Quiz 3 Practice Problems

10-606

October 4, 2025

1 Matrix Calculus and Optimization

1. Let $f(x) = (Ax+b)^\top(Ax+b)$ where $A \in \mathbb{R}^{m\times n}$, $x \in \mathbb{R}^n$, $b \in \mathbb{R}^m$.

   (a) Compute $\nabla_x f(x)$.

   (b) Compute the Hessian $H(f) = \nabla_x^2 f(x)$.

   Solution. (a) Notice that

   $$f(x) = (Ax+b)^\top(Ax+b) = x^\top A^\top Ax + 2b^\top Ax + b^\top b.$$

   Then $\nabla_x x^\top A^\top Ax = 2A^\top Ax$ and $\nabla_x b^\top Ax = A^\top b$. So

   $$\nabla_x f(x) = 2A^\top(Ax+b).$$

   (b) Differentiate again:

   $$\nabla_x^2 f(x) = 2A^\top A.$$

2. Let $X \in \mathbb{R}^{n\times m}$ and $Y \in \mathbb{R}^{m\times n}$.

   (a) The trace of a square matrix A, denote $\mathrm{Tr}(A)$, is the sum of its diagonal elements. Show that $\mathrm{Tr}(XY) = \sum_{i=1}^n \sum_{j=1}^n x_{ij}y_{ji}$.

   (b) Show that $\nabla_X \mathrm{Tr}(XY) = Y^\top$.

   Solution.

   (a) The entry (i,i) of XY is the i-th row of X multiplied by the i-th column of Y. That is, $(XY)_{ii} = \sum_{j=1}^n x_{ij}y_{ji}$. Therefore

   $$\mathrm{Tr}(XY) = \sum_{i=1}^n (XY)_{ii} = \sum_{i=1}^n \sum_{j=1}^n x_{ij}y_{ji}.$$

(b) From above,

$$\frac{\partial \mathrm{Tr}(XY)}{\partial x_{k\ell}} = \sum_i \sum_j \frac{\partial}{\partial x_{k\ell}} x_{ij} y_{ji} = y_{\ell k}.$$

(Note the indices on $x_{k\ell}$ and $y_{\ell k}$ are reversed!) Hence $\nabla_X \mathrm{Tr}(XY) = \frac{\partial \mathrm{Tr}(XY)}{\partial X} = Y^\top$.

3. For $\theta, y \in \mathbb{R}^K$ such that $\sum_{i=1}^K y_i = 1$, let

$$f(\theta) = -\sum_{i=1}^K y_i \log\left(\frac{e^{\theta_i}}{\sum_{j=1}^K e^{\theta_j}}\right).$$

What is $\nabla_\theta f(\theta)$?

Solution. Define

$$g_i(\theta) = \frac{e^{\theta_i}}{\sum_{j=1}^K e^{\theta_j}}.$$

Then

$$\nabla_\theta f(\theta) = -\sum_{i=1}^K y_i \nabla_\theta \log(g_i(\theta)) = -\sum_{i=1}^K \frac{y_i}{g_i(\theta)} \nabla_\theta g_i(\theta).$$

Now let's compute $\nabla_\theta g_i(\theta)$. Let u_i be the i-th unit vector, i.e., $u_i = (0, \ldots, 0, 1, 0, \ldots, 0)$, where the 1 is in the i-th position. And let $\delta_{i,k} = 1$ if $i = k$ and 0 otherwise. Then,

$$\frac{\partial g_i(\theta)}{\partial \theta_k} = \frac{(\sum_{j=1}^K e^{\theta_j}) e^{\theta_i} \delta_{i,k} - e^{\theta_i} e^{\theta_k}}{(\sum_{i=1}^K e^{\theta_j})^2} = g_i(\theta)\delta_{i,k} - \frac{g_i(\theta)}{\sum_{j=1}^K e^{\theta_j}} e^{\theta_k}.$$

Therefore,

$$\nabla_\theta g_i(\theta) = g_i(\theta) u_i - \frac{g_i(\theta)}{\sum_{j=1}^K e^{\theta_j}} e^\theta,$$

where $e^\theta = (e^{\theta_1}, \ldots, e^{\theta_K})^\top$. Plugging this back into $\nabla_\theta f(\theta)$ we get

$$\begin{aligned}
\nabla_\theta f(\theta) &= -\sum_{i=1}^K \left(y_i u_i - \frac{y_i e^\theta}{\sum_{j=1}^K e^{\theta_j}} \right) \\
&= -y + \frac{e^\theta}{\sum_{j=1}^K e^{\theta_j}} \sum_{i=1}^K y_i \\
&= \frac{e^\theta}{\sum_{j=1}^K e^{\theta_j}} - y.
\end{aligned}$$

4. Let $h(w) = \log\big(1 + \exp(-y\, w^\top x)\big)$ with $x \in \mathbb{R}^d$ and $y \in \{-1, 1\}$. Compute $\nabla_w h(w)$.

   Solution. Set $z = y\, w^\top x$. Then $h(w) = \log(1 + e^{-z})$ and $\frac{dh}{dz} = -\frac{e^{-z}}{1+e^{-z}} = -\sigma(-z)$, where $\sigma(t) = \frac{1}{1+e^{-t}}$. Also $\nabla_w z = yx$. So by the chain rule,

$$\nabla_w h(w) = -\sigma(-y\, w^\top x)\, y\, x \;=\; -\frac{y\, x}{1 + \exp(y\, w^\top x)}.$$

5. For a square matrix $A \in \mathbb{R}^{n\times n}$ and vector $x \in \mathbb{R}^n$, let $f(x) = \frac{x^\top Ax}{x^\top x}$. Compute $\nabla_x f(x)$.

   Solution. Set $g(x) = x^\top Ax$. Note that $\nabla_x(x^\top x) = 2x$ (why?) and $\nabla_x g(x) = (A + A^\top)x$. Then

$$\begin{aligned}\nabla_x f(x) &= \frac{(x^\top x)\nabla_x g(x) - g(x)2x}{(x^\top x)^2} \\ &= \frac{(A + A^\top)x}{x^\top x} - 2\frac{x^\top Ax}{(x^\top x)^2}x.\end{aligned}$$

6. Let $f_1, \ldots, f_n$ be convex functions from $\mathbb{R}^n$ to $\mathbb{R}$.

   (a) Show that αf_1 is convex for any scalar $\alpha \geq 0$.

   (b) Show that $\sum_i \alpha_i f_i$ is convex for nonnegative real numbers α_i.

   (c) Is $\sum_i \alpha_i f_i$ for arbitrary real numbers α_i?

   Solution.

   (a) We have $H(\alpha f_1) = \alpha H(f_1)$ where H is the Hessian. Since f_1 is convex, $H(f_1)$ is positive semidefinite. Thus so too is $\alpha H(f_1)$ since α is nonnegative.

   (b) A similar logic applies. We have $H(\sum_i \alpha_i f_i) = \sum_i \alpha_i H(f_i)$. A nonnegative weighted sums of positive semidefinite matrices is also positive semidefinite (why?), hence $H(\sum_i \alpha_i f_i)$ is positive semidefinite, proving convexity.

   (c) No, $-x^2$ is not convex, but $f_1(x) = x^2$ is.

7. Let $f(x) = \frac{1}{2}x^\top Qx - c^\top x$ with

$$Q = \begin{bmatrix} k & 1 \\ 1 & k \end{bmatrix}, \qquad c = \begin{bmatrix} 1 \\ 0 \end{bmatrix},$$

   for some $k \in \mathbb{R}$.

(a) Compute $\nabla f(x)$.

(b) From $x_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, take one gradient descent step with step size $\eta = 0.2$. What is x_1?

Solution. (a) For $f(x) = \frac{1}{2}x^\top Qx - c^\top x$, $\nabla f(x) = Qx - c$. (b) $\nabla f(x_0) = Qx_0 - c = -c = \begin{bmatrix} -1, & 0 \end{bmatrix}^\top$. Then

$$x_1 = x_0 - \eta \nabla f(x_0) = \begin{bmatrix} 0 \\ 0 \end{bmatrix} - 0.2 \begin{bmatrix} -1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0.2 \\ 0 \end{bmatrix}.$$

2 Probability

1. You roll two fair dice.

   (a) $P(\text{sum} = 7)$

   (b) $P(\text{at least one six})$

   (c) Are the events in (a) and (b) independent?

   Solution. (a) Favorable outcomes: $(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)$ $\Rightarrow$ $6/36 = 1/6$. (b) $1 - P(\text{no six}) = 1 - (5/6)^2 = 11/36$. (c) No, one event gives you information about the other event. Formally, $P(\text{sum} = 7 \ \cap \ \text{at least one six}) = 2/36 = 1/18$. Product $P(\text{sum} = 7)P(\text{at least one six}) = (1/6)(11/36) = 11/216 \neq 1/18$.

2. From a standard 52-card deck, draw two cards without replacement.

   (a) What is the probability both are hearts?

   (b) What is $P(\text{second is a heart} \mid \text{first is red})$?

   Solution. (a) $(13/52) \cdot (12/51) = \frac{1}{4} \cdot \frac{12}{51} = \frac{1}{17}$. (b) Let H_i be the event that the i-th card is a heart. Let D_1 be the event that the first card is a diamond (the other red suit). Then $P(H_2 \wedge \text{first red}) = P(H_1 \wedge H_2) + P(D_1 \wedge H_2) = \frac{13}{52}\frac{12}{51} + \frac{13}{52}\frac{13}{51} = \frac{25}{204}$. (Here "$\wedge$" means "and".) Since $P(\text{first red}) = \frac{1}{2}$, we get

   $$P(H_2 \mid \text{first red}) = \frac{P(H_2 \wedge \text{first red})}{P(\text{first red})} = \frac{25/204}{1/2} = \frac{25}{102}.$$

3. Let $X \sim \text{Bernoulli}(p)$ (i.e., X is 1 with probability p, 0 with probability $1 - p$) and define $Y = 1 - X$.

(a) Write the joint pmf $P(X, Y)$.

(b) Are X and Y independent?

Solution. (a) $P(X = 1, Y = 0) = p$, $P(X = 0, Y = 1) = 1 - p$, and $P(X = 0, Y = 0) = P(X = 1, Y = 1) = 0$. (b) Independence would require $p = P(X = 1, Y = 0) = P(X = 1)P(Y = 0) = p \cdot p = p^2$, so $p \in \{0, 1\}$. Thus X and Y are *not* independent except in the degenerate cases $p = 0$ or $p = 1$.

4. Flip a fair coin three times. Let X be the number of heads, and $Y = \mathbf{1}\{\text{first flip is H}\}$.

(a) Write the joint pmf $P(X, Y)$.

(b) Compute $P(X = 3 \mid Y = 1)$.

Solution. (a) If $Y = 1$, the remaining two flips have $0, 1, 2$ heads with probabilities $1/4, 1/2, 1/4$, giving

$$P(Y = 1, X = 1) = \tfrac{1}{8}, \quad P(Y = 1, X = 2) = \tfrac{1}{4}, \quad P(Y = 1, X = 3) = \tfrac{1}{8}.$$

If $Y = 0$, then $X \in \{0, 1, 2\}$ with the same $1/4, 1/2, 1/4$ and each multiplied by $1/2$:

$$P(Y = 0, X = 0) = \tfrac{1}{8}, \quad P(Y = 0, X = 1) = \tfrac{1}{4}, \quad P(Y = 0, X = 2) = \tfrac{1}{8}.$$

All other pairs have probability 0. For (b),

$$P(X = 3 \mid Y = 1) = P(\text{remaining two are HH}) = 1/4.$$

5. Show that

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)},$$

if $P(B) \neq 0$. This is known as Bayes' theorem.

Solution. By definition, $P(A|B) = P(A \cap B)/P(B)$ and $P(B|A) = P(A \cap B)/P(A)$. Solving for $P(\cap B)$ in the second equation and plugging it into the first gives

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)},$$

which is what we wanted.

6. An urn contains 5 red, 3 blue, and 2 green balls. Two balls are drawn *without replacement.* What is:

   (a) The probability both are red.

   (b) The probability the two balls are the same color.

   (c) The probability the second ball is blue, given the first was green.

   Solution.

   (a)
   $$P(\text{both red}) = \frac{5}{10} \cdot \frac{4}{9} = \frac{20}{90} = \frac{2}{9}.$$

   (b) Sum over colors (RR, BB, GG):
   $$P(\text{same color}) = \frac{5}{10} \cdot \frac{4}{9} + \frac{3}{10} \cdot \frac{2}{9} + \frac{2}{10} \cdot \frac{1}{9} = \frac{20+6+2}{90} = \frac{14}{45}.$$

   (c) After drawing a green first, the urn has 5 red, 3 blue, 1 green left; total 9 balls. Thus
   $$P(\text{second blue} \mid \text{first green}) = \frac{3}{9} = \frac{1}{3}.$$

7. This is a classic phenomenon known as the "birthday paradox." Suppose there are 23 people in a room. Let's assume their birthdays are uniformly distributed across the days of the year. We're going to show that the probability that two people share a birthday is more than 50%.

   (a) Let A be the event that some two people in the room share a birthday. Express this probability in terms of the chances that no two people share a birthday.

   (b) Label the 23 people from 1 to 23. Let E_k be the event that person k does not share a birthday with person 1 through $k-1$. What is $P(E_k | E_1 \cap E_2 \cap \cdots \cap E_{k-1})$?

   (c) Express $P(\cap_{i=1}^{23} E_i)$ as a product of terms that look like $P(E_k | \cap_{j<k} E_j)$. (Recall that $\cap_{j<k} E_j$ is just a concise way of writing $E_1 \cap E_2 \cap \cdots \cap E_{k-1}$.)

   (d) What is $P(A)$?

   Solution.

(a) $P(A) = 1 - P(A^c)$, where A^c is the event that no two people share a birthday.

(b) If the event $\cap_{j<k} E_j$ occurs, it implies that persons 1 through $k-1$ have birthdays on different days. Given that event, person k must have a birthday on one of the $365 - (k-1)$ remaining days of the year for E_k to occur. Therefore,

$$P(E_k | \cap_{j<k} E_j) = \frac{365 - (k-1)}{365}.$$

(c) Recursively use the definition of conditional expectation:

$$\begin{aligned} P(\cap_{k=1}^{23} E_k) &= P(E_{23} | \cap_{k=1}^{22} E_k) P(\cap_{k=1}^{22} E_k) \\ &= P(E_{23} | \cap_{k=1}^{22} E_k) P(E_{22} | \cap_{k=1}^{21} E_k) P(\cap_{k=1}^{21} E_k) \\ &= \ldots \\ &= P(E_{23} | \cap_{k=1}^{22} E_k) P(E_{22} | \cap_{k=1}^{21} E_k) \cdots P(E_2 | E_1) P(E_1). \end{aligned}$$

(d) Note that the event A^c is precisely the event $\cap_{k=1}^{23} E_k$. Therefore, using the expansion of $P(\cap_{k=1}^{23} E_k)$ above, we have

$$P(\cap_{k=1}^{23} E_k) = \frac{343}{365} \cdot \frac{344}{365} \cdots \frac{364}{365} \approx 0.492.$$

(Note that $P(E_1) = 1$.)

(e) $P(A) = 1 - P(A^c) \approx 1 - 0.492 = 0.508$.