Math 164 Homework 2

Jiaping Zeng

2/9/2021

1. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a smooth function. Prove that $\nabla^2 f(x) = D(\nabla f(x)) = \nabla(\nabla f(x))$.
Answer: By definition, the Hessian of f is the following:

$$\nabla^2 f(x) = \begin{bmatrix} \frac{\delta f}{\delta x_1^2} & \cdots & \frac{\delta f}{\delta x_1 x_n} \\ \cdots & \ddots & \cdots \\ \frac{\delta f}{\delta x_n x_1} & \cdots & \frac{\delta f}{\delta x_n^2} \end{bmatrix}.$$

We will first show that $\nabla^2 f(x) = D(\nabla f(x))$. Let $g(x) = \nabla f(x)$, i.e.

$$g(x) = \begin{bmatrix} g_1(x) \\ \vdots \\ g_n(x) \end{bmatrix} = \begin{bmatrix} \frac{\delta f}{\delta x_1} \\ \vdots \\ \frac{\delta f}{\delta x_n} \end{bmatrix} = \nabla f(x),$$

then $D(\nabla f(x)) = D(g(x)) = \begin{bmatrix} \frac{\delta g}{\delta x_1} & \cdots & \frac{\delta g}{\delta x_n} \end{bmatrix}$, which expands into

$$D(\nabla f(x)) = D(g(x)) = \begin{bmatrix} \frac{\delta}{\delta x_1} \begin{bmatrix} \frac{\delta f}{\delta x_1} \\ \vdots \\ \frac{\delta f}{\delta x_n} \end{bmatrix} & \cdots & \frac{\delta}{\delta x_n} \begin{bmatrix} \frac{\delta f}{\delta x_1} \\ \vdots \\ \frac{\delta f}{\delta x_n} \end{bmatrix} \end{bmatrix} = \begin{bmatrix} \frac{\delta f}{\delta x_1^2} & \cdots & \frac{\delta f}{\delta x_1 x_n} \\ \cdots & \ddots & \cdots \\ \frac{\delta f}{\delta x_n x_1} & \cdots & \frac{\delta f}{\delta x_n^2} \end{bmatrix}.$$

This is identical to the Hessian matrix shown above, so $\nabla^2 f(x) = D(\nabla f(x))$. Now we will show that $\nabla^2 f(x) = \nabla(\nabla f(x))$ in a similar process. Note that $\nabla(\nabla f(x)) = \nabla g(x) = \begin{bmatrix} \nabla g_1(x) & \cdots & \nabla g_n(x) \end{bmatrix}$, which expands into

$$\nabla(\nabla f(x)) = \nabla g(x) = \begin{bmatrix} \frac{\delta g_1}{\delta x_1} & \cdots & \frac{\delta g_n}{\delta x_1} \\ \cdots & \ddots & \cdots \\ \frac{\delta g_1}{\delta x_n} & \cdots & \frac{\delta g_n}{\delta x_n} \end{bmatrix} = \begin{bmatrix} \frac{\delta f}{\delta x_1^2} & \cdots & \frac{\delta f}{\delta x_1 x_n} \\ \cdots & \ddots & \cdots \\ \frac{\delta f}{\delta x_n x_1} & \cdots & \frac{\delta f}{\delta x_n^2} \end{bmatrix}.$$

This is again identical to the Hessian matrix, so $\nabla^2 f(x) = \nabla(\nabla f(x))$. Therefore $\nabla^2 f(x) = D(\nabla f(x)) =$

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

2. Find the gradient and Hessian of the following functions.

   (a) $f(x) = \frac{1}{2}x^T Ax$ where A is an $n \times n$ symmetric data matrix and $x \in \mathbb{R}^n$.
   Answer: We have $f(x+dx) = \frac{1}{2}(x+dx)^T A(x+dx) = \frac{1}{2}(x^T + dx^T)(Ax + Adx) = \frac{1}{2}x^T Ax + \frac{1}{2}(x^T Adx + dx^T A) + dx^T Adx = f(x) + (x^T A)[dx] + dx^T[A]dx$, therefore $Df(x) = x^T A$ and $D^2 f(x) = A$.

   (b) $f(x) = \frac{1}{2}\|y - Ax\|_2^2$ where $y \in \mathbb{R}^m$, $A \in \mathbb{R}^{m\times n}$ are data, and $x \in \mathbb{R}^n$.
   Answer: We have $f(x+dx) = \frac{1}{2}\|A(x+dx) - y\|_2^2 = \frac{1}{2}\|(Ax-y) + Adx\|_2^2 = \frac{1}{2}\|Ax-y\|_2^2 + \langle Ax-y, Adx\rangle + \frac{1}{2}\|Adx\|_2^2 = f(x) + \langle A^T(Ax-y), dx\rangle + \frac{1}{2}\langle Adx, Adx\rangle = f(x) + \langle A^T(Ax-y), dx\rangle + \frac{1}{2}\langle dx, A^T Adx\rangle$. Therefore $Df(x) = A^T(Ax-y)$ and $D^2 f(x) = A^T A$.

   (c) $f(x) = \frac{1}{2}\left\|X - xx^T\right\|_F^2$ where X is an $n\times n$ symmetric data matrix and $x \in \mathbb{R}^n$.
   Answer: We have $f(x) = \frac{1}{2}\|A - xx^t\|_F^2 = \frac{1}{2}\langle(A - xx^T), (A - xx^T)\rangle_F = \frac{1}{2}\langle A, A\rangle_F - \langle A, xx^T\rangle_F + \frac{1}{2}\langle xx^T, xx^T\rangle = \frac{1}{2}\langle A, A\rangle_F - \operatorname{tr}\left(A^T xx^T\right) + \operatorname{tr}\left(xx^T xx^T\right) = \frac{1}{2}\langle A, A\rangle_F - \operatorname{tr}\left(x^T[Ax]\right) + \|x\|^2 \operatorname{tr}\left(x^T x\right) = \frac{1}{2}\|A\|_F^2 - x^T Ax + \|x\|^4$, then $f(x+dx) = \frac{1}{2}\|A\|_F^2 - (x+dx)^T A(x+dx) + \|x+dx\|^4 = \frac{1}{2}\langle A, A\rangle_F + \langle A^T x, dx\rangle + \frac{1}{2}dx^T[A]dx + \|x\|^4 + 4\|x\|^2 x^T dx + \|x\|^2 dx^T dx = \frac{1}{2}\langle A, A\rangle_F + \langle A^T x, dx\rangle + \frac{1}{2}dx^T[A]dx + \|x\|^4 + \langle\|x\|^2 x, dx\rangle + dx^T(\|x\|^2 I)dx$. Therefore $Df(x) = A^T x + \|x\|^2 x$ and $D^2 f(x) = \|x\|I + A$.

   (d) $f(x,y) = \frac{1}{2}\left\|Y - xy^T\right\|_F^2$ where Y is an $m\times n$ symmetric data matrix and $x \in \mathbb{R}^m, y \in \mathbb{R}^n$.
   Answer: We have $Df(x,y) = \frac{1}{2}D\langle Y - xy^T, Y - xy^T\rangle = \frac{1}{2}[(Y - xy^T)^T D(Y - xy)^T + (Y - xy^T)^T D(Y - xy^T)] = [(Y - xy^T)^T(-y^T), (Y - xy^T)^T(-x)] = [(-Yy^T + xy^2)^T, (-Yx + x^2 y^T)^T]$ and $D^2 f(x,y) = D(Df(x,y)) = \begin{bmatrix} y^2 & -Y + 2xy \\ -Y + 2xy & x^2 \end{bmatrix}$.

   (e) $f(W) = \sum_{j=1}^N \left\|W_{x_j} - y_j\right\|_2^2$ where $x_j \in \mathbb{R}^n, y_j \in \mathbb{R}^m$ are given data and $W \in \mathbb{R}^{m\times n}$.
   Answer: We have $\frac{1}{2}f(W) = \frac{1}{2}\sum_{j=1}^N \left\|W_{x_j} - y_j\right\|_2^2 = \sum_{j=1}^n[\frac{1}{2}x_j^T W^T W x_j - y_j^T W x_j + \|y_j\|^2] = \frac{1}{2}\operatorname{tr}\Big(W^T W[\sum_{j=1}^n x_j x_j^T]\Big) - \operatorname{tr}\left(W\sum x_j x_j^T\right)$. Let $X = \sum_{j=1}^n x_j x_j^T]$ and $\tilde{X} = \sum x_j x_j^T$, then $\frac{1}{2}f(W) = \frac{1}{2}\langle W, WX\rangle_F - \langle\tilde{X}, W\rangle_F$. Therefore $f(W + dW) = \frac{1}{2}\langle W + dW, (W + dW)X\rangle_F - \langle\tilde{X}, W + dW\rangle_F$.

3. Let $\sigma(\cdot) : \mathbb{R} \to \mathbb{R}$ be a scalar function and apply elementwise to its input vectors. Denote $\sigma'(\cdot)$ as its derivative function. Assume $f(W) = \sum_{j=1}^N \|\sigma(Wx_j) - y_j\|_2^2$ where $x_j \in \mathbb{R}^n, y_j \in \mathbb{R}^m$ are given data and $W \in \mathbb{R}^{m\times n}$. Compute $\nabla f(W)$.
   Answer: $f(W + dW) = \sum_{j=1}^N \|\sigma((W + dW)x_j) - y_j\|_2^2 =$

4. In each of the following problems justify your answer using optimality conditions.

   (a) Compute the gradient $\nabla f(x)$ and Hessian $\nabla^2 f(x)$ of the Rosenbrock function

   $$f(x) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2$$

   Show that $x^\star = [1,1]^T$ is the only local minimizer of this function, and that the Hessian matrix at that point is positive definite.

Answer: We can first find the gradient and Hessian as follows:

$$\nabla f(x) = \begin{bmatrix} \frac{\delta f}{\delta x_1} \\ \frac{\delta f}{\delta x_2} \end{bmatrix} = \begin{bmatrix} 400x_1^3 - 400x_1x_2 + 2x_1 - 2 \\ 200x_2 - 200x_1^2 \end{bmatrix}$$

$$\nabla^2 f(x) = \begin{bmatrix} \frac{\delta^2 f}{\delta x_1^2} & \frac{\delta^2 f}{\delta x_1 \delta x_2} \\ \frac{\delta^2 f}{\delta x_2 \delta x_1} & \frac{\delta^2 f}{\delta x_2^2} \end{bmatrix} = \begin{bmatrix} 1200x_1^2 - 400x_2 + 2 & -400x_1 \\ -400x_1 & 200 \end{bmatrix}.$$

We can first solve $\nabla f(x) = 0$, which gives us two equations $400x_1^3 - 400x_1x_2 + 2x_1 - 2 = 0$ and $200x_2 - 200x_1^2 = 0$. The only solution to the two equations is $x_1 = x_2 = 1$, so $x^\star$ is the only critical point of $f(x)$ (i.e., there is no other minimizers or maximizers). In addition, since $f([0,0]^T) = 1 > f(x^\star)$, $f(x^\star)$ is a local minimum and $x^\star$ is therefore the only local minimizer. Since $x^\star$ is a local minimizer and f is smooth, $\nabla^2 f(x^\star)$ is positive semidefinite. We can substitute $x^\star$ into $\nabla^2 f(x)$ found above, which gives us the following nonzero matrix:

$$\nabla^2 f(x^\star) = \begin{bmatrix} 802 & -400 \\ -400 & 200 \end{bmatrix}.$$

Since $\nabla^2 f(x^\star)$ is positive semidefinite and is also nonzero, it is positive definite.

(b) Show that the function $f(x) = 8x_1 + 12x_2 + x_1^2 - 2x_2^2$ has only one stationary point, and that it is neither a maximum or minimum, but a saddle point.

Answer: We have

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

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

We can solve $\nabla f(x) = 0$ which gives us $2x_1 + 8 = 0, -4x_2 + 12 \implies x_1 = -4, x_2 = 3$, so $[-4, 3]^T$ is the only stationary point. Then, $(2 - \lambda)(-4 - \lambda) = 0$ gives us $\lambda = 2, -4$, so the Hessian has both positve and negative eigenvalues and is therefore a saddle point.

(c) Find all the critical points of the 2-dimensional function $f(x_1, x_2) = (x_1^2 - 1)^2 + x_2^2$. Which are global minima? Which are not local minima?

Answer: We have

$$\nabla f(x) = \begin{bmatrix} \frac{\delta f}{\delta x_1} \\ \frac{\delta f}{\delta x_2} \end{bmatrix} = \begin{bmatrix} 4x_1^3 - 4x_1 \\ 2x_2 \end{bmatrix}$$

$$\nabla^2 f(x) = \begin{bmatrix} \frac{\delta^2 f}{\delta x_1^2} & \frac{\delta^2 f}{\delta x_1 \delta x_2} \\ \frac{\delta^2 f}{\delta x_2 \delta x_1} & \frac{\delta^2 f}{\delta x_2^2} \end{bmatrix} = \begin{bmatrix} 12x_1^2 - 4 & 0 \\ 0 & 2 \end{bmatrix}.$$

Solving $\nabla f(x) = 0$ gives us $4x_1^3 - 4x_1 = 0, 2x_2 = 0 \implies x_1 = 0, \pm 1, x_2 = 0$, so the critical points are $[-1, 0]^T, [0, 0]^T$ and $[0, 0]^T$. The eigenvalues of $\nabla^2 f([\pm 1, 0]^T)$ are $\lambda = 2, 8$ (both positives), so $[-1, 0]^T$ and $[1, 0]^T$ are local minimizers. Since $f([-1, 0]^T) = 4$ and $f([1, 0]^T) = 0$, $f([1, 0]^T) = 0$ is the global minimizer. The eigenvalues of $\nabla^2 f([0, 0]^T)$ are $\lambda = -4, 2$, so it is a saddle point.

(d) Find all the critical points of the 2-dimensional function $f(x_1, x_2) = (x_1^2 - 1)^2 + (x_2^2 - 1)^2$. Which are global minima? Which are not global minima?

Answer: We have

$$\nabla f(x) = \begin{bmatrix} \frac{\delta f}{\delta x_1} \\ \frac{\delta f}{\delta x_2} \end{bmatrix} = \begin{bmatrix} 4x_1^3 - 4x_1 \\ 4x_2^3 - 4x_2 \end{bmatrix}$$

$$\nabla^2 f(x) = \begin{bmatrix} \frac{\delta^2 f}{\delta x_1^2} & \frac{\delta^2 f}{\delta x_1 \delta x_2} \\ \frac{\delta^2 f}{\delta x_2 \delta x_1} & \frac{\delta^2 f}{\delta x_2^2} \end{bmatrix} = \begin{bmatrix} 12x_1^2 - 4 & -0 \\ 0 & 12x_2^2 - 4 \end{bmatrix}.$$

Solving $\nabla f(x) = 0$ gives us $4x_1^3 - 4x_1 = 0, 2x_2 = 0 \implies x_1 = 0, \pm 1, x_2 = 0, \pm 1$. We will analyze each point as follows:

x_1	x_2	λ	$f([x_1, x_2]^T)$	type of point
-1	-1	8	0	global minimizer
-1	0	$-4, 8$	1	saddle point
-1	1	8	0	global minimizer
0	-1	$-4, 8$	1	saddle point
0	0	-4	2	global maximizer
0	1	$-4, 8$	1	saddle point
1	-1	8	0	global minimizer
1	0	$-4, 8$	1	saddle point
1	-1	8	0	global minimizer

(e) Show that the 2-dimensional function $f(x_1, x_2) = (x_2 - x_1^2)^2 - x_1^2$ has only one stationary point, which is neither a local maximum nor a local minimum.

Answer: We have

$$\nabla f(x) = \begin{bmatrix} \frac{\delta f}{\delta x_1} \\ \frac{\delta f}{\delta x_2} \end{bmatrix} = \begin{bmatrix} 4x_1^3 - 4x_1 x_2 - 2x_1 \\ -2x_1^2 + 2x_2 \end{bmatrix}$$

$$\nabla^2 f(x) = \begin{bmatrix} \frac{\delta^2 f}{\delta x_1^2} & \frac{\delta^2 f}{\delta x_1 \delta x_2} \\ \frac{\delta^2 f}{\delta x_2 \delta x_1} & \frac{\delta^2 f}{\delta x_2^2} \end{bmatrix} = \begin{bmatrix} 12x_1^2 - 4x_2 - 2 & -4x_1 \\ -4x_1 & 2 \end{bmatrix}.$$

$\nabla f(x) = 0 \implies 4x_1^3 - 4x_1 x_2 - 2x_1 = 0, -2x_1^2 + 2x_2 = 0$ has only one solution $x_1 = x_2 = 0$, therefore it only has one stationary point. By substitution, $\nabla^2 f([0,0]^T)$ is the following:

$$\begin{bmatrix} -2 & 0 \\ 0 & 2 \end{bmatrix}.$$

$(-2-\lambda)(2-\lambda)$ gives us $\lambda = \pm 2$, so $[0,0]^T$ is a saddle point.