https://www.overleaf.com/project/62539e5894d9794af78c867f

AI535 Deep Learing – Homework 1

Background Math, Autograd, and Backpropagation

Solution by: Bikram Pandit (panditb@oregonstate.edu)

April 12, 2022

1 Flexing our mathematical Muscles

▶ Q1 Convexity [10pts]. Suppose $f(x)$ is a convex function and let $a < b$. Show that:

$$\frac{f(x)-f(a)}{x-a} \leq \frac{f(b)-f(a)}{b-a} \leq \frac{f(b)-f(x)}{b-x} \tag{1}$$

for $x \in (a, b)$. Draw a sketch that illustrates this inequality.

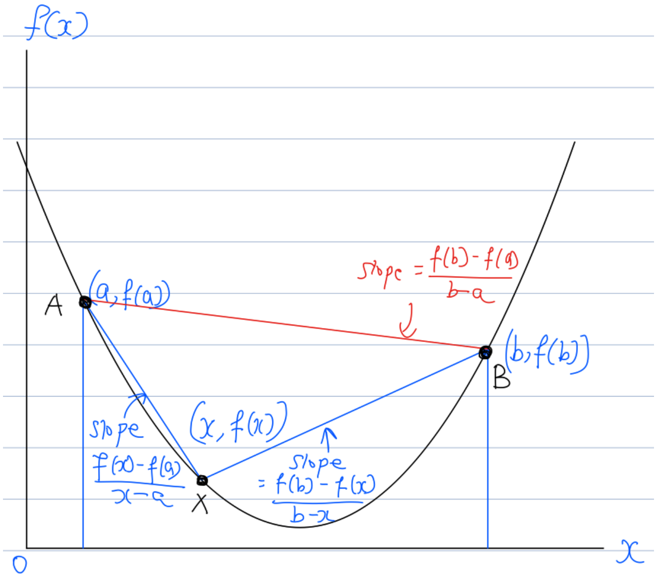
f(x)
A
(a, f(a))
Slope = $\frac{f(b)-f(a)}{b-a}$
(b, f(b))
B
Slope $\frac{f(x)-f(a)}{x-a}$
(x, f(x))
X
Slope = $\frac{f(b)-f(x)}{b-x}$
0
x

Figure 1: The point X exists in the bound of $x \in (a, b)$ which can move freely from point A to B as shown in the figure. Each term in three inequalities determines the slope of a line. It is clear from the figure that the slope of line AX is always less than AB which is the actually the meaning of first two inequalities given in the problem. Similarly, the slope of line BX is always greater than line AB which is the meaning of last two inequalities. This is visual representation of how inequalities hold true.

Proof:

First we will prove inequalities on left-side. Let $t = \frac{x-a}{b-a}$. Given that $x \in (a, b)$ and $a < b$, we can say $0 < t < 1$. Let's rewrite $f(x)$ as following:-

$$\begin{aligned}
f(x) &= f(a + x - a) \\
&= f(a + \frac{x-a}{b-a}(b-a)) \\
&= f(a + t(b-a)) \\
&= f(tb + a - ta) \\
&= f(tb + (1-t)a) \\
&\leq tf(b) + (1-t)f(a) \qquad \text{[According to the convexity property]} \\
&\leq tf(b) + f(a) - tf(a) \\
\Rightarrow f(x) - f(a) &\leq tf(b) - tf(a) \qquad \text{[Subtracting f(a) from both sides]} \\
\Rightarrow \frac{f(x)-f(a)}{x-a} &\leq \frac{tf(b) - tf(a)}{x-a} \qquad \text{[Diving both side by x - a]} \\
&\leq \frac{t}{x-a}(f(b) - f(a)) \\
&\leq \frac{f(b)-f(a)}{b-a} \qquad \text{[Because } t = \frac{x-a}{b-a}\text{]}
\end{aligned} \tag{2}$$

Similarly, we will prove inequalities on right-side. Let $t = \frac{b-x}{b-a}$. Given that $x \in (a, b)$ and $a < b$, we can say $0 < t < 1$. Let's rewrite $f(x)$ as following:-

$$\begin{aligned}
f(x) &= f(b + x - b) \\
&= f(b + \frac{x-b}{b-a}(b-a)) \\
&= f(b + \frac{b-x}{b-a}(a-b)) \\
&= f(b + t(a-b)) \\
&= f(ta + b - tb) \\
&= f(ta + (1-t)b) \\
&\leq tf(a) + (1-t)f(b) \qquad \text{[According to the convexity property]} \\
&\leq tf(a) + f(b) - tf(b) \\
\Rightarrow f(x) - f(b) &\leq tf(a) - tf(b) \qquad \text{[Subtracting f(b) from both sides]} \\
\Rightarrow f(b) - f(x) &\geq tf(b) - tf(a) \qquad \text{[Multiplying -1 on both sides. This changes inequality.]} \\
\Rightarrow \frac{f(b)-f(x)}{b-x} &\geq \frac{tf(b) - tf(a)}{b-x} \qquad \text{[Diving both side by b - x]} \\
&\geq \frac{t}{b-x}(f(b) - f(a)) \\
&\geq \frac{f(b)-f(a)}{b-a} \qquad \text{[Because } t = \frac{b-x}{b-a}\text{]} \\
\Rightarrow \frac{f(b)-f(a)}{b-a} &\leq \frac{f(b)-f(x)}{b-x}
\end{aligned} \tag{3}$$

Putting eq(2) and (3) proof together, we get,

$$\frac{f(x)-f(a)}{x-a} \leq \frac{f(b)-f(a)}{b-a} \leq \frac{f(b)-f(x)}{b-x} \tag{4}$$

▶ Q2 Linear Algebra [5 pts]. If A and B are positive definite matrices, prove that the matrix $\begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix}$ is also positive definite. 0's here represent matrices of zeros of sufficient size to fill.

Proof:

If A and B are positive definite then by definition we can say,

$$\begin{aligned} x^T Ax > 0 \forall x \in \mathbb{R}^n \\ y^T By > 0 \forall y \in \mathbb{R}^n \end{aligned} \tag{5}$$

Let z be formed by using x and y as follows,

$$\begin{aligned} \text{Let,} \qquad & z = \begin{bmatrix} x \\ y \end{bmatrix} \\ \text{Compute,} \qquad & z^T \begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix} z \\ & = \begin{bmatrix} x \\ y \end{bmatrix}^T \begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \\ & = x^T Ax + y^T By \\ & > 0 \qquad \text{[From equation 5]} \end{aligned} \tag{6}$$

It is clear that if $\forall x \in \mathbb{R}^n$ and $\forall y \in \mathbb{R}^n$ then $\forall z \in \mathbb{R}^{2\times n}$. Hence, given matrix is also a positive definite.

▶ Q3 Calculus I [10pts]. Compute the gradient $\nabla f(\mathbf{x})$ and Hessian $\nabla^2 f(x)$ for the function

$$f(\mathbf{x}) = (x_1 + x_2)(x_1 x_2 + x_1 x_2^2) \tag{7}$$

Find at least three stationary points of this function and compute the Hessian at each. Show that $\mathbf{x} = [3/8, -6/8]^T$ is a local maxima.

Proof:

Rewriting $f(x)$ and taking partial derivative with respect to x_1 and x_2 we get,

$$\begin{aligned} f(x) &= x_1^2 x_2 + x_1^2 x_2^2 + x_1 x_2^2 + x_1 x_2^3 \\ \Rightarrow \frac{\delta f}{\delta x_1} &= 2x_1 x_2 + 2x_1 x_2^2 + x_2^2 + x_2^3 \\ \Rightarrow \frac{\delta f}{\delta x_2} &= x_1^2 + 2x_1^2 x_2 + 2x_1 x_2 + 3x_1 x_2^2 \end{aligned} \tag{8}$$

Recomputing partial derivative of $\frac{\delta f}{\delta x_1}$ and $\frac{\delta f}{\delta x_2}$ with respect to x_1 and x_2 we get,

$$\begin{aligned} \Rightarrow \frac{\delta^2 f}{\delta x_1^2} &= 2x_2 + 2x_2^2 \\ \Rightarrow \frac{\delta^2 f}{\delta x_1 x_2} &= 2x_1 + 4x_1 x_2 + 2x_2 + 3x_2^2 \\ \Rightarrow \frac{\delta^2 f}{\delta x_2 x_1} &= 2x_1 + 4x_1 x_2 + 2x_2 + 3x_2^2 \\ \Rightarrow \frac{\delta^2 f}{\delta x_2^2} &= 2x_1 + 2x_1^2 + 6x_1 x_2 \end{aligned} \tag{9}$$

Therefore,

$$\begin{aligned} \nabla f(\mathbf{x}) &= \begin{bmatrix} \frac{\delta f}{\delta x_1} \\ \frac{\delta f}{\delta x_2} \end{bmatrix} \\ &= \begin{bmatrix} 2x_1 x_2 + 2x_1 x_2^2 + x_2^2 + x_2^3 \\ x_1^2 + 2x_1^2 x_2 + 2x_1 x_2 + 3x_1 x_2^2 \end{bmatrix} \end{aligned} \tag{10}$$

and Hessian is given by,

$$\begin{aligned}
Hessian \nabla^2 f(x) &= \left[\begin{array}{cc} \frac{\delta^2 f}{\delta x_1^2} & \frac{\delta^2 f}{\delta x_1 x_2} \\ \frac{\delta^2 f}{\delta x_2 x_1} & \frac{\delta^2 f}{\delta x_2^2} \end{array} \right] \\
Hessian \nabla^2 f(x) &= \left[\begin{array}{cc} 2x_2 + 2x_2^2 & 2x_1 + 4x_1x_2 + 2x_2 + 3x_2^2 \\ 2x_1 + 4x_1x_2 + 2x_2 + 3x_2^2 & 2x_1 + 2x_1^2 + 6x_1x_2 \end{array} \right]
\end{aligned} \tag{11}$$

To compute the stationary points of this function, let's equate gradient to zero and find x_1, x_2, and f(x) for each.

$$\begin{aligned}
&\nabla f(\mathbf{x}) = 0 \\
&\Rightarrow \left[\begin{array}{c} 2x_1x_2 + 2x_1x_2^2 + x_2^2 + x_2^3 \\ x_1^2 + 2x_1^2x_2 + 2x_1x_2 + 3x_1x_2^2 \end{array} \right] = 0
\end{aligned} \tag{12}$$

To solve this non-linear equation, let's multiply first and second equation by $3x_1$ and $-x_2$ respectively,

$$\begin{aligned}
&(2x_1x_2 + 2x_1x_2^2 + x_2^2 + x_2^3 = 0) \times 3x_1 \\
&(x_1^2 + 2x_1^2x_2 + 2x_1x_2 + 3x_1x_2^2 = 0) \times y \\
&\Rightarrow \\
&6x_1^2x_2 + 6x_1^2x_2^2 + 3x_1x_2^2 + 3x_1x_2^3 = 0 \\
&-x_1^2x_2 - 2x_1^2x_2^2 - 2x_1x_2^2 - 3x_1x_2^3 = 0 \\
&\overline{} \\
&5x_1^2x_2 + 4x_1^2x_2^2 + x_1^2 = 0 \\
&\Rightarrow x_1x_2(5x_1 + 4x_1x_2 + x_2) = 0
\end{aligned} \tag{13}$$

$$\begin{aligned}
&\Rightarrow x_1x_2 = 0 \\
or &\Rightarrow 5x_1 + 4x_1x_2 + x_2 = 0
\end{aligned} \tag{14}$$

From equation 14, we can say $x_1 = 0$ and putting that in equation 15, we get $x_2 = 0$ which is our first solution, multiplying first and second equation by $-1/x_2$ and $2/x_1$ respectively, we get,

$$\begin{aligned}
&(2x_1x_2 + 2x_1x_2^2 + x_2^2 + x_2^3 = 0) \times -1/x_2 \\
&(x_1^2 + 2x_1^2x_2 + 2x_1x_2 + 3x_1x_2^2 = 0) \times 2/x_1 \\
&\Rightarrow \\
&-2x_1 - 2x_1x_2 - x_2 - x_2^2 = 0 \\
&2x_1 + 4x_1x_2 + 4x_2 + 6x_2^2 = 0 \\
&\overline{} \\
&2x_1x_2 + 3x_2 + 5x_2 = 0 \\
&x_2 = \frac{-3 - 2x_1}{5}
\end{aligned} \tag{15}$$

Substituting x_2 in equation 15, we get,

$$\begin{aligned}
&5x_1 + 4x_1(\frac{-3 - 2x_1}{5}) + (-3 - 2x_1) = 0 \\
&8x_1 - 11x_1 + 3 = 0 \\
&(8x_1 - 3)(x_1 - 1) = 0 \\
&\Rightarrow x_1 = \frac{3}{8} \\
&\Rightarrow x_1 = 1
\end{aligned} \tag{16}$$

Putting $x_1 = 3/8$ and $x_1 = 1$ in equation 15, we get $x_2 = -6/8$ and $x_2 = -1$ respectively. Keeping all solution together, we have following,

x_1	x_2	$f(x)$	Hessian $\Delta^2 f(x)$	Det(Hessian$\Delta^2 f(x)$)
0	0	0	$\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$	0
$\frac{3}{8}$	$-\frac{6}{8}$	$\frac{27}{1024}$	$\begin{bmatrix} -3/8 & -3/16 \\ -3/16 & -21/32 \end{bmatrix}$	27/128
1	-1	0	$\begin{bmatrix} 0 & -1 \\ -1 & -2 \end{bmatrix}$	-1

From all stationary points, when $x_1 = 3/8$ and $y = -6/8$, the determinant of Hessian matrix provides the positive value so it must be either a local minima or maxima. Since corresponding $f_{xx}(x)$ (first row, first column of Hessian matrix) is negative, it is a local maxima because of negative concavity.

▶ Q4 Calculus II [10pts]. Show that the function

$$f(\mathbf{x}) = 8x_1 + 12x_2 + x_1^2 - 2x_2^2 \tag{17}$$

has only one stationary point, and that it is not a minimum nor a maximum, but a saddle point.

Proof:
To find a solution, we equate of $\nabla f(\mathbf{x})$ to zero. Therefore,

$$\begin{aligned} & \nabla f(\mathbf{x}) = 0 \\ \Rightarrow & \begin{bmatrix} 8 + 2x_1 \\ 12 - 4x_2 \end{bmatrix} = 0 \\ \Rightarrow & \\ & x_1 = -4, x_2 = 3, f(x) = 2 \end{aligned} \tag{18}$$

Computing Hessian $\Delta^2 f(x)$,

$$\begin{aligned} Hessian \nabla^2 f(x) &= \begin{bmatrix} \frac{\delta^2 f}{\delta x_1^2} & \frac{\delta^2 f}{\delta x_1 x_2} \\ \frac{\delta^2 f}{\delta x_2 x_1} & \frac{\delta^2 f}{\delta x_2^2} \end{bmatrix} \\ Hessian \nabla^2 f(x) &= \begin{bmatrix} 2 & 0 \\ 0 & -4 \end{bmatrix} \end{aligned} \tag{19}$$

We can see we have only one stationary point. If we compute determinant of Hessian $\Delta^2 f(x)$, we get -8, which is negative and it means it is a saddle point.

▶ Q5 Graphs [5pts]. A *topological order of a directed graph* $G = (V, E)$ is an ordering of its nodes such that for every edge $(v_i, v_j) \in E$, node i appears in the ordering before node j. Prove that every directed acyclic graph (DAG) has a topological order.

Hint: A DAG always has at least one node with no incoming edges.

Proof:
We can prove this by induction. Let G be DAG and $'n'$ represent number of nodes in G

- **Base case**:
  If n = 1, then G is topological ordering
- **Induction hypothesis**:
  G' has topological ordering for $n' \leq n$ where n' is number of nodes in G'
- **Induction step**:
  We shall prove why it holds true for G with n+1 nodes.
  We know that there is at least one node with no incoming edges in DAG. Let that node be v.
  $G - \{v\}$ is also a DAG because removing v from G does not create any cycles.
  Since, now the number of nodes in G becomes n, we can say G has topological ordering according to induction hypothesis.

This proves the hypothesis.

2 Automatic Differentiation with Dynamic Computation Graphs

Please refer to a code submission.

3 Debriefing

1. Approximately how many hours did you spend on this assignment?
   - I spent around 20 hours.

2. Would you rate it as easy, moderate, or difficult?
   - It was moderate.

3. Did you work on it mostly alone or did you discuss the problems with others?
   - I did it all alone.

4. How deeply do you feel you understand the material it covers (0%–100%)?
   - I would say 95%

5. Any other comments?
   - It was a good exercise in this assignment especially the math part. It's been a long time I looked back into what I studied in my undergrads. It never made sense 'where I would use this in my life' so it was easy to forgot. Realizing how important and empirical this is in deep learning and other engineering and science courses makes it interesting to learn.