DIFFERENTIAL CALCULUS

PAUL SCHRIMPF[1]
OCTOBER 25, 2013
UNIVERSITY OF BRITISH COLUMBIA
ECONOMICS 526

In this lecture, we will define derivatives for functions on vector spaces. We will show that all the familiar properties of derivatives — the mean value theorem, chain rule, etc — hold in any vector space. We will primarily focus on $\mathbb{R}^n$, but we also discuss infinite dimensional spaces (will we need to differentiate in them to study optimal control later). All of this material is also covered in chapter 4 of Carter. Chapter 14 of Simon and Blume and chapter 9 of Rudin's *Principles of Mathematical Analysis* cover differentiation on $\mathbb{R}^n$. Simon and Blume is better for general understanding and applications, but Rudin is better for proofs and rigor.

1. DERIVATIVES

1.1. **Partial derivatives.** We have discussed limits of sequences, but perhaps not limits of functions. To be complete, we define limits as follows.

Definition 1.1. Let X and Y be metric spaces and $f : X \to Y$.

$$\lim_{x \to x_0} f(x) = c$$

where x and $x_0 \in X$ and $c \in Y$, means that $\forall \epsilon > 0 \; \exists \delta > 0$ such that $d(x, x_0) < \delta$ implies $d(f(x), c) < \epsilon$.

Equivalently, we could say $\lim_{x \to x_0} f(x) = c$ means that for any sequence $\{x_n\}$ with $x_n \to x$, $f(x_n) \to c$.

You are probably already familiar with the derivative of a function of one variable. Let $f : \mathbb{R} \to \mathbb{R}$. f is differentiable at x_0 if

$$\lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h} = \frac{df}{dx}(x_0)$$

exists. Similiarly, if $f : \mathbb{R}^n \to \mathbb{R}$ we define its ith partial derivative as follows.

Definition 1.2. Let $f : \mathbb{R}^n \to R$. The ith **partial derivative** of f is

$$\frac{\partial f}{\partial x_i}(x_0) = \lim_{h \to 0} \frac{f(x_{01}, ..., x_{0i} + h, ...x_{0n}) - f(x_0)}{h}.$$

The ith partial derivative tells you how much the function changes as its ith argument changes.

[1]Thanks to Dana Galizia for corrections.

Example 1.1. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a production function. Then we call $\frac{\partial f}{\partial x_i}$ the **marginal product** of x_i. If f is Cobb-Douglas, $f(k,l) = Ak^\alpha l^\beta$, where k is capital and l is labor, then the marginal products of capital and labor are

$$\begin{aligned}\frac{\partial f}{\partial k}(k,l) &= A\alpha k^{\alpha-1} l^\beta \\ \frac{\partial f}{\partial l}(k,l) &= A\beta k^\alpha l^{\beta-1}.\end{aligned}$$

1.2. **Examples.**

Example 1.2. If $u : \mathbb{R}^n \to \mathbb{R}$ is a utility function, then we call $\frac{\partial u}{\partial x_i}$ the marginal utility of x_i. If u is CRRA,

$$u(c_1, ..., c_T) = \sum_{t=1}^{T} \beta^t \frac{c_t^{1-\gamma}}{1-\gamma}$$

then the marginal utility of consumption in period t is

$$\frac{\partial u}{\partial c_t} = \beta^t c_t^{-\gamma}.$$

Example 1.3. The price elasticity of demand is the percentage change in demand divided by the percentage change in its price. If $q_1 : \mathbb{R}^3 \to \mathbb{R}$ is a demand function with three arguments: own price p_1, the price of another good, p_2, and consumer income, y. The own price elasticity is

$$\epsilon_{q_1,p_1} = \frac{\partial q_1}{\partial p_1} \frac{p_1}{q_1(p_1, p_2, y)}.$$

The cross price elasticity is the percentage change in demand divided by the percentage change in the other good's price, i.e.

$$\epsilon_{q_1,p_2} = \frac{\partial q_1}{\partial p_2} \frac{p_2}{q_1(p_1, p_2, y)}.$$

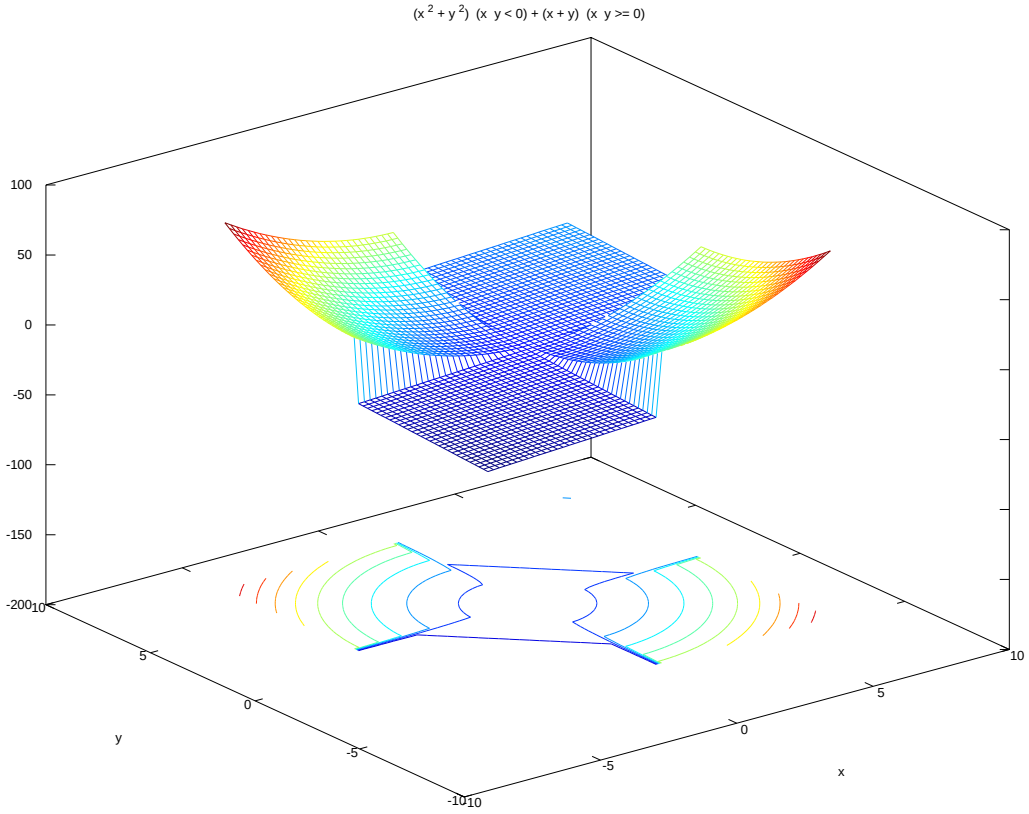
Similarly, the income elasticity of demand is

$$\epsilon_{q_1,y} = \frac{\partial q_1}{\partial y} \frac{y}{q_1(p_1, p_2, y)}.$$

1.3. **Total derivatives.** Derivatives of univariate functions have a number of useful properties that partial derivatives do not always share. Examples of useful properties include univariate derivatives giving the slope of a tangent line, the implicit function theorem, and Taylor series approximations. We would like the derivatives of multivariate functions to have these properties, but partial derivatives are not enough for this.

Example 1.4. Consider $f : \mathbb{R}^2 \to \mathbb{R}$,

$$f(x,y) = \begin{cases} x^2 + y^2 & \text{if } xy < 0 \\ x + y \text{ if } xy \geq 0 \end{cases}$$



The partial derivatives of this function at 0 are $\frac{\partial f}{\partial x}(0,0) = 1$ and $\frac{\partial f}{\partial y}(0,0) = 1$. However, there are points arbitrarily close to zero with $\frac{\partial f}{\partial x}(x,y) = 2x + 2y$. If we were to try to draw a tangent plane to the function at zero, we would find that we cannot. Although the partial derivatives of this function exist everywhere, it is in some sense not differentialable at zero (or anywhere with $xy = 0$).

Partially motivated by the preceding example, we define the total derivative (or just the derivative; we're saying "total" to emphasize the difference between partial derivatives and the derivative).

Definition 1.3. Let $f : \mathbb{R}^n \to \mathbb{R}$. The **derivative** (or total derivative or differential) of f at x_0 is a linear mapping, $Df_{x_0} : \mathbb{R}^n \to \mathbb{R}^1$ such that

$$\lim_{h \to 0} \frac{|f(x_0 + h) - f(x_0) - Df_{x_0}h|}{\|h\|} = 0.$$

The h in this definition is an n vector in $\mathbb{R}^n$. This is contrast to the h in the definition of partial derivatives, which was just a scalar. The fact that h is now a vector is important because h can approach 0 along any path. Partial derivatives only look at the limits as h approaches 0 along the axes. This allows partial derivatives to exist for strange functions like the one in example 1.4. We can see that the function from the example is not differentiable by letting h approach 0 along a path that switches from $xy < 0$ to $xy \geq 0$ infinitely

many times close to 0. The limit in the definition of the derivative does not exist along such a path, so the derivative does not exist.

Comment 1.1. In proofs, it will be useful to define $r(x,h) = f(x+h) - f(x) - Df_x h$. We will then repeatedly use the fact that $\lim_{h\to 0} \frac{|r(x,h)|}{\|h\|} = 0$.

If the derivative of f at x_0 exists, then so do the partial derivatives, and the total derivative is simply the $1 \times n$ matrix of partial derivatives.

Theorem 1.1. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable at x_0, then $\frac{\partial f}{\partial x_i}(x_0)$ exists for each i and*

$$Df_{x_0} h = \begin{pmatrix} \frac{\partial f}{\partial x_1}(x_0) & \cdots \frac{\partial f}{\partial x_n}(x_0) \end{pmatrix} h.$$

Proof. Since f is differentiable at x_0, we can make $h = e_i t$ for e_i the ith standard basis vector, and t a scalar. The definition of derivative says that

$$\lim_{t\to 0} \frac{|f(x_0 + e_i t) - f(x_0) - Df_{x_0}(e_i t)|}{\|e_i t\|} = 0.$$

Let

$$r_i(x_0, t) = f(x_0 + e_i t) - f(x_0) - t Df_{x_0} e_i$$

and note that $\lim_{t\to 0} \frac{|r_i(x_0,t)|}{|t|} = 0$. Rearranging and dividing by t,

$$\frac{f(x_0 + e_i t) - f(x_0)}{t} = Df_{x_0} e_i + \frac{r_i(x_0, t)}{t}$$

and taking the limit

$$\lim_{t\to 0} \frac{f(x_0 + e_i t) - f(x_0)}{t} = Df_{x_0} e_i$$

we get the exact same expression as in the definition of the partial derivative. Therefore, $\frac{\partial f}{\partial x_i} = Df_{x_0} e_i$. Finally, as when we first introduced matrices, we know that linear transformation Df_{x_0} must be represented by

$$Df_{x_0} h = \begin{pmatrix} \frac{\partial f}{\partial x_1}(x_0) & \cdots \frac{\partial f}{\partial x_n}(x_0) \end{pmatrix} h.$$

□

We know from example 1.4 that the converse of this theorem is false. The existence of partial derivatives is not enough for a function to be differentiable. However, if the partial derivatives exist and are continuous in a neighborhood, then the function is differentiable.

Theorem 1.2. *Let $f : \mathbb{R}^n \to \mathbb{R}$ and suppose its partial derivatives exist and are continuous in $N_\delta(x_0)$ for some $\delta > 0$. Then f is differentiable at x_0 with*

$$Df_{x_0} = \begin{pmatrix} \frac{\partial f}{\partial x_1}(x_0) & \cdots \frac{\partial f}{\partial x_n}(x_0) \end{pmatrix}.$$

Proof. Let $h = (h_1, ..., h_n)$ with $\|h\| < r$. Notice that

$$f(x_0+h) - f(x_0) = f(x_0+h_1e_1) - f(x_0) + f(x_0+h_1e_1+h_2e_2) - f(x_0+h_1e_1) + ... \quad (1)$$

$$+ f(x_0+h) - f\left(x_0 - \sum_{i=1}^{n-1} h_i e_i\right) \quad (2)$$

$$= \sum_{j=1}^{n} f\left(x_0 + \sum_{i=1}^{j} h_i e_i\right) - f\left(x_0 + \sum_{i=1}^{j-1} h_i e_i\right). \quad (3)$$

By the mean value theorem (1.5),

$$f\left(x_0 + \sum_{i=1}^{j} h_i e_i\right) - f\left(x_0 + \sum_{i=1}^{j-1} h_i e_i\right) = h_j \frac{\partial f}{\partial x_j}(x_0 + \sum_{i=1}^{j-1} h_i e_i + \bar{h}_j e_j)$$

for some $\bar{h}_j$ between 0 and h_j. The partial derivatives are continuous by assumption, so by making r small enough, we can make

$$\left| \frac{\partial f}{\partial x_j}(x_0 + \sum_{i=1}^{j-1} h_i e_i + \bar{h}_j e_j) - \frac{\partial f}{\partial x_j}(x_0) \right| < \epsilon/n,$$

for any $\epsilon > 0$. Combined with equation 3 now we have,

$$f(x_0+h) - f(x_0) = \sum_{j=1}^{n} h_j \left(\frac{\partial f}{\partial x_j}(x_0) + \frac{\epsilon}{n} \right) \quad (4)$$

$$\left| f(x_0+h) - f(x_0) - \sum_{j=1}^{n} h_j \frac{\partial f}{\partial x_j}(x_0) \right| = \left| \sum_{j=1}^{n} h_j \epsilon/n \right| \quad (5)$$

$$|f(x_0+h) - f(x_0) - Df_{x_0}h| \leq \epsilon \|h\| \quad (6)$$

Dividing by $\|h\|$ and taking the limit,

$$\lim_{h \to 0} \frac{|f(x_0+h) - f(x_0) - Df_{x_0}h|}{\|h\|} \leq \epsilon.$$

This is true for any $\epsilon > 0$, so the limit must be 0. □

A minor modification of this proof would show the stronger result that $f : \mathbb{R}^n \to \mathbb{R}$ has a continuous derivative on an open set $U \subseteq \mathbb{R}^n$ if and only if its partial derivatives are continuous on U. We call such a function **continuously differentiably** on U and denote the set of all such function as $C^1(U)$.

1.4. **Mean value theorem.** The mean value theorem in $\mathbb{R}^1$ says that $f(x+h) - f(x) = f'(\bar{x})h$ for some $\bar{x}$ between $x+h$ and x. The same theorem holds for multivariate functions. To prove it, we will need a couple of intermediate results. Recall the following from the midterm review.

Theorem 1.3. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuous and $K \subset \mathbb{R}^n$ be compact. Then $\exists x^* \in K$ such that $f(x^*) \geq f(x) \forall x \in K$.*

Simon and Blume call this Weierstrass's theorem (30.1).

Definition 1.4. Let $f : \mathbb{R}^n \to \mathbb{R}$. we say that f has a local maximum at x if $\exists \delta > 0$ such that $f(y) \leq f(x)$ for all $y \in N_\delta(x)$.

Next, we need a result that relates derivatives to maxima.

Theorem 1.4. *Let $f : \mathbb{R}^n \to \mathbb{R}$ and suppose f has a local maximum at x and is differentiable at x. Then $Df_x = 0$.*

Proof. Choose δ as in the definition of a local maximum. Since f is differentiable, we can write

$$\frac{f(x+h) - f(x)}{\|h\|} = \frac{Df_x h + r(x,h)}{\|h\|}$$

where $\lim_{h\to 0} \frac{|r(x,h)|}{\|h\|} = 0$. Let $h = tv$ for some $v \in \mathbb{R}^n$ with $\|v\| = 1$ and $t \in \mathbb{R}$. If $Df_x v > 0$, then for $t > 0$ small enough, we would have $\frac{f(x+tv)-f(x)}{|t|} = Df_x v + \frac{r(x,tv)}{|t|} > Df_x v/2 > 0$ and $f(x+tv) > f(x)$ in contradiction to x being a local maximum. Similary, if $Df_v v < 0$ then for $t < 0$ and small, we would have $\frac{f(x+tv)-f(x)}{|t|} = D - f_x v + \frac{r(x,tv)}{|t|} > -Df_x v/2 > 0$ and $f(x+tv) > f(x)$. Thus, it must be that $Df_x v = 0$ for all v, i.e. $Df_x = 0$. □

Now we can prove the mean value theorem.

Theorem 1.5 (mean value)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}^1$ be in $C^1(U)$ for some open U. Let $x, y \in U$ be such that the line connecting x and y, $\ell(x,y) = \{z \in \mathbb{R}^n : z = \lambda x + (1-\lambda)y, \lambda \in [0,1]\}$, is also in U. Then there is some $\bar{x} \in \ell(x,y)$ such that*

$$f(x) - f(y) = Df_{\bar{x}}(x - y).$$

Proof. Let $z(t) = y + t(x-y)$ for $t \in [0,1]$ (i.e. $t = \lambda$). Define

$$g(t) = f(y) - f(z(t)) + (f(x) - f(y))\, t$$

Note that $g(0) = g(1) = 0$. The set $[0,1]$ is closed and bounded, so it is compact. It is easy to verify that $g(t)$ is continuously differentiable since f is continuously differentiable . Hence, g must attain its maximum on $[0,1]$, say at $\bar{t}$. If $\bar{t} = 0$ or 1, then either g is constant, in which case any $\bar{t} \in (0,1)$ is also a maximum, or g must have an interior minimum, and we can look at the maximum of $-g$ instead. When $\bar{t}$ is not 0 or 1, then the previous theorem shows that $g'(\bar{t}) = 0$. Simple calculation shows that

$$g'(\bar{t}) = -Df_{z(\bar{t})}(x-y) + f(x) - f(y) = 0$$

so

$$Df_{\bar{x}}(x-y) = f(x) - f(y)$$

where $\bar{x} = z(\bar{t})$. □

1.5. **Functions from $\mathbb{R}^n \to \mathbb{R}^m$.** So far we have only looked at functions from $\mathbb{R}^n$ to $\mathbb{R}$. Functions to $\mathbb{R}^m$ work essentially the same way.

Definition 1.5. Let $f : \mathbb{R}^n \to \mathbb{R}^m$. The **derivative** (or total derivative or differential) of f at x_0 is a linear mapping, $Df_{x_0} : \mathbb{R}^n \to \mathbb{R}^m$ such that

$$\lim_{h\to 0} \frac{\|f(x_0+h) - f(x_0) - Df_{x_0} h\|}{\|h\|} = 0.$$

Theorems 1.1 and 2.1 sill hold with no modification. The total derivative of f can be represented by the m by n matrix of partial derivatives,

$$Df_{x_0} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(x_0) & \cdots & \frac{\partial f_1}{\partial x_n}(x_0) \\ \vdots & & \vdots \\ \frac{\partial f_m}{\partial x_1}(x_0) & \cdots & \frac{\partial f_m}{\partial x_n}(x_0) \end{pmatrix}.$$

This matrix of partial derivatives is often called the **Jacobian** of f.

The mean value theorem 1.5 holds for each of the component functions of $f : \mathbb{R}^n \to \mathbb{R}^m$. Meaning, that f can be written as $f(x) = \begin{pmatrix} f_1(x) & \cdots & f_m(x) \end{pmatrix}^T$ where each $f_j : \mathbb{R}^n \to \mathbb{R}$. The mean value theorem is true for each f_j, but the $\bar{x}$'s will typically differ with j.

Corollary 1.1 (mean value for $\mathbb{R}^n \to \mathbb{R}^m$)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be in $C^1(U)$ for some open U. Let $x, y \in U$ be such that the line connecting x and y, $\ell(x,y) = \{z \in \mathbb{R}^n : z = \lambda x + (1-\lambda)y, \lambda \in [0,1]\}$, is also in U. Then there are $\bar{x}_j \in \ell(x,y)$ such that*

$$f_j(x) - f_j(y) = Df_{j\bar{x}_j}(x - y)$$

and

$$f(x) - f(y) = \begin{pmatrix} Df_{1\bar{x}_1} \\ \vdots \\ Df_{m\bar{x}_m} \end{pmatrix} (x - y).$$

Slightly abusing notation, we might at times write $Df_{\bar{x}}$ instead of $\begin{pmatrix} Df_{1\bar{x}_1} & \cdots & Df_{m\bar{x}_m} \end{pmatrix}^T$ with the understanding that we mean the later.

1.6. **Chain rule.** For univariate functions, the chain rule says that the derivative of $f(g(x))$ is $f'(g(x))g'(x)$. The same is true for multivariate functions.

Theorem 1.6. *Let $f : \mathbb{R}^n \to \mathbb{R}^m$ and $g : \mathbb{R}^k \to \mathbb{R}^n$. Let g be continuously differentiable on some open set U and f be continuously differentiable on $g(U)$. Then $h : \mathbb{R}^k \to \mathbb{R}^m$, $h(x) = f(g(x))$ is continuously differentiable on U with*

$$Dh_x = Df_{g(x)} Dg_x$$

Proof. Let $x \in U$. Consider

$$\frac{\|f(g(x+d)) - f(g(x))\|}{\|d\|}.$$

Since g is differentiable by the mean value theorem, $g(x+d) = g(x) + Dg_{\bar{x}(d)}d$, so

$$\begin{aligned} \|f(g(x+d)) - f(g(x))\| &= \left\| f(g(x) + Dg_{\bar{x}(d)}d) - f(g(x)) \right\| \\ &\leq \|f(g(x) + Dg_x d) - f(g(x))\| + \epsilon \end{aligned}$$

where the inequality follows from the the continuity of Dg_x and f, and holds for any $\epsilon > 0$. f is differentiable, so

$$\lim_{Dg_x d \to 0} \frac{\left\| f(g(x) + Dg_x d) - f(g(x)) - Df_{g(x)} Dg_x d \right\|}{\|Dg_x d\|} = 0$$

Using the Cauchy-Schwarz inequality, $\|Dg_x d\| \leq \|Dg_x\| \, \|d\|$, we get

$$\frac{\left\|f(g(x)+Dg_x d)-f(g(x))-Df_{g(x)}Dg_x d\right\|}{\|Dg_x\| \, \|d\|} \leq \frac{\left\|f(g(x)+Dg_x d)-f(g(x))-Df_{g(x)}Dg_x d\right\|}{\|Dg_x d\|}$$

so

$$\lim_{d\to 0} \frac{\left\|f(g(x)+Dg_x d)-f(g(x))-Df_{g(x)}Dg_x d\right\|}{\|d\|} = 0.$$

□

1.7. **Higher order derivatives.** We can take higher order derivatives of multivariate functions just like of univariate functions. If $f : \mathbb{R}^n \to \mathbb{R}^m$, then is has nm partial first derivatives. Each of these has n partial derivatives, so f has $n^2 m$ partial second derivatives, written $\frac{\partial^2 f_k}{\partial x_i \partial x_j}$.

Theorem 1.7. *Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be twice continuously differentiable on some open set U. Then*

$$\frac{\partial^2 f_k}{\partial x_i \partial x_j}(x) = \frac{\partial^2 f_k}{\partial x_j \partial x_i}(x)$$

for all i, j, k and $x \in U$.

Proof. Using the definition of partial derivative, twice, we have

$$\begin{aligned}\frac{\partial^2 f}{\partial x_i \partial x_j} &= \lim_{t_j \to 0} \frac{\lim_{t_i \to 0} \frac{f(x+t_i e_i + t_j e_j) - f(x+t_j e_j)}{t_i} - \lim_{t_i \to 0} \frac{f(x+t_i e_i)-f(x)}{t_i}}{t_j} \\ &= \lim_{t_j \to 0} \lim_{t_i \to 0} \frac{f(x+t_j e_j + t_i e_i) - f(x+t_j e_j) - f(x+t_i e_i) + f(x)}{t_j t_i}\end{aligned}$$

from which it is apparent that we get the same expression for $\frac{\partial^2 f}{\partial x_j \partial x_i}$.[1] □

The same argument shows that in general the order of partial derivatives does not matter.

Corollary 1.2. *Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be k times continuously differentiable on some open set U. Then*

$$\frac{\partial^k f}{\partial x_1^{j_1} \cdots \partial x_n^{j_n}} = \frac{\partial^k f}{\partial x_{p(1)}^{j_{p(1)}} \cdots \partial x_{p(n)}^{j_{p(n)}}}$$

where $\sum_{i=1}^n j_i = k$ and $p : \{1, .., n\} \to \{1, ..., n\}$ is any permutation (i.e. reordering).

[1] This proof is not completely correct. We should carefully show that we can interchange the order of taking limits. Interchanging limits is not always possible, but the assumed continuity makes it possible here.

1.8. **Taylor series.** You have probably seen Taylor series for univariate functions before. A function can be approximated by a polynomial whose coefficients are the function's derivatives.

Theorem 1.8. *Let $f : \mathbb{R} \to \mathbb{R}$ be $k+1$ times continuously differentiable on some open set U, and let a, $a+h \in U$. Then*

$$f(a+h) = f(a) + f'(a)h + \frac{f^2(a)}{2}h^2 + ... + \frac{f^k(a)}{k!}h^k + \frac{f^{k+1}(\bar{a})}{(k+1)!}h^{k+1}$$

where $\bar{a}$ is between a and h.

The same theorem is true for multivariate functions.

Theorem 1.9. *Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be k times continuously differentiable on some open set U and $a, a+h \in U$. Then there exists a k times continuously differentiable function $r_k(a,h)$ such that*

$$f(a+h) = f(a) + \sum_{\sum_{i=1}^n j_i = 1}^{k} \frac{1}{k!} \frac{\partial^{\sum j_i} f}{\partial x_1^{j_1} \cdots \partial x_n^{j_n}}(a) h_1^{j_1} h_2^{j_2} \cdots h_n^{j_n} + r_k(a,h)$$

and $\lim_{h \to 0} \|r_k(a,h)\| \, \|h\|^k = 0$.

Proof. Follows from the mean value theorem. For $k = 1$, the mean value theorem says that

$$\begin{aligned} f(a+h) - f(a) =& Df_{\bar{a}} h \\ f(a+h) =& f(a) + Df_{\bar{a}} h \\ =& f(a) + Df_a h + \underbrace{(Df_{\bar{a}} - Df_a)h}_{r_1(a,h)} \end{aligned}$$

Df_a is continuous as a function of a, and as $h \to 0$, $\bar{a} \to a$, so $\lim_{h \to 0} r_1(a,h) = 0$, and the theorem is true for $k = 1$. For general k, suppose we have proven the theorem up to $k-1$. Then repeating the same argument with the $k-1$st derivative of f in place of f shows that theorem is true for k. The only complication is the division by $k!$. To see where it comes from, we will just focus on $f : \mathbb{R} \to \mathbb{R}$. The idea is the same for $\mathbb{R}^n$, but the notation gets messy. Suppose we want a second order approximation to f at a,

$$\hat{f}(h) = f(a) + f'(a)h + c_2 f^2(a) h^2$$

and pretend that we do not know c_2. Consider $f(a+h) = \hat{f}(h)$. Applying the mean value theorem to the difference of these functions twice, we have

$$\begin{aligned} f(a+h) - \hat{f}(h) =& f(a) - \underbrace{\hat{f}(0)}_{=f(a)} + \left[f'(a+\bar{h}_1) - \underbrace{\hat{f}'(\bar{h}_1)}_{=f'(a)} \right] h \\ =& f'(a) - \hat{f}'(0) + \left[f^2(a+\bar{h}_2) - \underbrace{\hat{f}^2(\bar{h}_2)}_{=2c_2 f^2(a)} \right] \bar{h}_1 h \\ =& f^2(a)(1-2c_2)\bar{h}_1 h + f^3(a+\bar{h}_3)\bar{h}_2 \bar{h}_1 h \end{aligned}$$

if we set $c_2 = \frac{1}{2}$, we can eliminate one term and

$$|f(a+h) - \hat{f}(h)| \leq |\underbrace{f^3(a+\bar{h}_3)h^3}_{=r_2(a,h)}|.$$

Repeating this sort of argument, we will see that setting $c_k = \frac{1}{k!}$ ensures that $\lim_{h\to 0} \|r_k(a,h)\| \|h\|^k = 0$. □

Example 1.5. The mean value theorem is used often in econometrics to show asymptotic normality. Many estimators can be written as

$$\hat{\theta}_n \in \arg\min_{\theta\in\Theta} Q^n(\theta)$$

where $Q^n(\theta)$ is some objective function that depends on the sampled data. Examples include least squares, maximum likelihood and the generalized method of moments. Suppose there is also a population version of the objective function, $Q^0(\theta)$ and $Q^n(\theta) \xrightarrow{p} Q^0(\theta)$ as $n\to\infty$. There is a true value of the parameter, θ_0, that satisfies

$$\theta_0 \in \arg\min_{\theta\in\Theta} Q^0(\theta).$$

For example for OLS,

$$Q^n(\theta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i\theta)^2$$

and

$$Q^0(\theta) = \mathrm{E}\left[(Y - X\theta)^2\right].$$

If Q^n is continuously differentiable[2] on Θ and $\hat{\theta}_n \in \mathrm{int}(\Theta)$, then from theorem 1.4,

$$DQ^n_{\hat{\theta}_n} = 0$$

Applying the mean value theorem,

$$\begin{aligned} 0 =& DQ^n_{\hat{\theta}_n} = DQ^n_{\theta_0} + D^2Q^n_{\bar{\theta}}(\hat{\theta}_n - \theta_0) \\ \hat{\theta}_n - \theta_0 =& -\left(D^2Q^n_{\bar{\theta}}\right)^{-1} DQ^n_{\theta_0}. \end{aligned}$$

Typically, some variant of the central limit theorem implies $\sqrt{n}DQ^n_{\theta_0} \xrightarrow{d} N(0,\Sigma)$. For example for OLS,

$$\sqrt{n}DQ^n_{\theta} = \frac{1}{\sqrt{n}}\sum_i 2(y_i - x_i\theta)\theta.$$

Also, typically $D^2Q^n_{\bar{\theta}} \xrightarrow{p} D^2Q^0_{\theta_0}$, so by Slutsky's theorem,[3]

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -\left(D^2Q^n_{\bar{\theta}}\right)^{-1}\sqrt{n}DQ^n_{\theta_0} \xrightarrow{d} N\left(0, \left(D^2Q^0_{\theta_0}\right)^{-1}\Sigma\left(D^2Q^0_{\theta_0}\right)^{-1}\right).$$

[2]Essentially the same argument works if you expand Q^0 instead of Q^n. This is sometimes necessary because there are some models, like quantile regression, where Q^n is not differentiable, but Q^0 is differentiable.

[3]Part of Slutsky's theorem says that if $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$, then $X_n/Y_n \xrightarrow{d} X/c$.

2. FUNCTIONS ON VECTOR SPACES

To analyze infinite dimensional optimization problems, we will need to differentiate functions on infinite dimensional vector spaces. We will come back to this point when we study optimal control and dynamic programming. Anyway, we can define the derivative of a function between any two vector spaces as follows.

Definition 2.1. Let $f : V \to W$. The Fréchet **derivative** of f at x_0 is a continuous[4] linear mapping, $Df_{x_0} : V \to W$ such that

$$\lim_{h \to 0} \frac{\|f(x_0 + h) - f(x_0) - Df_{x_0} h\|}{\|h\|} = 0.$$

Note that this definition is the same as the definition of total derivative.

Example 2.1. Let $V = \mathcal{L}^\infty(0,1)$ and $W = \mathbb{R}$. Suppose f is given by

$$f(x) = \int_0^1 g(x(\tau), (\tau)) d\tau$$

for some continuously differentiable function $g : \mathbb{R}^2 \to \mathbb{R}$. Then Df_x is a linear transformation from V to $\mathbb{R}$. How can we calculate Df_x? If V were $\mathbb{R}^n$ we would calculate the partial derivatives of f and then maybe check that they are continuous so that theorem holds. For an infinite dimensional space there are infinite partial derivatives, so we cannot possibly compute them all. However, we can look at directional derivatives.

Definition 2.2. Let $f : V \to W$, $v \in V$ and $x \in U \subseteq V$ for some open U. The **directional derivative** (or Gâteaux derivative when V is infinite dimensional) in direction v at x is

$$df(x; v) = \lim_{\alpha \to 0} \frac{f(x + \alpha v) - f(x)}{\alpha}.$$

where $\alpha \in \mathbb{R}$ is a scalar.

Analogs of theorems 1.1 and 2.1 relates the Gâteaux derivative to the Fréchet derivative.

Lemma 2.1. *If $f : V \to W$ is Fréchet differentiable at x, then the Gâteaux derivative, $df(x; v)$, exists for all $v \in V$, and*

$$df(x; v) = Df_x v.$$

The proof of theorem 2.1 relies on the fact that $\mathbb{R}^n$ is finite dimensional. In fact, in an infinite dimensional space it is not enough that all the directional derivatives be continuous on an open set around x for the function to be differentiable at x; we also require the directional derivatives to be linear in v. In finite dimensions we can always create a linear map from the partial derivatives by arranging the partial derivatives in a matrix. In infinite dimensions, we cannot do that.

[4] If V and W are finite dimensional, then all linear functions are continuous. In infinite dimensions, there can be discontinuous linear functions.

Lemma 2.2. *If $f : V \to W$ has Gâteaux derivatives that are linear in v and "continuous" in x in the sense that $\forall \epsilon > 0\ \exists \delta > 0$ such that if $\|x_1 - x\| < \delta$, then*

$$\sup_{v \in V} \frac{\|df(x_1; v) - df(x; v)\|}{\|v\|} < \epsilon$$

then f is Fréchet differentiable with $Df_{x_0} v = df(x; v)$.

Comment 2.1. This continuity in x is actually a very natural definition. If V and W are normed vector spaces, then the set of all bounded (or equivalently continuous) linear transformations is also a normed vector space with norm

$$\|A\| \equiv \sup_{v \in V} \frac{\|Av\|_W}{\|v\|_V}.$$

We are requiring $df(x; v)$ as a function of x to be continuous with respect to this norm.

Proof. This proof goes somewhat beyond the scope of the course. Note that

$$f(x + h) - f(x) = \int_0^1 df(x + th, h) dt$$

by the fundamental theorem of calculus (which we should really prove, but do not have time for, so we will take it as given). Then,

$$\begin{aligned} \|f(x + h) - f(x) - df(x; h)\| &= \left\| \int_0^1 df(x + th, h) - df(x, h) dt \right\| \\ &\leq \int_0^1 \|df(x + th, h) - df(x, h)\| \, dt \end{aligned}$$

By the definition of sup,

$$\|(df(x + th; h) - df(x; h))\| \leq \sup_{v \in V} \frac{\|(df(x + th; v) - df(x; v))\|}{\|v\|} \|h\|.$$

The "continuity" in x implies for any $\epsilon > 0\ \exists \delta > 0$ such that if $\|th\| < \delta$, then $\sup_{v \in V} \frac{\|(df(x+th;v) - df(x;v))\|}{\|v\|}$ ϵ. Thus,

$$\|f(x + h) - f(x) - df(x; h)\| < \int_0^1 \epsilon \|h\| \, dt = \epsilon \|h\|.$$

In other words, for any $\epsilon > 0\ \exists \delta > 0$ such that if $\|h\| < \delta$, then

$$\|f(x + h) - f(x) - df(x; h)\| \|h\| < \epsilon,$$

and we can conclude that $df(x; h) = Df_x h$. □

Example (Example 2.1 continued)**.** Motivated by lemmas 2.1 and 2.2, we can find the Fréchet derivative of f by computing its Gâteaux derivatives. Let $v \in V$. Remember that both x and v are functions in this example. Then,

$$f(x + \alpha v) = \int_0^1 g(x(\tau) + \alpha v(\tau), \tau) d\tau$$

and

$$\begin{aligned} df(x;v) &= \lim_{\alpha\to 0} \frac{\int_0^1 g(x(\tau)+\alpha v(\tau),\tau)d\tau}{\alpha} \\ &= \int_0^1 \frac{\partial g}{\partial x}(x(\tau),\tau)v(\tau)d\tau \end{aligned}$$

Now, we can either check that these derivatives are linear and continuous, or just guess and verify that

$$Df_x(v) = \int_0^1 \frac{\partial g}{\partial x}(x(\tau),\tau)v(\tau)d\tau.$$

Note that this expression is linear in v as it must be for it to be the derivative. Now, we check that the limit in the definition of the derivative is zero,

$$\begin{aligned} \lim_{h\to 0} \frac{\|f(x+h)-f(x)-Df_x(h)\|}{\|h\|} &= \lim_{h\to 0} \frac{\left|\int g(x(\tau)+h(\tau),\tau)-g(x(\tau),\tau)-\frac{\partial g}{\partial x}(x(\tau),\tau)h(\tau)d\tau\right|}{\|h\|} \\ &\leq \lim_{h\to 0} \frac{\int \left|g(x(\tau)+h(\tau),\tau)-g(x(\tau),\tau)-\frac{\partial g}{\partial x}(x(\tau),\tau)h(\tau)\right| d\tau}{\|h\|} \end{aligned}$$

where the inequality follows from the triangle inequality. To simplify, let us assume that g and $\frac{\partial g}{\partial x}$ are bounded. Then, by the dominated convergence, theorem, we can interchange the integral and the limit.[5] We then have

$$\leq \int \lim_{h\to 0} \frac{\left|g(x(\tau)+h(\tau),\tau)-g(x(\tau),\tau)-\frac{\partial g}{\partial x}(x(\tau),\tau)h(\tau)\right|}{\|h\|} d\tau$$

The definition of $\frac{\partial g}{\partial x}$ says that

$$\left|\frac{g(x(\tau)+h(\tau),\tau)-g(x(\tau),\tau)-\frac{\partial g}{\partial x}(x(\tau),\tau)h(\tau)}{h(\tau)}\right| \to 0$$

Also $\frac{|h(\tau)|}{\|h\|} \leq 1$ for all τ because in $\mathcal{L}^\infty(0,1)$, $\|h\| = \sup_{0\leq\tau\leq 1}|h(\tau)|$. Thus, we can conclude that

$$\begin{aligned} &\lim_{h\to 0} \frac{\left|g(x(\tau)+h(\tau),\tau)-g(x(\tau),\tau)-\frac{\partial g}{\partial x}(x(\tau),\tau)h(\tau)\right|}{\|h\|} = \\ &\quad = \lim_{h\to 0} \frac{\left|g(x(\tau)+h(\tau),\tau)-g(x(\tau),\tau)-\frac{\partial g}{\partial x}(x(\tau),\tau)h(\tau)\right|}{|h(\tau)|} \frac{|h(\tau)|}{\|h\|} = 0, \end{aligned}$$

so f is Fréchet differentiable at x with derivative Df_x.

[5] We have not covered the dominated convergence theorem. Unless specifically stated otherwise, on homeworks and exams you can assume that interchanging limits and integrals is allowed. However, do not forget that this is not always allowed. The issue is the order of taking limits. Integrals are defined in terms of limits (either Riemann sums or integrals of simple functions). It is not difficult to come up with examples where $a_{m,n}$ is a doubly indexed sequence and $\lim_{m\to\infty}\lim_{n\to\infty} a_{m,n} \neq \lim_{n\to\infty}\lim_{m\to\infty} a_{m,n}$.