CS 189/289A Introduction to Machine Learning
Spring 2021 Jonathan Shewchuk

HW2: I ♥ Math

Due Wednesday, February 10 at 11:59 pm

- Homework 2 is an entirely written assignment; no coding involved.
- We prefer that you typeset your answers using LaTeX or other word processing software. If you haven't yet learned LaTeX, one of the crown jewels of computer science, now is a good time! Neatly handwritten and scanned solutions will also be accepted.
- In all of the questions, **show your work**, not just the final answer.
- **Start early. This is a long assignment. Most of the material is prerequisite material not covered in lecture; you are responsible for finding resources to understand it.**

Deliverables:

1. Submit a PDF of your homework to the Gradescope assignment entitled "HW2 Write-Up". You may typeset your homework in LaTeX or Word (submit PDF format, **not** .doc/.docx format) or submit neatly handwritten and scanned solutions. **Please start each question on a new page.** If there are graphs, include those graphs in the correct sections. **Do not** put them in an appendix. We need each solution to be self-contained on pages of its own.
    - In your write-up, please state whom you had discussions with (not counting course staff) about the homework contents.
    - In your write-up, please copy the following statement and sign your signature next to it. (Mac Preview and FoxIt PDF Reader, among others, have tools to let you sign a PDF file.) We want to make it *extra* clear so that no one inadvertently cheats.

      "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."

1 Identities with Expectation

For this exercise, the following identity might be useful: for a probability event A, $\mathbb{P}(A) = \mathbb{E}[\mathbf{1}\{A\}]$, where $\mathbf{1}\{\cdot\}$ is the indicator function.

1. Let X be a random variable with density $f(x) = \lambda e^{-\lambda x}\,\mathbf{1}\{x > 0\}$. Show that $\mathbb{E}[X^k] = \frac{k!}{\lambda^k}$ for integer $k \geq 0$. *Hint*: One way is to do induction on k.

2. For any non-negative random variable X and constant $t > 0$, show that $\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}$. *Hint*: show that for $a, b > 0$, $\mathbf{1}\{a \geq b\} \leq \frac{a}{b}$.

3. For any non-negative random variable X, prove the identity

$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X \geq t)dt.$$

   You may assume that X admits a density to simplify.

4. For any non-negative random variable X with finite variance (i.e., $\mathbb{E}[X^2] < \infty$), prove that

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

   Hint: Use the Cauchy–Schwarz inequality $\langle u, v\rangle^2 \leq \langle u, u\rangle\langle v, v\rangle$. You have most likely seen it applied when the inner product is the real dot product; however, it holds for arbitrary inner products. Without proof, use the fact that the expectation $\mathbb{E}[UV]$ is a valid inner product of random variables U and V.
   (Note that by assumption we know $\mathbb{P}(X \geq 0) = 1$, so this inequality is indeed quite powerful.)

5. For a random variable X with finite variance and $\mathbb{E}[X] = 0$, prove that

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X^2]}{\mathbb{E}[X^2] + t^2} \text{ for any } t \geq 0$$

   Hint: Try using logic similar to Question 1.4 on $t - X$.

Solution:

1. **Moment Generating Function.** Calculate the MGF: $M_X(t) = \mathbb{E}[e^{tX}] = \int_{x>0} \lambda e^{(t-\lambda)x}dx = \frac{1}{1-(t/\lambda)}$ if $t < \lambda$ and undefined otherwise. Since the MGF is defined in a neighborhood of 0 (specifically $|t| < \lambda$), all moments $\mathbb{E}[X^k]$ exist. Furthermore, from properties of the MGF, $\frac{\mathbb{E}[X^k]}{k!}$ is the coefficient of t^k. Expanding $\frac{1}{1-(t/\lambda)}$ as $\sum_{k\geq 0} \frac{1}{\lambda^k} t^k$ completes the solution.

   Induction. Base case: $\mathbb{E}X^0 = 1$. Inductive hypothesis: for $k > 0$, $\mathbb{E}X^k = \frac{k}{\lambda}\mathbb{E}X^{k-1}$. Inductive step: $\mathbb{E}X^k = \int_0^\infty \lambda x^k e^{-\lambda x}dx$. Let $u = x^k$ and $dv = \lambda e^{-\lambda x}$, so $du = kx^{k-1}$ and $v = -e^{-\lambda x}$. Then $\int_0^\infty \lambda x^k e^{-\lambda x}dx = [-x^k e^{-\lambda x}]_0^\infty + \int_0^\infty kx^{k-1}e^{-\lambda x}dx = 0 + \frac{k}{\lambda}\int_0^\infty \lambda x^{k-1}e^{-\lambda x}dx = \frac{k}{\lambda}\mathbb{E}X^{k-1}$, where the last equality comes from the inductive hypothesis. So $\mathbb{E}X^k = \Pi_{i=0}^k \frac{i}{\lambda} = \frac{k!}{\lambda^k}$. Note that the trick of separating out the k $(= \frac{k\lambda}{\lambda})$ factor in the second-to-last equality represents a generally useful approach for solving problems: figure out what form you want the problem to "look like" and try to transform it as close as possible to that form. Since we know we're dealing with induction, we know we would like to somehow obtain $\mathbb{E}X^{k-1}$ during the inductive step. By our assumption, $\mathbb{E}X^{k-1} = \int_0^\infty \lambda x^{k-1} e^{\lambda x} dx$. By keeping this in mind and paying close attention, we realize we can move a constant $\frac{k}{\lambda}$ outside the integral in the second to last equality, leaving behind the needed λ factor.

[RUBRIC: There could be other ways to solve this. Any completely correct solution gets **(+2 points)**. Any partially correct or incomplete solution gets **(+1 point)**.]

2. When $X \geq t$, $\mathbf{1}\{X \geq t\} = 1 \leq \frac{X}{t}$. On the other hand, when $X < t$, $\mathbf{1}\{X \geq t\} = 0 \leq \frac{X}{t}$ since X is non-negative. Take expectations on both sides to complete.

   [RUBRIC: A completely correct solution gets **(+1 point)**.]

3. First see that $X = \int_{t\geq 0} \mathbf{1}\{X \geq t\}dt$. Take expectation and use linearity of expectation: $\mathbb{E}[X] = \mathbb{E}\left[\int_{t\geq 0} \mathbf{1}\{X \geq t\}dt\right] = \int_{t\geq 0} \mathbb{E}[\mathbf{1}\{X \geq t\}]dt = \int_0^\infty \mathbb{P}(X \geq t)dt$. Note that X need not have a density function for this solution.

   Assuming density $f(x)$:

$$\begin{aligned}\mathbb{E}[X] &= \mathbb{E}\left[\int_0^\infty \mathbf{1}\{X \geq t\}dt\right] = \int_0^\infty \int_0^\infty \mathbf{1}\{x \geq t\}dt\, f(x)dx = \int_0^\infty \int_0^\infty \mathbf{1}\{x \geq t\} f(x)dx\, dt \\ &= \int_0^\infty \mathbb{P}(X \geq t)dt.\end{aligned}$$

   Again assuming density $f(x)$:

$$\mathbb{E}[X] = \int_0^\infty x f(x)dx = \int_0^\infty \int_0^x f(x)dtdx = \int_0^\infty \int_t^\infty f(x)dxdt = \int_0^\infty P(X \geq t)dt$$

   [RUBRIC: A completely correct solution gets **(+1 point)**.]

4. Using the non-negativity of X, we have $\mathbb{E}X = \mathbb{E}[X\mathbf{1}\{X > 0\}]$. [RUBRIC: **(+1 point)**]

   Now use Cauchy–Schwarz applied to $U := X$ and $V := \mathbf{1}\{X > 0\}$ to conclude that

$$(\mathbb{E}X)^2 = (\mathbb{E}[X\mathbf{1}\{X > 0\}])^2 \leq \mathbb{E}[X^2]\mathbb{E}[\mathbf{1}\{X > 0\}^2] = \mathbb{E}[X^2]\mathbb{E}[\mathbf{1}\{X > 0\}] = \mathbb{E}[X^2]\mathbb{P}(X > 0).$$

   [RUBRIC: Correct application of Cauchy–Schwarz Inequality gets **(+1 point)**.]

   [RUBRIC: **Total (+2 points)**.]

5. Using the same idea as in the previous part,

$$\mathbb{E}[t - X] \leq \mathbb{E}[(t - X)\mathbf{1}\{t - X > 0\}] = \mathbb{E}[(t - X)\mathbf{1}\{X < t\}].$$

   [RUBRIC: using indicators correctly and arriving at $\mathbb{E}[t - X] \leq \mathbb{E}[(t - X)\mathbf{1}\{X < t\}]$ gets **(+1 point)**.]

Now apply Cauchy–Schwarz to get

$$(\mathbb{E}[t-X])^2 \le (\mathbb{E}[(t-X)\mathbf{1}\{X<t\}])^2 \le \mathbb{E}[(t-X)^2]\mathbb{E}[\mathbf{1}\{X<t\}]. \tag{1}$$

[RUBRIC: Applying Cauchy–Schwarz on $t-X$ correctly gets **(+1 point)**.]

Evaluate the terms on the right-hand side and left-hand side separately. The LHS is

$$(\mathbb{E}[t-X])^2 = t^2$$

because $\mathbb{E}X = 0$. The first term on the RHS is

$$\mathbb{E}[(t-X)^2] = t^2 - 2t\mathbb{E}X + \mathbb{E}[X^2] = t^2 + \mathbb{E}[X^2].$$

The second term on the RHS is

$$\mathbb{E}[\mathbf{1}\{X<t\}] = \mathbb{P}(X<t) = 1 - \mathbb{P}(X \ge t).$$

Plugging these expressions back into equation (1) gives $t^2 \le (t^2 + \mathbb{E}[X^2])(1-\mathbb{P}(X \ge t))$, which after some rearranging gives $\mathbb{P}(X \ge t) \le \frac{\mathbb{E}[X^2]}{\mathbb{E}[X^2]+t^2}$ as desired.

[RUBRIC: Correctly substituting of $\mathbb{E}[X] = 0$ and simplifying gets **(+1 point)**.]

[RUBRIC: **Total (+3 points).**]

2 Probability Potpourri

1. Recall the covariance of two random variables X and Y is defined as $\text{Cov}(X,Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$. For a multivariate random variable Z (i.e., each index of Z is a random variable), we define the covariance matrix Σ with entries $\Sigma_{ij} = \text{Cov}(Z_i, Z_j)$. Concisely, $\Sigma = \mathbb{E}[(Z-\mu)(Z-\mu)^\top]$, where μ is the mean value of the (column) vector Z. Show that the covariance matrix is always positive semidefinite (PSD).

2. The probability that an archer hits her target when it is windy is 0.4; when it is not windy, her probability of hitting the target is 0.7. On any shot, the probability of a gust of wind is 0.3. Find the probability that

   (i) on a given shot there is a gust of wind and she hits her target.

   (ii) she hits the target with her first shot.

   (iii) she hits the target exactly once in two shots.

   (iv) there was no gust of wind on an occasion when she missed.

3. An archery target is made of 3 concentric circles of radii $1/\sqrt{3}$, 1 and $\sqrt{3}$ feet. Arrows striking within the inner circle are awarded 4 points, arrows within the middle ring are awarded 3 points, and arrows within the outer ring are awarded 2 points. Shots outside the target are awarded 0 points.

   Consider a random variable X, the distance of the strike from the center (in feet), and let the probability density function of X be

$$f(x) = \begin{cases} \frac{2}{\pi(1+x^2)} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

   What is the expected value of the score of a single strike?

4. A random variable Z is said to be drawn from the Poisson distribution with parameter $\lambda > 0$ if it takes values in non-negative integers with probability $\mathbb{P}(Z = k) = \frac{\lambda^k e^{-\lambda}}{k!}$. Let X and Y be two independent Poisson random variables with parameters $\lambda > 0$ and $\mu > 0$ respectively. Derive an expression for $\mathbb{P}(X \mid X + Y = n)$. What well-known probability distribution is this? What are its parameters?

Solution:

1. For $v \in \mathbb{R}^n$, $v^\top \mathbb{E}[(X-\mu)(X-\mu)^\top]v = \Sigma_{i=1}^n \Sigma_{j=1}^n \mathbb{E}[(X_i-\mu_i)(X_j-\mu_j)]v_i v_j = \mathbb{E}[v^\top(X-\mu)(X-\mu)^\top v] = \mathbb{E}[((X-\mu)^\top v)^2] \geq 0$. Note that the second identity comes from linearity of expectation.

   [RUBRIC: A completely correct solution gets **(+1 point)**.]

2. Denote with H the event that she hits her target, and with W the event that there is a gust of wind. Then we know that: $P(H \mid W) = 0.4$, $P(H \mid W^c) = 0.7$ and $P(W) = 0.3$.

   (i) $P(H \cap W) = P(H \mid W)P(W) = 0.12$

(ii) $P(H) = P(H \mid W)P(W) + P(H \mid W^c)P(W^c) = 0.61$

(iii) This probability is $\binom{2}{1}P(H)P(H^c) = 0.4758$

(iv) $P(W^c \mid H^c) = \frac{P(H^c|W^c)P(W^c)}{P(H^c)} = 0.538$

[RUBRIC: A correct derivation & answer to a sub-part gets **(+0.5 point)**. Total **(+2 points)**.]

3. The expected value is

$$
\begin{aligned}
&\int_0^{1/\sqrt{3}} 4\frac{2}{\pi(1+x^2)}\,\mathrm{d}x + \int_{1/\sqrt{3}}^{1} 3\frac{2}{\pi(1+x^2)}\,\mathrm{d}x + \int_1^{\sqrt{3}} 2\frac{2}{\pi(1+x^2)}\,\mathrm{d}x \\
= \;& \frac{2}{\pi}\left[4\left(\arctan\frac{1}{\sqrt{3}} - \arctan 0\right) + 3\left(\arctan 1 - \arctan\frac{1}{\sqrt{3}}\right) + 2\left(\arctan\sqrt{3} - \arctan 1\right)\right] \\
= \;& \frac{13}{6}.
\end{aligned}
$$

[RUBRIC: A correct derivation and answer gets **(+1 point)**.]

4. To derive this conditional distribution, we can write

$$P(X = k|X + Y = n) = \frac{P(X = k \cap X + Y = n)}{P(X + Y = n)}$$

using the definition of conditional probability.

[RUBRIC: Correct application of Bayes Rule gets **(+1 point)**.]

The event $X = k \cap X + Y = n$ can equivalently be expressed as $X = k \cap Y = n - k$ and we can express this using independence,

$$
\begin{aligned}
P(X = k \cap Y = n - k) &= \frac{e^{-\lambda}\lambda^k}{k!}\frac{e^{-\mu}\mu^{n-k}}{(n-k)!} \\
&= \frac{1}{n!}e^{-(\lambda+\mu)}\binom{n}{k}\lambda^k\mu^{n-k}.
\end{aligned} \tag{2}
$$

[RUBRIC: Correct application of independence of X and Y gets **(+1 point)**.]

[RUBRIC: Correct derivation and answer of $P(X = k, Y = n - k)$ gets **(+1 point)**.]

Now, we note that we can use the law of total probability with the above to get an expression for the denominator,

$$
\begin{aligned}
P(X + Y = n) &= \sum_{k=0}^{n} P(X = k \cap Y = n - k) \\
&= \sum_{k=0}^{n} \frac{1}{n!}e^{-(\lambda+\mu)}\binom{n}{k}\lambda^k\mu^{n-k} \\
&= \frac{1}{n!}e^{-(\lambda+\mu)}\sum_{k=0}^{n}\binom{n}{k}\lambda^k\mu^{n-k} \\
&= \frac{1}{n!}e^{-(\lambda+\mu)}(\lambda+\mu)^n,
\end{aligned} \tag{3}
$$

where the last equality comes from binomial expansion.

[RUBRIC: Correct derivation and answer of $P(X+Y=n)$ using law of total probability gets **(+1 point)**.]

Lastly, we plug these in to obtain

$$P(X=k|X+Y=n) = \binom{n}{k}\frac{\lambda^k\mu^{n-k}}{(\lambda+\mu)^n}.$$

[RUBRIC: Correct final answer gets **(+1 point)**.]

This is exactly the PMF for a binomial distribution with parameters n and $p = \frac{\lambda}{\lambda+\mu}$.

[RUBRIC: Correct distribution name gets **(+0.5 point)**.]

[RUBRIC: Correct parameter set gets **(+0.5 point)**.]

[RUBRIC: **Total (+6 points)**.]

3 Properties of Gaussians

1. Prove that $\mathbb{E}[e^{\lambda X}] = e^{\sigma^2\lambda^2/2}$, where $\lambda \in \mathbb{R}$ is a constant, and $X \sim N(0, \sigma^2)$. As a function of λ, $\mathbb{E}[e^{\lambda X}]$ is also known as the *moment-generating function*.

2. *Concentration inequalities* are inequalities that place upper bounds on the likelihood that a random variable X is far away from its mean μ, written $\mathbb{P}(|X - \mu| \geq t)$, with a falling exponential function ae^{-bt^2} having constants $a, b > 0$. Such inequalities imply that X is very likely to be close to its mean μ. To make a tight bound, we want a to be as small and b to be as large as possible.

   For $t > 0$ and $X \sim N(0, \sigma^2)$, prove that $\mathbb{P}(X \geq t) \leq \exp\left(-t^2/2\sigma^2\right)$, then show that $\mathbb{P}(|X| \geq t) \leq 2\exp\left(-t^2/2\sigma^2\right)$.
   Hint: Consider using Markov's inequality and the result from Question 3.1.

3. Let $X_1, \ldots, X_n \sim N(0, \sigma^2)$ be i.i.d. (independent and identically distributed). Find a concentration inequality, similar to Question 3.2, for the average of n Gaussians: $\mathbb{P}(\frac{1}{n}\sum_{i=1}^n X_i \geq t)$? What happens as $n \to \infty$?
   Hint: Without proof, use the fact that linear combinations of i.i.d. Gaussian-distributed variables are also Gaussian-distributed. Be warned that summing two Gaussian variables does **not** mean that you can sum their probability density functions (no no no!).

4. Let $X \in \mathbb{R}^n \sim N(0, \sigma^2 I_n)$ be an n-dimensional Gaussian random variable, where I_n denotes the $n \times n$ identity matrix. You may interpret X as a (column) vector whose entries are i.i.d. real values drawn from the scalar Gaussian $N(0, \sigma^2)$. Given a constant (i.e., not random) matrix $A \in \mathbb{R}^{n\times n}$ and a constant vector $b \in \mathbb{R}^n$, derive the mean (which is a vector) and covariance matrix of $Y = AX + b$. Use the fact that any linear transformation of a Gaussian random variable is also a Gaussian random variable.

5. Let vectors $u, v \in \mathbb{R}^n$ be orthogonal (i.e., $\langle u, v\rangle = 0$). Let $X = (X_1, \ldots, X_n)$ be a vector of n i.i.d. standard Gaussians, $X_i \sim N(0, 1), \forall i \in [n]$. Let $u_x = \langle u, X\rangle$ and $v_x = \langle v, X\rangle$. Are u_x and v_x independent? If $X_1, \ldots, X_n$ are independently but not identically distributed, say $X_i \sim N(0, i)$, are u_x and v_x still independent?
   Hint: Two Gaussian random variables are independent if and only if they are uncorrelated.

6. Prove that $\mathbb{E}\left[\max_{1\leq i\leq n} |X_i|\right] \leq C\sqrt{\log(2n)}\sigma$ for some constant $C \in \mathbb{R}$, where $X_1, \ldots, X_n \sim N(0, \sigma^2)$ are i.i.d. (Interestingly, a similar lower bound holds: $\mathbb{E}\left[\max_{1\leq i\leq n} |X_i|\right] \geq C'\sqrt{\log(2n)}\sigma$ for some C'; but you don't need to prove the lower bound).
   Hint: Use Jensen's inequality: $f(\mathbb{E}[Y]) \leq \mathbb{E}[f(Y)]$ for any convex function f.

Solution:

1.

$$\mathbb{E}[e^{\lambda X}] = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} e^{\lambda x}e^{-x^2/2\sigma^2}dx = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{\lambda\sigma z}e^{-z^2/2}dz$$

$$= e^{\sigma^2\lambda^2/2}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-(z-\lambda\sigma)^2/2}dz = e^{\sigma^2\lambda^2/2}.$$

[RUBRIC: A completely correct solution gets **(+1 point)**.]

2. For any $\lambda > 0$,

$$\mathbb{P}(X \geq t) = \mathbb{P}(e^{\lambda X} \geq e^{\lambda t}) \leq e^{-\lambda t}\mathbb{E}[e^{\lambda X}] = e^{-\lambda t}e^{\sigma^2\lambda^2/2},$$

where the inequality applies Markov's inequality. Setting $\lambda = t/\sigma^2$ gives the claim.

[RUBRIC: A completely correct solution gets **(+1 point)**.]

3. From the hint we know that $\frac{1}{n}\sum_{i=1}^n X_i$ follows a Gaussian distribution, so we only need to determine its mean and variance. Its mean is clearly 0. Its variance is

$$\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^n X_i\right) = \frac{1}{n^2}\, n\mathrm{Var}(X_i) = \frac{\sigma^2}{n},$$

where we use the fact that the variance of a sum of uncorrelated variables separates into a sum of their variances.

[RUBRIC: Correct mean value gets **(+0.5 point)**.]

[RUBRIC: Correct variance value gets **(+0.5 point)**.]

Now we apply the concentration result of the previous part to conclude that

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i \geq t\right) \leq \exp\left(-nt^2/2\sigma^2\right).$$

[RUBRIC: A correct concentration result gets **(+1 point)**.]

As $n \to \infty$, the probability of the empirical mean $\frac{1}{n}\sum_{i=1}^n X_i$ not being zero vanishes. This is usually expressed as $\frac{1}{n}\sum_i X_i$ "converges in probability" to the constant 0. The phenomena that the empirical mean of i.i.d. random variables $\frac{1}{n}\sum_i X_i$ (not necessarily Gaussian) converges in probability to its true mean is called the *Weak Law of Large Numbers*.

[RUBRIC: Realizing that probability of empirical mean *being non-zero* converges to zero gets **(+1 point)**. Caution: probability of being a given non-zero value is always zero as it is a continuous variable.]

[RUBRIC: **Total (+3 points)**.]

4. By linearity of expectation, $\mathbb{E}[Y] = \mathbb{E}[AX + b] = A\mathbb{E}[X] + b = b$.

[RUBRIC: A correct derivation and answer for mean gets **(+0.5 point)**.]

For the covariance matrix, we want to calculate $\mathbb{E}[(AX + b - \mathbb{E}[Y])(AX + b - \mathbb{E}[Y])^\top] = \mathbb{E}[AX(AX)^\top] = \mathbb{E}[AXX^\top A^\top] = A\mathbb{E}[XX^\top]A^\top = A(\sigma^2 I_n)A^\top = \sigma^2 AA^\top$. Note we also used the linearity of the expectation here.

[RUBRIC: A correct derivation and answer for covariance gets **(+1 point)**.]

[RUBRIC: **Total (+1.5 point)**.]

5. We use the fact that Gaussian random variables are independent if and only if they are uncorrelated (again under some regularity which is satisfied). Therefore, we only need to compute the correlation of u_x and v_x,

$$\mathbb{E}[u_x v_x] = \mathbb{E}\left[\left(\sum_{i=1}^n u_i X_i\right)\left(\sum_{i=1}^n v_i X_i\right)\right] = \sum_{i=1}^n u_i v_i \mathbb{E}[X_i^2] = \langle u, v\rangle = 0.$$

[RUBRIC: Correct argument for u_x and v_x being independent for i.i.d. X_i gets **(+1 point)**.]

Therefore, u_x and v_x are independent. However, if $X_1, \ldots, X_n$ are not identically distributed, $\mathbb{E}[u_x v_x] = \sum_{i=1}^n u_i v_i \mathbb{E}[X_i^2] = \sum_{i=1}^n u_i v_i i$, not necessarily equal to 0. Therefore if the X_i's are not identically distributed, u_x and v_x are not necessarily independent.

[RUBRIC: Correct derivation and value of covariance between u_x and v_x when $X_i \sim N(0, i)$ gets **(+0,5 point)**.]

[RUBRIC: Arguing that u_x and v_x may not independent for non-iid X_i gets **(+0.5 point)**.]

[RUBRIC: **Total (+2 points)**.]

6. Let $\lambda > 0$. By Jensen's inequality,

$$\begin{aligned}\lambda\mathbb{E}\left[\max_{1\le i\le n}|X_i|\right] &\le \log\mathbb{E}[e^{\lambda\max_i|X_i|}] \le \log\sum_{i=1}^n \mathbb{E}[e^{\lambda|X_i|}] \le \log\sum_{i=1}^n\left(\mathbb{E}[e^{\lambda X_i}] + \mathbb{E}[e^{-\lambda X_i}]\right)\\ &\le \log\sum_{i=1}^n 2e^{\sigma^2\lambda^2/2} = \log 2ne^{\sigma^2\lambda^2/2} = \log(2n) + \frac{\sigma^2\lambda^2}{2}.\end{aligned}$$

Setting $\lambda = \frac{\sqrt{\log(2n)}}{\sigma}$ yields

$$\mathbb{E}\left[\max_{1\le i\le n}|X_i|\right] \le \sigma\sqrt{\log(2n)} + \frac{\sigma}{2}\sqrt{\log(2n)} = \frac{3}{2}\sigma\sqrt{\log(2n)}.$$

[RUBRIC: Any completely correct solution gets **(+2 points)**. Any partially correct or incomplete solution gets **(+1 point)**.]

4 Linear Algebra Review

1. First we review some basic concepts of rank and elementary matrix operations. Let $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times p}$. Let I_n denote the $n \times n$ identity matrix.

   (a) Perform elementary row and column operations to transform $\begin{bmatrix} I_n & 0 \\ 0 & AB \end{bmatrix}$ to $\begin{bmatrix} B & I_n \\ 0 & A \end{bmatrix}$.

   (b) Use part (a) to prove that $\text{rank}(A) + \text{rank}(B) - n \le \text{rank}(AB) \le \min\{\text{rank}(A), \text{rank}(B)\}$.

2. Let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix. Prove equivalence between these three different definitions of positive semi-definiteness (PSD).

   (a) For all $x \in \mathbb{R}^n$, $x^\top A x \ge 0$.

   (b) All the eigenvalues of A are non-negative.

   (c) There exists a matrix $U \in \mathbb{R}^{n\times n}$ such that $A = UU^\top$.

   Positive semi-definiteness will be denoted as $A \succeq 0$.

3. Now that we're equipped with different definitions of positive semi-definiteness, use them to prove the following properties of PSD matrices.

   (a) If A is PSD, all diagonal entries of A are non-negative: $A_{ii} \ge 0, \forall i \in [n]$.

   (b) If A is PSD, the sum of all entries of A is non-negative: $\sum_{j=1}^n \sum_{i=1}^n A_{ij} \ge 0$.

   (c) If A and B are PSD, then $\text{Tr}(AB) \ge 0$, where $\text{Tr}\, M$ denotes the *trace* of M.

   (d) If A and B are PSD, then $\text{Tr}(AB) = 0$ if and only if $AB = 0$.

4. If $M - N \succeq 0$ and both M and N are positive definite, is $N^{-1} - M^{-1}$ PSD? Show your work.

5. Let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix. Prove that the largest eigenvalue of A is

$$\lambda_{\max}(A) = \max_{\|x\|_2 = 1} x^\top A x.$$

Solution:

1. The operations for (a) are as follows.

$$\begin{bmatrix} I_n & 0 \\ 0 & AB \end{bmatrix} \implies \text{(Left multiply first row by } A \text{ and add it to second row)}$$

$$\begin{bmatrix} I_n & 0 \\ A & AB \end{bmatrix} \implies \text{(Right multiply first column by B and subtract it from second column)}$$

$$\begin{bmatrix} I_n & -B \\ A & 0 \end{bmatrix} \implies \text{(Exchange columns and multiply constants)}$$

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

[RUBRIC: A complete and correct set of of operations gets **(+2 points)**. A partially correct or incomplete set of operations gets **(+1 point)**.]

Using (a) and rearranging gives us the lower bound,

$$n + \text{rank}(AB) = \text{rank}\left(\begin{bmatrix} I_n & 0 \\ 0 & AB \end{bmatrix}\right) = \text{rank}\left(\begin{bmatrix} B & I_n \\ 0 & A \end{bmatrix}\right) \geq \text{rank}(A) + \text{rank}(B)$$

[RUBRIC: Correctly proving the lower bound gets **(+1 point)**.]

Let $\mathcal{R}(M)$ denote the range (column space) of a matrix M. Since $\mathcal{R}(AB) \subseteq \mathcal{R}(A)$, we have $\text{rank}(AB) \leq \text{rank}(A)$. Similarly, since $\mathcal{R}(B^\top A^\top) \subseteq \mathcal{R}(B^\top)$, we have $\text{rank}(AB) \leq \text{rank}(B)$. Thus $\text{rank}(AB) \leq \min\{\text{rank}(A), \text{rank}(B)\}$

[RUBRIC: Correctly proving the upper bound gets **(+1 point)**.]

[RUBRIC: **Total (+4 points)**.]

2. (a) $\Rightarrow$ (b): Let λ be an eigenvalue of A with corresponding eigenvector v. Then

$$v^\top A v = \lambda v^\top v = \lambda \|v\|^2.$$

By part (a), we know that $\lambda\|v\|^2 \geq 0$, so $\lambda \geq 0$.
(b) $\Rightarrow$ (c): Consider the eigendecomposition of A, $A = V\Lambda V^\top$, where Λ is a diagonal matrix with entries equal to the eigenvalues of A, $\lambda_1, \ldots, \lambda_n$. Define $U := V\sqrt{\Lambda}$, where $\sqrt{\Lambda}$ is diagonal with entries equal to $\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n}$; notice that this choice is justified because, by assumption, the eigenvalues are non-negative. Clearly, $A = UU^\top$.
(c) $\Rightarrow$ (a): Let $x \in \mathbb{R}^n$. Then

$$x^\top A x = x^\top U U^\top x = (U^\top x)^\top (U^\top x) = \|U^\top x\|^2 \geq 0.$$

[RUBRIC: Correctly proving any of the required three directions for equivalence gets **(+0.5 point)**.]

[RUBRIC: **Total (+1.5 points)**.]

3. (a) Fix $i \in [n]$. Take $x = e_i$ in the first definition of PSD, where e_i is a canonical vector, i.e., it has zeros everywhere but at coordinate i, where it is equal to 1. Then $e_i^\top A e_i = A_{ii} \geq 0$.
   (b) Take $x = \mathbf{1}$ to be the all-ones vector in the first definition of PSD. Then $\mathbf{1}^\top A \mathbf{1} = \sum_{i=1}^n \sum_{j=1}^n A_{ij} \geq 0$.
   (c) By the third definition of PSD, let $A = UU^\top$ and $B = VV^\top$. Then

$$\text{Tr}(AB) = \text{Tr}\Big(UU^\top VV^\top\Big) = \text{Tr}\Big(U^\top VV^\top U\Big) = \text{Tr}\Big(U^\top V(U^\top V)^\top\Big) \geq 0,$$

   which follows because $M \overset{\text{def}}{=} U^\top V(U^\top V)^\top$ is PSD by the third definition, and $\text{Tr}\, M \geq 0$ by part (b).

(d) If $AB = 0$, then clearly $\mathrm{Tr}(AB) = 0$. To prove the other direction, by the third definition of PSD, let $A = UU^\top$ and $B = VV^\top$ for some U and V. Then

$$\mathrm{Tr}(AB) = \mathrm{Tr}\Big(UU^\top VV^\top\Big) = \mathrm{Tr}\Big(V^\top UU^\top V\Big) = \mathrm{Tr}\Big((U^\top V)^\top U^\top V\Big).$$

Since $M \overset{\text{def}}{=} (U^\top V)^\top U^\top V$ is PSD, $\mathrm{Tr}\, M = \sum_i \lambda_i(M) = 0$ only if $\lambda_i(M) = 0$ for all $i \in [n]$. From the eigendecomposition of M, it follows that $M = 0$, and moreover this implies $U^\top V = 0$. With this, we have $AB = U(U^\top V)V^\top = U(0)V^\top = 0$.

[RUBRIC: Correctly proving any sub-part gets **(+0.5 point)**. **Total (+2 points)**.]

4. Yes, it is PSD. [RUBRIC: Correctly identifying it as PSD gets **(+1 point)**.]

   Proof:

$$M - N \succeq 0 \implies N^{-1/2}(M - N)N^{-1/2} \succeq 0 \implies N^{-1/2}MN^{-1/2} \succeq I$$

   This means that $N^{-1/2}MN^{-1/2}$ is invertible and its smallest eigenvalue is at least 1. Let $B = N^{-1/2}MN^{-1/2}$ and then $B^{-1} \preceq I$, implying $I - N^{1/2}M^{-1}N^{1/2} \succeq 0$. Left and right multiply by $N^{-1/2}$, yielding the desired result: $N^{-1} - M^{-1} \succeq 0$.

   [RUBRIC: Correct argument for proof of PSD gets **(+1 point)**.]

   [RUBRIC: **Total (+2 points)**.]

5. Let $A = V\,\mathrm{diag}(\lambda_1, \ldots, \lambda_n)V^\top$ be an eigendecomposition of A (given by the spectral theorem). Since $V^\top$ is invertible, for every $y \in \mathbb{R}^n$ there is an $x \in \mathbb{R}^n$ such that $V^\top x = y$. Moreover, V is orthonormal (a.k.a. orthogonal), so $\|Vy\|_2 = \|y\|_2$. Therefore,

$$\begin{aligned}\max_{\|x\|_2=1} x^\top Ax &= \max_{\|x\|_2=1} (V^\top x)^\top \mathrm{diag}(\lambda_1, \ldots, \lambda_n)(V^\top x) = \max_{\|Vy\|_2=1} y^\top \mathrm{diag}(\lambda_1, \ldots, \lambda_n)y \\ &= \max_{\|y\|_2=1} \sum_{i=1}^n y_i^2 \lambda_i\end{aligned}$$

   [RUBRIC: Correct derivation of $\max_{\|x\|_2=1} x^\top Ax = \max_{\|y\|_2=1} \sum_{i=1}^n y_i^2\lambda_i$ gets **(+1 point)**.]

   In the optimization problem $\max_{\|y\|_2=1} \sum_{i=1}^n y_i^2\lambda_i =$ our best choice is to place all weight on the coefficient y_i which corresponds to the largest eigenvalue of A.

   Here is a cute alternative solution. Write out the Lagrangian for this optimization problem.

$$\mathcal{L}(x, \nu) = x^\top Ax + \nu(1 - \|x\|_2^2)$$

   Then differentiate with respect to x and ν and set the derivatives equal to zero to write out the first-order optimality conditions.

$$\begin{aligned}\nabla_x \mathcal{L}(x, \nu) &= 2Ax - 2\nu x = 0, \\ \frac{\mathrm{d}}{\mathrm{d}\nu}\mathcal{L}(x, \nu) &= 1 - \|x\|_2^2 = 0.\end{aligned}$$

The first condition tells us that the optimal x^* is an eigenvalue of A, while the second one confirms its norm is 1. This immediately implies that the optimum is a unit eigenvector of A corresponding to its largest eigenvalue. Plugging this in to the objective, we immediately find it is $\lambda_{\max}$.

[RUBRIC: Correct argument for optimum being λ_{max} gets **(+1 point)**.]

[RUBRIC: **Total (+2 points)**.]

5 Gradients and Norms

1. Define the ℓ_p-norm as $\|x\|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}$, where $x \in \mathbb{R}^n$. Prove that the ℓ_1, ℓ_2, ℓ_∞ norms are all within a constant factor of one another. The Cauchy–Schwarz inequality is useful here.

2. Aside from norms on vectors, we can also impose norms on matrices, and the most common kind of norm on matrices is called the induced norm. Induced norms are defined to be

$$\|A\|_p = \sup_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_p}$$

where the notation $\|\cdot\|_p$ on the right-hand side denotes the vector ℓ_p-norm. Please give the closed-form (or the most simple) expressions for the following induced norms of $A \in \mathbb{R}^{m \times n}$.

   (a) $\|A\|_2$. (Hint: Similar to Question 4.5)

   (b) $\|A\|_\infty$.

3. (a) Let $\alpha = \sum_{i=1}^{n} y_i \ln \beta_i$ for $y, \beta \in \mathbb{R}^n$. What are the partial derivatives $\frac{\partial \alpha}{\partial \beta_i}$?

   (b) Let $\beta = \sinh \gamma$ for $\gamma \in \mathbb{R}^n$ (treat the sinh as an element-wise operation; i.e., $\beta_i = \sinh \gamma_i$). What are the partial derivatives $\frac{\partial \beta_i}{\partial \gamma_j}$?

   (c) Let $\gamma = A\rho + b$ for $b \in \mathbb{R}^n$, $\rho \in \mathbb{R}^m$, $A \in \mathbb{R}^{n \times m}$. What are the the partial derivatives $\frac{\partial \gamma_i}{\partial \rho_j}$?

   (d) Let $f(x) = \sum_{i=1}^{n} y_i \ln(\sinh(Ax + b)_i)$; $A \in \mathbb{R}^{n \times m}$, $y \in \mathbb{R}^n$, $b \in \mathbb{R}^n$ are given. What are the partial derivatives $\frac{\partial f}{\partial x_j}$?
   Hint: Use the chain rule.

4. Consider a linear decision function $f(x) = w \cdot x + \alpha$ and the hyperplane decision boundary $H = \{x : w \cdot x = -\alpha\}$. Prove that if w is a unit vector, then the *signed distance* (the ℓ_2-norm distance with an appropriate sign) from x to the closest point on H is $w \cdot x + \alpha$.

5. Let $X \in \mathbb{R}^{n \times d}$ be a data matrix, consisting of n samples, each of which has d features, and let $y \in \mathbb{R}^n$ be a vector of labels. We wish to find the *best linear approximation*, i.e., we want to find the w that minimizes the loss $L(w) = \|y - Xw\|_2^2$. Assuming X has full column rank, compute $w^* = \text{argmin}_w L(w)$ in terms of X and y.

Solution:

1. First, we show $\frac{1}{\sqrt{n}}\|x\|_2 \leq \|x\|_\infty$.

$$\frac{1}{\sqrt{n}}\|x\|_2 = \frac{1}{\sqrt{n}}\sqrt{\sum x_i^2} = \sqrt{\frac{1}{n}\sum x_i^2} \leq \sqrt{\frac{1}{n}\sum \max x_i^2} = \sqrt{\max x_i^2} = \max \sqrt{x_i^2} = \max \|x_i\| = \|x\|_\infty.$$

Next, $\|x\|_\infty \leq \|x\|_1$.

$$\|x\|_\infty = \max_i |x_i| \leq \sum_{1}^{n} |x_i| = \|x\|_1.$$

Lastly, $\|x\|_1 \le \sqrt{n}\|x\|_2$. From the Cauchy–Schwarz theorem, $|\langle x, y\rangle|^2 \le \|x\|_2^2\|y\|_2^2$. Let $y = \text{sign}(x)$. Then

$$|\langle x, \text{sign}(x)\rangle|^2 \le \|x\|_2^2\|\text{sign}(x)\|_2^2 \Leftrightarrow \left(\sum_i |x_i|\right)^2 \le \left(\sum_i {x_i}^2\right)\left(\sum_i \text{sign}(x_i)^2\right).$$

Since $\sum_i \text{sign}(x_i)^2 \le \sum_i 1 = n$,

$$\|x\|_1^2 \le n \cdot \|x\|_2^2 \Leftrightarrow \|x\|_1 \le \sqrt{n} \cdot \|x\|_2.$$

Thus, we have shown $\frac{1}{\sqrt{n}}\|x\|_2 \le \|x\|_\infty \le \|x\|_1 \le \sqrt{n}\|x\|_2$.

[RUBRIC: A complete and correct solution to any of the three required inequalities gets **(+1 point)**.]

[RUBRIC: **Total (+3 points)**.]

2. (a) Let $A = U\Sigma V^\top$ be an SVD of A. Then

$$\begin{aligned}
\sup_{x\neq 0} \frac{\|Ax\|_2}{\|x\|_2} &= \sup_{x\neq 0} \frac{\|U\Sigma V^\top x\|_2}{\|x\|_2} \\
&= \sup_{x\neq 0} \frac{\|\Sigma V^\top x\|_2}{\|x\|_2} \\
&= \sup_{y\neq 0} \frac{\|\Sigma y\|_2}{\|Vy\|_2} (\text{suppose } y = V^\top x) \\
&= \sup_{y\neq 0} \frac{\|\Sigma y\|_2}{\|y\|_2} \\
&= \sigma_1. \text{ (The largest singular value of } A.)
\end{aligned}$$

(b)

$$\begin{aligned}
\sup_{x\neq 0} \frac{\|Ax\|_\infty}{\|x\|_\infty} &= \sup_{x\neq 0} \frac{\max_i(|a_{i1}x_1 + \cdots + a_{in}x_n|)}{\max_i |x_i|} \\
&= \sup_{x\neq 0} \frac{\max_i(|a_{i1}x_1| + \cdots + |a_{in}x_n|)}{\max_i |x_i|} \\
&= \sup_{x\neq 0} \frac{\max_i(|a_{i1}|\,|x_1| + \cdots + |a_{in}|\,|x_n|)}{\max_i |x_i|} \\
&= \max_i \sum_{j=1}^n |a_{ij}| \text{ (the largest row sum)}
\end{aligned}$$

[RUBRIC: A complete derivation and correct answer for any of the two sub-part gets **(+1 point)**. **Total (+2 points)**.]

3. (a) $\frac{\partial \alpha}{\partial \beta_i} = \sum_{j=1}^n \frac{\partial(y_j \ln \beta_j)}{\partial \beta_i} = \frac{y_i}{\beta_i}/$

(b) $\frac{\partial \beta_i}{\gamma_j} = \begin{cases} 0 & i \neq j, \\ \cosh\left(\gamma_j\right) & i = j/ \end{cases}$

(c) $\frac{\partial \gamma_i}{\partial \rho_j} = A_{ij}$.

(d) Using the previous parts, we can apply the chain rule as $\frac{\partial f}{\partial x_j} = \Sigma_{k=1}^n \Sigma_{l=1}^n \frac{\partial f}{\partial \beta_k} \frac{\partial \beta_k}{\partial \gamma_l} \frac{\partial \gamma_l}{\partial x_j}$. This can be simplified using the result from (b) to see the partial derivative $\frac{\partial \beta_a}{\partial \gamma_b}$ is zero unless $k = l$. This yields $\Sigma_{k=1}^n \frac{\partial f_i}{\partial \beta_k} \frac{\partial \beta_k}{\partial \gamma_k} \frac{\partial \gamma_k}{\partial x_j}$. Then we can expand and substitute in to get $\Sigma_{k=1}^n \frac{y_k}{\sinh(Ax+b)_k} \cosh((Ax+b)_k) A_{kj} = A_j^T (y \circ \frac{\cosh(Ax+b)}{\sinh(Ax+b)}) = A_j^T (y \circ \coth(Ax+b))$.

[RUBRIC: A complete derivation and correct answer for any of the 4 sub-part gets **(+1 point)**. **Total (+4 points)**. Simplified answer for (a)–(c) is necessary, but not for (d).]

4. Let $\bar{x}$ be the closest point on H to x; the distance from x to $\bar{x}$ is given by $\|x - \bar{x}\|_2$. We know that from the projection theorem, $x - \bar{x}$ is orthogonal to H, and the direction orthogonal to H is given exactly by w as that is the definition of H. Therefore, we can say that $x - \bar{x} = \kappa w$ for some κ. We note that w is unit norm, so it now suffices to find κ to define the sought distance. We also know that $\bar{x}$ lies on H and so must satisfy $w \cdot \bar{x} = -\alpha$. We substitute $\bar{x} = x - \kappa w$ into that equation to get

$$w \cdot (x - \kappa w) = -\alpha.$$

We distribute using the fact that κ is a scaling factor, giving

$$w \cdot x - \kappa w \cdot w = -\alpha.$$

Now, we use that w is a unit vector, so

$$w \cdot x + \alpha = \kappa.$$

So we have shown the unsigned distance is $w \cdot x + \alpha$. We then sign appropriately.

[RUBRIC: There could be other ways to solve this. Any completely correct solution gets **(+2 points)**. Any partially correct or incomplete solution gets **(+1 point)**.]

5. We start by finding a stationary point of $L(w)$ by solving

$$\nabla_w L(w) = -2X^\top (y - Xw) = 0,$$

or in other words, $X^\top y = X^\top X w$. Since X has full column rank, $X^\top X$ is invertible, and so $w^* = (X^\top X)^{-1} X^\top y$.

[RUBRIC: Correct derivation of w^* gets **(+0.5 point)**.]

Next, we find the Hessian $\nabla_w^2 L(w) = 2X^\top X$ which is a constant PSD matrix. Hence, the unique stationary point is the global minimizer.

[RUBRIC: Correct argument that w^* is the global minimizer of loss gets **(+0.5 point)**.]

[RUBRIC: **Total (+1 point)**.]

6 Gradient Descent

Consider the optimization problem $\min_{x \in \mathbb{R}^n} \frac{1}{2} x^\top A x - b^\top x$, where $A \in \mathbb{R}^{n \times n}$ is a PSD matrix with $0 < \lambda_{\min}(A) \leq \lambda_{\max}(A) < 1$.

1. Find the optimizer x^*.

2. Solving a linear system directly using Gaussian elimination takes $O(n^3)$ time, which may be wasteful if the matrix A is sparse. For this reason, we will use gradient descent to compute an approximation to the optimal point x^*. Write down the update rule for gradient descent with a step size of 1 (i.e., taking a step whose length is the length of the gradient).

3. Show that the iterates $x^{(k)}$ satisfy the recursion $x^{(k)} - x^* = (I - A)(x^{(k-1)} - x^*)$.

4. Using Question 4.5, prove $\|Ax\|_2 \leq \lambda_{\max(A)} \|x\|_2$.
   Hint: Use the fact that, if λ is an eigenvalue of A, then λ^2 is an eigenvalue of A^2.

5. Using the previous two parts, show that for some $0 < \rho < 1$,

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

6. Let $x^{(0)} \in \mathbb{R}^n$ be the starting value for our gradient descent iterations. If we want a solution $x^{(k)}$ that is $\epsilon > 0$ close to x^*, i.e. $\|x^{(k)} - x^*\|_2 \leq \epsilon$, then how many iterations of gradient descent should we perform? In other words, how large should k be? Give your answer in terms of ρ, $\|x^{(0)} - x^*\|_2$, and ϵ.

Solution:

1. Since the objective is convex, the optimizer is a stationary point of the objective, i.e., it satisfies

$$Ax - b = 0,$$

   and since A is invertible, the optimizer is $x^* = A^{-1}b$.

   [RUBRIC: A correct derivation and answer gets **(+1 point)**.]

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

   [RUBRIC: Correct update law gets **(+1 point)**.]

3. We expand the gradient descent update to obtain

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

   In the third equality we used the stationary condition $Ax^* = b$.

   [RUBRIC: Correct argument gets **(+1 point)**.]

4. We can write $\|Ax\|_2^2 = x^\top A^2 x$. First assume x has unit length. By exercise 3 in Problem 3,

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

Now take any $x \neq 0$, not necessarily of unit length ($x = 0$ trivially satisfies the inequality). Then, we have proved that

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

Multiplying both sides by $\|x\|_2^2$ and taking the square root completes the proof of the identity.

[RUBRIC: Correct argument gets **(+1 point)**.]

5. Note that $I - A \succ 0$, because $\lambda_{\max}(A) < 1$. Therefore

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

Let $\rho = \lambda_{\max}(I - A) = 1 - \lambda_{\min}(A)$, which is in $(0, 1)$ because $\lambda_{\min}(A) \leq \lambda_{\max}(A) < 1$ and $\lambda_{\min}(A) > 0$. Then

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

[RUBRIC: Correct argument gets **(+1 point)**.]

6. Unrolling the recursion of part (d) gives us

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

[RUBRIC: Correct application of recursion for $\|x^{(k)} - x^*\|_2$ gets **(+1 point)**.]

Therefore a sufficient condition for $\|x^{(k)} - x^*\|_2 \leq \epsilon$ to hold true is

$$\rho^k \|x^0 - x^*\|_2 \leq \epsilon.$$

Taking logarithms and rearranging, this yields

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

[RUBRIC: Correct inequality for k gets **(+1 point)**.]

[RUBRIC: **Total (+2 points)**.]