AI 535 Deep Learning - Assignment #1

Hyuntaek Oh

ohhyun@oregonstate.edu

Due: Jan. 27, 2025

1. (Optimization) Compute the gradient $\nabla f(\mathbf{x})$ and Hessian $\nabla^2 f(\mathbf{x})$ of the function (5 points)

$$f(\mathbf{x}) = (x_1 + x_2)(x_1 x_2 + x_1 x_2^2)$$

Find at least 3 stationary points of this function (3 points). Show that $[3/8,\ -6/8]^T$ is a local maximum of this function (2 points).

(i) Compute $\nabla f(\mathbf{x})$*, which is* $\begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \end{bmatrix}$ *and Hessian* $\nabla^2 f(\mathbf{x})$*, which is* $\begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} \end{bmatrix}$

(i-1) $\begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \end{bmatrix}$

$$\frac{\partial f}{\partial x_1} = \frac{\partial}{\partial x_1}[(x_1 + x_2)(x_1 x_2 + x_1 x_2^2)]$$
$$= (x_1 x_2 + x_1 x_2^2) + (x_1 + x_2)(x_2 + x_2^2)$$
$$= x_1 x_2 + x_1 x_2^2 + x_1 x_2 + x_1 x_2^2 + x_2^2 + x_2^3$$
$$= 2x_1 x_2 + 2x_1 x_2^2 + x_2^2 + x_2^3$$

$$\frac{\partial f}{\partial x_2} = \frac{\partial}{\partial x_2}[(x_1 + x_2)(x_1 x_2 + x_1 x_2^2)]$$
$$= (x_1 x_2 + x_1 x_2^2) + (x_1 + x_2)(x_1 + 2x_1 x_2)$$
$$= x_1 x_2 + x_1 x_2^2 + x_1^2 + 2x_1^2 x_2 + x_1 x_2 + 2x_1 x_2^2$$
$$= 2x_1 x_2 + 3x_1 x_2^2 + x_1^2 + 2x_1^2 x_2$$

$$\nabla f(\mathbf{x}) = \begin{bmatrix} 2x_1 x_2 + 2x_1 x_2^2 + x_2^2 + x_2^3 \\ 2x_1 x_2 + 3x_1 x_2^2 + x_1^2 + 2x_1^2 x_2 \end{bmatrix}$$

(i-2) $\begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} \end{bmatrix}$

$$\frac{\partial}{\partial x_1}(2x_1 x_2 + 2x_1 x_2^2 + x_2^2 + x_2^3) = 2x_2 + 2x_2^2$$

$$\frac{\partial}{\partial x_2}(2x_1 x_2 + 2x_1 x_2^2 + x_2^2 + x_2^3) = 2x_1 + 4x_1 x_2 + 2x_2 + 3x_2^2$$

$$\frac{\partial}{\partial x_2}(2x_1 x_2 + 3x_1 x_2^2 + x_1^2 + 2x_1^2 x_2) = 2x_1 + 6x_1 x_2 + 2x_1^2$$

$$\nabla f(\mathbf{x}) = \begin{bmatrix} 2x_2 + 2x_2^2 & 2x_1 + 4x_1 x_2 + 2x_2 + 3x_2^2 \\ 2x_1 + 4x_1 x_2 + 2x_2 + 3x_2^2 & 2x_1 + 6x_1 x_2 + 2x_1^2 \end{bmatrix}$$

(ii) Find at least 3 stationary points of the function.
The stationary points of the function are when the derivative of the function is equal to zero.

$$2x_1x_2 + 2x_1x_2^2 + x_2^2 + x_2^3 = 0$$

$$2x_1x_2 + 3x_1x_2^2 + x_1^2 + 2x_1^2x_2 = 0$$

$$if\ x_1 = 0,\ x_2^2 + x_2^3 = 0\ \rightarrow x_2^2(1 + x_2) = 0 \Rightarrow x_2 = 0, -1$$

$$possible\ stationary\ points : (0, 0),\ (0, -1)$$

$if\ x_2 = 0$, *there are no new stationary points since the values are zeros.*

if neither x_1 *nor* x_2 *are zero,* $2x_1 + 2x_1x_2 + x_2 + x_2^2 = 0,$

$$2x_2 + 3x_2^2 + x_1 + 2x_1x_2 = 0$$

$$possible\ stationary\ points : (3/8, -3/4),\ (1, -1)$$

Thus, stationary points are:

$$(0, 0),\ (0, -1),\ (3/8, -3/4),\ (1, -1)$$

(iii) Show that $[3/8,\ -6/8]^T$ *is a local maximum of this function.*

$$\nabla f(\mathbf{x}) = \begin{bmatrix} 2(-\frac{6}{8}) + 2(-\frac{6}{8})^2 & 2(\frac{3}{8}) + 4(\frac{3}{8})(-\frac{6}{8}) + 2(-\frac{6}{8}) + 3(-\frac{6}{8})^2 \\ 2(\frac{3}{8}) + 4(\frac{3}{8})(-\frac{6}{8}) + 2(-\frac{6}{8}) + 3(-\frac{6}{8})^2 & 2(\frac{3}{8}) + 6(\frac{3}{8})(-\frac{6}{8}) + 2(\frac{3}{8})^2 \end{bmatrix}$$

$$\nabla f(\mathbf{x}) = \begin{bmatrix} -\frac{3}{8} & -\frac{3}{16} \\ -\frac{3}{16} & -\frac{21}{32} \end{bmatrix}$$

Eigenvalues formula:

$$det(A) = (-0.375)(-0.65625) - (-0.1875)^2$$

$$\lambda = \frac{-1.03125}{2} \pm \sqrt{(-\frac{1.03125}{2})^2 - 0.2109375}$$

$$\lambda_1 = -0.28125,\ \lambda_2 = -.75$$

Since both eigenvalues are negative, the Hessian matrix is negative definite.
Thus, $[3/8,\ -6/8]^T$ *is a local maximum.*

2. (Optimization) Show that the function $f(\mathbf{x}) = 8x_1 + 12x_2 + x_1^2 - 2x_2^2$ has only one stationary point (4 points), and that it is neither a minimum nor a maximum, but is a saddle point (4 points).
(i) show that the function has only one stationary point.

$$\nabla f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 8 + 2x_1 \\ 12 - 4x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$x_1 = -4,\ x_2 = 3 \Rightarrow \textit{stationary point} : (-4, 3)$$

(ii) show that it is neither a minimum nor a maximum, but is a saddle point.

$$\nabla^2 f(\mathbf{x}) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 0 & -4 \end{bmatrix}$$

The eigenvalues of the Hessian matrix are the diagonal entries:

$$\lambda_1 = 2,\ \lambda_2 = -4$$

Since the Hessian has both positive and negative eigenvalues, it implies that the stationary point (-4, 3) is a saddle point, meaning that it is neither a local minimum nor a maximum.

3. (Linear Algebra) If **A** and **B** are positive definite matrices, prove that the matrix $\begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix}$ is also positive definite (7 points).
*To prove it, we need to use contradiction, assuming that the matrix M is **not** positive definite:*

$$\mathbf{x}^T M \mathbf{x} \leq 0$$

$$\mathbf{x}^T M \mathbf{x} = [\mathbf{x}_1^T, \mathbf{x}_2^T]. \begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} = \mathbf{x}_1^T A \mathbf{x}_1 + \mathbf{x}_2^T B \mathbf{x}_2$$

$$\mathbf{x}_1^T A \mathbf{x}_1 + \mathbf{x}_2^T B \mathbf{x}_2 \leq 0$$

Since A and B are positive definite, we know that for any nonzero vectors:

$$\mathbf{x}_1^T A \mathbf{x}_1 > 0 \textit{ if } \mathbf{x}_1 \neq 0$$

$$\mathbf{x}_2^T B \mathbf{x}_2 > 0 \textit{ if } \mathbf{x}_2 \neq 0$$

Therefore, the sum $\mathbf{x}_1^T A \mathbf{x}_1 + \mathbf{x}_2^T B \mathbf{x}_2$ can only be non-positive if both terms are zero, meaning that $\mathbf{x}_1 = 0,\ \mathbf{x}_2 = 0$.
*However, this contradicts our original assumption that **x** is a nonzero vector.*
Thus, the matrix M is positive definite.

4. (Chain Rule Calculus) Consider this function: $f(\mathbf{x}) = \mathbf{w}_2^T sigmoid(\mathbf{W}_1 \mathbf{x})$, where $sigmoid(x) = \frac{1}{1+e^{-x}}$ applies to each entry of the vector, please compute the derivatives of $\frac{\partial f}{\partial \mathbf{w}_2}$, $\frac{\partial f}{\partial \mathbf{W}_1}$, $\frac{\partial f}{\partial \mathbf{x}}$ (15 points), $\mathbf{W}_1$ is $c \times d$, $\mathbf{x}$ is $d \times 1$, $\mathbf{w}_2$ is $c \times 1$.

The result of the calculation $\boldsymbol{W}_1\boldsymbol{x}$ *is* $c \times 1$ *since the matrix product of* $\boldsymbol{W}_1\boldsymbol{x} = (c \times d)\ (d \times 1)$ *is* $c \times 1$. *This means the sigmoid function in the* $f(\boldsymbol{x})$ *is vector. So, it can be expressed as:*

$$sigmoid(\boldsymbol{W}_1\boldsymbol{x}) = \begin{bmatrix} \sigma([\boldsymbol{W}_1\boldsymbol{x}]_1) \\ \sigma([\boldsymbol{W}_1\boldsymbol{x}]_2) \\ \sigma([\boldsymbol{W}_1\boldsymbol{x}]_3) \\ \dots \\ \sigma([\boldsymbol{W}_1\boldsymbol{x}]_c) \end{bmatrix}$$

Thus, the function $f(\boldsymbol{x})$ *is:*

$$f(\boldsymbol{x}) = [w_{21},\ w_{22},\ w_{23},\ \dots,\ w_{2c}] \begin{bmatrix} \sigma([\boldsymbol{W}_1\boldsymbol{x}]_1) \\ \sigma([\boldsymbol{W}_1\boldsymbol{x}]_2) \\ \sigma([\boldsymbol{W}_1\boldsymbol{x}]_3) \\ \dots \\ \sigma([\boldsymbol{W}_1\boldsymbol{x}]_c) \end{bmatrix}$$

$$= w_{21}\sigma([\boldsymbol{W}_1\boldsymbol{x}]_1) + w_{22}\sigma([\boldsymbol{W}_1\boldsymbol{x}]_2) + \dots + w_{2c}\sigma([\boldsymbol{W}_1\boldsymbol{x}]_c)$$

$$= \sum_{i=1}^{c} w_{2i} \cdot \sigma([\boldsymbol{W}_1\boldsymbol{x}]_i)$$

(i) $\frac{\partial f}{\partial \boldsymbol{w}_2}$:

$$\frac{\partial f}{\partial \boldsymbol{w}_2} = \sigma(\boldsymbol{W}_1\boldsymbol{x})$$

(ii) $\frac{\partial f}{\partial \boldsymbol{W}_1}$:

$$\frac{\partial f}{\partial \boldsymbol{W}_1} = \frac{\partial f}{\partial \sigma} \cdot \frac{\partial \sigma}{\partial \boldsymbol{W}_1} \quad (chain\ rule)$$

$$\frac{\partial f}{\partial \sigma} = \boldsymbol{w}_2^T$$

The derivative of the sigmoid vector $\sigma(\boldsymbol{W}_1\boldsymbol{x})$ *with respect to* $\boldsymbol{W}_1\boldsymbol{x}$ *is:*

$$\frac{\partial \sigma}{\partial \boldsymbol{W}_1} = diag(\sigma(\boldsymbol{W}_1\boldsymbol{x})(1 - \sigma(\boldsymbol{W}_1\boldsymbol{x})))$$

Then, the derivative with respect to $\boldsymbol{W}_1$ using the chain rule:

$$\frac{\partial(\boldsymbol{W}_1\mathbf{x})}{\partial \boldsymbol{W}_1} = \mathbf{x}^T$$

Finally, combining all these results:

$$\frac{\partial f}{\partial \boldsymbol{W}_1} = (diag(\sigma(\boldsymbol{W}_1\mathbf{x})(1-\sigma(\boldsymbol{W}_1\mathbf{x})))\mathbf{w}_2)\mathbf{x}^T$$

(iii) $\frac{\partial f}{\partial \mathbf{x}}$

$$\frac{\partial f}{\partial \mathbf{x}} = \frac{\partial f}{\partial \sigma} \cdot \frac{\partial \sigma}{\partial \mathbf{x}} \quad (chain\ rule)$$

$$\frac{\partial f}{\partial \sigma} = \mathbf{w}_2^T$$

$$\frac{\partial \sigma}{\partial \mathbf{x}} = diag(\sigma(\boldsymbol{W}_1\mathbf{x})(1-\sigma(\boldsymbol{W}_1\mathbf{x}))\boldsymbol{W}_1$$

Thus, the derivative with respect to $\mathbf{x}$ is:

$$\frac{\partial f}{\partial \mathbf{x}} = \boldsymbol{W}_1^T diag(\sigma(\boldsymbol{W}_1\mathbf{x})(1-\sigma(\boldsymbol{W}_1\mathbf{x})))\mathbf{w}_2$$

5. (High Dimensional Statistics ("Curse of Dimensionality")) Consider N data points independent and uniformly distributed in a p-dimensional unit ball B (for every $x \in B, ||x||^2 \leq 1$), centered at the origin. The median distance from the origin to the closest data point is given by the expression:
$d(p, N) = (1 - \frac{1}{2}^{\frac{1}{N}})^{\frac{1}{P}}$
Prove this expression (8 points). Compute the median distance $d(p, N)$ for $N = 10,000,\ p = 1,000$ (2 points).
(i) Prove the expression $d(p, N) = (1 - \frac{1}{2}^{\frac{1}{N}})^{\frac{1}{p}}$
For the unit ball where $R = 1$, it can be simplified to:

$$V_p(1) = \frac{\pi^{\frac{p}{2}}}{\Gamma(\frac{p}{2}+1)}$$

The volume of a ball in p-dimensional space that has radius r is:

$$V_P(r) = \frac{\pi^{\frac{p}{2}}}{\Gamma(\frac{p}{2}+1)} r^p$$

The probability that a randomly chosen point lies within a radius r is:

$$P(||\mathbf{x}|| \leq r) = \frac{V_p(r)}{V_p(1)} = r^p$$

We assume that the event that all N points are outside a given radius r to find the median distance of the closest point. Since points are uniformly distributed, the probability that all N points is outside a ball of radius r is:

$$(1 - r^p)^N$$

The median distance is the value of r such that at least one point lies inside the ball with probability 50% :

$$1 - (1 - r^p)^N = \frac{1}{2}$$

This can be solved for r:

$$(1 - r^p)^N = \frac{1}{2}$$

$$1 - r^p = (\frac{1}{2})^{\frac{1}{N}}$$

$$r^p = 1 - (\frac{1}{2})^{\frac{1}{N}}$$

$$r = (1 - (\frac{1}{2})^{\frac{1}{N}})^{\frac{1}{p}}$$

Thus, we get:

$$d(p, N) = (1 - \frac{1}{2}^{\frac{1}{N}})^{\frac{1}{P}}$$

(ii) Compute the distance $d(p, N)$ for $N = 10,000,\ p = 1,000$

$$d(1000,\ 10000) = (1 - \frac{1}{2}^{\frac{1}{10000}})^{\frac{1}{1000}}$$

For large N, the term using logarithms can be approximated:

$$(\frac{1}{2})^{\frac{1}{10{,}000}} = e^{\frac{ln(\frac{1}{2})}{10000}}$$

$$= e^{-\frac{ln\ 2}{10000}} \approx 1 - \frac{ln\ 2}{10000} \approx 1 - 0.0000693$$

$$1 - (\frac{1}{2})^{\frac{1}{10000}} \approx 0.0000693$$

Then, taking the $\frac{1}{1{,}000}$ root:

$$d(1000,\ 10000) \approx (0.0000693)^{\frac{1}{1000}}$$

With logarithm expansion:

$$\approx e^{\frac{ln(0.0000693)}{1000}} \approx e^{-0.00957} \approx 1 - 0.00957$$

Thus, the approximate value is:

$$d(1000, 10000) \approx 0.9904$$