

Lecture #6

Multivariable Calculus

Multivariate like univariate calculus has to do with the local approximation of functions by affine ones.

Definition: Let U be open subset of \mathbb{R}^N and let $f: U \rightarrow \mathbb{R}^M$. The function f is differentiable at $c \in U$ if there exists a linear transformation $Df(c): \mathbb{R}^N \rightarrow \mathbb{R}^M$, called the derivative of f at c , such that for every positive number ε , there exists a positive number δ such that

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

if $0 \leq \|x - c\| \leq \delta$. That is,

$$\lim_{\substack{x \rightarrow c \\ x \neq c}} \frac{\|f(x) - f(c) - Df(c)(x - c)\|}{\|x - c\|} = 0.$$

This definition says that the affine function $f(c) + Df(c)(x - c)$ of x approximates the function f locally near c .

Remark: It should be clear that if $f(x) = a + T(x)$, where $a \in \mathbb{R}^M$ and $T: \mathbb{R}^N \rightarrow \mathbb{R}^M$ is linear, then $Df(x) = T$, for all x . That is, the derivative of an affine function is its linear part.

Lemma 6.1: A function $f: U \rightarrow \mathbb{R}^M$, where U is an open subset of \mathbb{R}^N , has at most one derivative at a point.

Proof: Suppose that $S: \mathbb{R}^N \rightarrow \mathbb{R}^M$ and $T: \mathbb{R}^N \rightarrow \mathbb{R}^M$ are linear and satisfy the definition of a derivative of f at $c \in U$. If $S \neq T$, then there is a $v \in \mathbb{R}^N$ such that $\|v\| = 1$ and $\|S(v) - T(v)\| > 0$. Let $\varepsilon > 0$ and $\delta > 0$ be such that $\|f(x) - f(c) - S(x - c)\| \leq \varepsilon \|x - c\|$ and $\|f(x) - f(c) - T(x - c)\| \leq \varepsilon \|x - c\|$, if $0 \leq \|x - c\| \leq \delta$. Let t be a non-zero number such that $|t| < \delta$ and let $x = c + tv$. Then $0 < \|x - c\| < \delta$ and

$$\begin{aligned} 0 < |t| \|S(v) - T(v)\| &= \|S(tv) - T(tv)\| \\ &= \|-[f(x) - f(c)] + S(x - c) + [f(x) - f(c)] - T(x - c)\| \\ &\leq \|f(x) - f(c) - S(x - c)\| + \|f(x) - f(c) - T(x - c)\| \\ &\leq 2\varepsilon \|x - c\| = 2\varepsilon \|tv\| = 2\varepsilon |t|. \end{aligned}$$

Dividing by $|t|$, we see that $0 < \|S(v) - T(v)\| \leq 2\varepsilon$, for all positive numbers ε , which is impossible. ■

The next lemma implies that every linear transformation is continuous.

Lemma 6.2: If $T: \mathbb{R}^N \rightarrow \mathbb{R}^M$ is a linear transformation, then there exists a positive number b such that $\|T(v) - T(w)\| \leq b\|v - w\|$, for all v and w in \mathbb{R}^N .

Proof: Let A be the $M \times N$ matrix representing T with respect to the standard bases of \mathbb{R}^N and \mathbb{R}^M and let $a = \max_{\substack{1 \leq m \leq M, \\ 1 \leq n \leq N}} |a_{mn}|$. If $y = T(x)$, then by the Cauchy-Schwarz inequality,

$$|y_m| = \left| \sum_{n=1}^N a_{mn} x_n \right| \leq \|a_m\| \|x\|,$$

for $m = 1, \dots, M$, where a_m is the m^{th} row of A . Because

$$\|a_m\| = \sqrt{\sum_{n=1}^N a_{mn}^2} \leq \sqrt{Na^2} = a\sqrt{N},$$

it follows that

$$|y_m| \leq a\sqrt{N} \|x\|$$

and so

$$\|y\| = \sqrt{\sum_{m=1}^M y_m^2} \leq \sqrt{Ma^2N \|x\|^2} = a\sqrt{MN} \|x\|.$$

Let $b = a\sqrt{MN}$. Then $\|T(v) - T(w)\| = \|T(v - w)\| \leq b\|v - w\|$, for all v and w in \mathbb{R}^N , as was to be proved. ■

Lemma 6.3: Let $f: U \rightarrow \mathbb{R}^M$, where U is an open subset of \mathbb{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 positive number δ such that

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

if $\|x - c\| \leq \delta$. By lemma 6.2, there is a positive number b such that

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

Therefore

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

if $\|x - c\| \leq \delta$. Let $B = 1 + b$. ■

I next relate the derivative to directional derivatives and partial derivatives.

Definition: Let $f: U \rightarrow \mathbb{R}$, where U is an open subset of \mathbb{R}^N , and let $v \in \mathbb{R}^N$. The number $\nabla_v f(c)$ is said to be the directional derivative of f at c in the direction v if, for every positive number ε , there is a positive number δ such that if $0 < |t| \leq \delta$, then

$$\left| \frac{f(c + tv) - f(c)}{t} - \nabla_v f(c) \right| \leq \varepsilon.$$

That is,

$$\nabla_v f(c) = \lim_{\substack{t \rightarrow 0, \\ t \neq 0}} \frac{f(c + tv) - f(c)}{t}.$$

Remarks:

1) $\nabla_0 f(c) = 0$

2) $\nabla_v f(c) = \left. \frac{df(c + tv)}{dt} \right|_{t=0} = \frac{dg(0)}{dt}$, where $g(t) = f(c + tv)$.

Theorem 6.4: Suppose that $f: U \rightarrow \mathbb{R}$ is differentiable at $c \in U$, where U is an open subset of \mathbb{R}^N . If $v \in \mathbb{R}^N$, then $\nabla_v f(c)$ exists and $\nabla_v f(c) = Df(c)(v)$.

Proof: If $v = 0$, $Df(c)(v) = 0 = \nabla_v f(c)$. Let $v \neq 0$. By the definition of the differentiability of f , for any positive number ε , there is a positive number δ such that $|f(c + tv) - f(c) - Df(c)(tv)| \leq \frac{\varepsilon \|tv\|}{\|v\|}$, if $\|tv\| \leq \delta$. If $0 < |t| \leq \delta/\|v\|$, then

$$\begin{aligned} \left| \frac{f(c + tv) - f(c)}{t} - Df(c)(v) \right| &= \frac{1}{|t|} |f(c + tv) - f(c) - Df(c)(tv)| \\ &\leq \frac{\varepsilon \|tv\|}{|t| \|v\|} = \frac{\varepsilon |t| \|v\|}{|t| \|v\|} = \varepsilon. \end{aligned}$$

Therefore $Df(c)(v) = \nabla_v f(c)$, by the definition of $\nabla_v f(c)$. ■

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

Remark:
$$\frac{\partial f(c)}{\partial x_n} = \left. \frac{df(c_1, \dots, c_{n-1}, x_n, c_{n+1}, \dots, c_N)}{dx_n} \right|_{x_n = c_n}.$$

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

Example: If $f(x_1, x_2, x_3) = x_1 x_2^3 x_3^2$, the

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

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

Theorem 6.5: Let $f: U \rightarrow \mathbb{R}^M$, where U is an open subset of \mathbb{R}^N . If f is differentiable at c , then f_m is differentiable at c , for all m , and

$$Df(c) = \begin{pmatrix} Df_1(c) \\ \vdots \\ Df_m(c) \\ \vdots \\ Df_M(c) \end{pmatrix}.$$

Proof: Let ϵ and δ be positive numbers such that $\|f(x) - f(c) - Df(c)(x - c)\| \leq \epsilon \|x - c\|$, if $\|x - c\| \leq \delta$. The derivative $Df(c)$ is a linear transformation from \mathbb{R}^N to \mathbb{R}^M , so that

$$Df(c) = \begin{pmatrix} (Df(c))_1 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ (Df(c))_M \end{pmatrix},$$

where $(Df(c))_m : \mathbb{R}^N \rightarrow \mathbb{R}$ is linear, for all m , and is the m^{th} component of the function $Df(c)$. If $\|x - c\| \leq \delta$, then

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

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

$$Df(c) = \begin{pmatrix} Df_1(c) \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ Df_M(c) \end{pmatrix}.$$

■

Theorem 6.6: If $f: U \rightarrow \mathbb{R}^M$, where U is an open subset of \mathbb{R}^N , then the $M \times N$ matrix

$$\begin{pmatrix} \frac{\partial f_1(c)}{\partial x_1} & \dots & \frac{\partial f_1(c)}{\partial x_N} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \frac{\partial f_M(c)}{\partial x_1} & \dots & \frac{\partial f_M(c)}{\partial x_N} \end{pmatrix}$$

represents the linear transformation $Df(c): \mathbb{R}^N \rightarrow \mathbb{R}^M$ with respect to the standard bases of \mathbb{R}^N and \mathbb{R}^M .

Proof: Let $v = (v_1, \dots, v_N) \in \mathbb{R}^N$. Then $v = \sum_{n=1}^N v_n e_n$, where e_n is the n^{th} standard basis vector of \mathbb{R}^N . Therefore

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

$$= \begin{pmatrix} \frac{\partial f_1(c)}{\partial x_1} & \dots & \frac{\partial f_1(c)}{\partial x_N} \\ \vdots & & \vdots \\ \frac{\partial f_M(c)}{\partial x_1} & \dots & \frac{\partial f_M(c)}{\partial x_N} \end{pmatrix} \begin{pmatrix} v_1 \\ \vdots \\ v_N \end{pmatrix}$$

■

Theorem 6.7:

- 1) Let $f: U \rightarrow \mathbb{R}^M$ and $g: U \rightarrow \mathbb{R}^M$, where U is an open subset of \mathbb{R}^N . 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 any $v \in \mathbb{R}^N$, where $f \cdot g$ is the dot product or inner product of f and g . Its value at c is $f \cdot g(c) = f(c) \cdot g(c)$.
- 3) If $\phi: U \rightarrow \mathbb{R}$ and ϕ is differentiable at c and if f is as in part (1), then ϕf is differentiable at c and $D(\phi f)(c)(v) = D\phi(c)(v)f(c) + \phi(c)Df(c)(v)$.

Proof: I leave the proof of this theorem to the reader.

Parts (2) and (3) of this theorem generalize Leibniz's rule for differentiating products of functions. The equation in part (2) of the previous theorem may be written as

$$\begin{aligned} D(f^T g)(c)v &= (Df(c)v)^T g(c) + f(c)^T Dg(c)v \\ &= v^T (Df(c))^T g(c) + f(c)^T Dg(c)v = g(c)^T Df(c)v + f(c)^T Dg(c)v, \end{aligned} \tag{6.1}$$

where I treat $Df(c)$ and $Dg(c)$ as $M \times N$ matrices. The last equation holds because $v^T (Df(c))^T g(c)$ is a number and so equals its own transpose.

I now describe derivatives in some useful special cases. If $f(x) = a^T x$, where a is a constant vector in \mathbb{R}^N , 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 $Df(x) = A$.

Let $M = N$ in the previous theorem, let $f(x) = x$, and let $g(x) = Ax$, where A is an $N \times N$ matrix of constants. Then by equation 6.1,

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

where I is the $N \times N$ identity matrix. If in addition A is symmetric, so that $A^T = A$, then

$$D(x^T Ax)(c)(v) = 2c^T A v,$$

so that the matrix representation of this derivative is

$$D(x^T Ax)(c) = 2c^T A. \tag{6.2}$$

As in the case of differentiable functions of one variable, the derivative of a differentiable function of several variables is 0 at a local maximum or minimum. Let $f: U \rightarrow \mathbb{R}$, where U is an open subset of \mathbb{R}^N .

Definition: The function f has a relative or local maximum at $c \in U$, if for some positive number ε , $f(c) \geq f(x)$, for all $x \in B_\varepsilon(c)$. Similarly f has a relative or local minimum at c , if for some positive number ε , $f(x) \leq f(c)$, if $x \in B_\varepsilon(c)$.

Theorem 6.8: If $f: U \rightarrow \mathbb{R}$, where U is an open subset of \mathbb{R}^N and has a local maximum at $c \in U$ and if f is differentiable at c , 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 $v \in \mathbb{R}^N$. Since by theorem 6.4 $\nabla_v f(c) = Df(c)(v)$, it follows that $Df(c) = 0$. ■

I now apply these results to least squares estimation. Suppose that a variable y equals a linear function of K other variables, x_1, x_2, \dots, x_K plus an error term ε . Formally our model is

$$y = \sum_{k=1}^K \beta_k x_k + \varepsilon.$$

Suppose we do not know the coefficients $\beta_1, \beta_2, \dots, \beta_K$, but we do have N observations of the variables x_1, \dots, x_K and of the corresponding values of y . We can use these observations to estimate β_1, \dots, β_K . Let the observations of y be y_1, \dots, y_N , and let the corresponding observations of x_k be x_{1k}, \dots, x_{Nk} . A very commonly used estimator of β_1, \dots, β_K is the least

squares estimator b_1, \dots, b_K , which is the choice of b_1, \dots, b_K that minimizes the sum of squared errors, $\sum_{n=1}^N (y_n - \sum_{k=1}^K b_k x_{nk})^2$. We need some notation to develop a convenient formula for this estimator. Let

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

The least squares estimator is the value of b that minimizes

$$\begin{aligned} \sum_{n=1}^N (y_n - \sum_{k=1}^K b_k x_{nk})^2 &= (y - Xb) \cdot (y - Xb) = (y - Xb)^T (y - Xb) = (y^T - b^T X^T) (y - Xb) \\ &= y^T y - y^T Xb - b^T X^T y + b^T X^T Xb = y^T y - 2y^T Xb + b^T X^T Xb. \end{aligned}$$

The last equation follows from the fact that since $b^T X^T y$ is a number, it equals its own transpose.

In order to visualize the meaning of the least squares estimator, suppose that $k = 1$. In the next figure, the least estimate b minimizes the sum of squares of the vertical distances from the data points (x_n, y_n) to the line $y = bx$.

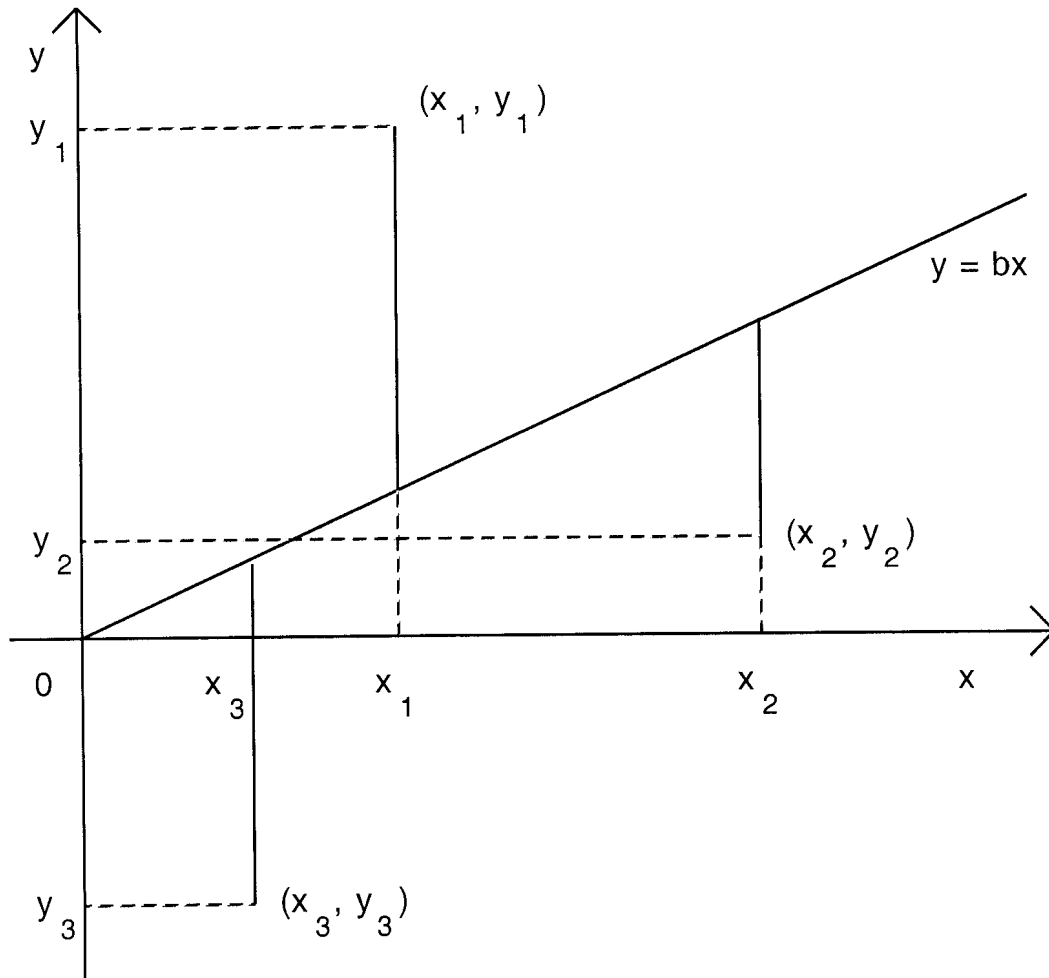
In order to calculate the least squares estimator, we set the derivative of $(y - Xb) \cdot (y - Xb) = y^T y - 2y^T Xb + b^T X^T Xb$ with respect to b equal to 0. Let D_b denote the derivative with respect to the vector b . Then

$$D_b (y - Xb) \cdot (y - Xb) = D_b y^T y - 2D_b y^T Xb + D_b b^T X^T Xb = 0 - 2y^T X + 2b^T X^T X,$$

where I have used the fact that the matrix $X^T X$ is symmetric. It 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 equation (6.2). Setting $D_b (y - Xb) \cdot (y - Xb)$ equal to 0, we obtain the equation

$$0 = -2y^T X + 2b^T X^T X,$$

which implies that $b^T X^T X = y^T X$.



Taking the transpose of both sides of this equation, we obtain

$$X^T X b = X^T y.$$

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.

The N-vector Xb is the projection of the N-vector y onto the linear span of the columns of the $N \times K$ matrix X . In order to see that this is so, we must verify that $y - Xb$ is orthogonal to the columns of X . This is clearly so, because

$$(y - Xb)^T X = (y^T - b^T X^T) X = y^T X - b^T X^T X = 0.$$

I now return to the basic theory of multivariable calculus. The next theorem generalizes the chain rule of differentiation (theorem 5.11) to the multivariate case. I give no proof.

Theorem (Chain Rule of Differentiation) 6.9: Suppose that $f: U \rightarrow V$, where U and V are open subsets of \mathbb{R}^N and \mathbb{R}^M , respectively, and that $g: V \rightarrow \mathbb{R}^K$. Assume that f is differentiable at $c \in U$ and that g is differentiable at $f(c)$. Let $h: U \rightarrow \mathbb{R}^K$ be defined by $h(x) = g(f(x)) = g \circ f(x)$. Then h is differentiable at c and $Dh(c) = Dg(f(c))Df(c)$.

The next theorem generalizes the mean value theorem of single variable calculus, theorem 5.7.

Mean Value Theorem 6.10: Let $f: U \rightarrow \mathbb{R}$, where U is an open subset of \mathbb{R}^N , and suppose that f is everywhere differentiable. Let a and b be points in U and suppose that the line segment from a to b (i.e. $\{(1-t)a + tb \mid 0 \leq t \leq 1\}$) 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 $\phi: [0, 1] \rightarrow \mathbb{R}$ be defined by $\phi(t) = f((1-t)a + tb)$. It is clear from the definition of ϕ that $\phi(0) = f(a)$ and $\phi(1) = f(b)$. By the chain rule, $d\phi(t)/dt = Df((1-t)a + tb)(b - a)$. By the mean value theorem for one variable (theorem 5.7), there exists a number t_0 such that $0 < t_0 < 1$ and $d\phi(t_0)/dt = \phi(1) - \phi(0) = f(b) - f(a)$. Let $c = (1 - t_0)a + t_0b$. Then $Df(c)(b - a) = d\phi(t_0)/dt = f(b) - f(a)$. ■

I do not give a proof of the next theorem, though it is quite important.

Theorem (Existence of a Derivative) 6.11: Let $f: U \rightarrow \mathbb{R}^M$, where U is an open subset of \mathbb{R}^N . If $\partial f_m(x) / \partial x_n$ exists and is continuous on U , for all m and n , then f is differentiable on U .

Definition: Let $f: U \rightarrow \mathbb{R}^M$, where U is an open subset of \mathbb{R}^N . If $\partial f_m(x) / \partial x_n$ exists and is continuous on U , for all m and n , then f is said to be continuously differentiable.

The following language is often used in referring to the derivative.

Terminology:

1) Let $f: U \rightarrow \mathbb{R}^M$, where U is an open subset of \mathbb{R}^N , be everywhere differentiable. For $x \in U$, the $M \times N$ matrix

$$\begin{pmatrix} \frac{\partial f_1(x)}{\partial x_1} & \dots & \frac{\partial f_1(x)}{\partial x_N} \\ \vdots & & \vdots \\ \frac{\partial f_M(x)}{\partial x_1} & \dots & \frac{\partial f_M(x)}{\partial x_N} \end{pmatrix}$$

is called the Jacobian matrix of f at x . This matrix represents the derivative $Df(x)$, and $Df(x)$ is referred to as the Jacobian of f at x .

2) If $f: U \rightarrow \mathbb{R}$, where U is an open subset of \mathbb{R}^N , the derivative $Df(x)$ is sometimes referred to as the gradient of f at x . Its matrix representation is $\left(\frac{\partial f(x)}{\partial x_1}, \dots, \frac{\partial f(x)}{\partial x_N} \right)$. This vector is sometimes denoted $\nabla f(x)$ and is also called the gradient of f at x .

3) The symbol ∇ also refers to the function that carries the differentiable function $f: U \rightarrow \mathbb{R}$ to the function $\nabla f: U \rightarrow \mathbb{R}^N$ defined by $\nabla f(x) = \left(\frac{\partial f(x)}{\partial x_1}, \dots, \frac{\partial f(x)}{\partial x_N} \right)$.

I now turn to the definition and interpretation of the second and higher derivatives of a function $f: U \rightarrow \mathbb{R}$, where U is an open subset of \mathbb{R}^N . Notice that the derivative $Df(x)$ may be interpreted as a function from U to \mathbb{R}^N , the point in \mathbb{R}^N being the gradient vector. By theorem 6.11, the function $Df(x)$ is itself differentiable if each of the components of the gradient in turn have partial derivatives that are continuous functions of x . A partial derivative of a component of the gradient, $\frac{\partial}{\partial x_n} \frac{\partial f(x)}{\partial x_m}$, is written as $\frac{\partial^2 f(x)}{\partial x_n \partial x_m}$ and is called a second partial derivative of f at x .

Theorem (Interchange of the Order of Partial Differentiation) 6.12: Let $f: U \rightarrow \mathbb{R}$, where U is an open subset of \mathbb{R}^N . If for all n and m , $\frac{\partial^2 f(x)}{\partial x_n \partial x_m}$ exists and is a continuous function of x ,

then $\frac{\partial^2 f(x)}{\partial x_m \partial x_n} = \frac{\partial^2 f(x)}{\partial x_n \partial x_m}$, for all m and n .

Remark: This theorem implies that if the second partial derivatives exist and are continuous functions of x , then the matrix

$$\begin{pmatrix} \frac{\partial^2 f(x)}{\partial x_1 \partial x_1} & \dots & \frac{\partial^2 f(x)}{\partial x_N \partial x_1} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \frac{\partial^2 f(x)}{\partial x_1 \partial x_N} & \dots & \frac{\partial^2 f(x)}{\partial x_N \partial x_N} \end{pmatrix}$$

is symmetric. By theorem 6.11, we know that $Df(x)$ is differentiable.

More Terminology: The matrix of second partial derivatives,

$$\begin{pmatrix} \frac{\partial^2 f(x)}{\partial x_1 \partial x_1} & \dots & \frac{\partial^2 f(x)}{\partial x_N \partial x_1} \\ \vdots & & \vdots \\ \frac{\partial^2 f(x)}{\partial x_1 \partial x_N} & \dots & \frac{\partial^2 f(x)}{\partial x_N \partial x_N} \end{pmatrix}$$

is called the Hessian matrix of f at x and is the Jacobian matrix of the gradient of f at x .

I now interpret the Hessian matrix and explain what the second derivative of $f: U \rightarrow \mathbb{R}$ is, where U is an open subset of \mathbb{R}^N . If $v \in \mathbb{R}^N$, $Df(x)(v) = \nabla_v f(x) = \sum_{n=1}^N v_n \frac{\partial f(x)}{\partial x_n}$ is the rate of change of f in the direction v . The rate of change of $Df(x)(v)$ in the direction $w \in \mathbb{R}^N$ is

$$\nabla_w \left(\sum_{n=1}^N v_n \frac{\partial f(x)}{\partial x_n} \right) = \sum_{m=1}^N w_m \frac{\partial}{\partial x_m} \left(\sum_{n=1}^N v_n \frac{\partial f(x)}{\partial x_n} \right) = \sum_{m=1}^N \sum_{n=1}^N v_n w_m \frac{\partial^2 f(x)}{\partial x_m \partial x_n} = v^T D^2 f(x) w,$$

where $D^2 f(x)$ is the Hessian matrix of f at x . The function $v^T D^2 f(x) w$ is a bilinear form on \mathbb{R}^N and may be written as $D^2 f(x)(v, w)$. This bilinear form is the second derivative of f at x . By theorem 6.12, it is symmetric if the second partial derivatives of f are continuous functions of x .

I now define the third derivative of f . The rate of change of $D^2 f(x)(v, w)$ at x in the direction $u \in \mathbb{R}^N$ is

$$D^3 f(x)(v, w, u) = \sum_{n=1}^N \sum_{m=1}^N \sum_{k=1}^N v_n w_m u_k \frac{\partial^3 f(x)}{\partial x_k \partial x_m \partial x_n}.$$

This is a trilinear form that is represented by a three dimensional matrix with typical entry

$$\frac{\partial^3 f(x)}{\partial x_k \partial x_m \partial x_n}.$$

$D^3 f(x)$ is the third derivative of f at x . Continuing inductively, for any positive integer r , we can define the r^{th} derivative at x of f to be the r -linear form

$$Df(x)(v_1, \dots, v_r) = \sum_{n_1=1}^N \dots \sum_{n_r=1}^N v_{1,n_1} \dots v_{r,n_r} \frac{\partial f(x)}{\partial x_{n_1} \dots \partial x_{n_r}},$$

where for each $s = 1, \dots, r$, $v_s = (v_{s1}, \dots, v_{sN}) \in \mathbb{R}^N$.

Remark: The matrix representation of $Df(x)$ is $\left(\frac{\partial f(x)}{\partial x_1}, \dots, \frac{\partial f(x)}{\partial x_N} \right)$. When we take

the derivative of $Df(x)$, we think of it as a function from U to \mathbb{R}^N and so write $Df(x)$ as the column vector

$$Df(x) = \begin{pmatrix} \frac{\partial f(x)}{\partial x_1} \\ \cdot \\ \cdot \\ \cdot \\ \frac{\partial f(x)}{\partial x_N} \end{pmatrix}.$$

The matrix representation of the derivative of this vector function is the Hessian matrix.

Appendix to Lecture #6, Example

Example: (An example of a function that has partial derivatives everywhere, but is not differentiable.) Let $f: \mathbb{R}^2 \rightarrow \mathbb{R}$ be defined by the equations

$$f(x, y) = \frac{x^2y + xy^2}{x^2 + y^2}, \text{ if } (x, y) \neq 0, \text{ and } f(0, 0) = 0.$$

Clearly all the partial derivatives of f exist for $(x, y) \neq 0$. Since $f(x, 0) = 0$ and $f(0, y) = 0$, it follows that

$$\frac{\partial f(0, 0)}{\partial x} = 0 \text{ and } \frac{\partial f(0, 0)}{\partial y} = 0,$$

so that f has both partial derivatives at $(x, y) = (0, 0)$ as well.

Notice that $f(tx, ty) = tf(x, y)$, for all real numbers t . Such a function is said to be homogeneous of degree 1. This equation implies that

$$\frac{df(tx, ty)}{dx} = t \frac{\partial f(x, y)}{\partial x},$$

where on the left-hand side, I consider $f(tx, ty)$ to be a function of x alone. Also by the chain rule,

$$\frac{df(tx, ty)}{dx} = t \frac{\partial f(tx, ty)}{\partial x},$$

where again on the left I consider $f(tx, ty)$ to be a function of x alone. These two equations imply that

$$t \frac{\partial f(x, y)}{\partial x} = t \frac{\partial f(tx, ty)}{\partial x}$$

and so

$$\frac{\partial f(x, y)}{\partial x} = \frac{\partial f(tx, ty)}{\partial x},$$

for all $t \neq 0$. It follows that

$$\frac{\partial f(t, t)}{\partial x} = \frac{\partial f(1, 1)}{\partial x} = \frac{1}{2},$$

for all $t \neq 0$, so that

$$\lim_{t \rightarrow 0} \frac{\partial f(t, t)}{\partial x} = \frac{1}{2} \neq 0 = \frac{\partial f(0, 0)}{\partial x} ..$$

Therefore, the partial derivatives of f are not continuous at $(0, 0)$.

The function f is not differentiable at $(0, 0)$. In order to see that this is so, notice that $f(t, t) = t$, for all t , so that

$$\left. \frac{df(t, t)}{dt} \right|_{t=0} = 1.$$

If f were differentiable at $(0, 0)$, then by the chain rule we would have the equation

$$\left. \frac{df(t, t)}{dt} \right|_{t=0} = \frac{\partial f(0, 0)}{\partial x} + \frac{\partial f(0, 0)}{\partial y},$$

which is impossible since the right-hand side of this equation equals 0 and the left-hand side equals 1.