


CS 289A – Spring 2023 – Homework 2

Colin Skinner, 

For this problem set referred to the Ed discussion, the unofficial course Discord, and office hours for general help. Additionally, I used various resources from the web, which I cite in the relevant problems. I also used ChatGPT occasionally to help me get started on a problem.

“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 Colin Skinner

Signature _____ Date 2/8/2023

1 Identities and Inequalities with Expectation

1.1

Lemma 1.1:

$$\lim_{x \rightarrow \infty} x^k e^{-\lambda x} = 0, \quad \forall k \geq 0$$

Proof by induction:

base case:

$$\begin{aligned} \lim_{x \rightarrow \infty} x e^{-\lambda x} &= \lim_{x \rightarrow \infty} \frac{x}{e^{\lambda x}} \\ &= \lim_{x \rightarrow \infty} \frac{1}{\lambda e^{\lambda x}}, \quad \text{By L'Hospital's rule} \\ &= 0 \end{aligned}$$

Assume for arbitrary k that

$$\lim_{x \rightarrow \infty} \frac{x^k}{e^{\lambda x}} = 0$$

Show $\lim_{x \rightarrow \infty} x^{k+1} e^{-\lambda x} = 0$:

$$\begin{aligned} \lim_{x \rightarrow \infty} \frac{x^{k+1}}{e^{\lambda x}} &= \lim_{x \rightarrow \infty} \frac{(k+1)x^k}{\lambda e^{\lambda x}} \\ &= \frac{k+1}{\lambda} \lim_{x \rightarrow \infty} \frac{x^k}{e^{\lambda x}} \\ &= 0 \end{aligned}$$

Q.E.D.

Next, show by induction that

$$\mathbb{E}[X^k] = \frac{k!}{\lambda^k}, \quad \forall k \geq 0$$

base case: $k=1$

$$\begin{aligned} \mathbb{E}[X] &= \int_{-\infty}^{\infty} x \lambda e^{-\lambda x} \mathbb{1}\{x < 0\} dx \\ &= \lambda \int_0^{\infty} x e^{-\lambda x} dx \end{aligned}$$

I.B.P

$$= \lambda \left[-\frac{xe^{-\lambda x}}{\lambda} \Big|_0^\infty + \frac{1}{\lambda} \int_0^\infty e^{-\lambda x} dx \right]$$

$$0 - \lim_{x \rightarrow \infty} xe^{-\lambda x} - \frac{1}{\lambda} e^{-\lambda x} \Big|_0^\infty$$

Second term is zero by lemma 1.1

$$-\frac{1}{\lambda} \left(\lim_{x \rightarrow \infty} e^{-\lambda x} - 1 \right)$$

$$= \frac{1}{\lambda}$$

Assume true for

$$\mathbb{E}[X^k] = \lambda \int_0^\infty x^k e^{-\lambda x} dx$$

$$= \frac{k!}{\lambda^k}$$

Show $\mathbb{E}[X^{k+1}] = \frac{(k+1)!}{\lambda^{k+1}}$

$$\mathbb{E}[X^{k+1}] = \lambda \int_0^\infty x^{k+1} e^{-\lambda x} dx$$

I.B.P

$$= \lambda \left[-\frac{x^{k+1}e^{-\lambda x}}{\lambda} \Big|_0^\infty + \frac{(k+1)}{\lambda} \int_0^\infty x^k e^{-\lambda x} dx \right]$$

$$= 0 - \lim_{x \rightarrow \infty} x^{k+1}e^{-\lambda x} + (k+1) \int_0^\infty x^k e^{-\lambda x} dx$$

Second term is zero by lemma 1.1

$$= (k+1) \frac{\mathbb{E}[X^k]}{\lambda}$$

$$= \frac{(k+1)}{\lambda} \frac{k!}{\lambda^k}$$

$$= \frac{(k+1)!}{\lambda^{k+1}}$$

Q.E.D.

1.2

Lemma 1.2

$$\mathbb{1}\{a \geq b\} \leq \frac{a}{b}$$

Proof:

Say $a \geq b$, then

$$\frac{a}{b} \geq 1$$

Note that $\mathbb{P}(a \geq b) \leq 1$. Then

$$\mathbb{P}(a \geq b) \leq 1 \leq \frac{a}{b}$$

and therefore

$$\mathbb{P}(a \geq b) \leq \frac{a}{b}$$

$$\mathbb{E}[\mathbb{1}\{a \geq b\}] \leq \frac{a}{b} = \mathbb{E}\left[\frac{a}{b}\right]$$

Therefore

$$\mathbb{1}\{a \geq b\} \leq \frac{a}{b}$$

Q.E.D.

Show

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}$$

By lemma 1.2 we can say

$$\mathbb{1}\{X \geq t\} \leq \frac{X}{t}, \quad \forall X, t > 0$$

$$\mathbb{E}[\mathbb{1}\{X \geq t\}] \leq \mathbb{E}\left[\frac{X}{t}\right]$$

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{\mathbb{E}[t]}$$

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}$$

Q.E.D.

1.3

For a non-negative random variable X

$$\begin{aligned}\mathbb{P}(X \geq t) &= \int_t^\infty f(x)dx \\ \int_0^\infty \mathbb{P}(X \geq t)dt &= \int_0^\infty \int_t^\infty f(x)dxdt\end{aligned}$$

Note that for $\mathbb{P}(X \geq t)$, t is bounded above by x , i.e. $0 \leq t \leq x$ so

$$\int_0^\infty \mathbb{P}(X \geq t)dt = \int_0^\infty \int_0^x f(x)dt dx$$

By Fubini's theorem, since the sum of probabilities is finite

$$\begin{aligned}&= \int_0^\infty x f(x)dx \\ &= \int_{-\infty}^\infty x f(x)dx\end{aligned}$$

Because X is non-negative, $f(x) = 0$ for all $x < 0$

$$= \mathbb{E}[X]$$

Q.E.D

1.4

For a non-negative r.v. X , according to Cauchy-Schwarz

$$|\mathbb{E}[X \mathbb{1}\{X > 0\}]|^2 \leq \mathbb{E}[X^2] \mathbb{E}[(\mathbb{1}\{X > 0\})^2]$$

$$\mathbb{E}[X \mathbb{1}\{X > 0\}]^2 \leq \mathbb{E}[X^2] \mathbb{E}[\mathbb{1}\{X > 0\}]$$

Because range of the indicator function is $\{0, 1\}$, which are equal to their squares (r.h.s.)

$$(\mathbb{E}[X])^2 \leq \mathbb{E}[X^2] \mathbb{P}(X > 0)$$

Because X is non-negative, $X \mathbb{1}\{X > 0\} = 0$ only when $X = 0$ and $= X$ otherwise, and also $\mathbb{E}[0] = 0$ (l.h.s)

$$\mathbb{P}(X > 0) \geq \frac{(\mathbb{E}[X])^2}{\mathbb{E}[X^2]}$$

Q.E.D.

1.5

For $t \geq 0$, according Cauchy-Schwarz

$$|\mathbb{E}[(t - X)\mathbb{1}\{t - X > 0\}]|^2 \leq \mathbb{E}[(t - X)^2]\mathbb{E}[(\mathbb{1}\{t - X > 0\})^2]$$

$$(\mathbb{E}[(t - X)\mathbb{1}\{t > X\}])^2 \leq \mathbb{E}[t^2 - 2tX + X^2]\mathbb{E}[\mathbb{1}\{t > X\}]$$

Because range of the indicator function is $\{0, 1\}$, which are equal to their squares (r.h.s.)

$$(\mathbb{E}[t - X])^2 \leq (\mathbb{E}[t^2] - 2t\mathbb{E}[X] + \mathbb{E}[X^2]) \mathbb{P}(t > X)$$

Because $(t - X)\mathbb{1}\{t > X\} = 0$ only when $t \leq X$ and $= t - X$ otherwise, and also $\mathbb{E}[0] = 0$ (l.h.s)

$$(\mathbb{E}[t] - \mathbb{E}[X])^2 \leq (\mathbb{E}[t^2] - 2t\mathbb{E}[X] + \mathbb{E}[X^2]) (1 - \mathbb{P}(X \geq t))$$

Because complementary probabilities sum to one (r.h.s)

$$t^2 \leq (t^2 + \mathbb{E}[X^2]) (1 - \mathbb{P}(X \geq t))$$

Because $\mathbb{E}[X] = 0$

$$t^2 \leq t^2 + \mathbb{E}[X^2] - (t^2 + \mathbb{E}[X^2]) \mathbb{P}(X \geq t)$$

$$(t^2 + \mathbb{E}[X^2]) \mathbb{P}(X \geq t) \leq t^2 + \mathbb{E}[X^2] - t^2$$

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X^2]}{\mathbb{E}[X^2] + t^2}$$

Q.E.D.

2 Probability Potpourri

2.1

Sources: Math StackExchange and YouTube: Brunei Math Club, "Simple proof that any covariance matrix is positive semi-definite"

Let $x \in \mathbb{R}^n$ be an arbitrary constant vector (assuming $\Sigma \in \mathbb{R}^{n \times n}$)

$$\begin{aligned} x^T \Sigma x &= x^T \mathbb{E}[(Z - \mu)(Z - \mu)^T] x \\ &= \mathbb{E}[x^T] \mathbb{E}[(Z - \mu)(Z - \mu)^T] \mathbb{E}[x] \end{aligned}$$

Because x is constant

$$\begin{aligned} &\mathbb{E}[x^T (Z - \mu)(Z - \mu)^T x] \\ &\mathbb{E}[(Z - \mu)^T x]^T [(Z - \mu)^T x] \end{aligned}$$

Because $(Z - \mu)^T x$ is just a scalar

$$\mathbb{E}[(Z - \mu)^T x]^2$$

$$\geq 0$$

Therefore Σ must be PSD
Q.E.D.

2.2

Let H and W be the random variables for the archer hitting her target and whether there is wind, respectively.

2.2.1 (i)

$$P(H \cap W) = P(H|W)P(W)$$

$$= \left(\frac{2}{5}\right) \left(\frac{3}{10}\right)$$

$$\boxed{= \frac{3}{25}}$$

2.2.2 (ii)

$$P(H \text{ on 1st shot}) = P(H) = P(H|W)P(W) + P(H|W^c)P(W^c)$$

$$= \frac{3}{25} + \left(\frac{7}{10}\right)\left(\frac{7}{10}\right)$$

$$\boxed{= \frac{61}{100}}$$

2.2.3 (iii)

Assume that shots are independent

$$P(\text{hit exactly once in two shots}) = 1 - (P(HH) + P(H^cH^c))$$

$$= 1 - (P(H)^2 + P(H^c)^2)$$

By independence $P(HH) = P(H)P(H) = P(H)^2$. Ditto for $P(H^cH^c)$.

$$= 1 - (P(H)^2 + (1 - P(H))^2)$$

$$= 1 - \left(\left(\frac{61}{100} \right)^2 + \left(\frac{39}{100} \right)^2 \right)$$

$$= 1 - \left(\frac{2621}{5000} \right)$$

$$\boxed{= \frac{2379}{5000}}$$

2.2.4 (iv)

$$P(W^c|H^c) = \frac{P(H^c|W^c)P(W^c)}{P(H^c|W^c)P(W^c) + P(H^c|W)P(W)}$$

$$= \frac{\left(\frac{3}{10}\right)\left(\frac{7}{10}\right)}{\left(\frac{3}{10}\right)\left(\frac{7}{10}\right) + \left(\frac{6}{10}\right)\left(\frac{3}{10}\right)}$$

$$\boxed{= \frac{7}{13}}$$

2.3

$$S(x) = \begin{cases} 4, & 0 \leq x < \frac{1}{\sqrt{3}} \\ 3, & \frac{1}{\sqrt{3}} \leq x < 1 \\ 2, & 1 \leq x < \sqrt{3} \\ 0, & \text{otherwise} \end{cases}$$

$$\mathbb{E}[S(x)] = \frac{2}{\pi} \left(\int_0^{\frac{1}{\sqrt{3}}} \frac{4}{1+x^2} dx + \int_{\frac{1}{\sqrt{3}}}^1 \frac{3}{1+x^2} dx + \int_1^{\sqrt{3}} \frac{2}{1+x^2} dx \right)$$

$$= \frac{2}{\pi} \left(4 \arctan(x) \Big|_0^{\frac{1}{\sqrt{3}}} + 3 \arctan(x) \Big|_{\frac{1}{\sqrt{3}}}^1 + 2 \arctan(x) \Big|_1^{\sqrt{3}} \right)$$

$$= \frac{2}{\pi} \left[4 \frac{\pi}{6} + 3 \left(\frac{\pi}{4} - \frac{\pi}{6} \right) + 2 \left(\frac{\pi}{3} - \frac{\pi}{4} \right) \right]$$

$$\boxed{= \frac{13}{6}}$$

2.4

Source: <https://llc.stat.purdue.edu/2014/41600/notes/prob1805.pdf>

Let $Z = X + Y$

$$P(Z = n) = \sum_{i=0}^k P(X = i)P(Y = n - i)$$

Because X and Y are independent.

$$\begin{aligned} &= \sum_{i=0}^k \frac{\lambda^i e^{-\lambda}}{i!} \frac{\mu^{n-i} e^{-\mu}}{(n-i)!} \\ &= e^{-(\lambda+\mu)} \sum_{i=0}^k \frac{\lambda^i \mu^{n-i}}{(n-i)! i!} \\ &= \frac{e^{-(\lambda+\mu)}}{n!} \sum_{i=0}^k \frac{n! \lambda^i \mu^{n-i}}{(n-i)! i!} \\ &= \frac{e^{-(\lambda+\mu)}}{n!} \sum_{i=0}^k \binom{n}{i} \lambda^i \mu^{n-i} \\ &= \frac{(\lambda + \mu)^n e^{-(\lambda+\mu)}}{n!} \end{aligned}$$

We see that $Z \sim \text{Poi}(\lambda + \mu)$

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

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

So we see that

$$X|Z = n \sim \text{Bin} \left(n, \frac{\lambda}{\lambda + \mu} \right)$$

3 Properties of the Normal Distribution (Gaussians)

3.1

$$\mathbb{E}[e^{\lambda X}] = \int_{-\infty}^{\infty} e^{\lambda x} f_X(x) dx$$

For $X \sim \mathcal{N}(0, \sigma^2)$

$$\begin{aligned} &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp(\lambda x) \exp\left(-\frac{x^2}{2\sigma^2}\right) dx \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2} (x^2 - 2\sigma^2\lambda x)\right) dx \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2} (x^2 - 2\sigma^2\lambda x + \sigma^4\lambda^2 - \sigma^4\lambda^2)\right) dx \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\left(-\frac{(x - \sigma^2\lambda)^2}{2\sigma^2} + \frac{\sigma^2\lambda^2}{2}\right) dx \\ &= \exp\left(\frac{\sigma^2\lambda^2}{2}\right) \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \sigma^2\lambda)^2}{2\sigma^2}\right) dx \\ &= e^{\frac{\sigma^2\lambda^2}{2}} \end{aligned}$$

Because the integrand is the PDF for $X \sim \mathcal{N}(\sigma^2\lambda, \sigma^2)$ summed over all possible values and probabilities sum to 1.

Q.E.D.

3.2

Source: Wikipedia

Use the Chernoff bound:

$$P(X \geq t) = P(e^{\lambda X} \geq e^{\lambda t}) \leq \mathbb{E}[e^{\lambda X}]e^{-\lambda t}$$

This is a valid application of Markov's inequality because $e^{\lambda X} > 0$, $\forall x$

$$= \exp \left\{ \sigma^2 \lambda^2 / 2 \right\} \exp \left\{ -\lambda t \right\}$$

by the result from 3.1. Want to find a lambda which gives the tightest bound for t.

$$= \exp \left\{ \frac{\sigma^2}{2} \left(\lambda^2 - \frac{2t}{\sigma^2} \lambda \right) \right\}$$

$$= \exp \left\{ \frac{\sigma^2}{2} \left(\lambda^2 - \frac{2t}{\sigma^2} \lambda + \frac{t^2}{\sigma^4} \right) - \frac{\sigma^2}{2} \frac{t^2}{\sigma^4} \right\}$$

$$= \exp \left\{ \frac{\sigma^2}{2} \left(\lambda - \frac{t}{\sigma^2} \right)^2 - \frac{t^2}{2\sigma^2} \right\}$$

$$= \exp \left\{ \frac{\sigma^2}{2} \left(\lambda - \frac{t}{\sigma^2} \right)^2 \right\} \exp \left\{ -\frac{t^2}{2\sigma^2} \right\}$$

To minimize the r.h.s. choose

$$\lambda = \frac{t}{\sigma^2}$$

And we get

$$P(X \geq t) \leq \mathbb{E}[e^{\frac{tX}{\sigma^2}}]e^{-\frac{t^2}{\sigma^2}}$$

$$= \exp \left\{ \frac{\sigma^2 (t/\sigma^2)^2}{2} \right\} \exp \left\{ -\frac{t^2}{\sigma^2} \right\}$$

$$= \exp \left\{ \frac{t^2}{2\sigma^2} \right\} \exp \left\{ -\frac{t^2}{\sigma^2} \right\}$$

$$= e^{-\frac{t^2}{2\sigma^2}}$$

Note that

$$P(|X| \geq t) = P(X \geq t) + P(X \leq -t)$$

and to find $P(X \leq -t)$ we can set

$$\lambda = -\frac{t}{\sigma^2}$$

as the upper-bound for the lower tail of the distribution for X , and by the symmetry of the normal distribution we will again get

$$P(X \leq -t) \leq e^{-\frac{t^2}{2\sigma^2}}$$

Therefore,

$$P(|X| \geq t) \leq 2e^{-\frac{t^2}{2\sigma^2}}$$

Q.E.D.

3.3

Let

$$\bar{S}_n = \frac{1}{n} \sum_i^n X_i$$

Which is itself a Gaussian r.v. with mean

$$\begin{aligned} \mathbb{E}[\bar{S}_n] &= \mathbb{E} \left[\frac{1}{n} \sum_i^n X_i \right] \\ &= \frac{1}{n} \sum_i^n \mathbb{E}[X_i] \\ &= 0 \end{aligned}$$

Because $X_1, \dots, X_n \sim \mathcal{N}(0, \sigma^2)$. It has a variance

$$\begin{aligned} \text{Var}(\bar{S}_n) &= \text{Var} \left(\frac{1}{n} \sum_i^n X_i \right) \\ &= \frac{1}{n^2} \sum_i^n \text{Var}(X_i) \end{aligned}$$

because X_1, \dots, X_n are independent

$$= \frac{\sigma^2}{n}$$

Thus $\bar{S}_n \sim \mathcal{N}(0, \sigma^2/n)$. To give an upper bound on the probability that \bar{S}_n is far from some $t > 0$, we can use the same Chernoff bound as in 3.2, replacing the variance of one Gaussian r.v. with that of \bar{S}_n .

$$P(\bar{S}_n \geq t) \leq \exp \left\{ -\frac{t^2}{\sigma^2/n} \right\}$$

$$\boxed{= e^{-\frac{nt^2}{\sigma^2}}}$$

And

$$\begin{aligned} &\lim_{n \rightarrow \infty} P(\bar{S}_n \geq t) \\ &= \boxed{\lim_{n \rightarrow \infty} e^{-\frac{nt^2}{\sigma^2}} = 0} \end{aligned}$$

3.4

$$\mathbb{E}[Y] = \mathbb{E}[AX + b]$$

$$= A\mathbb{E}[X] + \mathbb{E}[b]$$

$$= b$$

Because $\mathbb{E}[X_1] = E[X_2] = \dots = E[X_n] = 0$

$$\text{Var}(Y) = \mathbb{E} [((AX + b) - \mathbb{E}[Y])((AX + b) - \mathbb{E}[Y])^T]$$

$$= \mathbb{E}[(AX)(AX)^T]$$

$$= A\mathbb{E}[XX^T A^T]$$

$$= A\mathbb{E}[XX^T]\mathbb{E}[A^T]$$

$$= A\sigma^2 I_n A^T$$

$$= \sigma^2 AA^T$$

Because

$$\mathbb{E}[XX^T] = \mathbb{E} \left[\begin{bmatrix} X_1^2 & X_1X_2 & \dots & X_1X_n \\ X_2X_1 & X_2^2 & \dots & X_2X_n \\ \vdots & \vdots & \ddots & \vdots \\ X_nX_1 & X_nX_2 & \dots & X_n^2 \end{bmatrix} \right]$$

$$= \begin{bmatrix} \mathbb{E}[X_1^2] & 0 & \dots & 0 \\ 0 & \mathbb{E}[X_2^2] & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mathbb{E}[X_n^2] \end{bmatrix}$$

Because each X_i is i.i.d. and each $E[X_i] = 0$

$$= \sigma^2 I_n$$

Because

$$\text{Var}(X_i) = \mathbb{E}[X_i^2] - (E[X_i])^2$$

$$= \mathbb{E}[X_i^2]$$

$$= \sigma^2$$

Therefore

$$\boxed{\mathbb{E}[Y] = b, \quad \text{Var}(Y) = \sigma^2 A A^T}$$

3.5

Note that

$$\begin{aligned}\mathbb{E}[u_x] &= \mathbb{E}[u^T X] \\ &= u^T \mathbb{E}[X] \\ &= 0\end{aligned}$$

And

$$\mathbb{E}[v_x] = 0$$

by an analogous argument.

Also, note that for an i.i.d. standard normal random vector X

$$\begin{aligned}\mathbb{E}[X^2] &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \\ &= \text{Var}(X) \\ &= 1\end{aligned}$$

Note that

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

And this only equals zero when $\text{Cov}(X, Y) = 0$. Therefore, X and Y are independent if $\text{Cov}(X, Y) = 0$. For u_x and v_x

$$\begin{aligned}\text{Cov}(u_x, v_x) &= \mathbb{E}[(u_x - \mathbb{E}[u_x])(v_x - \mathbb{E}[v_x])] \\ &= \mathbb{E}[u_x v_x] \\ &= \mathbb{E}[\langle u, X \rangle \langle v, X \rangle] \\ &= \mathbb{E}[(u_1 X_1 + \dots + u_n X_n)(v_1 X_1 + \dots + v_n X_n)] \\ &= \mathbb{E}\left[\sum_i^n \sum_j^n u_i v_j X_i X_j\right]\end{aligned}$$

$$\begin{aligned}
&= \sum_i^n \sum_j^n u_i v_j \mathbb{E}[X_i X_j] \\
&= \sum_i^n u_i v_i \mathbb{E}[X_i^2]
\end{aligned}$$

Because $\mathbb{E}[X_i X_j] = \mathbb{E}[X_i] \mathbb{E}[X_j] = 0$ for all $i \neq j$

$$\begin{aligned}
&= \sum_i^n u_i v_i \\
&= u^T v \\
&= 0
\end{aligned}$$

Therefore u_x and v_x are independent. If instead each $X_i \sim \mathcal{N}(0, i)$ then

$$\begin{aligned}
\text{Cov}(u_x, v_x) &= \sum_i^n u_i v_i \mathbb{E}[X_i^2] \\
&= \sum_i^n u_i v_i i
\end{aligned}$$

which is not generally equal to zero and depends on the specific vectors u and v . So yes, the answer changes.

3.6

Sources: ChatGPT

Let $X = \max_{1 \leq i \leq n} |X_i|$. Also, since X is non-negative we can use Markov's inequality to say

$$P(X \geq t) \leq \frac{\mathbb{E}[X]}{t}$$

and

$$\mathbb{E}[X] \geq tP(X \geq t)$$

In 3.2 we found

$$P(|X| \geq t) \leq 2e^{\frac{-t^2}{2\sigma^2}}$$

for $X \sim \mathcal{N}(0, \sigma^2)$

We can set a convenient upper bound on this probability that makes t sufficiently "extreme" by letting it be $1/n$

$$2e^{\frac{-t^2}{2\sigma^2}} = \frac{1}{n}$$

$$\frac{-t^2}{2\sigma^2} = \log\left(\frac{1}{n}\right)$$

$$t = \sqrt{2\log(2n)}\sigma$$

So we have

$$\mathbb{E}[X] \geq \sqrt{2\log(2n)}\sigma \frac{1}{n}$$

4 Linear Algebra Review

4.1

4.1.1 (a)

Add in an A matrix

$$\begin{bmatrix} I_n & 0 \\ A & I_m \end{bmatrix} \begin{bmatrix} I_n & 0 \\ 0 & AB \end{bmatrix} = \begin{bmatrix} I_n & 0 \\ A & AB \end{bmatrix}$$

Permute the columns

$$\begin{bmatrix} I_n & 0 \\ A & AB \end{bmatrix} \begin{bmatrix} 0 & I_n \\ I_p & 0 \end{bmatrix} = \begin{bmatrix} 0 & I_n \\ AB & A \end{bmatrix}$$

Add -1 of column 1 to B of column 2

$$\begin{bmatrix} 0 & I_n \\ AB & A \end{bmatrix} \begin{bmatrix} -I_p & 0 \\ B & I_n \end{bmatrix} = \begin{bmatrix} B & I_n \\ 0 & A \end{bmatrix}$$

Together

$$\boxed{\begin{bmatrix} I_n & 0 \\ A & I_m \end{bmatrix} \begin{bmatrix} I_n & 0 \\ 0 & AB \end{bmatrix} \begin{bmatrix} 0 & I_n \\ I_p & 0 \end{bmatrix} \begin{bmatrix} -I_p & 0 \\ B & I_n \end{bmatrix} = \begin{bmatrix} B & I_n \\ 0 & A \end{bmatrix}}$$

4.1.2 (b)

Source: Linear Algebra 5e, Strang

Let

$$M = \begin{bmatrix} I_n & 0 \\ 0 & AB \end{bmatrix}$$

$$N = \begin{bmatrix} B & I_n \\ 0 & A \end{bmatrix}$$

We can see that

$$\text{rank}(M) = n + \text{rank}(AB)$$

because the identity columns are linearly independent and have full rank, and the AB columns have as many independent columns as AB . Also

$$\text{rank}(N) \geq \text{rank}(A) + \text{rank}(B)$$

Where the inequality comes from the fact that A can have at most n linearly independent columns, but the identity matrix above forces the columns containing A to be linearly independent. Also, elementary row operations preserve rank, therefore

$$\text{rank}(M) = \text{rank}(N)$$

and therefore

$$\text{rank}(A) + \text{rank}(B) \leq \text{rank}(N) = \text{rank}(M) = n + \text{rank}(AB)$$

$$\text{rank}(A) + \text{rank}(B) - n \leq \text{rank}(AB)$$

Consider a vector $u \in \mathbb{R}^p$

$$\dim((AB)u) = \text{rank}(AB)$$

and

$$\dim(A(Bu)) \leq \text{rank}(A)$$

because $Bu \in C(A)$, and because multiplication with B may be reducing the rank. Therefore

$$\text{rank}(AB) \leq \text{rank}(A)$$

Also

$$\dim(Bu) = \text{rank}(B)$$

and

$$\dim(A(Bu)) \leq \text{rank}(B)$$

because multiplication by A may reduce the rank, as $A(Bu)$ is a linear combination of the columns of A , of which, at most n are independent, which is the dimension of (Bu) . Therefore

$$\text{rank}(AB) \leq \text{rank}(B)$$

Because

$$\text{rank}(AB) \leq \text{rank}(A) \quad \& \quad \text{rank}(AB) \leq \text{rank}(B)$$

$$\text{rank}(AB) \leq \min\{\text{rank}(A), \text{rank}(B)\}$$

and therefore

$$\text{rank}(A) + \text{rank}(B) - n \leq \text{rank}(AB) \leq \min\{\text{rank}(A), \text{rank}(B)\}$$

Q.E.D.

4.1.3 (c)

For a matrix $A \in \mathbb{R}^{m \times n}$ with columns A_1, A_2, \dots, A_n and $v \in \mathbb{R}^n$

$$Av = v_1 A_1 + v_2 A_2 + \dots + v_n A_n$$

In other words, Av is just a linear combination of the columns of A , weighted by the components of v , and Av is therefore in the column space of A . Now consider

$$A^T A = \left[\begin{array}{c|c|c|c|c|c} & & & & & \\ A^T A_1 & A^T A_2 & \cdot & \cdot & \cdot & A^T A_n \\ & & & & & \end{array} \right]$$

We see this is a new matrix where each column is a linear combination of the columns of A^T (rows of A), each weighted by the components of the respective column vector. Thus, each column of $A^T A$ is in the row space of A , and thus $A^T A$ can only have as many linearly independent columns as the number of linearly independent rows of A .

In other words

$$C(A^T A) = C(A^T)$$

and therefore

$$\text{rank}(A^T A) = \text{rank}(A^T) = \text{rank}(A)$$

Because the column rank is equal to the row rank of a matrix.

4.2

For a PSD matrix $A \in \mathbb{R}^{n \times n}$ we have

$$\text{For all } x \in \mathbb{R}^n, x^T A x \geq 0$$

Then for every eigenvector $u_i \in \mathbb{R}^n$ with eigenvalue λ_i we know

$$\begin{aligned} u_i^T A u_i &= u_i^T \lambda_i u_i \\ &= \lambda_i u_i^T u_i \\ &= \lambda_i \|u_i\|^2 \end{aligned}$$

and

$$\lambda_i \|u_i\|^2 = u_i^T A u_i \geq 0$$

Which means all eigenvalues $\lambda_i \geq 0$. To show in the other direction, since A is a symmetric matrix, we know we can form a basis for the column space of A with an orthonormal set of eigenvectors. Let u_1, u_2, \dots, u_n be the orthonormal eigenvectors of A , with corresponding eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_n$ such that $\lambda_i \geq 0, \forall i \in \{1, \dots, n\}$. Then we can say for any vector x in the column space of A and $a_1, a_2, \dots, a_n \in \mathbb{R}$

$$x = a_1 u_1 + a_2 u_2 + \dots + a_n u_n$$

and

$$\begin{aligned} x^T A x &= \sum_{i=1}^n a_i u_i^T A \sum_{i=1}^n a_i u_i \\ &= \sum_{i=1}^n a_i u_i^T \sum_{i=1}^n a_i \lambda_i u_i \\ &= \sum_{i=1}^n a_i^2 \lambda_i \|u_i\|^2 + \sum_{i \neq j} a_i a_j \lambda_i u_i^T u_j \end{aligned}$$

But since $\|u_i\|^2 = 1, \forall i \in \{1, \dots, n\}$, and $u_i^T u_j = 0, \forall i \neq j$ because the vectors are orthonormal

$$\begin{aligned} \sum_{i=1}^n a_i^2 \lambda_i \|u_i\|^2 + \sum_{i \neq j} a_i a_j \lambda_i u_i^T u_j &= \sum_{i=1}^n a_i^2 \lambda_i \\ &\geq 0 \end{aligned}$$

Therefore

For all $x \in \mathbb{R}^n$, $x^T A x \geq 0 \iff$ All eigenvalues of A are nonnegative

Suppose we can write A as $A = U U^T$, then

$$x^T A x = x^T U U^T x$$

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

$$\|U^T x\|^2 \geq 0, \forall x \in \mathbb{R}^n$$

Therefore $A = U U^T \Rightarrow x^T A x \geq 0$. Now suppose

$$x^T A x \geq 0, \forall x \in \mathbb{R}^n$$

Since A is symmetric we can say

$$A = Q \Lambda Q^T$$

Where Q is an orthonormal matrix. We showed that $x^T A x \geq 0$ implies all positive eigenvalues. Since we have all positive eigenvalues, we can take their square roots. Let there be a matrix S such that

$$S^2 = \Lambda$$

and let there be an $n \times n$ matrix U defined as

$$U = Q S$$

Then

$$U U^T = Q S (Q S)^T$$

$$= Q S S^T Q^T$$

$$= QS^2Q^T$$

$$= Q\Lambda Q^T$$

$$= A$$

Therefore, $x^T Ax \geq 0$, $\forall x \in \mathbb{R}^n$ implies there exists U such that $A = UU^T$ and therefore

$$A = UU^T \iff x^T Ax \geq 0, \forall x \in \mathbb{R}^n \iff \text{all eigenvalues of } A \text{ are } \geq 0$$

Q.E.D.

4.3

4.3.1 (a)

Note that

$$xy^T = \begin{bmatrix} x_1y_1 & x_1y_2 & \dots & x_1y_n \\ x_2y_1 & x_2y_2 & \dots & x_2y_n \\ \cdot & \cdot & & \cdot \\ \cdot & & \cdot & \cdot \\ \cdot & & & \cdot \\ x_my_1 & x_my_2 & \dots & x_my_n \end{bmatrix}$$

And

$$A^T = \begin{bmatrix} a_{11} & a_{21} & \dots & a_{m1} \\ a_{12} & a_{22} & \dots & a_{m2} \\ \cdot & \cdot & & \cdot \\ \cdot & & \cdot & \cdot \\ \cdot & & & \cdot \\ a_{1n} & a_{2n} & \dots & a_{mn} \end{bmatrix}$$

By definition

$$\langle A, xy^T \rangle = \text{Tr}(A^T xy^T)$$

And

$$A^T xy^T = \begin{bmatrix} y_1 \sum_i^m a_{i1}x_i & y_2 \sum_i^m a_{i1}x_i & \dots & y_n \sum_i^m a_{i1}x_i \\ y_1 \sum_i^m a_{i2}x_i & y_2 \sum_i^m a_{i2}x_i & \dots & y_n \sum_i^m a_{i2}x_i \\ \cdot & \cdot & & \cdot \\ \cdot & & \cdot & \cdot \\ \cdot & & & \cdot \\ y_1 \sum_i^m a_{in}x_i & y_2 \sum_i^m a_{in}x_i & \dots & y_n \sum_i^m a_{in}x_i \end{bmatrix}$$

Therefore

$$\text{Tr}(A^T xy^T) = y_1 \sum_i^m a_{i1}x_i + y_2 \sum_i^m a_{i2}x_i + \dots + y_n \sum_i^m a_{in}x_i$$

$$= \begin{bmatrix} \sum_i^m a_{i1}x_i & \sum_i^m a_{i2}x_i & \dots & \sum_i^m a_{in}x_i \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ y_n \end{bmatrix}$$

$$= \begin{bmatrix} x^T A_1 & x^T A_2 & \dots & x^T A_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ y_n \end{bmatrix}$$

Where $A_i, \quad i \in \{1, 2, \dots, n\}$ are the columns of A

$$= x^T \begin{bmatrix} A_1 & A_2 & \dots & A_n \end{bmatrix} y$$

$$= x^T A y$$

Q.E.D.

4.3.2 (b)

Note that for a matrices $A, B \in \mathbb{R}^{m \times n}$ with columns (i.e. column vectors) A_1, A_2, \dots, A_n and B_1, B_2, \dots, B_n , where $A_i = (A_{1i}, A_{2i}, \dots, A_{mi})^T$ and $B_i = (B_{1i}, B_{2i}, \dots, B_{mi})^T$

$$\begin{aligned}
 A^T B &= \begin{bmatrix} -A_1^T - \\ -A_2^T - \\ \cdot \\ \cdot \\ \cdot \\ -A_n^T - \end{bmatrix} \begin{bmatrix} | & | & \dots & | \\ B_1 & B_2 & \dots & B_n \\ | & | & & | \end{bmatrix} \\
 &= \begin{bmatrix} A_1^T B_1 & A_1^T B_2 & \dots & A_1^T B_n \\ A_2^T B_1 & A_2^T B_2 & \dots & A_2^T B_n \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & & \cdot \\ A_n^T B_1 & A_n^T B_2 & \dots & A_n^T B_n \end{bmatrix}
 \end{aligned}$$

Where

$$A_i^T B_j = A_{1i} B_{1j} + A_{2i} B_{2j} + \dots + A_{mi} B_{mj}$$

The inner product of A and B is

$$\langle A, B \rangle = \text{Tr}(A^T B)$$

From the above expansion we get

$$= A_1^T B_1 + A_2^T B_2 + \dots + A_n^T B_n$$

$$= \sum_{i=1}^n A_i^T B_i$$

By Cauchy-Schwarz

$$\left(\sum_{i=1}^n A_i^T B_i \right)^2 \leq \sum_{i=1}^n \|A_i\|^2 \sum_{i=1}^n \|B_i\|^2$$

Or in other words

$$\text{Tr}(A^T B) \leq \sqrt{\sum_{i=1}^n \|A_i\|^2 \sum_{i=1}^n \|B_i\|^2}$$

Note that

$$\begin{aligned}
\|A\|_F &= \sqrt{\sum_{i=1}^m \sum_{j=1}^n |A_{ij}|^2} \\
&= \sqrt{\sum_{i=1}^m (|A_{i1}|^2 + |A_{i2}|^2 + \dots + |A_{in}|^2)} \\
&= \sqrt{\sum_{i=1}^m |A_{i1}|^2 + \sum_{i=1}^m |A_{i2}|^2 + \dots + \sum_{i=1}^m |A_{in}|^2} \\
&= \sqrt{\|A_1\|^2 + \|A_2\|^2 + \dots + \|A_n\|^2} \\
&= \sqrt{\sum_{i=1}^n \|A_i\|^2}
\end{aligned}$$

In other words the Frobenius norm is the sum of the squares of the elements, which is equivalently the sum of the squares of the column norms. Therefore we see

$$\begin{aligned}
\text{Tr}(A^T B) &\leq \sqrt{\sum_{i=1}^n \|A_i\|^2 \sum_{i=1}^n \|B_i\|^2} \\
&= \sqrt{\sum_{i=1}^n \|A_i\|^2} \sqrt{\sum_{i=1}^n \|B_i\|^2} \\
&= \|A\|_F \|B\|_F
\end{aligned}$$

Therefore

$$\langle A, B \rangle \leq \|A\|_F \|B\|_F$$

Q.E.D.

4.3.3 (c)

Sources: 189 Discord (CS 189 SP 23), Math StackExchange and Wikipedia

Because A is PSD (and symmetric) it is diagonalizable and can be decomposed into

$$A = U\Lambda U^T$$

where U and U^T are orthonormal and Λ is a diagonal matrix of the non-negative eigenvalues of A . Note that

$$\begin{aligned} (U\Lambda^{\frac{1}{2}}U^T)^2 &= U\Lambda^{\frac{1}{2}}U^T U\Lambda^{\frac{1}{2}}U^T \\ &= U\Lambda^{\frac{1}{2}}\Lambda^{\frac{1}{2}}U^T \\ &= U\Lambda U^T \\ &= A \end{aligned}$$

Therefore, A has a defined square root and we can say

$$\begin{aligned} \text{Tr}(AB) &= \text{Tr}(A^{\frac{1}{2}}A^{\frac{1}{2}}B) \\ &= \text{Tr}(A^{\frac{1}{2}}BA^{\frac{1}{2}}) \end{aligned}$$

by the cyclic property of the trace. Note that

$$\begin{aligned} x^T(A^{\frac{1}{2}}BA^{\frac{1}{2}})x &= (x^TA^{\frac{1}{2}})B(A^{\frac{1}{2}}x) \\ (A^{\frac{1}{2}}x)^TB(A^{\frac{1}{2}}x) &\geq 0 \end{aligned}$$

Because B is PSD, and therefore $A^{\frac{1}{2}}BA^{\frac{1}{2}}$ is PSD and has all non-negative eigenvalues. Say $A^{\frac{1}{2}}BA^{\frac{1}{2}}$ has eigenvalues μ_i , then

$$\begin{aligned} \text{Tr}(A^{\frac{1}{2}}BA^{\frac{1}{2}}) &= \sum_i \mu_i \\ &\geq 0 \end{aligned}$$

Therefore, if A and B are PSD

$$\text{Tr}(AB) \geq 0$$

Q.E.D.

4.3.4 (d)

Consider the matrix

$$\lambda_{\max}(A)I_n$$

which is PSD because

$$\begin{aligned} x^T \lambda_{\max}(A)I_n x &= \lambda_{\max}(A) \|x\|^2 \\ &\geq 0, \quad \forall x \end{aligned}$$

Also, note that since A is symmetric, by the spectral theorem it is diagonalizable and

$$\begin{aligned} x^T (\lambda_{\max}(A)I_n - A)x &= x^T \lambda_{\max}(A)I_n x - x^T A x \\ &= \lambda_{\max}(A) \|x\|^2 - x^T (Q \Lambda Q^T) x \end{aligned}$$

where Λ is the diagonal matrix containing the eigenvalues of A

$$= \lambda_{\max}(A) \|x\|^2 - (Q^T x)^T \Lambda (Q^T x)$$

Let $v = Q^T x$

$$= \lambda_{\max}(A) \|x\|^2 - (\lambda_{\max}(A)v_1^2 + \dots + \lambda_{\min}(A)v_n^2)$$

But

$$\begin{aligned} \|v\|^2 &= \|Q^T x\|^2 \\ &= (Q^T x)^T (Q^T x) \\ &= x^T Q Q^T x \\ &= \|x\|^2 \end{aligned}$$

Therefore, it is easy to see

$$\begin{aligned} \lambda_{\max}(A) \|x\|^2 &= \lambda_{\max}(A) \|v\|^2 \\ &\geq \lambda_{\max}(A)v_1^2 + \dots + \lambda_{\min}(A)v_n^2 \end{aligned}$$

Therefore

$$\lambda_{\max}(A) \|x\|^2 - x^T A x \geq 0, \quad \forall x$$

And therefore

$$\lambda_{\max}(A)I_n - A$$

is PSD. Now, consider

$$\text{Tr}((\lambda_{\max}(A)I_n - A)B) \geq 0$$

Which we proved in 4.3 (c) for the trace of two PSD matrices. Then

$$\text{Tr}(\lambda_{\max}(A)I_n B - AB) \geq 0$$

$$\text{Tr}(\lambda_{\max}(A)I_n B) - \text{Tr}(AB) \geq 0$$

$$\text{Tr}(AB) \leq \text{Tr}(\lambda_{\max}(A)I_n B)$$

$$\text{Tr}(A^T B) \leq \text{Tr}((\lambda_{\max}(A)I_n)^T B)$$

$$\langle A, B \rangle \leq \|\lambda_{\max}(A)I_n\|_F \|B\|_F$$

$$= |\lambda_{\max}(A)| \|I_n\| \|B\|_F$$

$$= \lambda_{\max}(A) \sqrt{\sum_j^n \sum_i^n |I_{ij}|^2} \|B\|_F$$

$$= \sqrt{n} \lambda_{\max}(A) \|B\|_F$$

Q.E.D.

4.4

Source: Math StackExchange, Linear Algebra 5e by Strang

Lemma 4.4.1:

Let A be a matrix with the appropriate dimensions. Then

$$A^T(M - N)A$$

is PSD if $M - N$ is PSD.

Proof: Let x be any vector with the correct dimensions. Then

$$x^T(A^T(M - N)A)x = (Ax)^T(M - N)(Ax)$$

$$(Ax)^T(M - N)(Ax) \geq 0$$

Because $M - N$ is PSD. Therefore,

$$A^T(M - N)A \succeq 0$$

Q.E.D.

Lemma 4.4.2

For matrices A and B which have inverses, $AB \sim BA$. Proof:

$$AB = ABI$$

$$AB = A(BA)A^{-1}$$

Let $A^{-1} = M$, then

$$AB = M^{-1}(BA)M$$

Q.E.D.

$N^{-1} - M^{-1}$ is PSD. Proof:

Since N is positive definite, N^{-1} is also positive definite because its eigenvalues are the reciprocals of the eigenvalues of N , which, by definition of positive definiteness, are all positive. Positive definite matrices have defined square roots since they are symmetric and thus diagonalizable. The same can be said for M since it is also positive definite. Consider then

$$(N^{-\frac{1}{2}})^T(M - N)N^{-\frac{1}{2}} \succeq 0$$

i.e. it is PSD by lemma 4.4.1. Note that

$$(N^{-\frac{1}{2}})^T(M - N)N^{-\frac{1}{2}} = N^{-\frac{1}{2}}(M - N)N^{-\frac{1}{2}}$$

because $N^{-\frac{1}{2}}$ is also symmetric. So

$$N^{-\frac{1}{2}}(M - N)N^{-\frac{1}{2}} \succeq 0$$

$$N^{-\frac{1}{2}}MN^{-\frac{1}{2}} - N^{-\frac{1}{2}}NN^{-\frac{1}{2}} \succeq 0$$

$$N^{-\frac{1}{2}}MN^{-\frac{1}{2}} - (N^{-\frac{1}{2}}N^{\frac{1}{2}})(N^{\frac{1}{2}}N^{-\frac{1}{2}}) \succeq 0$$

$$N^{-\frac{1}{2}}MN^{-\frac{1}{2}} - I \succeq 0$$

$$(N^{-\frac{1}{2}}M^{\frac{1}{2}})(M^{\frac{1}{2}}N^{-\frac{1}{2}}) - I \succeq 0$$

Temporarily let $A = N^{-\frac{1}{2}}M^{\frac{1}{2}}$ and $B = M^{\frac{1}{2}}N^{-\frac{1}{2}}$. Then we have

$$AB - I \succeq 0$$

By lemma 4.4.2, BA is similar to AB and similar matrices have the same eigenvalues. Therefore, if the above is PSD, then it is also true that

$$BA - I \succeq 0$$

$$(M^{\frac{1}{2}}N^{-\frac{1}{2}})(N^{-\frac{1}{2}}M^{\frac{1}{2}}) - I \succeq 0$$

$$M^{\frac{1}{2}}(N^{-\frac{1}{2}}N^{-\frac{1}{2}})M^{\frac{1}{2}} - I \succeq 0$$

$$M^{\frac{1}{2}}N^{-1}M^{\frac{1}{2}} - I \succeq 0$$

$$M^{\frac{1}{2}}N^{-1}M^{\frac{1}{2}} - (M^{\frac{1}{2}}M^{-\frac{1}{2}})(M^{-\frac{1}{2}}M^{\frac{1}{2}}) \succeq 0$$

$$M^{\frac{1}{2}}N^{-1}M^{\frac{1}{2}} - M^{\frac{1}{2}}(M^{-\frac{1}{2}}M^{-\frac{1}{2}})M^{\frac{1}{2}} \succeq 0$$

$$M^{\frac{1}{2}}N^{-1}M^{\frac{1}{2}} - M^{\frac{1}{2}}M^{-1}M^{\frac{1}{2}} \succeq 0$$

$$M^{\frac{1}{2}}(N^{-1} - M^{-1})M^{\frac{1}{2}} \succeq 0$$

Q.E.D.

4.5

Sources: Wikipedia (https://en.wikipedia.org/wiki/Operator_norm),

Ed discussion

ChatGPT

Note that

$$\sigma_{\max}(A) = \|A\|_{op}$$

Where

$$\|A\|_{op} = \inf\{c \geq 0 : \|Av\| \leq c\|v\| \quad \forall v \in V\}$$

Since $\|v\| = 1$

$$\|A\|_{op} = \inf\{c \geq 0 : \|Av\| \leq c \quad \forall v \in V\}$$

Thus, there is some $c = \sigma_{\max}(A)$ and by the condition for the operator norm, we know that

$$\|Av\|_{\|v\|=1} \leq \sigma_{\max}(A)$$

By Cauchy-Schwarz, we also know

$$u^T Av \leq \|u\| \|Av\|$$

$$= \|Av\|_{\|u\|=1}$$

because $\|u\| = 1$. Therefore,

$$u^T Av_{\|u\|=1} \leq \sigma_{\max}(A)$$

However, it is also true that

$$\sigma_{\max}(A) = \sup_{v \neq 0} \frac{\|Av\|_2}{\|v\|_2}$$

$$= \sup_{\|v\|=1} \|Av\|$$

because $\|v\|_2 = 1$. Putting it all together

$$u^T Av_{\|u\|=1} \leq \|Av\|_{\|v\|=1} \leq \sup_{\|v\|=1} \|Av\| = \sigma_{\max}(A)$$

There can only be equality with $u^T Av$ when

$$\sigma_{\max}(A) = \max_{u \in \mathbb{R}^m, v \in \mathbb{R}^n, \|u\|=1, \|v\|=1} (u^T A v)$$

Q.E.D.

5 Matrix/Vector Calculus and Norms

5.1

Sources: The Matrix Cookbook,
Wikipedia (https://en.wikipedia.org/wiki/Matrix_calculus).

Let

$$w = \sin(A_{11}^2 + e^{A_{11}+A_{22}}) + x^T Ay$$

and

$$\frac{\partial}{\partial A} = \begin{bmatrix} \frac{\partial}{\partial A_{11}} & \frac{\partial}{\partial A_{12}} \\ \frac{\partial}{\partial A_{21}} & \frac{\partial}{\partial A_{22}} \end{bmatrix}$$

Then

$$\frac{\partial w}{\partial A} = \frac{\partial}{\partial A} \sin(A_{11}^2 + e^{A_{11}+A_{22}}) + \frac{\partial}{\partial A} (x^T Ay)$$

$$= \begin{bmatrix} \frac{\partial}{\partial A_{11}} \sin(A_{11}^2 + e^{A_{11}+A_{22}}) & \frac{\partial}{\partial A_{12}} \sin(A_{11}^2 + e^{A_{11}+A_{22}}) \\ \frac{\partial}{\partial A_{21}} \sin(A_{11}^2 + e^{A_{11}+A_{22}}) & \frac{\partial}{\partial A_{22}} \sin(A_{11}^2 + e^{A_{11}+A_{22}}) \end{bmatrix} + xy^T$$

$$= \begin{bmatrix} (2A_{11} + e^{A_{11}+A_{22}}) \cos(A_{11}^2 + e^{A_{11}+A_{22}}) & 0 \\ 0 & (e^{A_{11}+A_{22}}) \cos(A_{11}^2 + e^{A_{11}+A_{22}}) \end{bmatrix} + \begin{bmatrix} x_1 y_1 & x_1 y_2 \\ x_2 y_1 & x_2 y_2 \end{bmatrix}$$

$$= \begin{bmatrix} x_1 y_1 + (2A_{11} + e^{A_{11}+A_{22}}) \cos(A_{11}^2 + e^{A_{11}+A_{22}}) & x_1 y_2 \\ x_2 y_1 & x_2 y_2 + (e^{A_{11}+A_{22}}) \cos(A_{11}^2 + e^{A_{11}+A_{22}}) \end{bmatrix}$$

5.2

5.2.1 (a)

Sources: Wikipedia (https://en.wikipedia.org/wiki/Rayleigh_quotient)

$$\begin{aligned}\|A\|_2 &= \sup_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2} \\ &= \sup_{x \neq 0} \frac{\sqrt{(Ax)^T(Ax)}}{\sqrt{x^T x}} \\ &= \sup_{x \neq 0} \sqrt{\frac{x^T A^T A x}{x^T x}}\end{aligned}$$

Because $A^T A = (A^T A)^T$ (i.e. it is symmetric) the Rayleigh quotient can be used i.e.

$$\frac{x^T A^T A x}{x^T x} = R(A^T A, x) \leq \lambda_{\max}(A^T A)$$

Therefore

$$\begin{aligned}\sup_{x \neq 0} \sqrt{\frac{x^T A^T A x}{x^T x}} &\leq \sup_{x \neq 0} \sqrt{\lambda_{\max}(A^T A)} \\ &= \sqrt{\lambda_{\max}(A^T A)} \\ &= \sqrt{\sigma_{\max}^2(A)}\end{aligned}$$

Where $\sigma_{\max}(A)$ is the largest singular value of A

$$= \boxed{\sigma_{\max}(A)}$$

5.2.2 (b)

$$\begin{aligned}\|A\|_\infty &= \sup_{x \neq 0} \frac{\|Ax\|_\infty}{\|x\|_\infty} \\ &= \sup_{x \neq 0} \frac{\max_{1 \leq i \leq m} \sum_j^n |A_{ij}x_j|}{\max_{1 \leq j \leq n} |x_j|}\end{aligned}$$

Note that

$$\begin{aligned}\sum_j^n |A_{ij}x_j| &= |A_{i1}x_1| + |A_{i2}x_2| + \dots + |A_{in}x_n| \\ &\leq (|A_{i1}| + |A_{i2}| + \dots + |A_{in}|) (|x_1| + |x_2| + \dots + |x_n|) \\ &= \sum_j^n |A_{ij}| \sum_j |x_j|\end{aligned}$$

Therefore

$$\sup_{x \neq 0} \frac{\max_{1 \leq i \leq m} \sum_j^n |A_{ij}x_j|}{\max_{1 \leq j \leq n} |x_j|} \leq \sup_{x \neq 0} \frac{\max_{1 \leq i \leq m} \left(\sum_j^n |A_{ij}| \sum_j |x_j| \right)}{\max_{1 \leq j \leq n} |x_j|}$$

Note that for a vector x , the sum of its components is a constant so

$$\begin{aligned}&\sum_j^n |x_j| \max_{1 \leq i \leq m} \sum_j^n |A_{ij}| \\ &= \sup_{x \neq 0} \frac{\sum_j^n |x_j| \max_{1 \leq i \leq m} \sum_j^n |A_{ij}|}{\max_{1 \leq j \leq n} |x_j|}\end{aligned}$$

Note that

$$\max_{1 \leq j \leq n} |x_j| \leq \sum_j^n |x_j|$$

The supremum is the least upper bound, which occurs when the fraction is the smallest, which occurs when $\max_{1 \leq j \leq n} |x_j| = \sum_j^n |x_j|$. Therefore

$$\begin{aligned}
 \sup_{x \neq 0} \frac{\sum_j^n |x_j| \max_{1 \leq i \leq m} \sum_j^n |A_{ij}|}{\max_{1 \leq j \leq n} |x_j|} &= \frac{\sum_j^n |x_j| \max_{1 \leq i \leq m} \sum_j^n |A_{ij}|}{\sum_j^n |x_j|} \\
 &= \boxed{\max_{1 \leq i \leq m} \sum_j^n |A_{ij}|}
 \end{aligned}$$

5.3

5.3.1 (a)

If we have

$$\alpha = \sum_{i=1}^n y_i \ln \beta_i \text{ for } y, \beta \in \mathbb{R}^n$$

Then for $i = k$ for $1 \leq k \leq n$

$$\begin{aligned} \frac{\partial \alpha}{\partial \beta_k} &= \frac{\partial}{\partial \beta_k} (y_1 \ln \beta_1 + y_2 \ln \beta_2 + \dots + y_k \ln \beta_k + \dots + y_n \ln \beta_n) \\ &= \frac{\partial y_1}{\partial \beta_k} \ln \beta_1 + \frac{\partial y_2}{\partial \beta_k} \ln \beta_2 + \dots + \left(\frac{y_k}{\beta_k} + \frac{\partial y_k}{\partial \beta_k} \ln \beta_k \right) + \dots + \frac{\partial y_n}{\partial \beta_k} \ln \beta_n \\ &= \frac{y_k}{\beta_k} + \frac{\partial y_k}{\partial \beta_k} \ln \beta_k + \sum_{j \neq k} \frac{\partial y_j}{\partial \beta_k} \ln \beta_j, \text{ for } 1 \leq j \leq n \end{aligned}$$

Therefore

$$\frac{\partial \alpha}{\partial \beta_i} = \frac{y_i}{\beta_i} + \frac{\partial y_i}{\partial \beta_i} \ln \beta_i + \sum_{j \neq i} \frac{\partial y_j}{\partial \beta_i} \ln \beta_j, \text{ for } 1 \leq j \leq n$$

However, if y is independent of β then this just simplifies to

$$\boxed{\frac{\partial \alpha}{\partial \beta_i} = \frac{y_i}{\beta_i}}$$

5.3.2 (b)

$$\gamma_i = A_{i,*}\rho + b_i$$

$$\frac{\partial \gamma_i}{\partial \rho_j} = \frac{\partial}{\partial \rho_j} A_{i,*}\rho + \frac{\partial b_i}{\partial \rho_j}$$

$$= \frac{\partial}{\partial \rho_j} (A_{i1}\rho_1 + A_{i2}\rho_2 + \dots + A_{im}\rho_m) + \frac{\partial b_i}{\partial \rho_j}$$

$$= \frac{\partial}{\partial \rho_j} A_{i1}\rho_1 + \dots + (A_{ij} + \rho_j \frac{\partial A_{ij}}{\partial \rho_j}) + \dots + \frac{\partial}{\partial \rho_j} A_{im}\rho_m + \frac{\partial b_i}{\partial \rho_j}$$

$$= A_{ij} + \rho_j \frac{\partial A_{ij}}{\partial \rho_j} + \sum_{k \neq j} \left(A_{ik} \frac{\partial \rho_k}{\partial \rho_j} + \rho_k \frac{\partial A_{ik}}{\partial \rho_j} \right) + \frac{\partial b_i}{\partial \rho_j}$$

However, if A, b are independent of ρ then this just simplifies to

$$\boxed{\frac{\partial \gamma_i}{\partial \rho_j} = A_{ij}}$$

5.3.3 (c)

$$y = \begin{bmatrix} f_1(x_1, x_2, \dots, x_n) \\ f_2(x_1, x_2, \dots, x_n) \\ \cdot \\ \cdot \\ \cdot \\ f_m(x_1, x_2, \dots, x_n) \end{bmatrix}$$

$$J_y(x) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdot & \cdot & \cdot & \frac{\partial f_1}{\partial x_n} \\ \cdot & \cdot & & & & \cdot \\ \cdot & & \cdot & & & \cdot \\ \cdot & & & & \cdot & \cdot \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \cdot & \cdot & \cdot & \frac{\partial f_m}{\partial x_n} \end{bmatrix}$$

$$z = \begin{bmatrix} g_1(y_1, y_2, \dots, y_m) \\ g_2(y_1, y_2, \dots, y_m) \\ \cdot \\ \cdot \\ \cdot \\ g_k(y_1, y_2, \dots, y_m) \end{bmatrix}$$

$$J_z(y) = \begin{bmatrix} \frac{\partial g_1}{\partial y_1} & \frac{\partial g_1}{\partial y_2} & \cdot & \cdot & \cdot & \frac{\partial g_1}{\partial y_m} \\ \cdot & \cdot & & & & \cdot \\ \cdot & & \cdot & & & \cdot \\ \cdot & & & & \cdot & \cdot \\ \frac{\partial g_k}{\partial y_1} & \frac{\partial g_k}{\partial y_2} & \cdot & \cdot & \cdot & \frac{\partial g_k}{\partial y_m} \end{bmatrix}$$

$$\left(\frac{\partial g_1}{\partial f_1} \frac{\partial f_1}{\partial x_1} + \frac{\partial g_1}{\partial f_2} \frac{\partial f_2}{\partial x_1} + \dots + \frac{\partial g_1}{\partial f_m} \frac{\partial f_m}{\partial x_1} \right)$$

$$J_z(x) = \begin{bmatrix} \left(\frac{\partial g_1}{\partial f_1} \frac{\partial f_1}{\partial x_1} + \frac{\partial g_1}{\partial f_2} \frac{\partial f_2}{\partial x_1} + \dots + \frac{\partial g_1}{\partial f_m} \frac{\partial f_m}{\partial x_1} \right) & \cdot & \cdot & \cdot & \left(\frac{\partial g_1}{\partial f_1} \frac{\partial f_1}{\partial x_n} + \frac{\partial g_1}{\partial f_2} \frac{\partial f_2}{\partial x_n} + \dots + \frac{\partial g_1}{\partial f_m} \frac{\partial f_m}{\partial x_n} \right) \\ \cdot & & & & \cdot \\ \cdot & & & \cdot & \cdot \\ \cdot & & & & \cdot \\ \left(\frac{\partial g_k}{\partial f_1} \frac{\partial f_1}{\partial x_1} + \frac{\partial g_k}{\partial f_2} \frac{\partial f_2}{\partial x_1} + \dots + \frac{\partial g_k}{\partial f_m} \frac{\partial f_m}{\partial x_1} \right) & \cdot & \cdot & \cdot & \left(\frac{\partial g_k}{\partial f_1} \frac{\partial f_1}{\partial x_n} + \frac{\partial g_k}{\partial f_2} \frac{\partial f_2}{\partial x_n} + \dots + \frac{\partial g_k}{\partial f_m} \frac{\partial f_m}{\partial x_n} \right) \end{bmatrix}$$

$$J_z(x) = \begin{bmatrix} \frac{\partial g_1}{\partial y_1} & \frac{\partial g_1}{\partial y_2} & \cdot & \cdot & \cdot & \frac{\partial g_1}{\partial y_m} \\ \cdot & \cdot & & & & \cdot \\ \cdot & & \cdot & & & \cdot \\ \cdot & & & \cdot & & \cdot \\ \frac{\partial g_k}{\partial y_1} & \frac{\partial g_k}{\partial y_2} & \cdot & \cdot & \cdot & \frac{\partial g_k}{\partial y_m} \end{bmatrix} \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdot & \cdot & \cdot & \frac{\partial f_1}{\partial x_n} \\ \cdot & \cdot & & & & \cdot \\ \cdot & & \cdot & & & \cdot \\ \cdot & & & \cdot & & \cdot \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \cdot & \cdot & \cdot & \frac{\partial f_m}{\partial x_n} \end{bmatrix}$$

5.3.4 (d)

Source: Math StackExchange

$$\nabla_x y^T z = (\nabla y) \cdot z + (\nabla z) \cdot y$$

$$= \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \cdot & \cdot & \cdot & \frac{\partial y_1}{\partial x_n} \\ \cdot & \cdot & & & \cdot \\ \cdot & & \cdot & & \cdot \\ \cdot & & & \cdot & \cdot \\ \frac{\partial y_m}{\partial x_1} & \cdot & \cdot & \cdot & \frac{\partial y_m}{\partial x_n} \end{bmatrix}^T \begin{bmatrix} z_1 \\ \cdot \\ \cdot \\ \cdot \\ z_m \end{bmatrix} + \begin{bmatrix} \frac{\partial z_1}{\partial x_1} & \cdot & \cdot & \cdot & \frac{\partial z_1}{\partial x_n} \\ \cdot & \cdot & & & \cdot \\ \cdot & & \cdot & & \cdot \\ \cdot & & & \cdot & \cdot \\ \frac{\partial z_m}{\partial x_1} & \cdot & \cdot & \cdot & \frac{\partial z_m}{\partial x_n} \end{bmatrix}^T \begin{bmatrix} y_1 \\ \cdot \\ \cdot \\ \cdot \\ y_m \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\partial y_1}{\partial x_1} z_1 + \dots + \frac{\partial y_m}{\partial x_1} z_m \\ \cdot \\ \cdot \\ \cdot \\ \frac{\partial y_1}{\partial x_n} z_1 + \dots + \frac{\partial y_m}{\partial x_n} z_m \end{bmatrix} + \begin{bmatrix} \frac{\partial z_1}{\partial x_1} y_1 + \dots + \frac{\partial z_m}{\partial x_1} y_m \\ \cdot \\ \cdot \\ \cdot \\ \frac{\partial z_1}{\partial x_n} y_1 + \dots + \frac{\partial z_m}{\partial x_n} y_m \end{bmatrix}$$

$$= \begin{bmatrix} \left(\frac{\partial y}{\partial x_1} \right)^T z + \left(\frac{\partial z}{\partial x_1} \right)^T y \\ \cdot \\ \cdot \\ \cdot \\ \left(\frac{\partial y}{\partial x_n} \right)^T z + \left(\frac{\partial z}{\partial x_n} \right)^T y \end{bmatrix}$$

$$\boxed{= \left(\frac{\partial y}{\partial x} \right)^T z + \left(\frac{\partial z}{\partial x} \right)^T y}$$

5.4

Sources: <https://people.math.sc.edu/josephcf/Teaching/142/Files/Worksheets/Estimation%20of%20the%20Taylor%20Remainder.pdf>

YouTube: https://www.youtube.com/watch?v=DP_pGQaNGdw&t=148s

Assuming f is twice differentiable within the sphere

$$f(x) = f(x^*) + f'(x^*)(x - x^*) + R_1(x)$$

$$f(x) = f(x^*) + R_1(x)$$

because by definition

$$f'(x^*) = 0$$

So

$$f(x) - f(x^*) = R_1(x)$$

By the Taylor remainder theorem

$$|f(x) - f(x^*)| \leq \frac{|x - x^*|^2}{2!} \max |f''(z)|$$

for some $z \in \mathcal{X}$

$$\frac{|x - x^*|^2}{2!} \max |f''(z)| \leq \frac{D}{2}$$

Because

$$\|x - x^*\| \leq D$$

and

$$f''(z) \leq 1, \quad \forall x, z \in \mathcal{X}$$

because

$$\lambda_{\max}(H(f(x))) = 1$$

Therefore

$$|f(x) - f(x^*)| \leq \frac{D}{2}$$

and since $D \geq 0$

$$f(x) - f(x^*) \leq \frac{D}{2}$$

Q.E.D.

5.5

Sources: The Matrix Cookbook

Note that

$$\begin{aligned}\frac{\partial L}{\partial w} &= \frac{\partial}{\partial w} \|y - Xw\|_2^2 \\ &= \frac{\partial}{\partial w} (y - Xw)^T (y - Xw) \\ &= -2X^T (y - Xw) \\ &= 2X^T Xw - 2X^T y\end{aligned}$$

We want

$$w^* = \arg \min_w L(w)$$

and

$$\arg \min_w L(w) = w^* \implies \frac{\partial L(w^*)}{\partial w} = 0$$

So

$$2X^T Xw^* - 2X^T y = 0$$

$$\boxed{w^* = (X^T X)^{-1} X^T y}$$

6 Gradient Descent

6.1

Sources: The Matrix Cookbook

Let

$$y = \frac{1}{2} x^T A x - b^T x$$

$$\frac{\partial y}{\partial x} = \frac{1}{2} \frac{\partial}{\partial x} (x^T A x) - \frac{\partial}{\partial x} (b^T x)$$

$$= \frac{1}{2} (A + A^T) x - b$$

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

Because A is PSD and therefore symmetric

$$= Ax - b$$

$$x^* = \min_{x \in \mathbb{R}^n} y(x) \implies \frac{\partial y(x^*)}{\partial x} = 0$$

$$Ax^* - b = 0$$

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

6.2

For a current position $x^{(k-1)}$ and next position $x^{(k)}$, with a step size = 1 our update function is

$$x^{(k)} = x^{(k-1)} - \frac{\partial y(x^{(k-1)})}{\partial x}$$

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

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

6.3

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

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

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

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

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

Q.E.D.

6.4

Source: Wikipedia page on Rayleigh quotient

Lemma 6.4

If A is PSD then $\lambda_{\max}(A^T A) = \lambda_{\max}^2(A)$. Proof:

Assume a matrix A is PSD and $A = A^T$. By the spectral theorem

$$A = U\Lambda U^T$$

Where U is orthonormal and Λ is a diagonal matrix containing the ordered eigenvalues of A . Then

$$\lambda_{\max}(A^T A) = \lambda_{\max}(U\Lambda U^T U\Lambda U^T)$$

$$= \lambda_{\max}(U\Lambda^2 U^T)$$

and

$$\Lambda^2 = \begin{bmatrix} \lambda_{\max}(A) & & & \\ & \ddots & & \\ & & \ddots & \\ & & & \lambda_{\min}(A) \end{bmatrix}^2$$

$$= \begin{bmatrix} \lambda_{\max}^2(A) & & & \\ & \ddots & & \\ & & \ddots & \\ & & & \lambda_{\min}^2(A) \end{bmatrix}$$

Therefore,

$$\lambda_{\max}(A^T A) = \lambda_{\max}^2(A)$$

Q.E.D.

Note that

$$\frac{\|Ax\|_2^2}{\|x\|_2^2} = \frac{(Ax)^T(Ax)}{x^T x}$$

$$= \frac{x^T A^T A x}{x^T x}$$

By the Rayleigh quotient

$$\frac{x^T A^T A x}{x^T x} \leq \lambda_{\max}(A^T A)$$

and because A is PSD

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

by lemma 6.4. Therefore

$$\frac{\|Ax\|_2^2}{\|x\|_2^2} \leq \lambda_{\max}^2(A)$$

$$\|Ax\|_2^2 \leq \lambda_{\max}^2(A) \|x\|_2^2$$

and

$$\sqrt{\|Ax\|_2^2} \leq \sqrt{\lambda_{\max}^2(A) \|x\|_2^2}$$

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

Q.E.D.

6.5

Lemma 6.5

For a matrix A with min eigenvalue $\lambda_{\min}(A)$, such that $0 < \lambda_{\min}(A) \leq \lambda_{\max}(A) < 1$, and associated eigenvector v_i

$$\lambda_{\max}(I - A) = 1 - \lambda_{\min}(A)$$

Proof:

$$(I - A)v = v - Av$$

$$= v - \lambda_{\min}(A)v$$

$$= (1 - \lambda_{\min}(A))v$$

and since $1 - \lambda_{\min}(A)$ is a constant, it must be an eigenvalue of $I - A$. Also

$$1 - 0 > 1 - \lambda_{\min}(A) \geq 1 - \lambda_{\max}(A) > 1 - 1$$

or

$$0 < 1 - \lambda_{\max}(A) \leq 1 - \lambda_{\min}(A) < 1$$

Therefore

$$\lambda_{\max}(I - A) = 1 - \lambda_{\min}(A)$$

Q.E.D.

From 6.3 we have

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

Then

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

By Cauchy-Schwarz

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

In 5.2 (a) we saw for any $m \times n$ matrix A

$$\|A\|_2 = \sigma_{\max}(A)$$

And from 4.5 we know

$$\sigma_{\max}^2(A) = \lambda_{\max}(A^T A)$$

and in 6.4 we saw that for PSD matrices A

$$\lambda_{\max}(A^T A) = \lambda_{\max}^2(A)$$

So since our A is PSD

$$\sigma_{\max}(A) = \lambda_{\max}(A)$$

and therefore

$$\|I - A\|_2 = \lambda_{\max}(I - A)$$

By Lemma 6.5

$$\|I - A\|_2 = 1 - \lambda_{\min}(A)$$

Putting it all together, we can say

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

becomes

$$\boxed{\|x^{(k)} - x^*\|_2 \leq (1 - \lambda_{\min}(A)) \|x^{(k-1)} - x^*\|_2}$$

Which has the form

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

for some $0 < \rho < 1$, since

$$0 < 1 - \lambda_{\min}(A) < 1$$

Therefore we see

$$\boxed{\rho = 1 - \lambda_{\min}(A)}$$

6.6

From 6.5 we see

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

Where $\rho = 1 - \lambda_{\min}(A)$. Let

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

Then

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

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

$$= \rho^3 \|x^{(k-2)} - x^*\|_2$$

.

.

.

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

Therefore, by induction

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

If we want,

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

for some $\epsilon > 0$, then we need

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

Which means we need

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

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

$$k \geq \frac{1}{\log(\rho)} \log \left(\frac{\epsilon}{\|x^{(0)} - x^*\|_2} \right)$$

$$k = \left\lceil \frac{1}{\log(\rho)} \log \left(\frac{\epsilon}{\|x^{(0)} - x^*\|_2} \right) \right\rceil$$