \\
&= e^{\frac{\sigma^2\lambda^2}{2}} \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}\left(x-\sigma^2\lambda\right)^2} dx \\
&= e^{\frac{\sigma^2\lambda^2}{2}}. \qquad (1)
\end{aligned}$$

Note that the expression within the integral in equation (1) is the probability density function of $N\left(\sigma^2\lambda, \sigma\right)$, so the integral evaluates to 1.

□

3.2 Part 2

Proof.

$$\begin{aligned} P(X \geq t) &= P(\lambda X \geq \lambda t) \\ &= P\left(e^{\lambda X} \geq e^{\lambda t}\right). \end{aligned}$$

Using Markov's Inequality and the result from Question 3.1,

$$\begin{aligned} P\left(e^{\lambda x} \geq e^{\lambda t}\right) &\leq \frac{E\left[e^{\lambda X}\right]}{e^{\lambda t}} \\ &\leq \frac{e^{\frac{\sigma^2\lambda^2}{2}}}{e^{\lambda t}} \\ &\leq e^{\frac{\sigma^2\lambda^2}{2} - \lambda t}. \end{aligned}$$

Therefore, we can set

$$\begin{aligned} e^{\frac{\sigma^2\lambda^2}{2} - \lambda t} &= e^{-\frac{t^2}{2\sigma^2}} \\ \therefore \frac{\sigma^2\lambda^2}{2} - \lambda t &= -\frac{t^2}{2\sigma^2} \\ \therefore \lambda &= \frac{t}{\sigma^2} \end{aligned}$$

By setting λ to this value, we get

$$\begin{aligned} P(X \geq t) &= P\left(e^{\lambda x} \geq e^{\lambda t}\right) \\ &\leq e^{\frac{\sigma^2\lambda^2}{2} - \lambda t} \\ &\leq e^{-\frac{t^2}{2\sigma^2}}. \end{aligned}$$

By X's symmetry around 0,

$$\begin{aligned} P(|X| \geq t) &= P(X \leq -t) + P(X \geq t) \\ &= 2P(X \geq t) \\ &\leq 2e^{-\frac{t^2}{2\sigma^2}}. \end{aligned}$$

□

3.3 Part 3

Let random variable $Z = \Sigma_{i=1}^{n} X_i$. This gives $Z \sim N\left(0, n\sigma^2\right)$. Therefore,

$$\begin{aligned} P\left(\frac{1}{n}\Sigma_{i=1}^{n} X_i \geq t\right) &= P\left(Z \geq nt\right) \\ &\leq e^{-\frac{(nt)^2}{2n\sigma^2}} \\ &\leq e^{-\frac{nt^2}{2\sigma^2}}. \end{aligned}$$

As $n \to \infty$, this value goes to 0.

3.4 Part 4

$$E\left[\vec{X}\right] = \vec{0}$$
$$\Sigma_{\vec{X}} = \sigma^2 I_n$$

$$\begin{aligned}\therefore E\left[\vec{Y}\right] &= E\left[A\vec{X} + \vec{b}\right] \\ &= AE\left[\vec{X}\right] + \vec{b} \\ &= A\vec{0} + \vec{b} \\ &= \vec{b}.\end{aligned}$$

$$\begin{aligned}\Sigma_{\vec{Y}} &= \Sigma_{A\vec{0}+\vec{b}} \\ &= A\Sigma_{\vec{X}} + 0_{nxn} \\ &= A\sigma^2 I_n \\ &= \sigma^2 A.\end{aligned}$$

3.5 Part 5

$$\begin{aligned} u_x \cdot v_x &= \langle u, X\rangle\langle v, X\rangle \\ &= \langle X, u\rangle\langle v, X\rangle \\ &= X u^T v X^T \\ &= X\langle u, v\rangle X^T \\ &= 0. \end{aligned}$$

$$\begin{aligned} Cov\left(u_x, v_x\right) &= E\left[u_x v_x\right] - E\left[u_x\right] E\left[v_x\right] \\ &= 0 - 0 \\ &= 0. \end{aligned}$$

Therefore, u_x and v_x are uncorrelated. This holds even if the X_i are not identically distributed, as the expected value of X_i is 0 and the inner product $\langle u, v\rangle$ will always cause $E\left[u_x v_x\right]$ to be 0.

3.6 Part 6

Proof. Using convex function $e^{\lambda x}$ for $\lambda > 0$,

$$\begin{aligned}
e^{\lambda E[\max_i |X_i|]} &\leq E\left[e^{\lambda \cdot \max_i |X_i|}\right] \\
&\leq E\left[\max_i e^{\lambda \cdot |X_i|}\right] \\
&\leq \Sigma_i E\left[e^{\lambda \cdot |X_i|}\right] \\
&\leq 2\Sigma_i E\left[e^{\lambda X_i}\right] \\
&\leq 2\Sigma_i e^{\frac{\sigma^2\lambda^2}{2}} \\
&\leq 2ne^{\frac{\sigma^2\lambda^2}{2}} \\
\therefore \lambda E\left[\max_i |X_i|\right] &\leq \ln 2n + \frac{\sigma^2\lambda^2}{2} \\
\therefore E\left[\max_i |X_i|\right] &\leq \frac{\ln 2n}{\lambda} + \frac{\sigma^2\lambda}{2}.
\end{aligned}$$

Maximizing this expected value with respect to λ gives $\lambda = \frac{\sqrt{2\ln 2n}}{\sigma}$ and

$$\begin{aligned}
E\left[\max_i |X_i|\right] &\leq \frac{\sigma\sqrt{\ln 2n}}{\sqrt{2}} + \frac{\sigma^2}{2} \cdot \frac{\sqrt{2\ln 2n}}{\sigma} \\
&\leq \sigma\sqrt{2\ln 2n} \\
\therefore C &= \sqrt{2}.
\end{aligned}$$

□

4 Linear Algebra Review

4.1 Part 1

4.1.1 Section (a)

$$\begin{aligned}
\begin{bmatrix} I_n & 0 \\ 0 & AB \end{bmatrix} &= \begin{bmatrix} 0 & AB \\ I_n & 0 \end{bmatrix} \\
&= \begin{bmatrix} 0 & I_n \\ AB & 0 \end{bmatrix} \\
&= \begin{bmatrix} 0 + I_n B & I_n \\ AB + 0B & 0 \end{bmatrix} \\
&= \begin{bmatrix} B & I_n \\ AB & 0 \end{bmatrix} \\
&= \begin{bmatrix} B & I_n \\ AB - AB & 0 - AI_n \end{bmatrix} \\
&= \begin{bmatrix} B & I_n \\ 0 & -A \end{bmatrix} \\
&= \begin{bmatrix} B & I_n \\ 0 & A \end{bmatrix}.
\end{aligned}$$

4.1.2 Section (b)

Proof. All the matrices shown in the previous section have the same rank, since elementary matrix operations preserve rank.

$$\begin{aligned}
rank\left(\begin{bmatrix} B & I_n \\ 0 & A \end{bmatrix}\right) &= rank\left(\begin{bmatrix} B - I_n & 0 \\ 0 & A \end{bmatrix}\right) \\
&= rank\,(A) + rank\,(B) \\
rank\left(\begin{bmatrix} I_n & 0 \\ 0 & AB \end{bmatrix}\right) &= n + rank\,(AB) \\
rank\left(\begin{bmatrix} B & I_n \\ 0 & A \end{bmatrix}\right) &= n + \min\left\{rank\,(A), rank\,(B)\right\}.
\end{aligned}$$

Therefore, since all these ranks are equal, we get

$$\begin{aligned}
&rank\,(A) + rank\,(B) = n + rank\,(AB) = n + \min\left\{rank\,(A), rank\,(B)\right\} \\
\therefore\; &rank\,(A) + rank\,(B) - n = rank\,(AB) = \min\left\{rank\,(A), rank\,(B)\right\} \\
\therefore\; &rank\,(A) + rank\,(B) - n \leq rank\,(AB) \leq \min\left\{rank\,(A), rank\,(B)\right\}.
\end{aligned}$$

□

4.1.3 Section (c)

First note that, because rank measures span of rows or columns, linear combinations do not affect rank. A^TA can be viewed as a linear transformation on either A^T or A, so the rank does not differ. In addition, the rank of the transpose is equal to the rank of the original, since, as previously mentioned, rank measures dimensions spanned by both columns and rows.

4.2 Part 2

Proof.

$$\begin{aligned} sign\left(x^T Ax\right) &= sign\left(x^T \lambda x\right) \\ &= sign\left(\lambda x^T x\right) \\ &= sign\left(\lambda\right). \end{aligned}$$

Therefore, since definition (a) implies definition (b) and definition (b) implies definition (a), definitions (a) and (b) are equivalent.

Symmetric matrices can be decomposed into $A = Q^T \Lambda Q$, where Λ is the diagonal matrix of eigenvalues. If the eigenvalues are non-negative, then Λ can be split into $\sqrt{\Lambda}\sqrt{\Lambda}$, where $\sqrt{\Lambda}$ has diagonal elements equal to the square root of the corresponding elements in Λ. Therefore, if the eigenvalues are non-negative, any symmetric matrix can be decomposed into

$$\begin{aligned} A &= Q^T \Lambda Q \\ &= Q^T \sqrt{\Lambda}\sqrt{\Lambda} Q \\ &= Q^T \sqrt{\Lambda}^T \sqrt{\Lambda} Q \\ &= \left(\sqrt{\Lambda} Q\right)^T \sqrt{\Lambda} Q \\ &= U^T U. \end{aligned}$$

Additionally, the converse holds, as the diagonal matrix Λ can only be split when the eigenvalues are non-negative, so the existence of matrix U implies the eigenvalues are non-negative. Therefore, since definition (c) implies definition (b) and vice versa, the two definitions are equivalent. In addition, since definition (a) and definition (b) are equivalent, definition (a) and definition (c) are also equivalent.

□

4.3 Part 3

4.3.1 Section (a)

Proof. In $x^T Ay$, x_i multiplies row i of matrix A and y_j multiplies column j of matrix A. This gives

$$x^T Ay = \Sigma_i \Sigma_j x_i y_j A_{ij}.$$

For $\langle A, xy^T \rangle$, we only need to consider the values on the diagonal of $A^T xy^T$. For these values, column i of matrix A^T is multiplied by x_i and row j of matrix A^T is multiplied by y_j. Converting to the transpose gives the

$$trace\left(A^T xy^T\right) == \Sigma_i \Sigma_j x_i y_j A_{ij},$$

which is equivalent to $x^T Ay$.

□

4.3.2 Section (b)

Proof. For any matrix Λ,

$$\begin{aligned} \|\Lambda\|_F &= \sqrt{\Sigma_i \Sigma_j |\Lambda_{ij}|^2} \\ &= \sqrt{\langle \Lambda, \Lambda \rangle}. \end{aligned}$$

Therefore, by applying the Cauchy-Schwarz Inequality,

$$\begin{aligned} (\langle A, B \rangle) &\leq \sqrt{\langle A, A \rangle \langle B, B \rangle} \\ &\leq \|A\|_F \|B\|_F. \end{aligned}$$

□

4.3.3 Section (c)

Proof. Since A and B are symmetric PSD matrices, they can be decomposed into $A = Q_A^T \Lambda_A Q_A$ and $B = Q_B^T \Lambda_B Q_B$, where each Q is an orthogonal matrix and each Λ is a diagonal matrix with the corresponding PSD matrix's eigenvalues on the diagonal. Therefore,

$$\begin{aligned} trace\,(AB) &= trace\left(Q_A^T \Lambda_A Q_A Q_B^T \Lambda_B Q_B\right) \\ &= trace\left(Q_A^T Q_A Q_B^T Q_B \Lambda_A \Lambda_B\right) \\ &= trace\,(I_n I_m \Lambda_A \Lambda_B) \\ &= \Sigma_i \lambda_{A,i} \lambda_{B,i}, \end{aligned}$$

where $\lambda_{X,i}$ is the ith eigenvalue of matrix X. Since A and B are both PSD, the last summation must be non-negative, so $trace\,(AB)$ must be non-negative.

□

4.3.4 Section (d)

Proof. Since A is a real, symmetric matrix, it can be decomposed into $A = Q^T \Lambda Q$, where Q is an orthogonal matrix and Λ is the diagonal eigenvalue matrix. Let λ_{max} be the largest eigenvalue of A. Then,

$$\begin{aligned}
\|A\|_F &= \|Q^T \Lambda Q\|_F \\
&\leq \|Q^T \lambda_{max} \cdot IQ\|_F \\
&\leq \lambda_{max} \|Q^T Q\|_F \\
&\leq \lambda_{max} \|I\|_F \\
&\leq \lambda_{max} \sqrt{\Sigma_i 1} \\
&\leq \lambda_{max} \sqrt{n},
\end{aligned}$$

where n is the dimension of A. Therefore, using the results from section (b),

$$\begin{aligned}
\langle A, B \rangle &\leq \|A\|_F \|B\|_F \\
&\leq \sqrt{n} \lambda_{max} \|B\|_F.
\end{aligned}$$

□

4.4 Part 4

Yes, $N^{-1} - M^{-1}$ is positive semi-definite.

Since M and N are positive definite matrices, they both have an inverse, due to its non-zero eigenvalues, and can be raised to any real power, due to its diagonalizability. Therefore, using the fact that $M - N$ is positive semi-definite and that multiplying a positive semi-definite matrix by any vector is still non-negative,

$$\begin{aligned} N^{\frac{1}{2}}(M-N)N^{\frac{1}{2}} &\succeq 0 \\ \therefore N^{-\frac{1}{2}}MN^{-\frac{1}{2}} &\succeq I \\ \therefore N^{-\frac{1}{2}}MN^{-\frac{1}{2}} &\preceq I \\ \therefore N^{\frac{1}{2}}M^{-1}N^{\frac{1}{2}} &\preceq I. \end{aligned}$$

$$\begin{aligned} M^{-1} &= N^{-\frac{1}{2}}N^{\frac{1}{2}}M^{-1}N^{\frac{1}{2}}N^{-\frac{1}{2}} \\ &\preceq N^{-\frac{1}{2}}IN^{-\frac{1}{2}} \\ &\preceq N \\ \therefore N^{-1} - M^{-1} &\succeq 0. \end{aligned}$$

4.5 Part 5

Proof. By the Spectral Theorem, A can be decomposed into $A = V\Lambda V^T$, where V is an orthonormal vector and Λ is the diagonal matrix containing the eigenvalues. Therefore,

$$\|u^T V\| = 1$$
$$\|V^T v\| = 1.$$

This means that the result from $u^T Av$ will be a linear combination of the eigenvalues of A, where the sum of the multipliers applied onto the eigenvalues is one. Therefore, the maximum possible value is given by $1 \cdot \lambda_{max} = \lambda_{max}$, where λ_{max} is the largest eigenvalue of A. In fact, to get this maximum possible value, set $u = v = vec_{max}$, where vec_{max} is the eigenvector corresponding to the largest eigenvalue.

□

5 Matrix/Vector Calculus and Norms

5.1 Part 1

Let

$$\begin{aligned}
\vec{x} &= \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \\
\vec{y} &= \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \\
f(A) &= \sin\left(A_{11}^2 + e^{A_{11}+A_{22}}\right) + \vec{x}^T A \vec{y} \\
&= \sin\left(A_{11}^2 + e^{A_{11}+A_{22}}\right) + \vec{x}^T \begin{bmatrix} A_{11}y_1 + A_{12}y_2 \\ A_{21}y_2 + A_{22}y_2 \end{bmatrix} \\
&= \sin\left(A_{11}^2 + e^{A_{11}+A_{22}}\right) + x_1 A_{11} y_1 + x_1 A_{12} y_2 + x_2 A_{21} y_2 + x_2 A_{22} y_2.
\end{aligned}$$

Therefore, taking the derivative of $f(A)$ with respect to each entry in A gives

$$\begin{aligned}
\frac{d}{dA} f(A) &= \begin{bmatrix} \frac{d}{dA_{11}} f(A) & \frac{d}{dA_{12}} f(A) \\ \frac{d}{dA_{21}} f(A) & \frac{d}{dA_{22}} f(A) \end{bmatrix} \\
&= \begin{bmatrix} \left(2A_{11} + e^{A_{11}+A_{22}}\right) \sin\left(A_{11}^2 + e^{A_{11}+A_{22}}\right) + x_1 y_1 & x_1 y_2 \\ x_2 y_1 & e^{A_{11}+A_{22}} \sin\left(A_{11}^2 + e^{A_{11}+A_{22}}\right) + x_2 y_2 \end{bmatrix}.
\end{aligned}$$

5.2 Part 2

5.2.1 Section (a)

$$\begin{aligned}
\|A\|_2 &= \sup_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2} \\
&= \max_{\|x\|=1} \|Ax\|_2 \\
&= \max_{\|x\|=1} \sqrt{\langle Ax, Ax \rangle} \\
&= \max_{\|x\|=1} \sqrt{x^T A^T A x} \\
&= \sigma_{max}(A). \qquad (1)
\end{aligned}$$

Note that the result (1) follows from Question (4.5).

5.2.2 Section (b)

$$\begin{aligned}
\|A\|_\infty &= \sup_{x \neq 0} \frac{\|Ax\|_\infty}{\|x\|_\infty} \\
&= \max_{\|x\|=1} \|Ax\|_\infty \\
&= \max_{\|x\|=1} (\Sigma_i |Ax|_i^\infty)^{\frac{1}{\infty}} \\
&= \max_{\|x\|=1} (\Sigma_i \Sigma_j |A_{i,j} x_i|^\infty)^{\frac{1}{\infty}} \\
&= \max_{\|x\|=1} (\Sigma_i |x_i|^\infty \Sigma_j |A_{i,j}|^\infty)^{\frac{1}{\infty}} \\
&= \max_i \Sigma_j |A_{i,j}|. \qquad (1)
\end{aligned}$$

Equation (1) results from the fact that vector x is a unit vector, so setting $x_i = 1$ on the row with the maximum sum of matrix A's entries and $x_i = 0$ everywhere else maximizes the result.

5.3 Part 3

5.3.1 Section (a)

$$\begin{aligned}\frac{\partial \alpha}{\partial \beta_i} &= \frac{\partial}{\partial \beta_i} \Sigma_j y_j \ln \beta_j \\ &= \frac{\partial}{\partial \beta_i} y_i \ln \beta_i \\ &= \frac{y_i}{\beta_i}.\end{aligned}$$

5.3.2 Section (b)

$$\begin{aligned}\frac{\partial \gamma_i}{\partial \rho_j} &= \frac{\partial}{\partial \rho_j} \Sigma_k A_{i,k} \rho_k + b_i \\ &= A_{i,j}.\end{aligned}$$

5.3.3 Section (c)

$$\begin{aligned}z &= g\left(f\left(x\right)\right) \\ \therefore \frac{dz}{dx} &= \frac{d}{dx} f\left(x\right) \frac{d}{df\left(x\right)} g\left(f\left(x\right)\right) \\ &= \mathbf{J}_f \mathbf{J}_g,\end{aligned}$$

where $\mathbf{J}_f$ is the Jacobian of function f and $\mathbf{J}_g$ is the Jacobian of function g.

5.3.4 Section (d)

Let $f_i(x)$ and $g_i(x)$ be the ith entry of $f(x)$ and $g(x)$ respectively. Then,

$$\begin{aligned}
\nabla_x y^T z &= \nabla_x f^T(x)\, g(x) \\
&= \nabla_x \Sigma_i f_i(x)\, g_i(x) \\
&= \begin{bmatrix} \frac{\partial}{\partial x_1} \Sigma_i f_i(x)\, g_i(x) \\ \frac{\partial}{\partial x_2} \Sigma_i f_i(x)\, g_i(x) \\ \cdots \\ \frac{\partial}{\partial x_n} \Sigma_i f_i(x)\, g_i(x) \end{bmatrix} \\
&= \begin{bmatrix} \frac{\partial}{\partial x_1} f_1(x)\, g_1(x) \\ \frac{\partial}{\partial x_2} f_2(x)\, g_2(x) \\ \cdots \\ \frac{\partial}{\partial x_n} f_n(x)\, g_n(x) \end{bmatrix} \\
&= \begin{bmatrix} \frac{\partial f_1(x)}{\partial x_1} g_1(x) + f_1(x) \frac{\partial g_1(x)}{\partial x_1} \\ \frac{\partial f_2(x)}{\partial x_2} g_2(x) + f_2(x) \frac{\partial g_2(x)}{\partial x_2} \\ \cdots \\ \frac{\partial f_n(x)}{\partial x_n} g_n(x) + f_n(x) \frac{\partial g_n(x)}{\partial x_n} \end{bmatrix}.
\end{aligned}$$

5.4 Part 4

By the Taylor's Theorem for multivariate functions,

$$
\begin{aligned}
f(x) &\leq f(x^*) + \frac{1}{2}(x - x^*)^T H(x^*)(x - x^*) \\
\therefore f(x) - f(x^*) &\leq \frac{1}{2}(x - x^*)^T H(x^*)(x - x^*) \\
&\leq \frac{1}{2}(x - x^*)^T \lambda_{max}(x - x^*) \\
&\leq \frac{1}{2}(x - x^*)^T 1(x - x^*) \\
&\leq \frac{1}{2}\|x - x^*\|^2 \\
&\leq \frac{D}{2}. \qquad (1)
\end{aligned}
$$

Note that equation (1) is only guaranteed to hold for any $x \in X$.

5.5 Part 5

$$\begin{aligned}
w^* &= arg\,min_w \|y - Xw\|^2 \\
&= arg\,min_w (y - Xw)^T (y - Xw) \\
&= arg\,min_w y^T y - y^T Xw - w^T X^T y + w^T X^T Xw \\
&= arg\,min_w y^T y - 2w^T X^T y + w^T X^T Xw \\
&= arg\,min_w - 2w^T X^T y + w^T X^T Xw.
\end{aligned}$$

Now, to find the w^* that minimizes the loss function,

$$\begin{aligned}
\frac{\partial w^*}{\partial w} &= \frac{\partial}{\partial w}\left(-2w^T X^T y + w^T X^T Xw\right) \\
&= -2X^T y + wX^T X^T + X^T Xw \\
&= -2X^T y + 2X^T Xw \\
\therefore X^T Xw^* &= X^T y \\
\therefore w^* &= \left(X^T X\right)^{-1} X^T y.
\end{aligned}$$

Note that $\left(X^T X\right)^{-1}$ is guaranteed to exist because X is assumed to have full column rank.

6 Gradient Descent

6.1 Part 1

$$
\begin{aligned}
x^* &= arg\,min_x \frac{1}{2} x^T A x - b^T x \\
\therefore \frac{\partial x^*}{\partial x} &= 2\frac{1}{2} x^{*T} A - b^T = 0 \\
\therefore x^{*T} A &= b^T \\
\therefore x^{*T} &= b^T A^{-1} \\
\therefore x^* &= {A^{-1}}^T b \\
&= A^{-1} b. \qquad (1)
\end{aligned}
$$

The result from line (1) comes from the assumption that A is a positive semi-definite matrix, and therefore symmetric.

6.2 Part 2

Since we are trying to minimize $\frac{1}{2}x^T Ax - b^T x$, we always move in the direction opposite of the gradient at each iteration, as this gives the largest possible decrease of this objective function in one step. Therefore,

$$\begin{aligned} x^{(n+1)} &= x^{(n)} - \nabla_{x^{(n)}} \left(\frac{1}{2} x^T Ax - b^T x \right) \\ &= x^{(n)} - \left(Ax^{(n)} - b \right). \end{aligned}$$

6.3 Part 3

Proof. Using the results from the previous parts, expanding the left side gives

$$\begin{aligned} x^{(k)} - x^* &= x^{(k-1)} - \left(Ax^{(k-1)} - b\right) - A^{-1}b \\ &= x^{(k-1)} - Ax^{(k-1)} + b - A^{-1}b. \end{aligned}$$

Similarly, expanding the right side gives

$$\begin{aligned} (I - A)\left(x^{(k-1)} - x^*\right) &= x^{(k-1)} - x^* - Ax^{(k-1)} + Ax^* \\ &= x^{(k-1)} - A^{-1}b - Ax^{(k-1)} + b, \end{aligned}$$

which is equal to the left side's expanded form. Therefore, we conclude that

$$x^{(k)} - x^* = (I - A)\left(x^{(k-1)} - x^*\right).$$

□

6.4 Part 4

Proof.

$$\begin{aligned}
\|Ax\|_2 &= \sqrt{\langle Ax, Ax\rangle} \\
&= \sqrt{x^T A^T A x} \\
&\leq \sqrt{x^T \lambda_{max}^2 x} \\
&\leq \lambda_{max}\sqrt{x^T x} \\
&\leq \lambda_{max}\|x\|_2.
\end{aligned}$$

□

6.5 Part 5

Proof. First note that the maximum eigenvalue of $I - A$ is given by $1 - \lambda_{min}$, where λ_{min} is the smallest eigenvalue of A. Therefore, by using the results from the previous parts,

$$\begin{aligned} \|x^{(k)} - x^*\|_2 &= \|(I - A)\left(x^{(k-1)} - x^*\right)\|_2 \\ &\leq (1 - \lambda_{min})\,\|x^{(k-1)} - x^*\|_2, \end{aligned} \tag{1}$$

where the result from line (1) comes from Question 6.4. Therefore, for $\rho = 1 - \lambda_{min}$,

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

□

6.6 Part 6

$$\begin{aligned} \|x^{(k)} - x^*\|_2 = \rho^k \|x^{(0)} - x^*\|_2 \geq \epsilon \\ \therefore \rho^k \|x^{(0)} - x^*\|_2 \geq \epsilon \\ \therefore \rho^k \geq \frac{\epsilon}{\|x^{(0)} - x^*\|_2} \\ \therefore k \geq \log_\rho \frac{\epsilon}{\|x^{(0)} - x^*\|_2}. \end{aligned}$$