CS 189 Introduction to Machine Learning
Summer 2019 Marc Khoury & Brijen Thananjeyan

HW2

Due: Wednesday, February 13 at 11:59 pm

Deliverables:

1. Submit a PDF of your homework, **with an appendix listing all your code**, 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 with whom you worked on the homework.
    - 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 inadverdently 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, recall the following useful identity: 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 pdf $f(x) = \lambda e^{-\lambda x}$ for $x > 0$ (and zero everywhere else). Use induction on k to show that for $k \in \mathbb{Z}$, $\mathbb{E}[X^k] = \frac{k!}{\lambda^k}$.
   Hint: use integration by parts.

2. Assume that X is a non-negative real-valued random variable. Prove the following identity:
$$\mathbb{E}X = \int_0^\infty \mathbb{P}(X \geq t)dt.$$
   If you prefer, assume that X has a density $f(x)$ and a CDF $F(x)$; this might simplify notation.

3. Again assume $X \geq 0$, but now additionally let $\mathbb{E}[X^2] < \infty$. Prove the following:
$$\mathbb{P}(X > 0) \geq \frac{(\mathbb{E}X)^2}{\mathbb{E}[X^2]}.$$
   Note that by assumption we know $\mathbb{P}(X \geq 0) = 1$, so this inequality is indeed quite powerful.
   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 a valid inner product on the set of random variables is given by $\mathbb{E}(UV)$, for random variables U and V.

4. Now assume $\mathbb{E}[X^2] < \infty$, and additionally assume $\mathbb{E}X = 0$ (X no longer has to be non-negative). Prove the following inequality:
$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X^2]}{\mathbb{E}[X^2] + t^2}, \text{ for any } t \geq 0$$
   There is no typo — compared to the previous part, the inequality is flipped.
   Hint: Use similar logic as in the previous part, and think of how to apply Cauchy–Schwarz. Use the fact that $t - X \leq (t - X)\mathbf{1}\{t - X > 0\}$.

Solution:

1. 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 "look like" and try to transform it to 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.

2.

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

3. Using non-negativity of X, we have $\mathbb{E}X = \mathbb{E}[X\mathbf{1}\{X > 0\}]$. Now use Cauchy–Schwarz applied to $U := X$ and $V := \mathbf{1}\{X > 0\}$ to conclude:

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

4. Using the same idea as in the previous part:

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

Now apply Cauchy-Schwarz to get:

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

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:

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

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

2 Properties of Gaussians

1. Prove that $\mathbb{E}[e^{\lambda X}] = e^{\sigma^2\lambda^2/2}$, where $\lambda \in \mathbb{R}$ is a fixed 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. Prove that $(X \geq t) \leq \exp\left(-t^2/2\sigma^2\right)$, and conclude that $(|X| \geq t) \leq 2\exp\left(-t^2/2\sigma^2\right)$.
Hint: Consider using Markov's inequality in combination with the result of the previous part.

3. Let $X_1, \ldots, X_n \sim N(0, \sigma^2)$ be iid. Can you prove a similar concentration result for the average of n Gaussians: $(\frac{1}{n}\sum_{i=1}^n X_i \geq t)$? What happens as $n \to \infty$?
Hint: Without proof use the fact that (under some regularity, which is satisfied for iid Gaussians) linear combinations of Gaussians are also Gaussian.

4. Give an example of two Gaussian random variables X and Y, such that there exists a linear combination $\alpha X + \beta Y$, for some $\alpha, \beta \in \mathbb{R}$, which is *not* Gaussian. Note that examples of the kind $X \sim N(0,1), Y = -X$ and their linear combination $X + Y = 0$ *will not* be valid solutions; we will consider constant random variables as Gaussians with variance equal to 0.

5. Take two orthogonal vectors $u, v \in \mathbb{R}^n$, $u \perp v$, and let $X = (X_1, \ldots, X_n)$ be a vector of n iid 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?
Hint: First try to see if they are correlated; you may use the fact that jointly normal random variables are independent iff. they are uncorrelated.

6. Prove that $\mathbb{E}\left[\max_{1 \le i \le n} |X_i|\right] \le C\sqrt{\log(2n)}\sigma$, where $X_1, \ldots, X_n \sim N(0, \sigma^2)$ are iid. In fact, a stronger version of this claim holds - $\mathbb{E}\left[\max_{1 \le i \le n} |X_i|\right] \ge C'\sqrt{\log(2n)}\sigma$ for some C' (you don't need to prove the lower bound).
Hint: Use Jensen's inequality, which says that $f(\mathbb{E}[Y]) \le \mathbb{E}[f(Y)]$, for any convex function f. Take $f(Y) = e^Y$, and use exercise 1 of this Problem.

Solution:

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

2. For any $\lambda > 0$, we have:
$$\mathbb{P}(X \ge t) = \mathbb{P}(e^{\lambda X} \ge e^{\lambda t}) \le e^{-\lambda t}\mathbb{E}[e^{\lambda X}] = e^{-\lambda t} e^{\sigma^2\lambda^2/2},$$
where the inequality applies Markov's inequality. Setting $\lambda = t/\sigma^2$ gives the claim.

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, on the other hand, can be computed as follows:
$$\left(\frac{1}{n}\sum_{i=1}^n X_i\right) = \frac{1}{n^2} n(X_i) = \frac{\sigma^2}{n},$$
where we use the fact that the variance of a sum of uncorrelated variables separates into a sum of their variances. Now we can apply the concentration result of the previous part to conclude:
$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i \ge t\right) \le \exp\left(-nt^2/2\sigma^2\right).$$
As $n \to \infty$, the probability of the average $\frac{1}{n}\sum_{i=1}^n X_i$ being away from 0 vanishes; this result is a special case of the Weak Law of Large Numbers.

4. Let $X \sim N(0,1)$ and $Y = \xi X$, where ξ takes values in $\{-1, 1\}$ with equal probability. The sum $X + Y$ is not Gaussian, even though both X and Y are Gaussian.

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:

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

Therefore, u_x and v_x are independent. Notice that this is a somewhat paradoxical conclusion, given that both u_x and v_x were computed using the same Gaussian vector X.

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

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

Set $\lambda = \frac{\sqrt{\log(2n)}}{\sigma}$:

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

3 Linear Algebra Review

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

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

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

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

   Mathematically, we write positive semi-definiteness as $A \succeq 0$.

2. Now that we're equipped with different definitions of positive semi-definiteness, prove the following properties of PSD matrices:

   (a) If A and B are PSD, then $2A + 3B$ is PSD.

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

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

   (d) If A and B are PSD, then $\mathrm{Tr}(AB) \ge 0$.

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

3. 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. (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:

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

2. (a) $x^\top(2A + 3B)x = 2x^\top Ax + 3x^\top Bx \geq 0$.

(b) 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 $e_i^\top A e_i = A_{ii} \geq 0$.

(c) Take $x = \mathbf{1}$ to be the all-ones vector in the first definition of PSD. Then $\mathbf{1}^\top A \mathbf{1} = \sum_{i=1}^n \sum_{j=1}^n A_{ij} \geq 0$.

(d) 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 $M := U^\top V(U^\top V)^\top$ is PSD by the third definition, and $\text{Tr}(M) \geq 0$ by part (b).

(e) If $AB = 0$, then clearly $\text{Tr}(AB) = 0$. 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:

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

Since $M := (U^\top V)^\top U^\top V$ is PSD, $\text{Tr}(M) = \sum_i \lambda_i(M) = 0$ 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 $AB = U(U^\top V)V^\top = U(0)V^\top = 0$.

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

$$\max_{\|x\|_2=1} x^\top A x = \max_{\|x\|_2=1} (V^\top x)^\top \text{diag}(\lambda_1, \ldots, \lambda_n)(V^\top x) = \max_{\|Vy\|_2=1} y^\top \text{diag}(\lambda_1, \ldots, \lambda_n) y$$

$$= \max_{\|y\|_2=1} \sum_{i=1}^{n} y_i^2 \lambda_i = \lambda_{\max}(A).$$

The last equality follows because 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.

4 Gradients and Norms

1. Define ℓ_p norms as $\|x\|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}$, where $x \in \mathbb{R}^n$. Prove that $\|x\|_2 \leq \|x\|_1 \leq \sqrt{n}\|x\|_2$.
Hint: For the second inequality, consider applying the Cauchy-Schwarz inequality.

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

3. Let $X, A \in \mathbb{R}^{n \times n}$ (not necessarily symmetric). Compute $\nabla_X \operatorname{Tr}(A^\top X)$.

4. 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 $0 < \lambda_{\min}(A) \leq \lambda_{\max}(A) < 1$.

   (a) Find the optimizer x^*.

   (b) 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.

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

   (d) Using exercise 3 in Problem 3, 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.

   (e) Using the previous two parts, show that for some $0 < \rho < 1$,

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

   (f) 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 $\rho, \|x^{(0)} - x^*\|_2$, and ϵ.

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 outcomes. For example, each row of X could have information about a house on the market, like its area, number of floors, number of bathrooms/bedrooms, etc., and each entry of y could be the price of that house. We are interested in building a model that predicts house prices from the set of its features, as listed above. Suppose that domain knowledge tells us that the relationship between the features and outcomes is linear; ideally, there exists a set of parameters $\theta \in \mathbb{R}^d$ such that $X\theta = y$. However, n is large and there is noise in the acquisition of X and y, so this system is overdetermined. Still, we wish to find the *best linear approximation*, i.e. we want to find the θ that minimizes the loss $L(\theta) = \|y - X\theta\|_2^2$. Assuming X has full column rank, compute $\theta^* = \arg\min_\theta L(\theta)$ in terms of X and y.

Solution:

1. To prove the first inequality, observe that:

$$\left(\sum_{i=1}^n |x_i|\right)^2 = \sum_{i=1}^n x_i^2 + 2\sum_{i=1}^n \sum_{j>i}^n |x_i x_j| \geq \sum_{i=1}^n x_i^2.$$

   Taking the square root of both sides preserves the inequality due to this map being increasing, and that gives exactly $\|x\|_1 \geq \|x\|_2$. The second inequality follows from the Cauchy-Schwarz inequality:

$$\|x\|_1 = |\langle x, \mathbf{1}\rangle| \leq \|x\|_2 \cdot \|\mathbf{1}\|_2 = \|x\|_2 \sqrt{n}.$$

2. (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) $\frac{\partial \beta_i}{\gamma_j} = \begin{cases} 0 & i \neq j \\ \cosh(\gamma_j) & i = j \end{cases}$

   (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 $\Sigma_{k=1}^n \frac{y_k}{\sinh(Ax+b)_k} \cosh((Ax+b)_k) A_{kj} = A_j^T (y \circ \frac{\cosh(Ax+b)}{\sinh(Ax+b)}) = A_j^T (y \circ \coth(Ax+b))$.

3.

$$\mathrm{Tr}\left(A^\top X\right) = \sum_{i=1}^n \sum_{j=1}^n x_{ij} A_{ij} \quad \Rightarrow \quad \frac{\partial \,\mathrm{Tr}(A^\top X)}{\partial x_{ij}} = A_{ij}$$

$$\Rightarrow \quad \nabla_X \left(\mathrm{Tr}\left(A^\top X\right)\right) = A$$

4. (a) 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 $x^* = A^{-1}b$.

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

(c) We expand the gradient descent update to get:
$$\begin{aligned} x^{(k)} - x^* &= x^{(k-1)} - (Ax^{(k-1)} - b) - x^* = (I - A)x^{(k-1)} + b - x^* \\ &= (I - A)x^{(k-1)} - (I - A)x^* = (I - A)(x^{(k-1)} - x^*), \end{aligned}$$
where in the third equality we used the stationarity condition $Ax^* = b$.

(d) We can write $\|Ax\|_2^2 = x^\top A^2 x$. First assume x has unit length. By exercise 3 in Problem 3, we have:
$$\|Ax\|_2^2 = x^\top A^2 x \leq (\lambda_{\max}(A))^2.$$
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.

(e) Note that $I - A > 0$, 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:
$$\|x^{(k)} - x^*\|_2 = \|(I - A)(x^{(k-1)} - x^*)\|_2 \leq \rho\|x^{(k-1)} - x^*\|_2$$

(f) Unrolling the recursion of part (d) we have:
$$\|x^{(k)} - x^*\|_2 \leq \rho^k\|x^0 - x^*\|_2$$
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:
$$k \geq \frac{1}{\log \frac{1}{\rho}} \log\left(\frac{\|x^0 - x^*\|_2}{\epsilon}\right)$$

5. The loss is convex, so we find the optimizer θ^* by finding a stationary point of $L(\theta)$. This gives:
$$\nabla_\theta L(\theta) = -2X^\top(y - X\theta) = 0,$$
or in other words $X^\top y = X^\top X\theta$. Since X has full column rank, $X^\top X$ is invertible, and so $\theta^* = (X^\top X)^{-1}X^\top y$.

5 Covariance Practice

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 such that $\Sigma_{ij} = \text{Cov}(Z_i, Z_j)$. Concisely, $\Sigma = \mathbb{E}[(Z - \mu)(Z - \mu)^\top]$. Prove that the covariance matrix is always PSD.
   Hint: Use linearity of expectation.

2. Let X be a multivariate random variable (recall, this means it is a vector of random variables) with mean vector $\mu \in \mathbb{R}^n$ and covariance matrix $\Sigma \in \mathbb{R}^{n \times n}$. Let Σ have one zero eigenvalue. Prove the space where X takes values with non-zero probability (this space is called the support of X) has dimension $n - 1$. How could you construct a new $\tilde{X}$ so that no information is lost from the original distribution but the covariance matrix of $\tilde{X}$ has no zero eigenvalues? What would $\tilde{X}$ look like if Σ has $m \leq n$ zero eigenvalues?
   Hint: use the identity $\text{Var}(\sum_{i=1}^{n} Y_i) = \sum_{i=1}^{n} \sum_{j=1}^{n} \text{Cov}(Y_i, Y_j)$.

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 the second equality comes from linearity of expectation.

2. Let the eigenvector corresponding to the zero eigenvalue be v. Because $v^T \Sigma v = 0$, $\sum_{i=1}^{n} \sum_{j=1}^{n} v_i v_j \text{Cov}(X_i, X_j) = 0$. Using the hint, the variance of $v^T X = \sum_{i=1}^{n} v_i X_i$ is zero and hence $v^T X$ is a constant equal to $\mu^T v$. Therefore for any non-zero index i of v, there exists a deterministic relation $X_i = \frac{\mu^T v}{v_i} - \sum_{j \neq i} \frac{v_j}{v_i} X_j$. X has zero probability density over the space of values that violate this relation; this relation leaves $n - 1$ free variables and hence the space where X takes values with non-zero probability has dimension $n - 1$. For $\tilde{X}$, we can simply remove index i from X. If there are more zero eigenvalues, we can remove as many indices as the nullity of Σ, which is m.