
DIFFERENTIAL CALCULUS

PAUL SCHRIMPF

OCTOBER 31, 2018

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 $f : X \rightarrow Y$.

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

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 $\lim_{x \rightarrow x_0} f(x) = c$ means that for any sequence $\{x_n\}$ with $x_n \rightarrow x$, $f(x_n) \rightarrow c$.

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

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

exists. Similarly, if $f : V \rightarrow W$ we define its i th partial derivative as follows.

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

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

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

¹This work is licensed under a [Creative Commons Attribution-ShareAlike 4.0 International License](https://creativecommons.org/licenses/by-sa/4.0/)

Example 1.1. Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ 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 $u : \mathbb{R}^n \rightarrow \mathbb{R}$ is a utility function, then we call $\frac{\partial u}{\partial x_i}$ the marginal utility of x_i . If u is CRRA,

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

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 $q_1 : \mathbb{R}^3 \rightarrow \mathbb{R}$ 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.

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

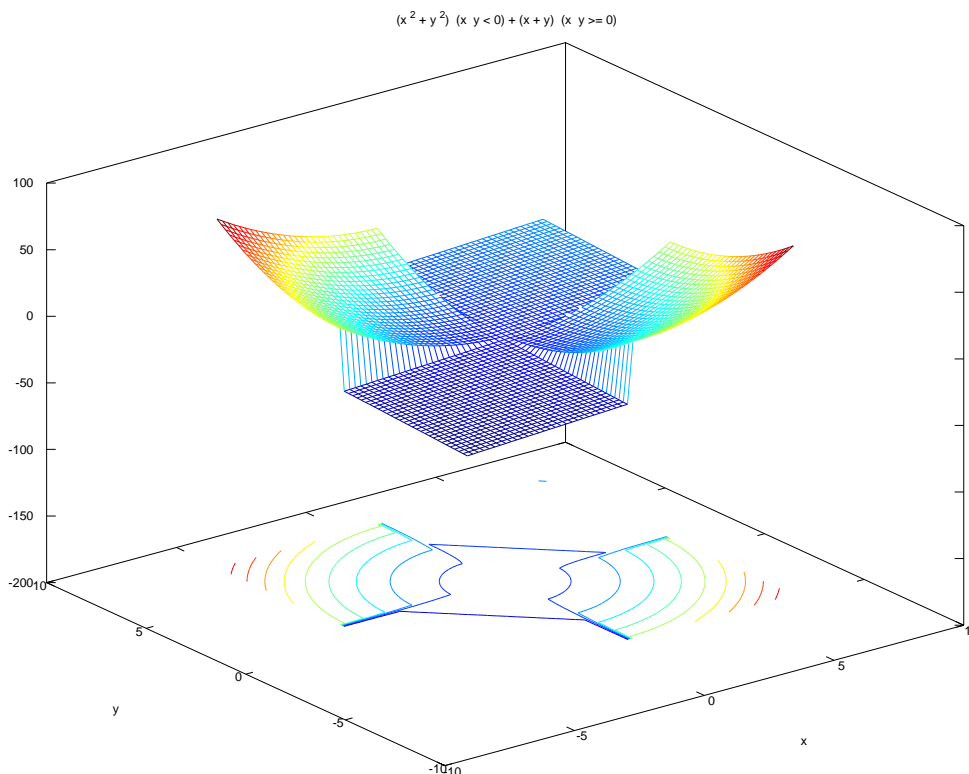
Similarly, the income elasticity of demand is

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

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 $f : \mathbb{R}^2 \rightarrow \mathbb{R}$,

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



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

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

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 $\lim_{h \rightarrow 0} \frac{|r(x, h)|}{\|h\|} = 0$.

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 $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be differentiable at x_0 , then $\frac{\partial f}{\partial x_i}(x_0)$ exists for each i and

$$Df_{x_0}h = \left(\frac{\partial f}{\partial x_1}(x_0) \quad \cdots \quad \frac{\partial f}{\partial x_n}(x_0) \right) h.$$

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

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

Let

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

and note that $\lim_{t \rightarrow 0} \frac{|r_i(x_0, t)|}{|t|} = 0$. 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

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

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

$$Df_{x_0}h = \left(\frac{\partial f}{\partial x_1}(x_0) \quad \cdots \quad \frac{\partial f}{\partial x_n}(x_0) \right) h.$$

□

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 $f : \mathbb{R}^n \rightarrow \mathbb{R}$ 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

$$Df_{x_0} = \left(\frac{\partial f}{\partial x_1}(x_0) \quad \cdots \quad \frac{\partial f}{\partial x_n}(x_0) \right).$$

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

$$f(x_0 + h) - f(x_0) = f(x_0 + h_1 e_1) - f(x_0) + f(x_0 + h_1 e_1 + h_2 e_2) - f(x_0 + h_1 e_1) + \dots \quad (1)$$

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

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

By the mean value theorem (1.5),

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

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

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

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,

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

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

A minor modification of this proof would show the stronger result that $f : \mathbb{R}^n \rightarrow \mathbb{R}$ 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 differentiable** 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.

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

Proof. In the last set of notes. \square

Definition 1.4. Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$. 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 $f : \mathbb{R}^n \rightarrow \mathbb{R}$ 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 $\lim_{h \rightarrow 0} \frac{|r(x,h)|}{\|h\|} = 0$. 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. Similarly, if $Df_x v < 0$ then for $t < 0$ and small, we would have $\frac{f(x+tv) - f(x)}{|t|} = Df_x v + \frac{r(x,tv)}{|t|} > -Df_x v / 2 > 0$ and $f(x+tv) > f(x)$. Thus, it must be that $Df_x v = 0$ for all v , i.e. $Df_x = 0$. \square

Now we can prove the mean value theorem.

Theorem 1.5 (mean value). Let $f : \mathbb{R}^n \rightarrow \mathbb{R}^1$ be in continuously differentiable on some open set U (i.e. $f \in C^1(U)$). 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 $\bar{x} = z(\bar{t})$. \square

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

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

Theorems 1.1 and 2.1 still 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 $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$. Meaning, that f can be written as $f(x) = (f_1(x) \cdots f_m(x))^T$ where each $f_j : \mathbb{R}^n \rightarrow \mathbb{R}$. 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 $\mathbb{R}^n \rightarrow \mathbb{R}^m$). *Let $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ 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 $(Df_{1\bar{x}_1} \cdots Df_{m\bar{x}_m})^T$ 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 $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ and $g : \mathbb{R}^k \rightarrow \mathbb{R}^n$. Let g be continuously differentiable on some open set U and f be continuously differentiable on $g(U)$. Then $h : \mathbb{R}^k \rightarrow \mathbb{R}^m$, $h(x) = f(g(x))$ 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

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

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

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

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

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

so

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

□

1.7. Higher order derivatives. We can take higher order derivatives of multivariate functions just like of univariate functions. If $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$, then f 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 $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ 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

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

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 $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ 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 $p : \{1, \dots, n\} \rightarrow \{1, \dots, n\}$ is any permutation (i.e. reordering).

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

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

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

The same theorem is true for multivariate functions.

Theorem 1.9. Let $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ 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

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

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

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 $h \rightarrow 0$, $\bar{a} \rightarrow a$, so $\lim_{h \rightarrow 0} r_1(a, h) = 0$, 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 $f : \mathbb{R} \rightarrow \mathbb{R}$. 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 ,

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

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

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 $n \rightarrow \infty$. 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

$$Q^0(\theta) = E[(Y - X\theta)^2].$$

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

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

Applying the mean value theorem,

$$0 = DQ_{\hat{\theta}_n}^n = DQ_{\theta_0}^n + D^2Q_{\hat{\theta}}^n(\hat{\theta}_n - \theta_0)$$

$$\hat{\theta}_n - \theta_0 = - \left(D^2Q_{\hat{\theta}}^n \right)^{-1} DQ_{\theta_0}^n.$$

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

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

Also, typically $D^2Q_{\hat{\theta}}^n \xrightarrow{p} D^2Q_{\theta_0}^0$, so by Slutsky's theorem,^b

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = - \left(D^2Q_{\hat{\theta}}^n \right)^{-1} \sqrt{n}DQ_{\theta_0}^n \xrightarrow{d} N \left(0, \left(D^2Q_{\theta_0}^0 \right)^{-1} \Sigma \left(D^2Q_{\theta_0}^0 \right)^{-1} \right).$$

^aEssentially 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.

^bPart 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 need to differentiate functions on infinite dimensional vector spaces. We already did this when studying optimal control, but we glossed over the details. Anyway, we can define the derivative of a function between any two vector spaces as follows.

Definition 2.1. Let $f : V \rightarrow W$. The Fréchet **derivative** of f at x_0 is a continuous² linear mapping, $Df_{x_0} : V \rightarrow W$ such that

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

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 $g : \mathbb{R}^2 \rightarrow \mathbb{R}$. 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.

²If V and W are finite dimensional, then all linear functions are continuous. In infinite dimensions, there can be discontinuous linear functions.

Definition 2.2. Let $f : V \rightarrow W$, $v \in V$ 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

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

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

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

Lemma 2.1. *If $f : V \rightarrow W$ 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.

Lemma 2.2. *If $f : V \rightarrow W$ 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. 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

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

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

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

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

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

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,

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

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.^a We then have

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

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

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

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

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

so f is Fréchet differentiable at x with derivative Df_x .

^aWe 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 $\lim_{m \rightarrow \infty} \lim_{n \rightarrow \infty} a_{m,n} \neq \lim_{n \rightarrow \infty} \lim_{m \rightarrow \infty} a_{m,n}$.

With this definition of the derivative, almost everything that we proved above for functions from $\mathbb{R}^n \rightarrow \mathbb{R}^m$ also holds for functions on Banach spaces. In particular, the chain rule, Taylor's theorem, the implicit function theorem, and the inverse function theorem hold. The proofs of these theorems on Banach spaces are essentially the same as above, so we will not go over them. If you find this interesting, you may want to go through the proofs of all these claims, but it is not necessary to do so.

The mean value theorem is slightly more delicate. It still holds for functions $f : V \rightarrow \mathbb{R}^m$, but must be modified when the target space of f is infinite dimensional. We start with the mean value theorem for $f : V \rightarrow \mathbb{R}$.

Theorem 2.1 (Mean value theorem (onto \mathbb{R})). *Let $f : V \rightarrow \mathbb{R}^1$ be continuously differentiable on some open set 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. Identical to the proof of 1.5. □

This result can then be generalized to $f : V \rightarrow \mathbb{R}^n$ in the same way as theorem 1.1.

Now, let's consider a first order Taylor expansion for $f : V \rightarrow W$:

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

To show that first order Taylor expansions have small approximation error, we need to show that $(Df_{\bar{x}} - Df_x)h$ is uniformly small for all possible h . When W is \mathbb{R}^1 , there is a single \bar{x} . When $W = \mathbb{R}^m$, then \bar{x} is different for each dimension (as in theorem 1.1). However, this is okay because we just take the maximum over the finitely many different \bar{x} to get an upper bound on $\|(Df_{\bar{x}} - Df_x)h\|$. When W is infinite dimensional, this argument does not work. Instead, we must use a result called the Hahn-Banach extension theorem, which is closely related to the separating hyperplane theorem.

Theorem 2.2 (Hahn-Banach theorem). *Let V be a vector space and $g : V \rightarrow \mathbb{R}$ be convex. Let $S \subseteq V$ be a linear subspace. Suppose $f_0 : S \rightarrow \mathbb{R}$ is linear and*

$$f_0(x) \leq g(x)$$

for all $x \in S$. Then f_0 can be extended to $f : V \rightarrow \mathbb{R}$ such that

$$f_0(x) = f(x)$$

for all $x \in S$ and

$$f(x) \leq g(x)$$

for all $x \in V$.

Proof. We will apply the separating hyperplane theorem to two sets in $V \times \mathbb{R}$. Let

$$A = \{(x, a) : a > g(x), x \in V, a \in \mathbb{R}\}.$$

Since g is a convex function, A is a convex set. Let

$$B = \{(x, a) : a = f_0(x), x \in S, a \in \mathbb{R}\}$$

Since f_0 is linear, B is convex. Clearly $A \cap B = \emptyset$. Also, $\int A \neq \emptyset$ because A is open, so $\int A = A$ and $A \neq \emptyset$ since e.g. $(x, g(x) + 1) \in A$.

Therefore, by the separating hyperplane theorem, $\exists \xi \in (V \times \mathbb{R})^*$ and $c \in \mathbb{R}$ such that

$$\xi z_A > c \geq \xi z_B$$

for all $z_A \in A$ and $z_B \in B$. B is a linear subspace, so it must be that $\xi z_B = 0$ for all $z_B \in B$, so we can take $c = 0$.

Note that since $(0, 1) \in A$, $\xi(0, 1) > 0$. Let $f : V \rightarrow \mathbb{R}$ be given by $f(x) = -\frac{\xi(x, 0)}{\xi(0, 1)}$. To conclude, we must show that $f(x) = f_0(x)$ for all $x \in S$ and $f(x) \leq g(x)$ for all $x \in V$. To show that $f(x) = f_0(x)$, by linearity of ξ we have

$$\xi(x, y) = \xi(x, 0) + y\xi(0, 1)$$

$$\frac{\xi(x, y)}{\xi(x, 0)} = -f(x) + y$$

If $x \in S$, then $\xi(x, f_0(x)) = 0$, so $f(x) = f_0(x)$ for all $x \in S$. Similarly for any $y > g(x)$, then $(x, y) \in A$, so $\frac{\xi(x,y)}{\xi(x,0)} = -f(x) + y > 0$, and $y > f(x)$. Therefore, $f(x) \leq g(x)$. \square

With this result in hand, we now can now prove the mean value theorem for arbitrary Banach spaces.

Theorem 2.3 (Mean value theorem (onto \mathbb{R})). *Let $f : V \rightarrow W$ be in continuously differentiable on some open set 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*

$$\|f(x) - f(y)\|_W \leq \|Df_{\bar{x}}(x - y)\|_W \leq \|Df_{\bar{x}}\|_{BL(V,W)} \|x - y\|_V$$

Proof. Step 1: use the Hahn Banach theorem to show that $\exists \phi \in W^*$ with $\|\phi\| = 1$ and $\|f(x) - f(y)\| = \phi(f(x) - f(y))$. The relevant f_0 is $f_0(\alpha(f(x) - f(y))) = \alpha \|f(x) - f(y)\|$.

Define $g : [0, 1] \rightarrow \mathbb{R}$ by $g(t) = \phi(f(tx) - f((1-t)y))$ and apply the mean value theorem on \mathbb{R} to get

$$\|f(x) - f(y)\| = \phi Df_{t^*x+(1-t^*)y}(x - y)$$

\square

3. OPTIMIZATION IN VECTOR SPACES

Recall from earlier that there are many sets of functions that are vector spaces. We talked a little bit about \mathcal{L}^p spaces of functions. The set of all continuous functions and the sets of all k times continuously differentiable functions are also vector spaces. One of these vector spaces of functions will be appropriate for finding the solution to optimal control problems. Exactly which vector space is a slightly technical problem dependent question, so we will not worry about that for now (and we may not worry about it at all). Similarly, there are vector spaces of infinite sequences. Little ℓ^p is similar to big \mathcal{L}^p , but with sequences instead of functions

$$\ell^p = \{\{x_t\}_{t=1}^\infty : \left(\sum_{t=1}^\infty |x_t|^p\right)^{1/p} < \infty\}$$

There are others as well. Again, the right choice of vector space depends on the problem being considered, and we will not worry about it too much.

Given two Banach spaces (complete normed vector spaces), V and W , let $BL(V, W)$ denote the set of all linear transformations from V to W . The derivative of $f : V \rightarrow W$ will be in $BL(V, W)$. We can show that $BL(V, W)$ is a vector space, and we can define a norm on $BL(V, W)$ by

$$\|D\|_{BL(V,W)} = \sup_{\|v\|_V=1} \|Dv\|_W$$

where $\|\cdot\|_V$ is the norm on V and $\|\cdot\|_W$ is the norm on W . Moreover, we could show that $BL(V, W)$ is complete. Thus, $BL(V, W)$ is also a Banach space. Viewed as function of x , Df_x is a function from V to $BL(V, W)$. As we just said, there are both Banach spaces, so can differentiate Df_x with respect to x . In this way, we can define the second and higher derivatives of $f : V \rightarrow W$.

In the previous section we saw that differentiation for functions on Banach spaces is the same as for functions on finite dimensional vector spaces. All of our proofs of first and second order conditions only relied on Taylor expansions and some properties of linear transformations. Taylor expansions and linear transformations are the same on Banach spaces as on finite dimensional vector spaces, so our results for optimization will still hold. Let's just state for the first order condition for equality constraints. The other results are similar, but stating them gets to be slightly cumbersome.

Theorem 3.1 (First order condition for maximization with equality constraints). *Let $f : U \rightarrow \mathbb{R}$ and $h : U \rightarrow W$ be continuously differentiable on $U \subseteq V$, where V and W are Banach spaces. Suppose $x^* \in \text{interior}(U)$ is a local maximizer of f on U subject to $h(x) = 0$. Suppose that $Dh_{x^*} : V \rightarrow W$ is onto. Then, there exists $\mu^* \in BL(W, \mathbb{R})$ such that for*

$$L(x, \mu) = f(x) - \mu h(x).$$

we have

$$\begin{aligned} D_x L(x^*, \mu^*) &= Df_{x^*} - \mu^* Dh_{x^*} = 0_{BL(V, \mathbb{R})} \\ D_\mu L(x^*, \mu^*) &= h(x^*) = 0_W \end{aligned}$$

There are a few differences compared to the finite dimensional case that are worth commenting on. First, in the finite dimensional case, we had $h : U \rightarrow \mathbb{R}^m$, and the condition that $\text{rank} Dh_{x^*} = m$. This is the same as saying that the $Dh_{x^*} : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is onto. Rank is not well-defined in infinite dimension, so we now state this condition as Dh_{x^*} being onto instead of being rank m .

Secondly, previously $\mu \in \mathbb{R}^m$, and the Lagrangian was

$$L(x, \mu) = f(x) - \mu^T h(x).$$

Viewed as a 1 by m matrix, μ^T is a linear transformation from \mathbb{R}^m to \mathbb{R} . Thus, in the abstract case, we just say $\mu \in BL(W, \mathbb{R})$, which as when we defined transposes, is called the dual space of W and is denoted W^* .

Finally, we have subscripted the zeroes in the first order condition with $BL(V, \mathbb{R})$ and W to emphasize that the first equation is for linear transformations from V to \mathbb{R} , and the second equation is in W . Df_{x^*} is a linear transformation from V to \mathbb{R} . Dh_{x^*} goes from V to W . μ goes from W to \mathbb{R} , so μ composed with Dh_{x^*} , which we just denoted by μDh_{x^*} is a linear transformation from V to \mathbb{R} .

4. INVERSE FUNCTIONS

Suppose $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$. If we know $f(x) = y$, when can we solve for x in terms of y ? In other words, when is f invertible? Well, suppose we know that $f(a) = b$ for some $a \in \mathbb{R}^n$ and $b \in \mathbb{R}^m$. Then we can expand f around a ,

$$f(x) = f(a) + Df_a(x - a) + r_1(a, x - a)$$

where $r_1(a, x - a)$ is small. Since r_1 is small, we can hopefully ignore it then $y = f(x)$ can be rewritten as a linear equation:

$$\begin{aligned} f(a) + Df_a(x - a) &= y \\ Df_a x &= y - f(a) + Df_a a \end{aligned}$$

we know that this equation has a solution if $\text{rank} Df_a = \text{rank} (Df_a \ y - f(a) + Df_a a)$. It has a solution for any y if $\text{rank} Df_a = m$. Moreover, this solution is unique if $\text{rank} Df_a = n$. This discussion is not entirely rigorous because we have not been very careful about what r_1 being small means. The following theorem makes it more precise.

Theorem 4.1 (Inverse function). *Let $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ be continuously differentiable on an open set E . Let $a \in E$, $f(a) = b$, and Df_a be invertible. Then*

- (1) *there exist open sets U and V such that $a \in U$, $b \in V$, f is one-to-one on U and $f(U) = V$, and*
- (2) *the inverse of f exists and is continuously differentiable on V with derivative $(Df_{f^{-1}(x)})^{-1}$.*

The open sets U and V are the areas where r_1 is small enough. The continuity of f and its derivative are also needed to ensure that r_1 is small enough. The proof of this theorem is a bit long, but the main idea is the same as the discussion preceding the theorem.

Comment 4.1. The proof uses the fact that the space of all continuous linear transformations between two normed vector spaces is itself a vector space. I do not think we have talked about this before. Anyway, it is a useful fact that already came up in the proof that continuous Gâteaux differentiable implies Fréchet differentiable last lecture. Let V and W be normed vector spaces with norms $\|\cdot\|_V$ and $\|\cdot\|_W$. Let $BL(V, W)$ denote the set of all continuous (or equivalently bounded) linear transformations from V to W . Then $BL(V, W)$ is a normed vector space with norm

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

This is sometimes called the operator norm on $BL(V, W)$. Last lecture, the proof that Gâteaux differentiable implies Fréchet differentiable required that the mapping from V to $BL(V, W)$ defined by Df_x as a function of $x \in V$ had to be continuous with respect to the above norm.

We will often use the inequality,

$$\|Av\|_W \leq \|A\|_{BL} \|v\|_V,$$

which follows from the definition of $\|\cdot\|_{BL}$. We will also use the fact that if V is finite dimensional and $f(x, v) : V \times V \rightarrow W$, is continuous in x and v and linear in v for each x , then $f(x, \cdot) : V \rightarrow BL(V, W)$ is continuous in x with respect to $\|\cdot\|_{BL}$.

Proof. For any $y \in \mathbb{R}^n$, consider $\varphi^y(x) = x + Df_a^{-1}(y - f(x))$. By the mean value theorem for $x_1, x_2 \in U$, where $a \in U$ and U is open,

$$\varphi^y(x_1) - \varphi^y(x_2) = D\varphi_{\bar{x}}^y(x_1 - x_2)$$

Note that

$$\begin{aligned} D\varphi_{\bar{x}}^y &= I - Df_a^{-1}Df_{\bar{x}} \\ &= Df_a^{-1}(Df_a - Df_{\bar{x}}). \end{aligned}$$

Since Df_x is continuous (as a function of x) if we make U small enough, then $Df_a - Df_{\bar{x}}$ will be near 0. Let $\lambda = \frac{1}{2\|Df_a^{-1}\|_{BL}}$. Choose U small enough that $\|Df_a - Df_x\| < \lambda$ for all $x \in U$. From above, we know that

$$\begin{aligned} \|\varphi^y(x_1) - \varphi^y(x_2)\| &= \|Df_a^{-1}(Df_a - Df_{\bar{x}})(x_1 - x_2)\| \\ &\leq \|D\varphi_x^y\|_{BL} \|Df_a - Df_x\|_{BL} \|x_1 - x_2\| \\ &\leq \frac{1}{2} \|x_1 - x_2\| \end{aligned} \tag{7}$$

For any $y \in f(U)$ we can start with an arbitrary $x_1 \in U$, then create a sequence by setting

$$x_{i+1} = \varphi^y(x_i).$$

From (7), this sequence satisfies

$$\|x_{i+1} - x_i\| \leq \frac{1}{2} \|x_i - x_{i-1}\|.$$

Using this it is easy to verify that x_i form a Cauchy sequence, so it converges. The limit satisfy $\varphi^y(x) = x$, i.e. $f(x) = y$. Moreover, this x is unique because if $\varphi^y(x_1) = x_1$ and $\varphi^y(x_2) = x_2$, then we have $\|x_1 - x_2\| \leq \frac{1}{2} \|x_1 - x_2\|$, which is only possible if $x_1 = x_2$.³ Thus for each $y \in f(U)$, there is exactly one x such that $f(x) = y$. That is, f is one-to-one on U . This proves the first part of the theorem and that f^{-1} exists.

We now show that f^{-1} is continuously differentiable with the stated derivative. Let $y, y+k \in V = f(U)$. Then $\exists x, x+h \in U$ such that $y = f(x)$ and $y+k = f(x+h)$. With φ^y as defined above, we have

$$\begin{aligned} \varphi^y(x+h) - \varphi^y(x) &= h + Df_a^{-1}(f(x) - f(x+h)) \\ &= h - Df_a^{-1}k \end{aligned}$$

By 7, $\|h - Df_a^{-1}k\| \leq \frac{1}{2} \|h\|$. It follows that $\|Df_a^{-1}k\| \geq \frac{1}{2} \|h\|$ and

$$\|h\| \leq 2\|Df_a^{-1}\|_{BL} \|k\| = \lambda^{-1} \|k\|.$$

³Functions like φ^y that have $d(\phi(x), \phi(y)) \leq cd(x, y)$ for $c < 1$ are called contraction mappings. The x with $x = \phi(x)$ is called a fixed point of the contraction mapping. The argument in the proof shows that contraction mappings have at most one fixed point. It is not hard to show that contraction mappings always have exactly one fixed point.

Importantly as $k \rightarrow 0$, we also have $h \rightarrow 0$. Now,

$$\begin{aligned} \frac{\|f^{-1}(y+k) - f^{-1}(y) - Df_x^{-1}k\|}{\|k\|} &= \frac{\|-Df_x^{-1}(f(x+h) - f(x) - Df_x h)\|}{\|k\|} \\ &\leq \|Df_x\|^{-1} \lambda \frac{\|f(x+h) - f(x) - Df_x h\|}{\|h\|} \\ \lim_{k \rightarrow 0} \frac{\|f^{-1}(y+k) - f^{-1}(y) - Df_x^{-1}k\|}{\|k\|} &\leq \lim_{k \rightarrow 0} \|Df_x\|_{BL}^{-1} \lambda \frac{\|f(x+h) - f(x) - Df_x h\|}{\|h\|} = 0 \end{aligned}$$

Finally, since Df_x is continuous, so is $(Df_{f^{-1}(y)})^{-1}$, which is the derivative of f^{-1} . \square

The proof of the inverse function theorem might be a bit confusing. The important idea is that if the derivative of a function is nonsingular at a point, then you can invert the function around that point because inverting the system of linear equations given by the mean value expansion around that point nearly gives the inverse of the function.

5. IMPLICIT FUNCTIONS

The implicit function theorem is a generalization of the inverse function theorem. In economics, we usually have some variables, say x , that we want to solve for in terms of some parameters, say β . For example, x could be a person's consumption of a bundle of goods, and b could be the prices of each good and the parameters of the utility function. Sometimes, we might be able to separate x and β so that we can write the conditions of our model as $f(x) = b(\beta)$. Then we can use the inverse function theorem to compute $\frac{\partial x_i}{\partial \beta_j}$ and other quantities of interest. However, it is not always easy and sometimes not possible to separate x and β onto opposite sides of the equation. In this case our model gives us equations of the form $f(x, \beta) = c$. The implicit function theorem tells us when we can solve for x in terms of β and what $\frac{\partial x_i}{\partial \beta_j}$ will be.

The basic idea of the implicit function theorem is the same as that for the inverse function theorem. We will take a first order expansion of f and look at a linear system whose coefficients are the first derivatives of f . Let $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$. Suppose f can be written as $f(x, y)$ with $x \in \mathbb{R}^k$ and $y \in \mathbb{R}^{n-k}$. x are endogenous variables that we want to solve for, and y are exogenous parameters. We have a model that requires $f(x, y) = c$, and we know that some particular x_0 and y_0 satisfy $f(x_0, y_0) = c$. To solve for x in terms of y , we can expand f around x_0 and y_0 .

$$f(x, y) = f(x_0, y_0) + D_x f_{(x_0, y_0)}(x - x_0) + D_y f_{(x_0, y_0)}(y - y_0) + r(x, y) = c$$

In this equation, $D_x f_{(x_0, y_0)}$ is the m by k matrix of first partial derivatives of f with respect to x evaluated at (x_0, y_0) . Similarly, $D_y f_{(x_0, y_0)}$ is the m by $n - k$ matrix of first partial derivatives of f with respect to y evaluated at (x_0, y_0) . Then, if $r(x, y)$ is small enough, we have

$$\begin{aligned} f(x_0, y_0) + D_x f_{(x_0, y_0)}(x - x_0) + D_y f_{(x_0, y_0)}(y - y_0) &\approx c \\ D_x f_{(x_0, y_0)}(x - x_0) &\approx \left(c - f(x_0, y_0) - D_y f_{(x_0, y_0)}(y - y_0) \right) \end{aligned}$$

This is just a system of linear equations with unknowns $(x - x_0)$. If $k = m$ and $D_x f_{(x_0, y_0)}$ is nonsingular, then we have

$$x \approx x_0 + \left(D_x f_{(x_0, y_0)} \right)^{-1} \left(c - f(x_0, y_0) - D_y f_{(x_0, y_0)}(y - y_0) \right)$$

which gives x approximately as function of y . The implicit function says that you can make this approximation exact and get $x = g(y)$. The theorem also tells you what the derivative of $g(y)$ is in terms of the derivative of f .

Theorem 5.1 (Implicit function). *Let $f : \mathbb{R}^{n+m} \rightarrow \mathbb{R}^n$ be continuously differentiable on some open set E and suppose $f(x_0, y_0) = c$ for some $(x_0, y_0) \in E$, where $x_0 \in \mathbb{R}^n$ and $y_0 \in \mathbb{R}^m$. If $D_x f_{(x_0, y_0)}$ is invertible, then there exists open sets $U \subset \mathbb{R}^n$ and $W \subset \mathbb{R}^{n-k}$ with $x_0 \in U$ and $y_0 \in W$ such that*

- (1) *For each $y \in W$ there is a unique x such that $(x, y) \in U$ and $f(x, y) = c$.*
- (2) *Define this x as $g(y)$. Then g is continuously differentiable on W , $g(y_0) = x_0$, $f(g(y), y) = c$ for all $y \in W$, and $Dg_{y_0} = - \left(D_x f_{(x_0, y_0)} \right)^{-1} D_y f_{(x_0, y_0)}$*

Proof. We will show the first part by applying the inverse function theorem. Define $F : \mathbb{R}^{n+m} \rightarrow \mathbb{R}^{n+m}$ by $F(x, y) = (f(x, y), y)$. To apply the inverse function theorem we must show that F is continuously differentiable and $DF_{(x_0, y_0)}$ is invertible. To show that F is continuously differentiable, note that

$$\begin{aligned} F(x+h, y+k) - F(x, y) &= (f(x+h, y+k) - f(x, y), k) \\ &= (Df_{(\bar{x}, \bar{y})}(h, k), k) \end{aligned}$$

where the second line follows from the mean value theorem. It is then apparent that

$$\lim_{(h, k) \rightarrow 0} \frac{\left\| F(x+h, y+k) - F(x, y) - \begin{pmatrix} D_x f_{(x, y)} & D_y f_{(x, y)} \\ 0 & I_m \end{pmatrix} \begin{pmatrix} h \\ k \end{pmatrix} \right\|}{\|(h, k)\|} = 0.$$

So, $DF_{(x, y)} = \begin{pmatrix} D_x f_{(x, y)} & D_y f_{(x, y)} \\ 0 & I_m \end{pmatrix}$, which is continuous since $Df_{(x, y)}$ is continuous. Also, $DF_{(x_0, y_0)}$ can be shown to be invertible by using the partitioned inverse formula because $D_x f_{(x_0, y_0)}$ is invertible by assumption. Therefore, by the inverse function theorem, there exists open sets U and V such that $(x_0, y_0) \in U$ and $(c, y_0) \in V$, and F is one-to-one on U .

Let W be the set of $y \in \mathbb{R}^m$ such that $(c, y) \in V$. By definition, $y_0 \in W$. Also, W is open in \mathbb{R}^m because V is open in \mathbb{R}^{n+m} .

We can now complete the proof of 1. If $y \in W$ then $(c, y) = F(x, y)$ for some $(x, y) \in U$. If there is another (x', y) such that $f(x', y) = c$, then $F(x', y) = (c, y) = F(x, y)$. We just showed that F is one-to-one on U , so $x' = x$.

We now prove 2. Define $g(y)$ for $y \in W$ such that $(g(y), y) \in U$ and $f(g(y), y) = c$, and

$$F(g(y), y) = (c, y).$$

By the inverse function theorem, F has an inverse on U . Call it G . Then

$$G(c, y) = (g(y), y)$$

and G is continuously differentiable, so g must be as well. Differentiating the above equation with respect to y , we have

$$D_y G_{(c,y)} = \begin{pmatrix} Dg_y \\ I_m \end{pmatrix}$$

On the other hand, from the inverse function theorem, the derivative of G at (x_0, y_0) is

$$\begin{aligned} DG_{(x_0,y_0)} &= \left(DF_{(x_0,y_0)} \right)^{-1} \\ &= \begin{pmatrix} D_x f_{(x_0,y_0)} & D_y f_{(x_0,y_0)} \\ 0 & I_m \end{pmatrix}^{-1} \\ &= \begin{pmatrix} D_x f_{(x_0,y_0)}^{-1} & -D_x f_{(x_0,y_0)}^{-1} D_y f_{(x_0,y_0)} \\ 0 & I_m \end{pmatrix} \end{aligned}$$

In particular,

$$D_y G_{(c,y_0)} = \begin{pmatrix} -D_x f_{(x_0,y_0)}^{-1} D_y f_{(x_0,y_0)} \\ I_m \end{pmatrix} = \begin{pmatrix} Dg_{y_0} \\ I_m \end{pmatrix}$$

so $Dg_{y_0} = -D_x f_{(x_0,y_0)}^{-1} D_y f_{(x_0,y_0)}$. □

6. CONTRACTION MAPPINGS

One step of the proof the of the inverse function theorem was to show that

$$\|\varphi^y(x_1) - \varphi^y(x_2)\| \leq \frac{1}{2} \|x_1 - x_2\|.$$

This property ensures that $\varphi(x) = x$ has a unique solution. Functions like φ^y appear quite often, so they have name.

Definition 6.1. Let $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$. f is a **contraction mapping** on $U \subseteq \mathbb{R}^n$ if for all $x, y \in U$,

$$\|f(x) - f(y)\| \leq c \|x - y\|$$

for some $0 \leq c < 1$.

If f is a contraction mapping, then an x such that $f(x) = x$ is called a **fixed point** of the contraction mapping. Any contraction mapping has at most one fixed point.

Lemma 6.1. Let $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ be a contraction mapping on $U \subseteq \mathbb{R}^n$. If $x_1 = f(x_1)$ and $x_2 = f(x_2)$ for some $x_1, x_2 \in U$, then $x_1 = x_2$.

Proof. Since f is a contraction mapping,

$$\|f(x_1) - f(x_2)\| \leq c \|x_1 - x_2\|.$$

$f(x_i) = x_i$, so

$$\|x_1 - x_2\| \leq c \|x_1 - x_2\|.$$

Since $0 \leq c < 1$, the previous inequality can only be true if $\|x_1 - x_2\| = 0$. Thus, $x_1 = x_2$. □

Starting from any x_0 , we can construct a sequence, $x_1 = f(x_0)$, $x_2 = f(x_1)$, etc. When f is a contraction, $\|x_n - x_{n+1}\| \leq c^n \|x_1 - x_0\|$, which approaches 0 as $n \rightarrow \infty$. Thus, $\{x_n\}$ is a Cauchy sequence and converges to a limit. Moreover, this limit will be such that $x = f(x)$, i.e. it will be a fixed point.

Lemma 6.2. *Let $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ be a contraction mapping on $U \subseteq \mathbb{R}^n$, and suppose that $f(U) \subseteq U$. Then f has a unique fixed point.*

Proof. Pick $x_0 \in U$. As in the discussion before the lemma, construct the sequence defined by $x_n = f(x_{n-1})$. Each $x_n \in U$ because $x_n = f(x_{n-1}) \in f(U)$ and $f(U) \subseteq U$ by assumption. Since f is a contraction on U , $\|x_{n+1} - x_n\| \leq c^n \|x_1 - x_0\|$, so $\lim_{n \rightarrow \infty} \|x_{n+1} - x_n\| = 0$, and $\{x_n\}$ is a Cauchy sequence. Let $x = \lim_{n \rightarrow \infty} x_n$. Then

$$\begin{aligned} \|x - f(x)\| &\leq \|x - x_n\| + \|f(x) - f(x_{n-1})\| \\ &\leq \|x - x_n\| + c \|x - x_{n-1}\| \end{aligned}$$

$x_n \rightarrow x$, so for any $\epsilon > 0 \exists N$, such that if $n \geq N$, then $\|x - x_n\| < \frac{\epsilon}{1+c}$. Then,

$$\|x - f(x)\| < \epsilon$$

for any $\epsilon > 0$. Therefore, $x = f(x)$. □