CSE 151: Introduction to Machine Learning	Winter 2021
Problem Set 1	
Instructor: Kamalika Chaudhuri	**Due on:** never

Problem 1 (10 points)

Let u_1 and u_2 be vectors such that $\|u_1\| = \|u_2\| = 1$, and $\langle u_1, u_2\rangle = 0$. For any vector x, we define $P(x)$ as the vector $P(x) = \langle x, u_1\rangle u_1 + \langle x, u_2\rangle u_2$.

1. How would you geometrically interpret $P(x)$? (Hint: Think about projections)
2. Show that: $\|P(x)\|^2 = \langle x, u_1\rangle^2 + \langle x, u_2\rangle^2$.
3. Using parts (1) and (2), show that $\|P(x)\| \leq \|x\|$. When is $\|P(x)\| = \|x\|$?

Solutions

1. $P(x)$ is the projection of x onto the subspace spanned by u_1 and u_2.

   Let V be the subspace spanned by u_1 and u_2. $P(x)$ is the projection of x onto subspace V if $x - P(x)$ is orthogonal to V. We first show that $x - P(x) \perp u_1$ and $x - P(x) \perp u_2$.

$$\begin{aligned}
\langle x - P(x), u_1\rangle &= \langle x - \langle x, u_1\rangle u_1 - \langle x, u_2\rangle u_2, u_1\rangle \\
&= \langle x, u_1\rangle - \langle\langle x, u_1\rangle u_1, u_1\rangle - \langle\langle x, u_2\rangle u_2, u_1\rangle \\
&= \langle x, u_1\rangle - \langle x, u_1\rangle\langle u_1, u_1\rangle - \langle x, u_2\rangle\langle u_2, u_1\rangle \\
&= \langle x, u_1\rangle - \langle x, u_1\rangle \cdot 1 - \langle x, u_2\rangle \cdot 0 \\
&= 0, \\
\langle x - P(x), u_2\rangle &= \langle x - \langle x, u_1\rangle u_1 - \langle x, u_2\rangle u_2, u_2\rangle \\
&= \langle x, u_2\rangle - \langle\langle x, u_1\rangle u_1, u_2\rangle - \langle\langle x, u_2\rangle u_2, u_2\rangle \\
&= \langle x, u_2\rangle - \langle x, u_1\rangle\langle u_1, u_2\rangle - \langle x, u_2\rangle\langle u_2, u_2\rangle \\
&= \langle x, u_2\rangle - \langle x, u_1\rangle \cdot 0 - \langle x, u_2\rangle \cdot 1 \\
&= 0.
\end{aligned}$$

   Since $x - P(x) \perp u_1$, $x - P(x) \perp u_2$ and u_1, u_2 are linearly independent, $x - P(x)$ is orthogonal to any vector in subspace V, which means that $x - P(x)$ is orthogonal to V. Therefore, $P(x)$ is the projection of x onto the subspace spanned by u_1 and u_2.

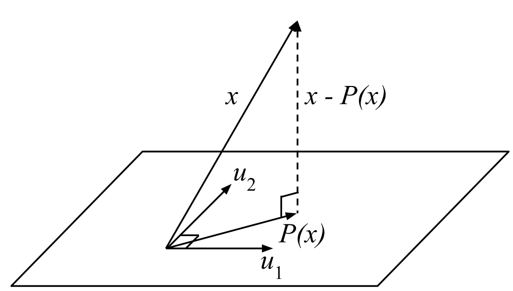

Figure 1: Visualization of $P(x)$, when $x, u_1, u_2 \in \mathbb{R}^3$

2. We show $\|P(x)\|^2 = \langle x, u_1\rangle^2 + \langle x, u_2\rangle^2$ by expanding $\|P(x)\|^2$.

$$\begin{aligned}
\|P(x)\|^2 &= \langle P(x), P(x)\rangle \\
&= \langle \langle x, u_1\rangle u_1 + \langle x, u_2\rangle u_2, \langle x, u_1\rangle u_1 + \langle x, u_2\rangle u_2\rangle \\
&= \langle \langle x, u_1\rangle u_1, \langle x, u_1\rangle u_1\rangle + \langle \langle x, u_1\rangle u_1, \langle x, u_2\rangle u_2\rangle + \langle \langle x, u_2\rangle u_2, \langle x, u_1\rangle u_1\rangle + \langle \langle x, u_2\rangle u_2, \langle x, u_2\rangle u_2\rangle \\
&= \langle x, u_1\rangle^2 \langle u_1, u_1\rangle + \langle x, u_1\rangle\langle x, u_2\rangle\langle u_1, u_2\rangle + \langle x, u_2\rangle\langle x, u_1\rangle\langle u_2, u_1\rangle + \langle x, u_2\rangle^2\langle u_2, u_2\rangle \\
&= \langle x, u_1\rangle^2 \cdot 1 + \langle x, u_1\rangle\langle x, u_2\rangle \cdot 0 + \langle x, u_2\rangle\langle x, u_1\rangle \cdot 0 + \langle x, u_2\rangle^2 \cdot 1 \\
&= \langle x, u_1\rangle^2 + \langle x, u_2\rangle^2
\end{aligned}$$

3. Since $P(x) \perp x - P(x)$, we have $\|x\|^2 = \|P(x)\|^2 + \|x - P(x)\|^2$. Or, from part (1), we have $\langle u_1, x - P(x)\rangle = 0$ and $\langle u_2, x - P(x)\rangle = 0$, thus

$$\begin{aligned}
\|x\|^2 &= \langle x, x\rangle \\
&= \langle P(x) + (x - P(x)), P(x) + (x - P(x))\rangle \\
&= \langle P(x), P(x)\rangle + \langle P(x), x - P(x)\rangle + \langle x - P(x), P(x)\rangle + \langle x - P(x), x - P(x)\rangle \\
&= \|P(x)\|^2 + 2\langle P(x), x - P(x)\rangle + \|x - P(x)\|^2 \\
&= \|P(x)\|^2 + 2(\langle \langle x, u_1\rangle u_1 + \langle x, u_2\rangle u_2, x - P(x)\rangle) + \|x - P(x)\|^2 \\
&= \|P(x)\|^2 + 2(\langle \langle x, u_1\rangle u_1, x - P(x)\rangle + \langle \langle x, u_2\rangle u_2, x - P(x)\rangle) + \|x - P(x)\|^2 \\
&= \|P(x)\|^2 + 2(\langle x, u_1\rangle\langle u_1, x - P(x)\rangle + \langle x, u_2\rangle\langle u_2, x - P(x)\rangle) + \|x - P(x)\|^2 \\
&= \|P(x)\|^2 + 2(\langle x, u_1\rangle \cdot 0 + \langle x, u_2\rangle \cdot 0) + \|x - P(x)\|^2 \\
&= \|P(x)\|^2 + \|x - P(x)\|^2.
\end{aligned}$$

Therefore, $\|P(x)\|^2 \leq \|x\|^2$. Since $\|P(x)\| \geq 0$ and $\|x\| \geq 0$, we have $\|P(x)\| \leq \|x\|$.

When $\|x - P(x)\|^2 = 0$, i.e. $x = P(x)$ or x itself is in the subspace spanned by u_1 and u_2, we have $\|P(x)\| = \|x\|$.

Problem 2 (10 points)

Given two column vectors x and y in d-dimensional space, the outer product of x and y is defined to be the $d \times d$ matrix $x \circ y = xy^\top$.

1. Show that for any x and y, $x^\top(x \circ y)y = \|x\|^2\|y\|^2$. When is this equal to $x^\top\langle x, y\rangle y$?

2. Show that for any non-zero x and y, the outer product $x \circ y$ always has rank 1.

3. Let $x_1, \ldots, x_n$ be n $d \times 1$ data vectors, and let X be the $n \times d$ data matrix whose i-th row is the row vector $x_i^\top$. Show that:

$$X^\top X = \sum_{i=1}^{n} x_i \circ x_i$$

Solutions

We know that for any vector x, $x^\top x = \|x\|^2$. Thus,

$$x^\top(x \circ y)y = x^\top(xy^\top)y = (x^\top x)(y^\top y) = \|x\|^2\|y\|^2$$

Also, $x^\top\langle x, y\rangle y = \langle x, y\rangle(x^\top y) = \langle x, y\rangle\langle x, y\rangle = \langle x, y\rangle^2 = (\|x\|\|y\|\cos\theta)^2 = \|x\|^2\|y\|^2\cos^2\theta$. This quantity is equal to $\|x\|^2\|y\|^2$ when $\theta = 0°$ or $180°$. This means that the two quantities are equal when the vectors x and y are collinear.

Let x_i be the ith element of vector x and y_i be the ith element of vector y. Thus,

$$x \circ y = xy^\top = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} [y_1, y_2, \cdots, y_d] = \begin{bmatrix} x_1y_1 & x_1y_2 & \cdots & x_1y_d \\ x_2y_1 & x_2y_2 & \cdots & x_2y_d \\ \vdots & \vdots & & \vdots \\ x_dy_1 & x_dy_2 & \cdots & x_dy_d \end{bmatrix}$$

Notice that every row is a scalar multiple of the first row of the above matrix. Therefore, when this matrix is reduced to a row echelon form, it will contain only one non-zero row. Therefore, the outer product $x \circ y$ always has rank 1.

Let $Y = X^\top X$. Therefore, Y is a $d \times d$ matrix. Let x_{ij} be the jth element of vector x_i. Therefore, X can be written as

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{bmatrix}$$

Therefore,

$$Y = X^T X = \begin{bmatrix} x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & x_{22} & \cdots & x_{n2} \\ \vdots & \vdots & & \vdots \\ x_{1d} & x_{2d} & \cdots & x_{nd} \end{bmatrix} \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{bmatrix} =$$

This works out to give $Y_{ij} = \sum_{k=1}^{n} x_{ki}x_{kj}$ where $i,j = 1, 2...d$. Now we work out the right side of the equation.

$$\sum_{k=1}^{n} x_k \circ x_k = \sum_{k=1}^{n} x_k x_k^\top = \sum_{k=1}^{n} \begin{bmatrix} x_{k1} \\ x_{k2} \\ \vdots \\ x_{kd} \end{bmatrix} [x_{k1}, x_{k2} \cdots x_{kd}] = \sum_{k=1}^{n} \begin{bmatrix} x_{k1}^2 & x_{k1}x_{k2} & \cdots & x_{k1}x_{kd} \\ x_{k2}x_{k1} & x_{k2}^2 & \cdots & x_{k2}x_{kd} \\ \vdots & \vdots & & \vdots \\ x_{kd}x_{k1} & x_{kd}x_{k2} & \cdots & x_{kd}^2 \end{bmatrix}$$

Thus, the right side of the equation equals Y.

Problem 3 (10 points)

Suppose A and B are $d \times d$ matrices which are symmetric (in the sense that $A_{ij} = A_{ji}$ and $B_{ij} = B_{ji}$ for all i and j) and positive semi-definite. Also suppose that u is a $d \times 1$ vector such that $\|u\| = 1$. Which of the following matrices are always positive semi-definite, no matter what A, B and u are? Justify your answer.

1. $10A$.
2. $A + B$.
3. $uu^\top$.
4. $A - B$.
5. $I - uu^\top$ (Hint: Write down $x^\top(I - uu^\top)x$ in terms of some dot-products, and try usng Cauchy-Schwartz.)

Solutions

A general strategy for solving this problem is to first try to prove that the matrix M is positive semi-definite; if you fail, then try to find a counter-example to disprove the claim. For the latter, you need find out a *specific vector* x for which $x^\top Mx < 0$.
By the definition of positive semi-definite matrices, for all $d \times 1$ vector x,

$$x^\top Ax \geq 0, x^\top Bx \geq 0$$

1. For the matrix $10A$, for all $d \times 1$ vector x,

$$x^\top (10A)x = 10(x^\top Ax) \geq 0$$

thus it is positive semidefinite.
2. For the matrix $A + B$, for all $d \times 1$ vector x,

$$x^\top (A + B)x = (x^\top Ax) + (x^\top Bx) \geq 0,$$

as both $x^\top Ax$ and $x^\top Bx$ are ≥ 0. Thus it is positive semidefinite.
3. For the matrix $uu^\top$, for all $d \times 1$ vector x,

$$x^\top (uu^\top)x = (x^\top u)(u^\top x) = (\langle x, u\rangle)(\langle u, x\rangle) = (\langle x, u\rangle)^2 \geq 0$$

thus it is positive semidefinite.
4. The matrix $A - B$ is not always positive semi-definite. As a concrete counter-example, take $d = 2$, $A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, and $B = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$. Then $A - B = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}$. There exists a 2×1 vector $x = [1, 0]^\top$ such that

$$x^\top (A - B)x = -1$$

which proves that $A - B$ is in fact not positive semi-definite.
5. For the matrix $I - uu^\top$, for all $d \times 1$ vector x,

$$x^\top (I - uu^\top)x = x^\top x - (\langle x, u\rangle)^2$$

Now applying Cauchy-Schwarz to $(\langle x, u\rangle)$ and using the fact that $\|u\| = 1$, we find that

$$(\langle x, u\rangle)^2 \leq \|x\|^2\|u\|^2 = \|x\|^2 = x^\top x$$

Thus, we conclude

$$x^\top (I - uu^\top)x \geq 0$$

This establishes the fact that $(I - uu^\top)$ is positive semi-definite.

Problem 4 (10 points)

In class, we discussed how to define a *norm* or a *length* for a vector. It turns out that one can also define a norm or a length for a matrix. Two popular matrix norms are the Frobenius norm and the spectral norm. The Frobenius norm of a $m \times n$ matrix A, denoted by $\|A\|_F$ is defined as:

$$\|A\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij}^2}$$

The spectral norm of a $m \times n$ matrix A, denoted by $\|A\|$ is defined as:

$$\|A\| = \max_x \frac{\|Ax\|}{\|x\|}$$

where x is a $n \times 1$ vector.

1. Let I be the $n \times n$ identity matrix. What is its Frobenius norm? What is its spectral norm? Justify your answer.
2. Suppose $A = uv^\top$ where u is a $m \times 1$ vector and v is a $n \times 1$ vector. Write down the Frobenius norm of A as function of $\|u\|$ and $\|v\|$. Justify your answer.
3. Write down the spectral norm of A in terms of $\|u\|$ and $\|v\|$. Justify your answer.

Solutions

Since I is an $n \times n$ identity matrix, therefore it has n elements along the diagonal which are 1 and all the remaining elements are 0. Therefore, the Frobenius norm of I is given by

$$\|I\|_F = \sqrt{n}$$

The spectral norm of I is given by

$$\|I\| = \max_x \frac{\|Ix\|}{\|x\|} = \max_x \frac{\|x\|}{\|x\|} = 1$$

Let $u = [u_1, u_2 \ldots u_m]^\top$ and $v = [v_1, v_2 \ldots v_n]^\top$. Since $A = uv^\top$, therefore

$$A = \begin{bmatrix} u_1v_1 & u_1v_2 & \cdots & u_1v_n \\ u_2v_1 & u_2v_2 & \cdots & u_2v_n \\ \vdots & \vdots & & \vdots \\ u_mv_1 & u_mv_2 & \cdots & u_mv_n \end{bmatrix}$$

The Frobenius norm of A is given by

$$\|A\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} u_i^2 v_j^2} = \sqrt{\sum_{i=1}^{m} u_i^2 \sum_{j=1}^{n} v_j^2} = \sqrt{\|u\|^2\|v\|^2} = \|u\|\|v\|$$

In order to find the spectral norm of A, observe that for any $n \times 1$ x,

$$\|Ax\| = \|u\langle v, x\rangle\| = \|u\||\langle v, x\rangle| = \|u\|\|v\|\|x\||\cos\theta|$$

where θ is the angle between v and x.

$|\cos\theta|$ attains a maximum value of 1 at $\theta = 0$ or 180. Therefore, $\|A\| = \|u\|\|v\|$.

Problem 5 (10 points)

Let x be a $d \times 1$ vector. Let y_i be constants, z_i be $d \times 1$ constant vectors, and β_i be $d \times 1$ constant vectors for $1 \leq i \leq n$. Write down the gradients for each of the following multivariate functions with respect to x. Given the other parameters describing the function, what is the time required to compute the gradient at a specific value of x?

1. $F(x) = \sum_{i=1}^n \log(1 + e^{-y_i x^\top z_i})$.
2. $G(x) = \sum_{i=1}^n (x^\top \beta_i - y_i)^2$.
3. $H(x) = \sum_{i=1}^d x_i \log \frac{1}{x_i}$.
4. $J(x) = \log(\sum_{i=1}^d e^{2x_i})$.

Solutions

1. We use a special case of the multivariate chain rule: if $g : \mathbb{R}^d \to \mathbb{R}$ and $f : \mathbb{R} \to \mathbb{R}$, then $\frac{\partial}{\partial x_j} f(g(x)) = \frac{\partial}{\partial x_j} g(x) \frac{\partial}{\partial g(x)} f(g(x))$. Differentiating $F(x)$ with respect to x_j, we have

$$
\begin{aligned}
\frac{\partial}{\partial x_j} F(x) &= \sum_{i=1}^{n} \frac{\partial}{\partial x_j} \log(1 + e^{-y_i x^\top z_i}) \\
&= \sum_{i=1}^{n} \left(\frac{\partial}{\partial x_j} (1 + e^{-y_i x^\top z_i}) \right) \frac{1}{1 + e^{-y_i x^\top z_i}}
\end{aligned}
$$

Now, we know

$$
\begin{aligned}
\frac{\partial}{\partial x_j} (1 + e^{-y_i x^\top z_i}) &= \frac{\partial}{\partial x_j} e^{-y_i x^\top z_i} \\
&= \left(\frac{\partial}{\partial x_j} (-y_i x^\top z_i) \right) \left(e^{-y_i x^\top z_i} \right) \\
&= (-y_i (z_i)_j) \left(e^{-y_i x^\top z_i} \right).
\end{aligned}
$$

where $(z_i)_j$ is the jth element of z_i. Thus, the answer is

$$
\frac{\partial}{\partial x_j} F(x) = \sum_{i=1}^{n} \frac{-(z_i)_j y_i e^{-y_i x^\top z_i}}{1 + e^{-y_i x^\top z_i}}.
$$

We can compute the gradient of F as follows: first, store $x^\top z_i$ for $1 \leq i \leq n$, taking $O(nd)$ operations. Then, compute $\frac{\partial}{\partial x_j} F(x)$ for $1 \leq j \leq d$, taking $O(n)$ operations for each j. This takes $O(nd) + O(nd) = O(nd)$ operations in total.

2. Differentiating $G(x)$ with respect to x_j, we have

$$
\begin{aligned}
\frac{\partial}{\partial x_j} G(x) &= \sum_{i=1}^{n} \frac{\partial}{\partial x_j} (x^\top \beta_i - y_i)^2 \\
&= \sum_{i=1}^{n} \left(\frac{\partial}{\partial x_j} (x^\top \beta_i - y_i) \right) 2(x^\top \beta_i - y_i) \\
&= \sum_{i=1}^{n} 2(\beta_i)_j (x^\top \beta_i - y_i)
\end{aligned}
$$

where $(\beta_i)_j$ is the jth element of β_i. We can compute the gradient of G as follows: first, store $x^\top \beta_i$ for $1 \leq i \leq n$, taking $O(nd)$ operations. Then, compute $\frac{\partial}{\partial x_j} G(x)$ for $1 \leq j \leq d$, taking $O(n)$ operations for each j. This takes $O(nd) + O(nd) = O(nd)$ operations in total.

3. Differentiating $H(x)$ with respect to x_j, we have

$$
\begin{aligned}
\frac{\partial}{\partial x_j} H(x) &= \sum_{i=1}^{d} \frac{\partial}{\partial x_j} \left(x_i \log \frac{1}{x_i} \right) \\
&= \frac{\partial}{\partial x_j} \left(x_j \log \frac{1}{x_j} \right).
\end{aligned}
$$

This is because $\frac{\partial}{\partial x_j}\left(x_i \log \frac{1}{x_i}\right) = 0$ for $i \neq j$. Finally,

$$\begin{aligned} \frac{\partial}{\partial x_j}\left(x_j \log \frac{1}{x_j}\right) &= \log \frac{1}{x_j} + x_j \left(\frac{1}{1/x_j}\right)\left(-\frac{1}{x_j^2}\right) \\ &= \log \frac{1}{x_j} - 1. \end{aligned}$$

We can compute $\frac{\partial}{\partial x_j} H(x)$ for each $1 \leq j \leq d$ in $O(1)$ time, resulting in $O(d)$ total operations.

4. Differentiating $J(x)$ with respect to x_j, we have

$$\begin{aligned} \frac{\partial}{\partial x_j} J(x) &= \left(\frac{\partial}{\partial x_j} \sum_{i=1}^{d} e^{2x_i}\right) \frac{1}{\sum_{i=1}^{d} e^{2x_i}} \\ &= \left(\frac{\partial}{\partial x_j} e^{2x_j}\right) \frac{1}{\sum_{i=1}^{d} e^{2x_i}} \end{aligned}$$

This is because $\frac{\partial}{\partial x_j} e^{2x_i} = 0$ for $i \neq j$. Finally,

$$\frac{\partial}{\partial x_j} e^{2x_j} = 2e^{2x_j}.$$

Thus,

$$\frac{\partial}{\partial x_j} J(x) = \frac{2e^{2x_j}}{\sum_{i=1}^{d} e^{2x_i}}$$

We can compute $\frac{\partial}{\partial x_j}$ for $1 \leq j \leq d$ by storing $\sum_{i=1}^{d} e^{2x_i}$, using $O(d)$ operations. Computing $\frac{\partial}{\partial x_j}$ for a particular j takes $O(1)$ operations, and thus the total number of operations is $O(d) + O(d) = O(d)$.