

MATH CAMP: Lecture 8

Lemma: Let $f : U \rightarrow R^M$, where U is an open subset of R^N . If f is differentiable at $c \in U$, then there exist positive numbers δ and B such that $\|f(x) - f(c)\| \leq B \|x - c\|$, if $\|x - c\| < \delta$. In particular, f is continuous at c .

Proof: Since f is differentiable at c , there exists a $\delta > 0$ such that

$$\|f(x) - f(c) - Df(c)(x - c)\| \leq \|x - c\|,$$

if $\|x - c\| < \delta$. Therefore,

$$\begin{aligned} \|f(x) - f(c)\| &= \|f(x) - f(c) - Df(c)(x - c) + Df(c)(x - c)\| \\ &\leq \|f(x) - f(c) - Df(c)(x - c)\| + \|Df(c)(x - c)\| \\ &\leq \|x - c\| + \|Df(c)(x - c)\|, \end{aligned}$$

if $\|x - c\| < \delta$. By the last lemma of the previous lecture, there is a $b > 0$ such that

$$\|Df(c)(x - c)\| \leq b\|x - c\|.$$

Therefore,

$$\|f(x) - f(c)\| \leq (1 + b)\|x - c\|$$

if $\|x - c\| < \delta$. ■

Definition: Let $f : U \rightarrow R$, where $U \subset R^N$ is open, and let $v \in R^N$. The vector $\nabla_v f(c)$ is said to be the *directional derivative* of f in the direction v if for every $\varepsilon > 0$, there is $\delta > 0$ such that if $0 < |t| < \delta$, then $|\frac{1}{t}[f(c + tv) - f(c)] - \nabla_v f(c)| < \varepsilon$. That is $\nabla_v f(c) = \lim_{t \rightarrow 0, t \neq 0} \frac{1}{t}[f(c + tv) - f(c)]$.

Remarks:

1. $\nabla_0 f(c) = 0$.
2. $\nabla_v f(c) = \frac{d}{dt} f(c + v) \Big|_{t=0} = \frac{d}{dt} f(c + tv) \Big|_{t=0} = \frac{d}{dx} g(0)$, where $g(t) = f(c + tv)$.

Theorem: Suppose that $f : U \rightarrow R$ is differentiable at $c \in U$, where U is an open subset of R^N . If $v \in R^N$, then $\nabla_v f(c)$ exists and $\nabla_v f(c) = Df(c)(v)$.

Proof: By the definition of the differentiability of f , for any $\varepsilon > 0$, there is a $\delta > 0$ such that $|f(c + tv) - f(c) - Df(c)(tv)| \leq \varepsilon \|tv\|$, if $\|tv\| < \delta$. If $v = 0$, $Df(c)(v) = 0 = \nabla_v f(c)$. If $v \neq 0$ and $0 < |t| < \frac{\delta}{\|v\|}$, then $|\frac{1}{t}[f(c + tv) - f(c)] - Df(c)(v)| \leq \varepsilon \|v\|$. Therefore, $Df(c)(v) = \nabla_v f(c)$, by the definition of $\nabla_v f(c)$. ■

Definition: If $f : U \rightarrow R$, where $U \subset R^N$ and U is open, then $\nabla_{e_n} f(c)$ is called the n th partial derivative of f at c and is written as $\partial f(c)/\partial x_n$, where e_n is the n th standard basis vector of R^N .

Remark:

$$\frac{\partial f}{\partial x_n}(c) = \frac{d}{dx_n} f(c_1, \dots, c_{n-1}, x_n, c_{n+1}, \dots, c_N) \Big|_{x_n=c_n}.$$

That is, all variables of but the n th are held constant at their values in the vector c . The result is a function of the single variable x_n . The derivative of this function at $x_n = c_n$ equals $\frac{\partial f}{\partial x_n}(c)$.

Example:

$$f(x_1, x_2, x_3) = x_1 x_2^3 x_3^2$$

$$\frac{\partial f(2, 4, 5)}{\partial x_2} = 2(3)(4^2)(5^2) = 6(16)(25) = 2400.$$

If $f : U \rightarrow R^M$ where $U \subset R^N$ and U is open, let $f_m : U \rightarrow R$ be the m th component of f , for $m = 1, \dots, M$.

Theorem: Let $f : U \rightarrow R^M$, where U is an open subset of R^N . If f is differentiable at c , then f_m is differentiable at c , for all m , and $Df(c) = \begin{pmatrix} Df_1(c) \\ \vdots \\ Df_M(c) \end{pmatrix}$.

Proof: Let $\varepsilon > 0$ and let $\delta > 0$ be such that $\|f(x) - f(c) - Df(c)(x - c)\| \leq \varepsilon \|x - c\|$, if $\|x - c\| < \delta$. $Df(c) : R^N \rightarrow R^M$ is a linear transformation, so that $Df(c) = \begin{pmatrix} (Df(c))_1 \\ \vdots \\ (Df(c))_M \end{pmatrix}$, where $(Df(c))_m : R^N \rightarrow R$ is linear, for all m , and is the m th component function of $Df(c)$. If $\|x - c\| < \delta$, then

$$\begin{aligned} & |f_m(x) - f_m(c) - (Df(c))_m(x - c)| \\ & \leq \|f(x) - f(c) - Df(c)(x - c)\| \leq \varepsilon \|x - c\|, \end{aligned}$$

since $f_m(x) - f_m(c) - (Df(c))_m(x - c)$ is the m th component of

$$f(x) - f(c) - Df(c)(x - c).$$

Therefore, by the definition of a derivative, $(Df(c))_m$ is the derivative of f_m at c .

That is, $(Df(c))_m = Df_m(c)$. Therefore, $Df(c) = \begin{pmatrix} Df_1(c) \\ \vdots \\ Df_m(c) \end{pmatrix}$. ■

Theorem: The matrix $\begin{pmatrix} \frac{\partial f_1(c)}{\partial x_1} & \cdots & \frac{\partial f_1(c)}{\partial x_N} \\ \vdots & & \vdots \\ \frac{\partial f_M(c)}{\partial x_1} & \cdots & \frac{\partial f_M(c)}{\partial x_N} \end{pmatrix}$ represents $Df(c)$, if f is differentiable at c and $f : U \rightarrow R^M$, where $U \subset R^N$ and U is open.

Proof: Let $v = (v_1, \dots, v_N) \in R^N$. Then, $v = \sum_{n=1}^N v_n e_n$, so that

$$\begin{aligned} Df(c)(v) &= Df(c) \left(\sum_{n=1}^N v_n e_n \right) = \sum_{n=1}^N v_n Df(c)(e_n) \\ &= \sum_{n=1}^N v_n \begin{pmatrix} Df_1(c) \\ \vdots \\ Df_M(c) \end{pmatrix} (e_n) = \sum_{n=1}^N v_n \begin{pmatrix} Df_1(c)(e_n) \\ \vdots \\ Df_M(c)(e_n) \end{pmatrix} \\ &= \sum_{n=1}^N v_n \begin{pmatrix} \nabla_{e_n} f_1(c) \\ \vdots \\ \nabla_{e_n} f_M(c) \end{pmatrix} = \sum_{n=1}^N v_n \begin{pmatrix} \frac{\partial f_1(c)}{\partial x_n} \\ \vdots \\ \frac{\partial f_M(c)}{\partial x_n} \end{pmatrix} \\ &= \begin{pmatrix} \frac{\partial f_1(c)}{\partial x_1} & \cdots & \frac{\partial f_1(c)}{\partial x_N} \\ \vdots & & \vdots \\ \frac{\partial f_M(c)}{\partial x_1} & \cdots & \frac{\partial f_M(c)}{\partial x_N} \end{pmatrix} \begin{pmatrix} v_1 \\ \vdots \\ v_N \end{pmatrix}. \end{aligned}$$

■

Theorem:

1. Let $f : U \rightarrow R^M$ and $g : U \rightarrow R^M$, where $U \subset R^N$ and U is open. If f and g are differentiable at $c \in U$ and a and b are numbers, then $af + bg$ is differentiable at c and $D(af + bg)(c) = aDf(c) + bDg(c)$.
2. If f and g are as in part (1), then $f \cdot g$ is differentiable at c and $D(f \cdot g)(c)(v) = Df(c)(v) \cdot g(c) + f(c) \cdot Dg(c)(v)$, for $v \in R^N$.
3. If $\varphi : U \rightarrow R$ and φ is differentiable at c and if f is as in part (1), then φf is differentiable at c and $D(\varphi f)(c)(v) = D\varphi(c)(v)f(c) + \varphi(c)Df(c)(v)$.

Parts 2 and 3 of this theorem generalize Leibniz's rule for differentiation.

Some background facts about matrix transposition:

1. If A is an $M \times N$ matrix and B is an $N \times K$ matrix, then $(AB)^T = B^T A^T$.
2. If A and B are $M \times N$ matrices and a and b are numbers, then $(aA + bB)^T = aA^T + bB^T$.

3. If A is a matrix, $(A^T)^T = A$.
4. If x and y are N -vectors, $x \cdot y = x^T y$.

I let you verify these assertions.

The equation in part 2 of the previous theorem may be written as

$$\begin{aligned} D(f^T(c)g(c))v &= (Df(c)v)^T g(c) + (f(c))^T Dg(c)v \\ &= v^T (Df(c))^T g(c) + (f(c))^T Dg(c)v, \end{aligned}$$

where I treat $Df(c)$ and $Dg(c)$ as $M \times N$ matrices.

Now, I do some useful special cases.

If $f(x) = a^T x$, where $a \in R^N$ is a constant vector, then $Df(x) = D(a^T x) = a^T$, since $a^T x$ is a linear function of x . Similarly, if $f(x) = Ax$, where A is an $M \times N$ matrix, then $Df(x) = A$, since Ax is linear.

Now, let $M = N$ in the previous theorem. Let $f(x) = x$ and let $g(x) = Ax$, where A is an $N \times N$ constant matrix. Then by part 2 of the theorem,

$$\begin{aligned} D(x^T Ax)(c)(v) &= D(f \cdot g)(c)(v) \\ &= [Df(c)v] \cdot g(c) + f(c) \cdot Dg(c)v \\ &= (Df(c)v)^T g(c) + f(c)^T Dg(c)v \\ &= v^T (Df(c))^T g(c) + f(c)^T Dg(c)v \\ &= v^T IAc + c^T Av = v^T Ac + c^T Av \\ &= c^T A^T v + c^T Av, \end{aligned}$$

since $v^T Ac$ is a number, so that $v^T Ac = (v^T Ac)^T = c^T A^T v$. If in addition A is symmetric, so that $A^T = A$, then $D(x^T Ax)(c)(v) = c^T A^T v + c^T Av = 2c^T Av$. Therefore, the matrix representation of $D(x^T Ax)(c)$ is $2c^T A$, if A is symmetric. That is, if A is symmetric, we may write $D(x^T Ax)(c) = 2c^T A$.

Let U be an open subset of R^N and $f : U \rightarrow R$.

Definition: f has a local maximum at $c \in U$ if for some $\varepsilon > 0$, $f(c) \geq f(x)$, for all $x \in B_\varepsilon(c)$.

Theorem: If f has a local maximum at $c \in U$, then $Df(c) = 0$.

Proof: The restriction of f to any line through c has a local maximum at c . Therefore, $\nabla_v f(c) = 0$, for all $v \in R^N$. In particular, $\partial f(c)/\partial x_n = 0$, for all n . Therefore, $Df(c) = 0$. ■

Similarly, you can define a local minimum for f , and $Df(c) = 0$ if f has a local minimum at a .

Application (Least Squares Estimator):

$$\text{Model} \quad y = \sum_{n=1}^K \beta_k x_k + e, \quad e = \text{error.}$$

We don't know the β_k . Suppose we have N observations,

$$\begin{array}{cccc} y_1 & x_{11} & \dots & x_{1K} \\ \vdots & \vdots & & \vdots \\ y_N & x_{N1} & \dots & x_{NK} \end{array} .$$

The least square estimator is (b_1, \dots, b_K) that minimizes $\sum_{n=1}^N (y_n - \sum_{k=1}^K b_k x_{nk})^2 =$ sum of squared errors. Let,

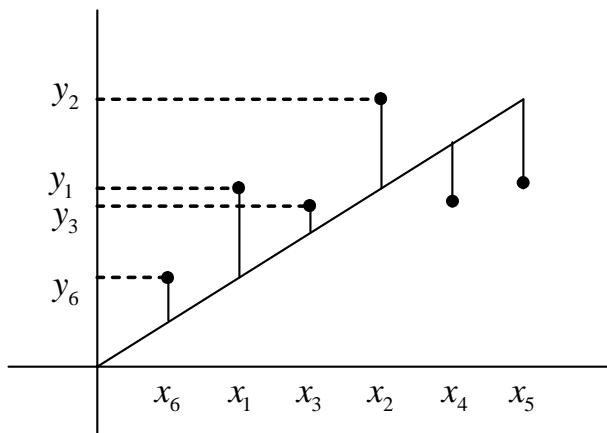
$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}, \quad X = \begin{pmatrix} x_{11} & \dots & x_{1K} \\ \vdots & & \vdots \\ x_{N1} & \dots & x_{NK} \end{pmatrix}, \quad b = \begin{pmatrix} b_1 \\ \vdots \\ b_K \end{pmatrix}$$

So, we wish to choose b so as to minimize:

$$\begin{aligned} (y - Xb) \cdot (y - Xb) &= (y - Xb)^T (y - Xb) \\ &= (y^T - b^T X^T)(y - Xb) \\ &= y^T y - y^T Xb - b^T X^T y + b^T X^T Xb \\ &= y^T y - 2y^T Xb + b^T X^T Xb, \end{aligned}$$

where I have used the rules for matrix transposition and I have used the fact that since $b^T X^T y$ is a number, $b^T X^T y = (b^T X^T y)^T = y^T X b^T = y^T X b$. The b that minimizes $(y - Xb)^T (y - Xb)$ is called the least squares estimator.

If $k = 1$, we have the following:



The least squares estimate, b , minimizes the sum of the squares of the vertical distances from the data points (x_n, y_n) to the line $y = bX$.

In order to calculate the least squares estimator, we set the derivative of $(y - Xb) \cdot (y - Xb) = y^T y - 2y^T Xb + b^T X^T Xb$ with respect to b equal to zero. Let D_b denote the derivative with respect to the vector b .

$$\begin{aligned} D_b(y - Xb) \cdot (y - Xb) &= D_b[y^T y - 2y^T Xb + b^T X^T Xb] \\ &= D_b y^T y - 2D_b y^T Xb + D_b b^T X^T Xb \\ &= 0 - 2y^T X + 2b^T X^T X, \end{aligned}$$

where I have used the fact that the matrix $X^T X$ is symmetric. $X^T X$ is symmetric because $(X^T X)^T = X^T X^{TT} = X^T X$. Since $X^T X$ is symmetric, $D_b b^T X^T Xb = 2b^T X^T X$, by a formula proved earlier. The equation $0 = -2y^T X + 2b^T X^T X$ implies that $b^T X^T X = y^T X$. Taking the transpose of both sides of this equation, we obtain $X^T Xb = X^T y$. If the matrix $X^T X$ is invertible, then $b = (X^T X)^{-1} X^T y$. This is the formula for the least squares estimator.

Theorem (The Chain Rule): Suppose that $f : U \rightarrow V$, where U and V are open subsets of R^N and R^M , respectively, and that $g : V \rightarrow R^K$. Suppose that f is differentiable at $c \in U$ and that g is differentiable at $b = f(c)$. Let $h : U \rightarrow R^K$ be defined by $h(x) = g(f(x)) = g \circ f(x)$. Then, h is differentiable at c and $Dh(c) = Dg(f(c)) \circ Df(c)$.

Mean Value Theorem: Let $f : U \rightarrow R$, where U is an open subset of R^N . Suppose that f is differentiable on U . Let $a \in U$ and $b \in U$ and suppose that the line segment from a to b ($= \{(1-t)a + tb \mid 0 \leq t \leq 1\}$) is contained in U . Then, there exists a point c on this line segment such that $f(b) - f(a) = Df(c)(b - a)$.

Proof: Let $\varphi : [0, 1] \rightarrow R$ be defined by $\varphi(t) = f((1-t)a + tb)$. $\varphi(0) = f(a)$. $\varphi(1) = f(b)$. By the chain rule, $d\varphi(t)/dt = Df((1-t)a + tb)(b - a)$. By the mean value theorem for one variable, there exists a t_0 such that $0 < t_0 < 1$ and $d\varphi(t_0)/dt = \varphi(1) - \varphi(0) = f(b) - f(a)$. Let $c = (1-t_0)a + t_0b$. Then $Df(c)(b - a) = f(b) - f(a)$.

Theorem: Existence of a Derivative: Let $f : U \rightarrow R^M$, where $U \subset R^N$ is open. Suppose that $\partial f_m(x)/\partial x_n$ exists and is continuous on U , for all n and m . Then, f is differentiable on U .